Optical Transmitters and Receivers

Wolfgang Freude*
Institute of Photonics and Quantum Electronics (IPQ)
(Institut für Photonik und Quantenelektronik)
http://www.ipq.kit.edu
Karlsruhe Institute of Technology (KIT)
http://www.kit.edu

Karlsruhe School of Optics & Photonics (KSOP)
http://ksop.idschools.kit.edu

WS 2015/2016

February 26, 2016

 mailto:w.freude@kit.edu

Contents

Preface

Lightwave technology developed over the last 40 years has greatly influenced our needs for communication. Resources made accessible in the World Wide Web (WWW) have changed our attitude towards information acquisition, which is being regarded as an everyday's necessity, and even as a natural right for everybody.

This course concentrates — after a brief introduction to optical communications as such — on basic communication concepts including a review of modulation formats, on optical transmitters including light sources and modulators, and on optical receivers including photodiodes and electronic circuitry.

Emphasis is on physical understanding. A selection of topics is presented, on which the questioning during the oral examination will be based. Especially the laser and photodiode sections, which have some overlap with the course "Optoelectronic Components (OC / IPQ)" were included for completeness' sake, but will not be treated in full detail. The same is true for the Appendices "Linear and nonlinear fibre properties", "Sampling, quantizing and discrete Fourier transform", and "Coherent signal and noise". — Some minimal background is required: Calculus, differential equations, linear systems, Fourier transform, and pn-junction physics. For further reading, the following list provides some material. References on more specialized topics are cited in the text.

Texbooks: Grau, G.; Freude, W.: **Optische Nachrichtentechnik**, 3. Ed. Berlin: Springer-Verlag 1991. In German. Since 1997 out of print. Corrected reprint 2005, available in electronic form via W. F. (`w.freude@kit.edu`). **Further material is found in:** Agrawal, G. P.: **Fiber-optic communication systems**. Chichester: John Wiley & Sons 1997 — Agrawal, G. P.: Lightwave technology. Vol. 1: Components and devices. **Vol. 2: Telecommunication systems**. Hoboken: John Wiley & Sons 2004 — H. Venghaus, N. Grote (Eds.): **Fibre optic communication — Key devices**. Heidelberg: Springer-Verlag 2012 — Hecht, E.: Optics, 2. Ed. Reading: Addison-Wesley 1974 — Hecht, J.: Understanding fiber optics, 4. Ed. Upper Saddle River: Prentice Hall 2002 — Iizuka, K.: Elements of photonics, Vol. I and II. New York: John Wiley & Sons 2002 — Jahns, J.: Photonik. Grundlagen, Komponenten und Systeme. München: Oldenburg 2001. In German — Leuthold, J.; Freude, W.: **Optical OFDM and Nyquist multiplexing**. In: Kaminow, I. P.; Li, Tingye; Willner, A. E. (Eds.): Optical Fiber Telecommunications VI B. Systems and Networks, 6th Ed. Elsevier (Imprint: Academic Press), Amsterdam 2013, Chapter 9, pp. 381–432 — Liu, M. M.-K.: **Principles and applications of optical communications**. Chicago: McGraw-Hill 1996 — Singh, J.: Physics of semiconductors and their heterostructures. New York: McGraw-Hill 1993. — Sze, S. M.: Physics of semiconductor devices. New York: John Wiley & Sons 1985 — Voges, E.; Petermann, K. (Eds.): Optische Kommunikationstechnik. Handbuch für Wissenschaft und Industrie (Handbook of optical communications). Springer-Verlag, Berlin 2002, pp. 214–260. In German

---

There are other courses on Optical Communications, which cover the material either with a broader view like "Optoelectronic Components (OC / IPQ)", or in more detail like "Optical Waveguides and Fibres (OWF / IPQ)", and "Field Propagation and Coherence (FPC / IPQ)". A course on "Nonlinear Optics (NLO / IPQ)" discusses nonlinear phenomena, which become increasingly important.

Many figures of this compuscript were taken from "Optische Nachrichtentechnik" (see above) carrying German lettering and decimal commas. Appropriate translations are given in the figure captions.

Until 2013, this lecture was jointly held with Prof. Jürg Leuthold, now with Institute of Electromagnetic Fields (Institut für elektromagnetische Felder, IEF), Swiss Federal Institute of Technology (Eidgenössische Technische Hochschule, ETH), Zürich. From his part of the lecture at IPQ and from a newly designed lecture held at IEF materials and graphs were used in some places with his kind permission[†].

[†] J. Leuthold: Optical communication systems — Part 2: Transmitters and modulation formats. Institute of Photonics and Quantum Electronics (IPQ) at Karlsruhe Institute of Technology (KIT), Compuscript 08.02.2013

J. Leuthold: Optical communication fundamentals. ETH Zürich, Institute of Electromagnetic Fields (IEF). Compuscript, November 2013

Directions

At first sight, these lecture notes might be scaring: When you browse through the material from the beginning of Chapter 1 on Page 1 to the end of Chapter 5 on Page 175 you might perceive an overwhelming mass of formulas, graphs and text. How can all these details on transmitters and receivers be taught and understood during a single compact course? I think it is possible, if a few rules are observed:

- Scan the text for physical explanations which could help you developing a pictorial view of the problem.
- Do try to understand the contents and the associated assumptions of important formulae in the given physical context. Best is to put the meaning into words.
- Study the graphical display of major findings. Begin with the axis labels, read the caption, and look carefully at the graph and its parameters.
- Do not start with deriving formulas, and do not learn them by heart (a few exceptions will be named during the lecture). Deriving a relation may be delayed until you have some basic understanding and become curious to learn more about the assumptions and implications. As long as you are fighting with the physical interpretation of a relation, the mathematical details should be of no concern for you, and you will not be questioned on them — it could be different if you aim at an outstanding examination mark.
- Do not practice for the exam during the exam itself. The only proof to have understood the physics is the ability to explain the topic to somebody else *before* you meet me in an examination. Best is you work in groups.

During the lectures and especially during the tutorial, which to attend actively I do recommend strongly, there will be time to answer your questions which may arise from studying the script. If you are then able to explain the matter to fellow students, you are on firm grounds and are well prepared for the examination.

The presentation and the examination will concentrate on the aforementioned points, while the lecture notes (which will be made available to you during the examination) provide a more complete background for your reference. If more information is to be found in the lecture notes than was presented during the lectures, it is intended to make the notes self-consistent, but these additions may be safely skipped when preparing for the examination.

You can also download the lecture slides which serve as a reminder of the actually presented material. Many slides are hidden, so sometimes the slide numbers increment unevenly. Remember also that there are right-pointing arrow-shaped links (mostly in the upper-right corner), which when clicking on them carry you forward. The skipped pages were also omitted during the lectures, but are kept to satisfy your curiosity (in case you have time for such a thing). *Once more: The omitted material is not relevant for the examination.*

Studying the lecture notes on paper is fine, but because all cross-references are linked in the electronic *p*ortable *d*ocument *f*ormat, it may be helpful to read the pdf-version in parallel on-screen. A click on a link carries you immediately to the target[1], and you can navigate at will. Acrobat Reader[2] or Foxit Reader[3] allow you to search the document for text. You can mark and comment certain lines with an electronic text marker, you can store your comments, and you can retrieve this information later on.

[1] On rare occasions, the target page for "floating" objects like figures or tables is wrong by one (e. g., you arrive on Page 14 instead on Page 15), however, the page number printed in the originating text is always correct.

[2] `http://www.adobe.de`

[3] `http://www.foxitsoftware.com/pdf/rd_intro.php` — A very lean application, no installation is required. Better refrain from establishing Foxit as the default reader if you consider using Acrobat in parallel.

Chapter 1

Introduction

1.1 The nature of light

According to Maxwell[1], light propagates as a wave having a wavelength λ. In vacuum, the speed of light is $c = 2.997\,924\,58 \times 10^8\,\mathrm{m/s}$. However, Planck[2] found that the energy of light radiated from a hot black body is emitted in quanta, the energy of which is in proportion to the observed frequency $f = c/\lambda$, so that each quantum or "photon" has an energy $W = hf = \hbar\omega$; Planck's constant is $h = 6.626\,075\,5 \times 10^{-34}\,\mathrm{Ws}$ with $\hbar = h/(2\pi)$, and the angular frequency is $\omega = 2\pi f$. Further, it was shown by de Broglie[3] that each particle having momentum p may be associated with a wavelength $\lambda = h/p$. This statement can be also reverted: Each wave with wavelength λ has a mechanical momentum $p = h/\lambda$ (in vacuum: $p = \hbar k_0$ with free-space propagation constant $k_0 = 2\pi/\lambda = \omega/c$). Obviously, the nature of light is ambiguous. Einstein[4] formulated[5]: "Light is like the French philosopher Voltaire[6]. Voltaire was born catholic, converted as a young man to Protestantism, and returned to Catholicism shortly before his death." Therefore Einstein concludes: "Light is born as a particle, lives as a wave, and dies as a photon when being absorbed."

1.2 Communication with light

An optical communication system uses lightwaves in a vacuum wavelength range $0.6\,\mu\mathrm{m} \ldots 1.2\,\mu\mathrm{m} \leq \lambda \leq 1.6\,\mu\mathrm{m}$ corresponding to carrier frequencies $f = c/\lambda$ of $500\,\mathrm{THz} \ldots 250\,\mathrm{THz} \geq f \geq 190\,\mathrm{THz}$. A communication *system* is referred to as a point-to-point transmission link. When many transmission links are interconnected with multiplexing or switching functions, they are called a communication *network*. The principle of an optical transmission link is shown in Fig. 1.1.

A semiconductor device (*l*aser *d*iode LD, *l*ight-*e*mitting *d*iode LED) emitting light near a wavelength λ is excited by an electric current, thereby converting the electrical signal information to light. This subsystem represents a simple *transmitter* (Tx). The signal can be transmitted simply as an analogue or a digital modulation of the light power $P(t)$ (unit W) or intensity $I(t) = P(t)/F$ (unit $\mathrm{W/m^2}$, power P

[1] James Clerk Maxwell, mathematician and physicist, ⋆ Edinburgh 13.6.1831, † Cambridge 5.11.1879. Professor in Cambridge, UK

[2] Max Planck, physicist, ⋆ Kiel 23.4.1858, † Göttingen 4.10.1947. Professor in Kiel and Berlin. Nobel prize in physics 1918

[3] Louis Victor, 7. Duke of Broglie (since 1960), named Louis de Broglie (the family name is pronounced [dəˈbʁɔj], but the town Broglie is prounounced [ˈbʁɔgli]), physicist, ⋆ Dieppe 15.8.1892, † Louveciennes (Département Yvelines) 19.3.1987. Nobel prize in physics 1929 (together with O. W. Richardson)

[4] Albert Einstein, physicist, ⋆ Ulm 14.3.1879, † Princeton (NJ) 18.4.1955. "Technical expert 3rd class" at the patent office in Bern (1902–09). Professor at the University of Zurich and Prague (1911/12) and at the Swiss Federal Institute of Technology (ETH) in Zurich. Emigration to the USA in 1933. Professor at the Institute for Advanced Study in Princeton (NJ). American citizen since 1940. Formulated in 1905 (1914–16) the special (general) theory of relativity. Nobel prize in physics 1921

[5] Jahns, J.: Photonik. Grundlagen, Komponenten und Systeme. München: Oldenbourg-Verlag 2001. Page 9

[6] Pseudonym or pen-name of François Marie Arouet, philosopher and writer, ⋆ Paris 21.11.1694, † Paris 30.5.1778

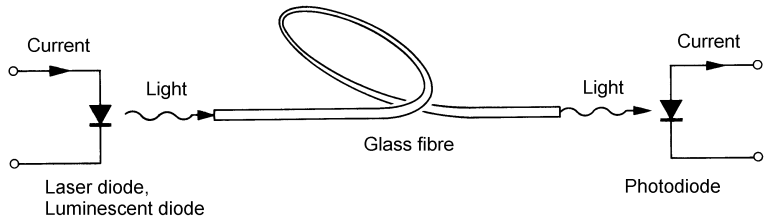


Fig. 1.1. Optical point-to-point transmission link with an intensity-modulated carrier centered at a wavelength λ and direct (incoherent) detection

per area F) as a function of time t. The classical power results from an average over a few optical cycles. More advanced modulation formats will be introduced at a later point of time.

The light is transported through a dielectric *l*ight *w*ave*g*uide (LWG), consisting of a low-refractive index cladding and a high-index core, which confines and guides the light in a cross-sectional area F. For long-distance communication, optical quartz glass fibres are used. Glass-based fibre waveguides are very thin, immune to electromagnetic interference, have low loss and guide the light over thousands of kilometers. In special cases, also free-space optical communication may be considered, for instance between a satellite and an Earth terminal. The dielectric waveguide or simply air represent the transmission *channel*, which we understand as "the medium used to transmit the signal from transmitter to receiver", following the definition by Shannon[7] in his seminal paper[8].

At the end of the channel, a *receiver* (Rx) evaluates the transmitted signal. In the simplest form of a Rx, a *p*hoto*d*etector (PD) with cross-section F and sensitivity S (unit A / W, also named responsivity) reconverts light with power $P = FI$ and photon energy hf to an electrical photocurrent i,

$$i(t) = SP(t), \qquad S = \frac{\eta e}{hf}, \qquad \frac{S}{\mathrm{A/W}} = \eta\,\frac{\lambda/\,\mu\mathrm{m}}{1.24} = 0.806 \times \eta\,\frac{\lambda}{\mu\mathrm{m}}\,, \qquad \frac{i}{e} = \eta\,\frac{P}{hf}\,. \tag{1.1}$$

The relation can be physically interpreted by observing that $i/\,e$ is the rate (unit: 1/ s) of photo-generated electrons, and $P/\,(hf)$ the rate (unit: 1/ s) of incident photons. Equation (1.1) then tells us that the number of electrons generated per time equals the number of incident photons per time reduced by the factor of the quantum efficiency $\eta \leq 1$, because on average a photon produces an electron only with probability η. The more the wavelength increases, i. e., the smaller the photon energy is, the larger the sensitivity S (and the photocurrent i) becomes[9], because for a constant optical power P more photons are available for generating electrons.

This very straightforward type of reception in Fig. 1.1 is called "direct" or incoherent, as opposed to coherent reception, where so-called heterodyne, intradyne or homodyne mixing with a *l*ocal laser *o*scillator (LO) is employed.

1.2.1 Modulation

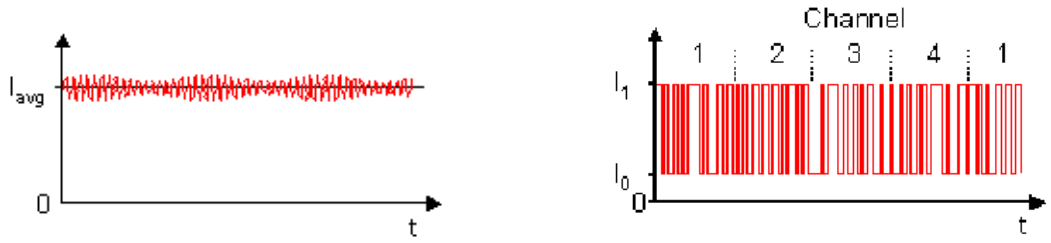
For encoding the signal information, the transmitted light must be altered ("modulated") in some way. The physical quantity to be modulated could be the frequency (as in *f*requency *m*odulation or FM broadcast), the phase, the electric field amplitude (as in *a*mplitude *m*odulation or AM broadcast), the polarization of the optical field, or, most simply, the optical intensity I (*i*ntensity *m*odulation IM) as depicted in Fig. 1.1. There are two major methods of modulation — analogue and digital.

Analogue modulation, Fig.1.2(a), uses less bandwidth and is simpler than digital modulation, which however provides a better signal quality at the expense of larger bandwidth requirements and a more

[7] Claude Elwood Shannon, engineer and physicist, ⋆ Gaylord (Michigan) 30.4.1916, † Medford (Massachusetts) 24.2.2001. Seminal papers on information theory in 1948. Professor at Massachusetts Institute of Technology since 1956

[8] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, 623–656

[9] For $hf = 1\,\mathrm{eV}$ ($f = 242\,\mathrm{THz}$, $\lambda = 1.24\,\mu\mathrm{m}$) and a quantum efficiency $\eta = 1$ the sensitivity amounts to $S = 1\,\mathrm{A/W}$.



(a) Analogue intensity modulation (b) Digital intensity modulation with TDM

Fig. 1.2. Modulation formats (a) Analogue intensity modulation around an operating point I_{avg} (b) Digital intensity modulation between an off (I_0) and an on value (I_1). For a 4-channel *t*ime *d*ivision *m*ultiplexing scheme (TDM) individual transmission time slots $1 \ldots 4$ are assigned to each data source

complicated circuitry. Present-day communications are controlled and initiated by digital computers, so it is natural to use a digital modulation format for transmission.

However, due to the strong increase of Internet traffic, bandwidth limitations of the transmitting fibre have become important. Therefore more advanced modulation schemes were developed, where both amplitude and phase of the optical field are modulated such that they take a number of discrete levels. This type of modulation has been named *q*uadrature *a*mplitude *m*odulation (QAM). Because the transmitted symbols are not just binary, more information per symbol can be transmitted without additional bandwidth requirements. As a disadvantage, the signal quality must be significantly better for an error-free discrimination between the various amplitude and phase levels.

The most common digital modulation format is pulse code modulation (PCM). Here, the value of an analogue signal $v(t)$ is sampled, and the values are then converted into a binary code. If the signal has a maximum bandwidth B (unit Hz) then sampling the analogue signal at equidistant time increments $1/(2B)$ with the so-called Nyquist[10] sampling rate $2B$ allows an exact reconstruction of the analogue signal from its samples, if these samples are properly interpolated. This is known as Nyquist-Shannon's sampling theorem[11,12,13,14].

The sampled values (e. g., the numbers $1, 4, 2, 5, 9 \ldots$) are then converted into a form suitable for transmission. For a binary format only two states are physically discriminated, light "off" or I_0 in Fig. 1.2(b) corresponding to a logical "0", and light "on" or I_1 corresponding to a logical "1". The decimal numbers $1, 4, 2, 5, 9$ would first be converted into binary numbers 0001, 0100, 0010, 1001, and then transmitted as temporal sequences (I_1, I_0, I_0, I_0), (I_0, I_0, I_1, I_0), (I_0, I_1, I_0, I_0), (I_1, I_0, I_0, I_1) of low and high optical intensities by switching the control current of the laser diode in Fig. 1.1 on and off. After transmission, an optical receiver converts the light impulses back into an electrical signal of low (i_0) and high currents (i_1). Finally, a digital-to-analogue converter reconstructs the original signal $v(t)$.

[10] Harry Nyquist (correct Swedish pronounciation is ['nʏkvɪst], not ['naɪkwɪst]), physicist and electrical and communications engineer, ⋆ Nilsby (Sweden) 7.2.1889, † Harlingen (Texas) 4.4.1976, a prolific inventor who made fundamental theoretical and practical contributions to telecommunications. — Nyquist moved to the United States in 1907. He earned a B. S. (1914) and an M. S. (1915) in electrical engineering from the University of North Dakota. In 1917, after earning a Ph. D. in physics from Yale University, he joined the American Telephone and Telegraph Company (AT&T). There he remained until his retirement in 1954, working in the research department and then (from 1934) at Bell Laboratories. Nyquist continued to serve as a government consultant on military communications well after his retirement.

His 1928 paper "Certain topics in telegraph transmission theory" refined his earlier results and established the principles of sampling continuous signals to convert them to digital signals. The Nyquist sampling theorem showed that the sampling rate must be at least twice the highest frequency present in the signal in order to reconstruct the original signal.

[11] H. Nyquist: Certain factors affecting telegraph speed. Bell Syst. Tech. J. 3 (1924) 324–346

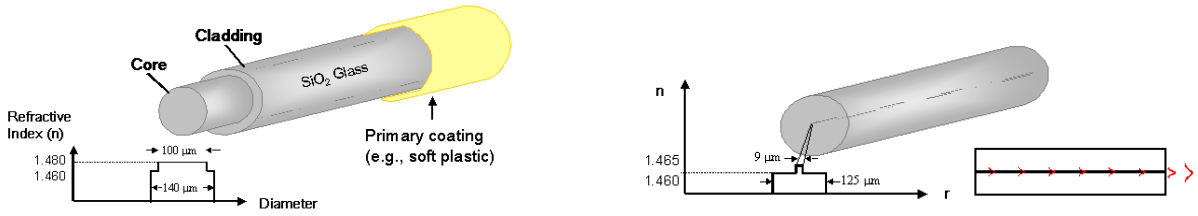
[12] H. Nyquist: Certain topics in telegraph transmission theory. Trans. Am. Inst. Electrical Engineers 47 (1928) 617–644. http://dx.doi.org/10.1109/T-AIEE.1928.5055024

[13] The papers from 1924 and 1928 by Nyquist are cited in Claude Shannon's classic 1948 essay (see Footnote 8 on Page 2), where Nyquist's seminal role in the development of information theory is acknowledged.

[14] C. E. Shannon: Communication in the presence of noise. Proc. IRE 37 (1949) 10–21. Reprinted in: Proc. IEEE 86 (1998) 447–457

1.2.2 Fibres

There are two types of fibers, multimode fibres with typical core diameters of 50 μm, 65 μm, 100 μm, 200 μm, 1 000 μm and 3 000 μm, Fig. 1.3(a), and single-mode fibres with a core diameter of about 9 μm, Fig. 1.3(b). For the multimode fibre, coupling light from the transmitter into the fiber core is easier



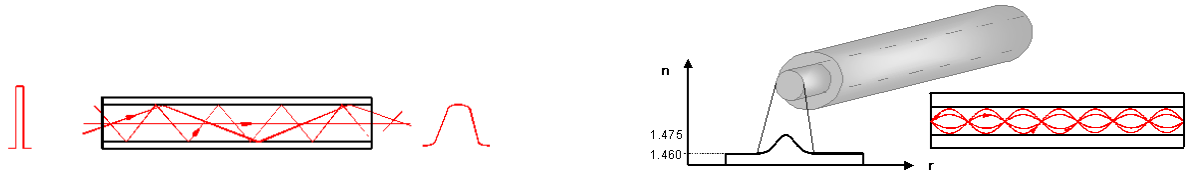
(a) Multimode fibre with step-index profile

(b) Single-mode fibre with step-index profile

Fig. 1.3. Fibre types with step-shaped refractive index profile comprising a higher-index core and a lower-index cladding (a) Fat-core step-index multimode fibre with a relative refractive index difference $\Delta \approx 1.3\,\%$ (b) Long-haul step-index single-mode communication fibre with $\Delta \approx 0.33\,\%$

than coupling light into the much smaller core of a single-mode fibre. The disadvantage is the stronger light impulse distortion of signals propagating in multimode fibres. A standard cladding diameter for single-mode communication fibres is 125 μm.

Intermodal dispersion Variation in propagation time among different modes creates intermodal dispersion, i. e., group delay differences, which are caused by optical path differences in a step-index multimode fibre. The effect on a light impulse entering the multimode fibre is shown in Fig. 1.4(a). The output impulse is broadened because it is composed of many smaller impulses arriving at different instances of time. If the group velocity in the outer-core regions could be increased, the group delay of these rays following longer geometrical paths could be made the same as for rays propagating on shorter geometrical paths. This is achieved by gradually reducing the refractive index away from the fibre axis. Such a waveguide is dubbed a graded-index fibre.



(a) Multimode fibre with step-index profile

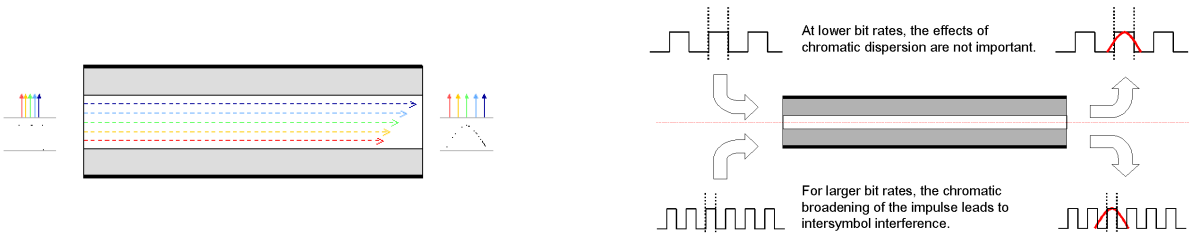
(b) Multimode fibre with graded-index profile

Fig. 1.4. Intermodal dispersion for multimode fibres. (a) Step-index profile with significant group delay differences (b) Graded-index profile, where geometrical path length differences are compensated by radial variations in the refractive index

Chromatic dispersion Dispersion in an optical fibre is not limited to intermodal dispersion. Even a single-mode fibre suffers from different group delays depending on the spectral content of the optical signal. This again leads to output impulse broadening or intramodal dispersion, Fig. 1.5(a). At higher bit rates, the broadened impulses spill into neighbouring time slots (intersymbol interference), and it becomes increasingly difficult to decide between a logical "0" and a logical "1", Fig. 1.5(b). Therefore, the bit error probability BER (bit error *ratio*; frequently, but wrongly named "bit error *rate*") increases, such limiting the maximum transmission rate.

Transmission bands

Various transmission bands are designated with the following letters: Extended *s*hort-wavelength band "S+", *s*hort-wavelength band "S", *c*onventional or *c*entral band "C", *l*ong-wavelength band "L", extended



(a) Chromatic dispersion in a single-mode fibre

(b) Group delay dispersion and intersymbol interference

Fig. 1.5. Group delay dispersion and bit error probability (bit error rate, BER). (a) Different wavelengths ("colours", therefore "chromatic") inside the same mode propagate with different velocities, thereby increasing the output impulse width (b) Broadening of the transmitted impulse leads to bit detection errors

*l*ong-wavelength band "L+". The attributed wavelengths (unit μm) are seen from Table 1.1. A typical DWDM ITU-T channel grid[15] for the C band is specified in Table 1.2.

Designation of 40-nm bands ($\lambda/\,\mu$m) at $\lambda = 1.550\,\mu$m										
S+		**S**		**C**		**L**		**L+**		
1.450	**1.470**	1.490	**1.510**	1.530	**1.550**	1.570	**1.590**	1.610	**1.630**	1.650

Table 1.1. Designation of bands at $\lambda = 1.550\,\mu$m

Wavelength table for the C band (DWDM ITU-T grid)					
$\lambda_{\mathrm{ITU}}/$ nm	$\lambda_{\mathrm{ITU}}/$ nm	$\lambda_{\mathrm{ITU}}/$ nm	$\lambda_{\mathrm{ITU}}/$ nm	$\lambda_{\mathrm{ITU}}/$ nm	$\lambda_{\mathrm{ITU}}/$ nm
1 527.99	1 534.25	1 540.56	1 546.92	1 553.33	1 559.78
1 528.77	1 535.04	1 541.35	1 547.72	1 554.13	1 560.61
1 529.55	1 535.82	1 542.14	1 548.51	1 554.94	1 561.42
1 530.33	1 536.61	1 542.94	1 549.32	1 555.75	1 562.23
1 531.12	1 537.40	1 543.73	1 550.12	1 556.55	1 563.05
1 531.90	1 538.19	1 544.53	1 550.92	1 557.36	$\Delta = 0.79$
1 532.68	1 538.98	1 545.32	1 551.72	1 558.17	**all:**
1 533.47	1 539.77	1 546.12	1 552.52	1 558.98	$\pm$**0.1**

Table 1.2. DWDM ITU-T grid at $\lambda = 1.550\,\mu$m. Channel spacing corresponds to frequency grid $\Delta f = 100\,$GHz

Attenuation

Losses in optical fibres arise through scattering and absorption. Therefore, the power of a guided wave decreases in z-direction from its initial value P_0 according to

$$P(z) = P_0\,\mathrm{e}^{-\alpha z}, \qquad a = 10\lg\frac{P_0}{P(z)} = \alpha z \times 10\lg\mathrm{e} = 4.34 \times \alpha z. \tag{1.2}$$

[15] Fujitsu: Lightwave Components & Modules Databook. (1998) p. 38. DFB lasers are commercially available with these wavelength gradings.

The power attenuation constant α (unit km^{-1}) is usually expressed by specifying the attenuation[16] a ("unit" dB).

1.2.3 Wavelength division multiplexing

The capacity of transmission links can be greatly extended by employing more than one optical carrier in a *w*avelength *d*ivision *m*ultiplexing scheme (WDM)[17,18], Figure 1.6. A number N of laser diodes are modulated in intensity and emit light at wavelengths $\lambda_1, \lambda_2, \ldots, \lambda_i, \ldots, \lambda_N$ near $\lambda = 1.55\,\mu\mathrm{m}$. An information channel when realized with a modulated optical carrier is termed an "optical channel" or a "wavelength channel", and referred to as "channel" for brevity. In addition to WDM, channel multiplexing can be done in the time domain, leading to (optical) time-division multiplexing (OTDM).

For a WDM scheme, the optical carriers are separated by, e. g., 25 GHz, 50 GHz, 100 GHz ($\Delta\lambda = 0.78\,\mathrm{nm}$ @ $\lambda = 1.55\,\mu\mathrm{m}$) or 200 GHz (e. g., 100 channels at $\lambda = 1.55\,\mu\mathrm{m}$ and 40 Gbit/s are commercially available, see Chapter 4). Thus, the total capacity of one single-mode fibre amounts to 4 Tbit/s if the channels are separated in frequency by, e. g., $\Delta f = 100\,\mathrm{GHz}$. An optical *mu*ltiple*x*er (MUX) spatially concentrates these modulated carriers to propagate as wavelength channels in one single fibre. An *o*ptical *a*mplifier (OA) in combination with a *equal*izer (EQUAL) for equalizing the gain in all channels amplifies the signals, which are then transmitted through the single-mode transport fibre. Optical am-

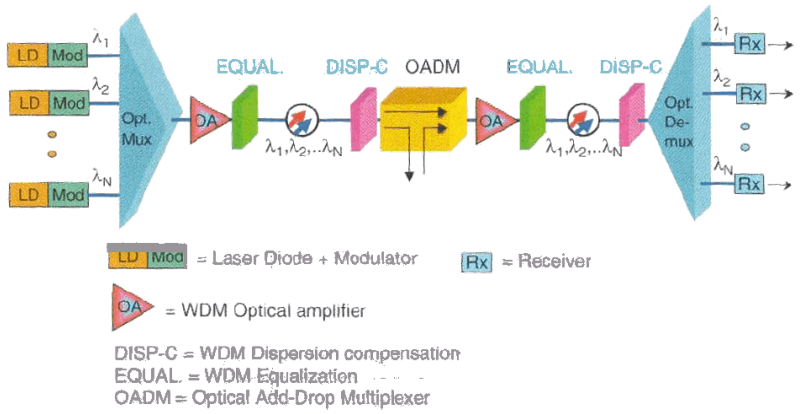


Fig. 1.6. Wavelength division multiplexing transmission scheme. The path from LD MOD(λ_i) to Rx(λ_i) corresponds to the simplified point-to-point transmission depicted in Fig. 1.1. [after Reference 18 on Page 6 (Fig. iii on Page xxiv)]

plifiers overcome the power loss in very long communication links. They have bandwidths in the order of $\Delta f = 5 \ldots 10\,\mathrm{THz}$ centred at $\lambda = 1.3\,\mu\mathrm{m}$ and $\lambda = 1.55\,\mu\mathrm{m}$, and remove the speed bottleneck from electronics by optics implementation. There are two primary types of OA, *s*emiconductor *o*ptical *a*mplifiers (SOA), and *d*oped *f*ibre *a*mplifiers (DFA). Among all DFA, Er^{3+}-*d*oped *f*ibre *a*mplifiers (EDFA) that amplify light around $\lambda = 1.55\,\mu\mathrm{m}$ are the most mature.

A *disp*ersion *c*ompensator (DISP-C) compensates the wavelength-dependent delay times inside each channel. At the end of the first span, an *o*ptical *a*dd-*d*rop *m*ultiplexer (OADM) adds or drops selectively certain wavelength channels. Then, after a possibly repeated sequence of such spans, an optical *demu*ltiple*x*er (DEMUX) finally separates all optical channels spatially, and optical receivers (Rx) attached to the DEMUX outputs receive the optical signals. The path from LD MOD(λ_i) to Rx(λ_i) corresponds to the simplified point-to-point transmission depicted in Fig. 1.1.

[16] Instead of choosing a new symbol a for the attenuation one could also write $\alpha_{\mathrm{dB}} = 10\lg(P_0/P(z)) = \alpha z \times 10\lg\mathrm{e}$, and specify α_{dB}/z in units of dB/km. However, giving the quantity α in units of dB/km would be misleading, because it implies the nonsensical expression $\alpha/z = \alpha \times 10\lg\mathrm{e}$.

[17] Agrawal, G. P.: Fiber-optic communication system. Chichester: John Wiley & Sons 1997. Chapter 7 p. 284

[18] Kartalopoulos, S. V.: DWDM — Networks, devices, and technology. John Wiley & Sons 2003

1.2.4 Advantages and shortcomings of optical communications

Data transmission capacity

Obviously, optical communication systems can replace conventional electrical systems only, if there is some advantage to be gained, which justifies the additional expenses of a twofold conversion current-light and light-current. Some important advantages of optical signal transport are:

- Large transmission capacity because of high carrier frequency near $f_O = 200\,\mathrm{THz}$, large fibre bandwidth in the order of $(250 \ldots 190)\,\mathrm{THz} = 60\,\mathrm{THz}$
- Low fibre loss, about $2.2, 0.35, 0.15\,\mathrm{dB}\,/\,\mathrm{km}$ at $\lambda = 0.85, 1.3, 1.55\,\mu\mathrm{m}$, i. e., down to 3 dB loss for a fibre length of $L = 20\,\mathrm{km}$ corresponding to a power attenuation by a factor of only 2
- Immunity to interference because of the high carrier frequency, and because of the strong confinement of the light inside the fibre

Three milestones of lightwave technology are especially noteworthy. Following an earlier suggestion,[19,20] the first *low-loss fibres* were produced[21] in 1970 reducing the loss from $1\,000\,\mathrm{dB}\,/\,\mathrm{km}$ to below $20\,\mathrm{dB}\,/\,\mathrm{km}$. Further progress[22] resulted by 1979 in a loss of only $0.2\,\mathrm{dB}\,/\,\mathrm{km}$ near $\lambda = 1.55\,\mu\mathrm{m}$. The ultimate low loss[23] of $0.154\,\mathrm{dB}\,/\,\mathrm{km}$ for fibres with a silica (SiO_2) core and a F-doped cladding is limited only by the amorphous structure of silica (Rayleigh scattering) and was reached in 1986.

Although *semiconductor lasers* were first made[24] in 1962, their use became practical only after 1970 when GaAs lasers operating continuously at room temperature were available[25].

Finally, it was only after the invention and perfection of the *Er-doped fibre amplifier*[26] (EDFA) in 1986 that optical communication became so powerful as it is today.

In recent years, with the increasing demand in transmission capacity, new frontiers have opened by the re-invention of *coherent optical communications*, which had formerly been regarded as too complicated and as obsolete in view of the availability of EDFA.

Reception sensitivity

Optical communications has also shortcomings as compared to electrical transmission. Electrical reception is limited by thermal noise with a power $P_v = kT_0B$ (Boltzmann's constant k, room temperature $T_0 = 293\,\mathrm{K}$, signal bandwidth B), see Eq. (5.54) on Page 124. Optical systems, however, are limited by quantum noise with an equivalent noise power $P_{\mathrm{qu}} = 2hf_OB$ (Planck's constant h, optical carrier frequency f_O), see Eq. (5.81) on Page 132. With the received electrical and optical signal powers P_{el} and P_{op}, we find the respective signal-to-noise power ratios $\mathrm{SNR}_{\mathrm{el}}$ and $\mathrm{SNR}_{\mathrm{op}}$. For equal SNR we see that electrical reception is by far more sensitive ($kT_0 = 25\,\mathrm{meV}$, $hf_O = 1\,\mathrm{eV}$ at $f_O = 242\,\mathrm{THz}$, $\lambda_O = 1.24\,\mu\mathrm{m}$),

$$\mathrm{SNR}_{\mathrm{el}} = \frac{P_{\mathrm{el}}}{kT_0B}\,, \quad \mathrm{SNR}_{\mathrm{op}} = \frac{P_{\mathrm{op}}}{2hf_OB}\,, \quad \frac{P_{\mathrm{el}}}{P_{\mathrm{op}}} = \frac{kT_0}{2hf_O} \ll 1\,, \quad \left.\frac{P_{\mathrm{el}}}{P_{\mathrm{op}}}\right|_{\mathrm{dB}} \approx 10\lg\frac{25\,\mathrm{meV}}{2\,\mathrm{eV}} = -19\,\mathrm{dB}\ . \quad (1.3)$$

Transmission span

Practical spans without amplification are about $L = 70\,\mathrm{km}$, because attenuation in the transmitting fibre causes the power of the optical signal to decay exponentially with the transmission distance L according

[19] Kao, K. C.; Hockham, G. A.: Proc. IEE 113 (1966) 1151
[20] Werts, A.: Onde Electr. 45 (1966) 967
[21] Kapron, F. P.; Keck, D. .B; Maurer, A. D.: Appl. Phys. Lett. 17 (1970) 423
[22] Miya, T.; Terunuma, Y.; Hosaka, T.; Miyashita, T.: Ultimate low-loss single-mode fibre at 1.55 μm. Electron. Lett. 15 (1979) 106–108
[23] Kanamori, H.; Yokota, H.; Tanaka, G.; Watanabe, M.; Ishiguro, Y.; Yoshida, I.; Kakii, T.; Itoh, S.; Asano, Y.; Tanaka, S.: Transmission characteristics and reliability of pure-silica-core single-mode fibers. IEEE J. Lightw. Technol. LT-4 (1986) 1144–1149
[24] Nasledov, D. N.; Rogachev, A. A.; Ryvkin, S. M.; Tsarenkov, B. V.: Fiz. Tverd. Tela. 4 (1962) 1062 (Soviet Phys. Solid State 4 (1962) 782
[25] Alferov, Z.: IEEE Sel. Topics Quantum Electron. 6 (2000) 832
[26] Poole, S. B.; Payne, D. N.; Mears, R. J.; Fermann, M. E.; Laming, R. E.: J. Lightw. Technol. 4 (1986) 870

to $\exp(-\alpha L)$, see Eq. (1.2) on Page 5. This transmission span seems to be astonishingly small. To explain why this is so, we discuss a practical example[27].

A transatlantic transmission from New York to London experiences an attenuation of about 1 400 dB (7 000 km @ 0.2 dB / km). Thus, for receiving one photon in London we have to inject 10^{140} photons into the optical fibre end in New York. If all the mass of our sun ($m_{\mathrm{sun}} = 3 \times 10^{33}$ g) having an energy equivalent of $W_{\mathrm{sun}} = mc^2 = 1.8 \times 10^{47}$ Ws could be converted into photons with a photon energy $hf = 6 \times 10^{-34}\,\mathrm{Ws}^2 \times 200\,\mathrm{THz} = 1.2 \times 10^{-19}$ Ws, we had generated 1.5×10^{66} photons at a wavelength of 1.55 μm ($f \approx 200$ THz), and could bridge a span with 660 dB loss, corresponding to a transmission distance of 3 300 km only. For a direct transmission New York – London we thus had to evaporate $10^{140}/10^{66} = 10^{74}$ suns.

This is quite a bit. The (observable) universe is estimated to have an extension of 1.4×10^{10} light years. Its mean density[28] is supposed to be $3 \times 10^{-30}\,\mathrm{g/cm^3}$. Therefore, the universe's mass (comprising not only suns) is $m_{\mathrm{univ}} = 7 \times 10^{54}$ g, and its energy equivalent is $W_{\mathrm{univ}} = m_{\mathrm{univ}}c^2 = 6 \times 10^{68}$ Ws corresponding to 4.7×10^{87} photons at a wavelength of 1.55 μm. If we are able to receive from these 4.7×10^{87} photons at least one photon, then the maximum span will be 877 dB /(0.2 dB / km) = 4 385 km. As a consequence, for bridging the distance New York – London in one go, we had to burn $10^{140}/10^{87} = 10^{53}$ universes!

This sounds absurd, because such an enormous explosion in New York could be definitely seen in London (at least for a *very* short moment). However, for wireless transmission in free space a different attenuation law holds. For an isotropic antenna, spherical waves are radiated having constant power on a phase surface. Therefore the intensity decreases only in proportion to the square of the transmission distance L^{-2}, and not according to $\mathrm{e}^{-\alpha L}$. Consequently, there is a break-even transmission distance L_{BE}: For $L < L_{\mathrm{BE}}$, *guided waves* propagate with smaller loss, while for $L > L_{\mathrm{BE}}$, *free-space propagation* has the longer reach.

For optical fibre transmission we therefore keep the spans short enough, and compensate the unavoidable loss with optical amplifiers. Common types are erbium-doped fibre amplifiers (EDFA) with an average output saturation power of about (20 ... 30) dBm, or linear semiconductor optical amplifiers (SOA) that are peak-power limited and saturate at output powers of about (0 ... 3) dBm. For wireless transmission, amplification would be also possible by using terrestrial relay stations, but it is naturally excluded for deep-space communication. Still, fibre communication has the advantage of extremely high carrier frequencies (e. g., 200 THz) which allow using a very broad spectrum for data transmission (e. g., 2 THz). The relative transmission bandwidth, however, remains small (e. g., 1 %). With wireless carrier frequencies even as high as 200 GHz, a relative bandwidth of 1 % means an absolute data bandwidth of 2 GHz only, which is three orders of magnitude less. Therefore, terrestrial broadband data communication calls for photonics.

1.3 Mathematical definitions and relations

In Table 1.3 on Page 9 a number of mathematical definitions and relations as listed. They are used throughout the text and are referred to only where required for understanding. Note that temporal and spatial Fourier transforms and their inverse transforms assume the *positive* time dependency $\exp(\mathrm{j}\,\omega t)$ as is common in electrical engineering (ee). Physicists (phys) prefer using the symbol $\mathrm{i} = \sqrt{-1}$ for the imaginary unit and work with a *negative* time dependency $\exp(-\mathrm{i}\,\omega t)$. Naturally, both notations describe the same physical situation. The formulations are complex conjugate[29] to each other, as can be seen for the example of a plane wave propagating along the $+z$-direction,

$$\Psi_{\mathrm{ee}}(t,z) = A(t)\,\mathrm{e}^{\mathrm{j}(\omega t - \beta z)}, \quad \psi_{\mathrm{phys}}(z,t) = a(t)\,\mathrm{e}^{\mathrm{i}(\beta z - \omega t)}, \quad \Psi_{\mathrm{ee}}(t,z) = \psi^*_{\mathrm{phys}}(z,t)\,. \tag{1.4}$$

[27] Calculations stimulated by an oral presentation of N. J. Doran (S. K. Turitsyn, M. P. Fedoruk, N. J. Doran and W. Forysiak: Optical soliton transmission in fiber lines with short-scale dispersion management. 25th European Conference on Optical Communication (ECOC'99), Nice, France, September 26–30, 1999). — Universe's mass calculations and web address contributed by Dipl.-Phys. Jan Brückner, DFG Research Training Group 786 "Mixed Fields and Nonlinear Interactions", Karlsruhe University, Germany, June 23, 2005

[28] `http://curious.astro.cornell.edu/question.php?number=342`

[29] Translating physics notation to electrical engineering and vice versa: "Take the complex conjugate of all quantities."

Notation and formulae		
Time	t	(1)
Frequency f, wavelength λ, vacuum speed of light c	$f = \frac{c}{\lambda}\,, \qquad c = 2.997\,924\,58 \times 10^8\ \mathrm{m/s}$	(2)
Angular frequency ω, vacuum propagation constant k_0	$\omega = 2\pi f\,, \qquad k_0 = \frac{\omega}{c} = \frac{2\pi}{\lambda}$	(3)
Cartesian spatial coordin. & spatial (angular) frequencies	$x, y, z \quad \& \quad \xi, \eta, \zeta \quad (k_x = 2\pi\xi,\ k_y = 2\pi\eta,\ k_z = 2\pi\zeta)$	(4)
Imaginary unit, complex conjugate u^* of u	$\mathrm{j} = \sqrt{-1}\,, \qquad u = p + \mathrm{j}\,q\,, \quad u^* = p - \mathrm{j}\,q \quad (p, q \text{ real})$	(5)
Adding to u its complex conjugate u^*	$u + u^* = 2\Re\{u\}\,, \quad u - u^* = \mathrm{j}\,2\Im\{u\}$	(6)
Plane wave propagating in medium with refractive index n, vacuum speed of light c	$\exp[\mathrm{j}\,(\omega t - (k_x x + k_y y + k_z z))]\,,$ $k_x^2 + k_x^2 + k_z^2 = n^2 k_0^2 = n^2\,\omega^2 / c^2$	(7)
Plane wave propagating in $+z$-direction, propagation constant $\beta \geq 0$, and effective index n_e	$\exp[\mathrm{j}\,(\omega t - \beta z)]\,, \qquad n_e = \frac{\beta}{k_0}$	(8)
Kronecker symbol $\delta_{m\,m'},\ m, m' \in \mathbb{Z}$	$\delta_{m\,m'} = \begin{cases} 1 \text{ for } m = m' \\ 0 \text{ else} \end{cases}$	(9)
Dirac function $\delta(t)$	$\Psi(0) = \int_{-\infty}^{+\infty} \delta(t)\,\Psi(t)\,\mathrm{d}t, \quad \delta(t) = \lim_{k\to\infty} \int_{-k}^{+k} \mathrm{e}^{\pm \mathrm{j}\,2\pi f t}\,\mathrm{d}f$ $\delta(t) = 0 \text{ for } t \neq 0$	(10)
Heaviside function $H(t)$	$\int_0^{+\infty} \Psi(t)\,\mathrm{d}t = \int_{-\infty}^{+\infty} H(t)\,\Psi(t)\,\mathrm{d}t,$ $H(t) = \begin{cases} 1 \text{ for } t > 0 \\ 0 \text{ for } t < 0 \end{cases}$	(11)
rect-function $\mathrm{rect}\left(\frac{t}{T}\right)$	$\int_{-T/2}^{+T/2} \Psi(t)\,\mathrm{d}z = \int_{-\infty}^{+\infty} \mathrm{rect}\left(\frac{t}{T}\right)\Psi(t)\,\mathrm{d}t,$ $\mathrm{rect}\left(\frac{t}{T}\right) = \begin{cases} 1 \text{ for } \lvert t\rvert < T/2 \\ 0 \text{ for } \lvert t\rvert > T/2 \end{cases}$	(12)
sinc-function $\mathrm{sinc}\left(\frac{t}{T}\right)$	$\mathrm{sinc}\left(\frac{t}{T}\right) = \begin{cases} 1 & \text{for } \ t = 0 \\ \frac{\sin(\pi t/T)}{\pi t/T} & \text{else} \end{cases}$	(13)
triang-function $\mathrm{triang}\left(\frac{t}{T}\right) = \frac{1}{T}\,\mathrm{rect}\left(\frac{t}{T}\right) * \mathrm{rect}\left(\frac{t}{T}\right)$	$\mathrm{triang}\left(\frac{t}{T}\right) = \begin{cases} 1 - \lvert t\rvert/T & \text{for } \ \lvert t\rvert \leq T \\ 0 & \text{else} \end{cases}$	(14)
Continuous Fourier transform (FT, $\breve{\Psi}(f) = \mathcal{F}\{\Psi(t)\}$)	$\breve{\Psi}(f) = \int_{-\infty}^{+\infty} \Psi(t)\,\mathrm{e}^{-\mathrm{j}\,2\pi f t}\,\mathrm{d}t$, if $\Psi(t)$ real: $\breve{\Psi}(f) = \breve{\Psi}^*(-f)$	(15)
Continuous inverse FT (IFT, $\Psi(t) = \mathcal{F}^{-1}\{\breve{\Psi}(f)\}$)	$\Psi(t) = \int_{-\infty}^{+\infty} \breve{\Psi}(f)\,\mathrm{e}^{+\mathrm{j}\,2\pi f t}\,\mathrm{d}f$	(16)
Power spectrum $\Theta_\Psi(f) := \mathcal{F}\{\vartheta_\Psi(t)\}$ Autocorrelation function (ACF) $\vartheta_\Psi(t) := \mathcal{F}^{-1}\{\Theta_\Psi(f)\}$	$\Theta_\Psi(f) = \lvert\breve{\Psi}(f)\rvert^2$, one-sided power spectrum: $2\Theta_\Psi(f)$ for $f > 0$ and real $\Psi(t)$	(17)
Continuous spatial Fourier transform (SFT)	$\widetilde{\Psi}(\xi, \eta) = \iint_{-\infty}^{+\infty} \Psi(x, y) \exp[+\mathrm{j}\,(\xi x + \eta y)]\,\mathrm{d}x\,\mathrm{d}y$	(18)
Continuous spatial inverse FT (SIFT)	$\Psi(x, y) = \iint_{-\infty}^{+\infty} \widetilde{\Psi}(\xi, \eta) \exp[-\mathrm{j}\,(\xi x + \eta y)]\,\mathrm{d}\xi\,\mathrm{d}\eta$	(19)
FT of rect-function	$\int_{-\infty}^{+\infty} \mathrm{rect}\left(\frac{t}{T}\right)\mathrm{e}^{-\mathrm{j}\,2\pi f t}\,\mathrm{d}t = T\,\mathrm{sinc}\,(fT)$	(20)
FT of sinc-function	$\int_{-\infty}^{+\infty} \mathrm{sinc}\left(\frac{t}{T}\right)\mathrm{e}^{-\mathrm{j}\,2\pi f t}\,\mathrm{d}t = T\,\mathrm{rect}\,(fT)$	(21)
FT of triang-function	$\int_{-\infty}^{+\infty} \mathrm{triang}\left(\frac{t}{T}\right)\mathrm{e}^{-\mathrm{j}\,2\pi f t}\,\mathrm{d}t = T\,\mathrm{sinc}^2\,(fT)$	(22)
Inner product	$(\Psi_1 \cdot \Psi_2) \equiv \langle \Psi_1 \mid \Psi_2 \rangle = \int_{-\infty}^{+\infty} \Psi_1^*(t')\,\Psi_2(t')\,\mathrm{d}t'$	(23)
Convolution	$(\Psi_1 * \Psi_2)(t) := \int_{-\infty}^{+\infty} \Psi_1(t')\,\Psi_2(t - t')\,\mathrm{d}t'$ $= \int_{-\infty}^{+\infty} \breve{\Psi}_1(f)\,\breve{\Psi}_2(f)\,\mathrm{e}^{\mathrm{j}\,2\pi f t}\,\mathrm{d}f$	(24)
Cross-correlation function $\vartheta_{\Psi_1\Psi_2}(t) := (\Psi_1 \otimes \Psi_2)(t)$ Cross power spectrum $\Theta_{\Psi_1\Psi_2}(f) := \mathcal{F}\left\{\vartheta_{\Psi_1\Psi_2}(t)\right\}$	$(\Psi_1 \otimes \Psi_2)(t) := \int_{-\infty}^{+\infty} \Psi_1(t')\,\Psi_2^*(t' - t)\,\mathrm{d}t'$ $= \int_{-\infty}^{+\infty} \breve{\Psi}_1(f)\,\breve{\Psi}_2^*(f)\,\mathrm{e}^{\mathrm{j}\,2\pi f t}\,\mathrm{d}f$	(25)
$\cos x + \cos y \qquad \&\ \sin x + \sin y$	$2\cos\frac{x-y}{2}\cos\frac{x+y}{2} \qquad \&\ 2\cos\frac{x-y}{2}\sin\frac{x+y}{2}$	(26)
$\cos(x \pm y) \qquad \&\ \sin(x \pm y)$	$\cos x \cos y \mp \sin x \sin y \qquad \&\ \sin x \cos y \pm \cos x \sin y$	(27)
$\cos x \cos y\ (+),\ \sin x \sin y\ (-) \quad \&\ \sin x \cos y$	$\frac{1}{2}[\cos(x-y) \pm \cos(x+y)]\ \&\ \frac{1}{2}[\sin(x-y) + \sin(x+y)]$	(28)
$a\cos x + b\sin x = \Re\left\{\sqrt{a^2+b^2}\,\mathrm{e}^{-\mathrm{j}\arctan(b/a)}\,\mathrm{e}^{\mathrm{j}\,x}\right\}$	$\sqrt{a^2+b^2}\cos\left(x - \arctan\frac{b}{a}\right) = \sqrt{a^2+b^2}\sin\left(x + \arctan\frac{a}{b}\right)$	(29)
Logarithms and their bases, $\log_a x = \frac{\log_b x}{\log_b a}$	$\lg x = \log_{10} x, \quad \ln x = \log_{\mathrm{e}} x, \quad \mathrm{lb}\,x = \log_2 x$	(30)
Power and amplitude ratios a and $b = \sqrt{a}$ in dB	$a_{\mathrm{dB}} = 10\lg a = 20\lg b$	(31)

Table 1.3. Mathematical definitions and relations that are used throughout the text, usually without a specific reference

1.4 Content overview

In the following, transmitters and receivers are discussed as well as their function in feeding and sinking data streams to and from an optical channel. In this context, Chapter 2 explains some basic communications concepts. We treat sampling, conversion between the analogue and the digital domain, an abstraction of the optical channel, preliminary information on statistical signal perturbations (noise), Shannon's channel capacity, modulator concepts and modulation formats.

Transmitters are presented in Chapter 3. This includes light sources for *e*lectro-*o*ptic (EO) conversion as well as modulators. Very briefly we mention in Chapter 4 semiconductor optical amplifiers, and refer to doped fibre amplifiers.

The functioning of receivers is illustrated in Chapter 5. We start by explicating the properties of pin photodiodes as *o*pto-*e*lectronic (OE) converters, give details on the optical and on the electrical subsystems for incoherent and coherent reception, derive optical and electrical signal-to-noise power ratios without and with an optical pre-amplifier, and explain various metrics for signal quality.

Finally, in Chapter 6, we list a few transmission impairments, discuss concatenated optical amplifier links, and mention signal shaping.

Several appendices complete the text: Appendix A on linear and nonlinear fibre properties, Appendix B on sampling, quantizing and on the discrete Fourier transform, and Appendix C on the rectification of coherent carriers embedded in noise.

Chapter 2

Optical communication concepts

In the following, we present the basic principles of communication systems with a focus an digital communications. Particularly important are the physical characteristics of the channel through which the information is transmitted because the channel determines the properties of the basic building blocks which complete the communication system[1]. Figure 2.1 displays the schematic[2] of a communication system. The building blocks are described following largely the text in Proakis' book[3]. The source may be either an analog signal, such as an audio or video signal, or a digital signal, such as the output of a teletype machine that is discrete in time and has a finite number of output characters. In a digital communication system, the analogue messages produced by the source enter a signal conditioning unit,

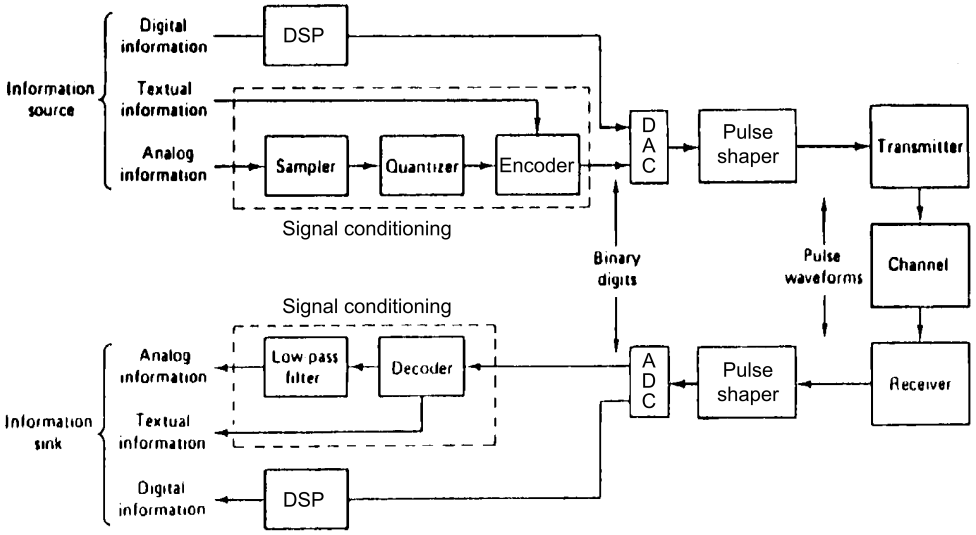


Fig. 2.1. Elements of a communications system. The block named "pulse shaper" is a filter that shapes the transmitted pulses in an appropriate fashion. At the receiver side, the corresponding block serves a similar purpose. The filter shapes the received pulses such that an optimum detection becomes possible. For instance, if by dispersion in the fibre channel the pulse has unduly broadened, this receiver filter can undo the broadening. Both filters can be part of the transmitter or the receiver, respectively, or they can be separated as in the figure. [Modified from Ref. 2 on Page 11. Slide 10]

[1] J. G. Proakis: Digital communications, 4th Ed. New York: McGraw-Hill 2001

[2] Wireless Information Transmission System Lab: Introduction to digital communications system. Institute of Communications Engineering, National Sun Yat-sen University, Taiwan.
`http://wits.ice.nsysu.edu.tw/course/pdfdownload/95_2%5C%E7%84%A1%E7%B7%9A%E9%80%9A%E8%A8%8A%E5%9F%BA%E9%A0%BB%E8%A8%8A%E8%99%9F%E8%99%95%E7%90%86%E8%88%87%E7%B3%BB%E7%B5%B1%E8%A8%AD%E8%A8%88/BB-03-DigitalComm.pdf`
This address looks weird: The symbols %E8 etc. represent Chinese characters.

[3] See Ref. 1 on Page 11, Sect. 1-1 and 1-2

where they are sampled and converted into a sequence of binary digits by the quantizer. Ideally, we should like to represent the source output (message) by as few binary digits as possible. In other words, we seek an efficient representation of the source output that results in little or no redundancy. The process of efficiently converting the output of either an analog or digital source into a sequence of binary digits is called source encoding.

The sequence of binary digits from the quantizer, which we call the information sequence, is passed to the encoder ("coder" for short). The purpose of the encoder is to introduce, in a controlled manner, some redundancy in the binary information sequence that can be used at the receiver to overcome the effects of noise and interference encountered in the transmission of the signal through the channel. In effect, redundancy in the information sequence aids the receiver in decoding the desired information sequence. For example, a (trivial) form of encoding of the binary information sequence is simply to repeat each binary digit m times, where m is some positive integer. More sophisticated (nontrivial) encoding involves taking k information bits at a time and mapping each k-bit sequence into a unique n-bit sequence, called a code word or symbol. The amount of redundancy introduced by encoding the data in this manner is measured by the encoder ratio $r_c = n/\,k$.

For originally digital information, the encoder would be part of a digital signal processing unit (DSP). The binary sequence at the output of the encoder or the DSP unit is passed to a digital-to-analogue converter (DAC) which provides a physical quantity (e. g., a voltage) corresponding to the binary symbol at its input. Finally, a pulse shaper serves as the interface to transmitter and communication channel.

The transmitter comprises a modulator that modulates the optical carrier with the physical representation of a symbol. In its simplest form the carrier could be just switched on and off (OOK, on-off keying). More complicated symbols could represent r coded information bits at a time by using $M = 2^r$ distinct waveforms $s_m\,(t)$, $m = 0, 1, 2, \ldots, M-1$, one waveform for each or the 2^r possible r-bit sequences. We call this M-ary modulation ($M > 2$). Hence, when the channel bit rate R_b is fixed, the amount of time available to transmit one of the M waveforms corresponding to a r-bit sequence (one "symbol") is r times the time period $1/\,R_b$ in a system that uses binary modulation, because a new r-bit sequence enters the modulator only after a symbol period $T_s = r/\,R_b$.

The communication channel is the physical medium that is used to send the signal from the transmitter to the receiver. In wireless transmission, the channel may be the atmosphere (free space). On the other hand, telephone channels usually employ a variety of physical media, including wire lines, optical fiber cables, and wireless (microwave radio). Whatever the physical medium used for transmission of the information, the essential feature is that the transmitted signal is corrupted in a random manner by a variety of possible mechanisms, such as additive thermal noise generated by electronic devices, man-made noise, e. g., automobile ignition noise, and atmospheric noise, e. g., electrical lightning discharges during thunderstorms.

As mentioned in Sect. 1.2.4 and Eq. (1.3) on Page 7, quantum phenomena play a critical role in optical communications, because the quantum energy $hf_{\text{opt}} = 1\,\text{eV}$ of an optical carrier photon at frequency $f_{\text{opt}} = 242\,\text{THz}$ ($\lambda_{\text{opt}} = 1.24\,\mu\text{m}$) is much larger than the thermal energy $kT_0 = 25\,\text{meV}$ at room temperature $T_0 = 293\,\text{K}$. Therefore, with a given power P, the granularity of the photon flux is much more noticeable at optical frequencies, and the optical signal power must be much larger than the electrical signal power for the same signal-to-noise power ratio at the receiver. As we had seen, electrical systems are significantly more sensitive than optical transmission systems, however, their limitations in bandwidth call for photonics.

At the receiving end of a digital communication system, the demodulator as part of the receiver processes the channel-corrupted transmitted waveform. After pulse shaping for optimal reception, the received waveform passes an analogue-to-digital converter (ADC). Its output enters either a DSP unit which does the required signal processing all-digitally, or it is input to the signal conditioning unit. This signal conditioning unit takes the sequence of numbers from the ADC and passes it to the channel decoder, which attempts to reconstruct the original information sequence from knowledge of the code used by the channel encoder and the redundancy contained in the received data.

A measure of how well the demodulator and decoder perform is the frequency with which errors occur in the decoded sequence. More precisely, the average probability of a bit error at the output of the

decoder is a measure of the performance of the demodulator-decoder combination. In general, the bit error probability (BER, bit error ratio) is a function of the code characteristics, the types of waveforms used to transmit the information over the channel, the transmitter power, the characteristics of the channel, i. e., the amount of noise, the nature of the interference, and the method of demodulation and decoding.

As a final step, when an analog output is desired, a low-pass filter interpolates the received sampled data for reconstructing the original message sent by the source. Due to unavoidable errors, the received message is an approximation to the originally sent message. The difference or some function of the difference between the original signal and the reconstructed signal is a measure of the distortion introduced by the digital communication system.

As discussed previously, optical fibres offer the communications system designer a channel bandwidth that is several orders of magnitude larger than coaxial cable channels. Optical fiber cables have been developed that have a very low signal attenuation, and highly reliable photonic devices are available for signal generation and signal detection.

The transmitter in a fiber optic communication system is a light source, mostly a semiconductor laser diode (LD), occasionally also a light-emitting diode (LED) for bridging short transmission distances. Information can transmitted most simply by modulating the intensity of the light source with the message, see Fig. 1.2 on Page 3. The resulting signal propagates through the fiber as a lightwave and is amplified periodically to compensate the fibre attenuation, see Sect. “Transmission span” on Page 7 ff. In the case of digital transmission, the signal can be also detected and regenerated by baseband repeaters at larger distances along the transmission path to undo any distortion. At the receiver, the light intensity is detected by a photodiode, whose output is an electrical signal that varies in direct proportion to the power of the light impinging on the photodiode, Eq. (1.1) on Page 2. Prominent sources of noise in fiber optic channels are light sources, optical amplifiers, photodiodes and electronic amplifiers.

2.1 Signal conditioning

Our focus lies on digital communication systems, so we need first to convert any analogue signal to the digital domain, Fig. 2.1 on Page 11. This is done in the signal conditioning unit by sampling, quantization, and coding. After transmission, the original analogue signals must be reconstructed by converting the received digital signals back to the analogue domain in another signal conditioning unit. Alternatively, if digital data only are to be transmitted and received, the digital signal processing (DSP) block provides all the required signal conditioning.

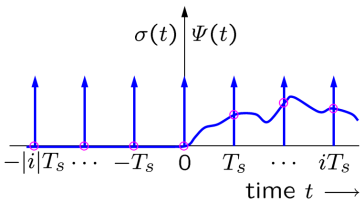
Appendix B on Page 183 ff. treats some important aspects of digital signal processing (DSP) like sampling with finite temporal bin sizes, quantization noise, effective number of bits, and properties of the discrete Fourier transform.

2.1.1 Sampling

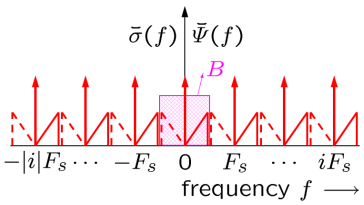
A real signal $\Psi(t)$ with a spectrum $\breve{\Psi}(f)$ that is limited to a bandwidth B can be reconstructed from samples $\Psi(iT_s)$ $(i = 0, \pm1, \pm2, \ldots)$, if $T_s = 1/F_s \leq 1/(2B)$ holds, i. e., if the sampling frequency F_s is as large or larger than the signal's bandwidth B. Real samples with a rate of at least $F_s = 2B$ have to be recorded. A sampled pulse amplitude modulated (PAM) signal $\Psi_s(t)$ results from multiplying the signal $\Psi(t)$ with the sampling function $\sigma(t)$. For the temporal functions $\Psi(t)$, $\sigma(t)$, $\Psi_s(t)$ and for their spectra $\breve{\Psi}(f)$, $\breve{\sigma}(f)$, $\breve{\Psi}_s(f)$ we find

$$\begin{aligned} \sigma(t) &= T_s \sum_{i=-\infty}^{+\infty} \delta(t - iT_s) = \sum_{i=-\infty}^{+\infty} \mathrm{e}^{\mathrm{j}\,2\pi i t/T_s}, \quad & \breve{\sigma}(f) &= \sum_{i=-\infty}^{+\infty} \delta(f - iF_s), \qquad & F_s &= \frac{1}{T_s}, \\ \Psi_s(t) &= \Psi(t)\,\sigma(t), & \breve{\Psi}_s(f) &= \breve{\Psi}(f) * \breve{\sigma}(f) = \sum_{i=-\infty}^{+\infty} \breve{\Psi}(f - iF_s). \end{aligned} \tag{2.1}$$

The sampling function $\sigma(t)$ is self-reciprocal in Fourier space. For calculating $\breve{\sigma}(f)$ one uses the identity $T_s \sum_{i=-\infty}^{+\infty} \exp(-\mathrm{j}\,2\pi f i T_s) = \sum_{i'=-\infty}^{+\infty} \delta(f - i'/T_s)$ resulting from a Fourier expansion of the periodically



(a) Sampling function $\sigma(t)$ and signal function $\Psi(t)$. Sampled values $\Psi_s(iT_s)$ indicated by circles (○)



(b) Fourier transform of sampling function $\breve{\sigma}(f)$ and baseband signal spectrum $\breve{\Psi}(f)$ with bandwidth B.

Fig. 2.2. Sampling of a real bandlimited signal. Arrows represent a comb of Dirac functions $\delta(t)$ and $\delta(f)$ in time and frequency domain, respectively. (a) Sampling function $\sigma(t)$ with period T_s and signal function $\Psi(t)$ with sampled values $\Psi(iT_s)$. (b) Fourier transform $\breve{\sigma}(f)$ of sampling function $\sigma(t)$ and signal spectrum $\breve{\Psi}(f)$, limited to the Nyquist bandwidth $B = F_s/2 = 1/(2T_s)$. This schematic baseband spectrum with upper (solid lines) and lower sidebands (broken lines) is periodically repeated at multiples of the sampling frequency F_s due to the sampling. For reconstructing the original signal $\Psi(t)$, the spectrum of the sampled time function must be filtered with a rectangular ("brick wall") filter to remove the so-called image spectra.

repeated function $\delta(f)$, which itself represents a δ-"comb". — Sampling with a finite window size is treated in Appendix B.1 on Page 183 ff.

Nyquist sampling The sampling process is illustrated in Fig. 2.2. The arrows represent combs of Dirac functions $\delta(t)$ and $\delta(f)$ in time and frequency domain, respectively. Figure 2.2(a) shows the sampling function $\sigma(t)$ with period T_s, the continuous real signal function $\Psi(t)$, and its sampled values $\Psi(iT_s)$. In Fig. 2.2(b) the self-reciprocal Fourier transform $\breve{\sigma}(f)$ of the sampling function is displayed. The signal spectrum $\breve{\Psi}(f)$ is assumed to be limited to the bandwidth B and sampled with the so-called Nyquist rate $F_s = 2B$. As a consequence of the sampling process (mixing of $\Psi(t)$ with $\sigma(t)$), the baseband spectrum $\breve{\Psi}(f)$ is periodically repeated. The repetitions are centred at multiples of the sampling frequency (rate) F_s. It is obvious that an inverse Fourier transform of the spectrum $\breve{\Psi}_s(f)$ of the sampled signal cannot reproduce the original continuous signal $\Psi(t)$. For a true reconstruction, $\breve{\Psi}_s(f)$ must be filtered with a rectangular ("brick wall") lowpass filter with a *one-sided width* B to remove the so-called image spectra.

Undersampling If the sampling frequency $F_s < 2B$ is smaller than the real signal's doubled bandwidth, adjacent upper and lower signal sidebands overlap. In this case, even an ideal lowpass with one-sided bandwidth B cannot prevent signal perturbations. Because the perturbation comes from spectra centred at "other" neighbouring positions, this is called an "aliasing error"[4].

Oversampling To prevent aliasing with non-ideal filters having finite filter slopes, the sampling frequency $F_s > 2B$ should be larger than the real signal's doubled bandwidth. Adjacent upper and lower sidebands are then separated by a spectral guard band of width $F_s - 2B$ that accommodates real-world filter slopes.

Reconstruction example for a real signal A real signal $\Psi(t)$ with bandwidth B can be reconstructed from its Nyquist-sampled data $\Psi(iT_s)$ by filtering with a brick wall filter having a transfer function $\breve{h}_{\mathrm{BW}}(f)$ and an impulse response $h_{\mathrm{BW}}(t)$, see Table 1.3 on Page 9,

$$\breve{h}_{\mathrm{BW}}(f) = \mathrm{rect}\left(\frac{f}{B}\right), \qquad h_{\mathrm{BW}}(t) = \frac{1}{T_s}\,\mathrm{sinc}\left(\frac{t}{T_s}\right), \qquad T_s = \frac{1}{2B} \quad \text{(real Nyquist sampling)}. \tag{2.2}$$

[4] The noun "alias" (pronounced ['eɪlɪəs]) denotes a false or assumed identity: "A spy operating under the alias H21" (H21 is better known by her stage name Mata Hari, a Dutch exotic dancer, courtesan, and convicted spy, who was executed by firing squad in France under charges of espionage for Germany during World War I.) — The noun "aliasing" means a misidentification of a signal frequency, introducing distortion or error.

ORIGIN late Middle English: from Latin *alias* (*sc. partes, acc. pl. f. of alius*), 'at another time, otherwise, else'

Because filtering means multiplying a signal spectrum with the filter's transfer function, we have to perform the operation $\breve{\Psi}_s(f)\,\breve{h}_{\mathrm{BW}}(f)$, and because this is equivalent to a convolution $\Psi_s(t) * h_{\mathrm{BW}}(t)$ in the time domain, see Table 1.3, we find from combining Eq. (2.1) and (2.2) the following recipe to interpolate between the known sampling points:

$$\Psi(t) = \int_{-\infty}^{+\infty} \Psi_s(t')\, h_{\mathrm{BW}}(t'-t)\,\mathrm{d}t' = \sum_{i=-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Psi(t')\,\delta(t'-iT_s)\,\mathrm{sinc}\left(\frac{t'-t}{T_s}\right)\mathrm{d}t'$$
$$= \sum_{i=-\infty}^{+\infty} \Psi(iT_s)\,\frac{\sin(\pi(t/T_s - i))}{\pi(t/T_s - i)} = \sum_{i=-\infty}^{+\infty} \Psi(iT_s)\,\mathrm{sinc}(t/T_s - i)\ . \tag{2.3}$$

With Nyquist sampling, all sinc-functions in Eq. (2.3) but one are zero at the sampling points iT_s. As can be easily seen, any other interpolation procedure, e. g., a linear interpolation, would not suffice for reconstructing $\Psi(t)$. If the same signal as before was oversampled by a factor $q > 1$ so that the samples are positioned at times $t = iT_s/q$ instead of $t = iT_s$, then all occurrences of i in the arguments of the sum Eq. (2.3) have to be replaced by i/q.

More on in-between zero padding, on end zero padding, and on interpolation can be found in Appendix B.3 on Page 193 ff.

Complex signals For a complex signal, each sampling point contains double the information compared to a real signal with equal information content. Upper and lower sidebands in Fig. 2.2(b) are no longer correlated, and the bandwidth B now describes the *full width* of the shaded area. The period for simultaneous Nyquist sampling of real and imaginary part doubles compared to Eq. (2.2) for real samples,

$$T_s = \frac{1}{B}\,, \qquad F_s = B \quad \text{(complex Nyquist sampling)}. \tag{2.4}$$

2.1.2 Quantization and coding

After the sampling process, the resulting time-discrete signal still covers a continuum of possible values $\Psi(iT_s)$. However, from a signal integrity point of view it is advisable to transmit a digital representation of the signal, i. e., a set of numbers representing a finite count of so-called quantization levels. For an average electrical signal power P_S the effective signal amplitude is $\sqrt{P_S}$, thus representing the span of the samples $\Psi(iT_s)$. However, there is also uncorrelated electrical noise[5] with power P_R and an effective noise amplitude $\sqrt{P_R}$.

Quantization For quantization, the effective signal span $\sqrt{P_S}$ has to be covered with a number of M discrete quantization levels (not necessarily equidistantly spaced) such that one can assign one out of M discrete values to each sampling point. Clearly, this procedure leads to additional inaccuracies named quantization noise[6], see Appendix B.2.3 on Page 192 ff., Eq. (B.42b). The logarithmic signal-to-noise power ratio due to quantizing a sinusoidal signal by an analogue-to-digital converter (ADC) with r bit and $M = 2^r \gg 1$ levels is approximately, according to Appendix B.2.3, Eq. (B.43) on Page 193,

$$\mathrm{SNR}_{q,\mathrm{dB}}^{(\sin)} = 6.02\,r + 1.76\,, \qquad r_e = \mathrm{ENOB} = \frac{\mathrm{SNDR}_{q,\mathrm{dB}}}{6.02} - 0.293\,. \tag{2.5}$$

The presence of noise decreases the effective number of bits (ENOB) from the physical value r to a smaller effective value $r_e = \mathrm{ENOB}$. Again, Eq. (2.5) can be used, if $\mathrm{SNR}_{q,\mathrm{dB}}^{(\sin)}$ is now interpreted as the signal-to-noise-and-distortion power ratio $\mathrm{SNDR}_{q,\mathrm{dB}}$ at the input of the ADC, and r is replaced by $r_e = \mathrm{ENOB}$, see Appendix B, Eq. (B.44) on Page 193. Figure 2.3 (lower row) shows that increasing the physical number

[5] Subscript "R" from German "Rauschen" (noise), in conformity with already existing texts and figures, which otherwise would have to be re-designed.

[6] Valley, C. G.: Photonic analog-to-digital converters. Opt. Express 15 (2007) 1955–1982

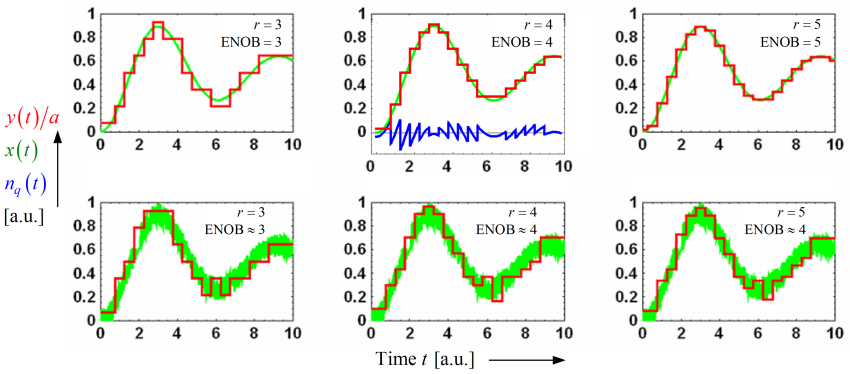


Fig. 2.3. Quantization, noise and effective number of bits (ENOB) in an analogue-to-digital converter (ADC). Input $x(t)$, down-scaled quantized output $y(t)/a$ and quantization error (quantization noise) $n_q(t)$ as a function of time t. The quantity a is the average slope of the ADC, $\mathbf{y} = a\,\mathbf{x}$. **Upper row:** Quantization of noiseless signal $x(t)$ with $r = \{3, 4, 5\}$ bit and $M = 2^r$ levels. The larger r becomes, the better the down-scaled quantized output $y(t)/a$ approximates the input $x(t)$ (clearly visible for $t = \{0 \ldots 1, 5 \ldots 7, 9 \ldots 10\}$). **Lower row:** Quantization of the signal $x(t)$ superimposed by noise. For larger r, the ENOB does not increase accordingly, due to the input noise of the ADC. [modified from Fig. 1 and 2 in Ref. 6 on Page 15]

of quantizing bits does not increase the effective number of bits above a certain level, which is fixed by $\mathrm{SNDR}_{q,\mathrm{dB}}$ of Eq. (2.5)

It makes no sense to choose $M = 2^r$ so large that the resulting quantization noise with RMS value σ_{n_q}, see Eq. (B.42b), becomes much smaller then the noise which comes along with the signal. If the step size $q \approx \sqrt{P_R}$ between levels is chosen to be of the order of the signal's noise $\sqrt{P_R}$, the quantization noise $\sigma_{n_q} = q/\sqrt{12} \approx \frac{1}{3}q$ is of the order of the signal's noise $\sqrt{P_R}$. Therefore a coarse estimate of the proper number of levels is $M = 1 + \sqrt{P_S}/\sqrt{P_R}$ (one more level than intervals), a number which is intimately connected to the signal-to-noise power ratio (SNR),

$$\gamma \equiv \mathrm{SNR} := \frac{P_S}{P_R}, \qquad \gamma_{\mathrm{dB}} = 10\lg(\gamma), \qquad M = 1 + \sqrt{\gamma}, \quad M^2 = 1 + \gamma + 2\sqrt{\gamma} \approx 1 + \gamma. \tag{2.6}$$

Coding Having associated the samples $\Psi(iT_s)$ with discrete values, these numbers are expressed by code words, mostly choosing binary numbers with r places (voice signals can be encoded with $r = 8$). Each digit of a binary code word (bit, *b*inary dig*it*) assumes $b_l = 2$ logical levels zero (0) or one (1). Also codes with $b_l > 2$ are used. The number of bits[7] r needed to code each sample is related to the number of quantized signal levels M of Eq. (2.6),

$$M = b_l^r, \qquad r = \log_{b_l} M \qquad \Longrightarrow \qquad r \approx \frac{10\lg\sqrt{1+\gamma}}{10\lg 2} \approx 3.32 \times \frac{\gamma_{\mathrm{dB}}}{2} \quad \text{for } b_l = 2,\ \gamma \gg 1\,. \tag{2.7}$$

The resulting temporal bit sequence of logical 1 and 0 is known as a binary pulse code modulation (PCM). For transmission, the PCM signal has to be encoded in a sequence of symbols made up of discrete values of a physical quantity, e. g., a sequence of binary impulses $p(t)$ with amplitudes $a_n = 0, 1$. For $b_l = 2$, the PCM bit rate R_b is a multiple r of the sampling rate F_s,

$$R_b = rF_s, \quad R_b = F_s \log_2 M \qquad \text{for } b_l = 2. \tag{2.8}$$

An example for binary PCM is telephone voice transmission with $B = 3.4\,\mathrm{kHz}$, where $r = 8$ and $F_s = 8\,\mathrm{kHz}$ with $R_b = 8\,F_s = 64\,\mathrm{kbit/s}$ are common. The circuitry responsible for converting analog electrical signals to digital data and vice versa is known as *co*der / *dec*oder (CODEC). A CODEC translates each sampled value into its binary representation.

[7] Changing the base of the logarithm of x from a to b:
$\log_a x = y \log_b x, \quad x = a^{y \log_b x}, \quad \log_b x = \log_b\left(a^{y \log_b x}\right) = y \log_b x \log_b a \quad \rightarrow \quad y = 1/\log_b a,$
$\log_b x = \log_b a \log_a x.$ Example: $\log_{10} x = \log_{10} 2 \log_2 x = 0.301 \times \log_2 x, \quad \log_2 x = 3.32 \times \log_{10} x.$

Forward error correction (FEC) With increasing computing power, a redundancy transmission scheme becomes possible, where codes are transmitted which allow an error correction at the receiver side. For an optical communication channel at 40 Gbit/s the bit error probability (BER) performance of moderate-length nonbinary low-density parity-check (LDP) codes[8,9,10,11] is as follows (RS stands for Reed-Solomon):

First generation FEC Hard-decision block code, typically RS(255, 239) with a 6.69 % overhead. For an output BER = 10^{-13}, the RS code yields a net coding gain (see below) of about 6 dB.

Second generation FEC Hard-decision concatenated codes combined with interleaving and iterative decoding techniques to improve the FEC capability. The ITU-T G.975.1 standard has defined eight second-generation FEC algorithms with 6.69 % overhead. As an example, an LDP(29136, 27 315) code[12] achieves a coding gain of 9.4 dB at an output BER = 10^{-15}, starting from a pre-FEC BER = 4.45×10^{-3}.

Third generation FEC Soft-decision[13] FEC (SD-FEC) with turbo product and LDP codes are especially necessary for 100G long-haul transmission equipment. Coherent receiving technology in optical communication systems and the rapid growth in computing power enables soft-decision FEC. For an output BER = 10^{-15} with 15 ... 20 % overhead, soft-decision FEC yields a net coding gain of 11 dB. An FEC scheme[14] with 15 % overhead and an input pre-FEC BER = $(1.8 \ldots 2) \times 10^{-2}$ effectively prevents line errors.

Without going into coding details, these examples demonstrates the potential of the technique: A coding gain of about 10 dB means that 10 dB less power can be received for a final BER = 10^{-15} than without FEC. This allows a raw BER of the order of 10^{-4}. The price is that a data rate of, e. g., 40 Gbit/s increases to a line rate of 43 Gbit/s with a redundancy overhead of about 7 %.

Bit rate The minimum PCM bit rate ($b_l = 2$) resulting from analogue-to-digital conversion of a bandwidth-limited real-valued signal (sampling rate $F_s = 2B$, number of quantized signal levels M) for a signal-to-noise power ratio γ as in Eq. (2.6) is[15,16]

$$R_b = rF_s = 2B \log_2 M = B \log_2 \left(M^2\right) \quad \text{(for a real-valued band-limited signal)},$$
$$R_b \approx B \log_2 (1+\gamma) = B \frac{10 \log_{10} (1+\gamma)}{10 \log_{10} 2} \approx 3.32 \times B\, \gamma_{\text{dB}} \quad \text{for } \gamma \gg 1\,. \tag{2.9}$$

If a real signal with bandwidth B is sampled with a rate $F_s = 2B$, quantized with multiple levels M, and encoded with symbols representing r bit each, the symbol rate[17] equals the sampling rate, $R_s = F_s$ (unit

[8] Djordjevic, I. B.; Vasic, B.: Nonbinary LDPC codes for optical communication systems. IEEE Phot. Technol. Lett. 17 (2005) 2224–2226

[9] F. Chang, K. Onohara, T. Mizuochi: Forward error correction for 100G transport networks. IEEE Comm. Mag. 48 (2010) S48–S55

[10] Fujitsu Network Communications: Soft-Decision FEC Benefits for 100G. White Paper (2012)
`http://www.fujitsu.com/downloads/TEL/fnc/whitepapers/Soft-Decision-FEC-Benefits-or-100G-wp.pdf`

[11] Zhu Xiao-yu: A brief analysis of SD-FEC. ZTE Technol. 15 (2012) 23–24
`http://wwwen.zte.com.cn/endata/magazine/ztetechnologies/2012/no5/`

[12] See Ref. 9 on Page 17. Fig. 3

[13] Soft-decision decoding uses the waveform information that is output by channels. A real number is output by a matched filter, and the demodulator sends this to a soft-decision decoder. The decoder needs not only 0 or 1 code streams but also soft information to indicate the reliability of these input code streams. The further the code value is from the decision threshold, the more reliable the signal is, and vice versa. Because a soft-decision decoder has more channel information than a hard-decision decoder, it can use the information through probability decoding and obtain higher coding gains than a hard-decision decoder. [Cited after Ref. 11 on Page 17]

[14] See Ref. 11 on Page 17

[15] See Ref. 11 on Page 3

[16] R. V. L. Hartley: Transmission of information. Bell Syst. Techn. J. 7 (1928) 535–563.

[17] Jean-Maurice-Émile Baudot, ⋆ Magneux (France) 11.9.1845, † Sceaux (France) 28.3.1903, engineer who, in 1874, received a patent on a telegraph code that by the mid-20th century had supplanted Morse code as the most commonly used telegraphic alphabet. In Baudot's code, each letter was represented by a five-unit combination of current-on or current-off signals of equal duration.

Bd)[18], and is smaller than the bit rate R_b by a factor of r. Depending on the shape of the pulses, the occupied spectral passband width $\mathcal{B}$ around an optical carrier changes. When signalling with a sequence of real-valued sinc-shaped so-called Nyquist pulses $\sum_{i=-\infty}^{+\infty} \Psi(iT_s)\,\mathrm{sinc}(t/T_s - i)$ as in Eq. (2.3), the required spectral passband width $\mathcal{B} = 2B$ is minimum as compared to other pulse shapes. If the Nyquist pulses are in addition complex-valued, the required spectral passband width is $\mathcal{B} = B$, and we find

$$R_s = F_s = \frac{R_b}{r}, \quad \text{sinc-pulses with spectral width } \mathcal{B} = F_s = \begin{cases} 2B & \text{(real symbols)} \\ B & \text{(complex symbols)} \end{cases}. \tag{2.10}$$

2.2 Optical fibre channel

Shannon[19]defined: "The channel is merely the medium used to transmit the signal from transmitter to receiver. It may be a pair of wires, a coaxial cable, a band of radio frequencies, a beam of light, etc." Here, we follow this definition and describe the *physical channel* first by the greatly simplified model of a linear lossless weakly guiding optical single-mode fibre (that in fact supports two orthogonally polarized modes), specifying in Sect. 2.2.1 its impulse response and its transfer function, respectively.

We start with the scalar form of Maxwell's equations Eq. (A.2) on Page 175 of Appendix A, and show interest only in the fundamental modal field $\Psi(t,z) := \Psi(t,\vec{r})$, which we represent by an equivalent plane wave propagating in $+z$-direction with the propagation constant $\beta(\omega) = n_e(\omega)\,k_0$ (propagation constant in vacuum $k_0 = \omega/c$, equivalent modal refractive index $n_e = \beta/k_0 = \sqrt{\epsilon_{r\,\mathrm{eff}}}$). The propagation constant must be calculated from an eigenfunction analysis.

Our solution ansatz consists of a carrier wave $\exp(\mathrm{j}\,\omega_0 t)$ with angular frequency $\omega_0 = 2\pi f_0$, which is modulated with a complex amplitude $a(t)$ that varies slowly on the scale of the optical carrier's period $1/f_0$. Consequently, Eq. (A.2) is solved by $\Psi(t,z)$, the Fourier transform of which is denoted by $\breve{\Psi}(f,z)$,

$$\Psi(t,z) = a(t)\,\mathrm{e}^{\mathrm{j}[\omega_0 t - \beta(\omega_0) z]}, \qquad \breve{\Psi}(f,z) = \breve{a}(f - f_0)\,\mathrm{e}^{-\mathrm{j}\,\beta(\omega_0) z}. \tag{2.11}$$

The impact of fibre nonlinearities is described in Sect. 2.2.2 with the help of the nonlinear material polarization introduced in Sect. A.3.2 of Appendix A on Page 177 ff.

After that we discuss the *logical channel* in Sect. 2.2.3 on Page 20 and formulate its data carrying capacity in terms of the signal-to-noise power ratio SNR.

2.2.1 Propagation in a linear fibre

The fundamendal-mode transfer function $\breve{h}_c(f) := \breve{h}(f,L)$ of a weakly guiding fibre with length L is defined by the ratio of the Fourier transforms of the fields at output $z = L$ and input $z = 0$ of the fibre. The (analytic) transfer function $\breve{h}_c(f)$ and the associated (causal) real impulse response $h_c(t)$ are

$$\breve{h}_c(f) := \breve{h}(f,L) = \frac{\breve{\Psi}(f,L)}{\breve{\Psi}(f,0)} = \mathrm{e}^{-\mathrm{j}\,\beta(\omega) L}, \quad \beta(\omega) = -\beta(-\omega); \quad h_c(t) = \int_{-\infty}^{+\infty} \breve{h}_c(f)\,\mathrm{e}^{\mathrm{j}\,2\pi f t}\,\mathrm{d}f. \tag{2.12}$$

As mentioned before, the propagation constant β is often replaced by the effective modal refractive index n_e. The normalized frequency V combines the quantities core radius a, optical angular frequency ω, core refractive index n_1, cladding refractive index n_2, and relative refractive index difference Δ,

$$n_e = \frac{\beta}{k_0}, \qquad k_0 = \frac{\omega}{c} = \frac{2\pi}{\lambda}, \qquad V = a k_0 n_1 \sqrt{2\Delta}, \qquad \Delta = \frac{n_1^2 - n_2^2}{2 n_1^2} \underset{\Delta \ll 1}{\approx} \frac{n_1 - n_2}{n_1}. \tag{2.13}$$

[18] In telecommunication and electronics, baud (pronounced ['bɔːd], unit symbol Bd) is synonymous to symbols per second, the unit of the symbol rate. Sometimes a symbol is denoted as a "baud", so that symbol rate and baud rate would be the same. However, if the meaning of the unit Bd is agreed upon, to talk of a *baud rate* (literally meaning $(\text{symbols}/\text{s})/\text{s} = \text{Bd}/\text{s}$) instead of naming it a symbol rate does not make sense.

[19] See Ref. 8 on Page 2. First page, first column, statement *3) The Channel*

For narrow-banded optical spectra it is useful to expand $\beta(\omega)$ in a Taylor series around the carrier frequency $f_0 = c/\lambda_0 = \omega_0/(2\pi)$ and retain terms up to the third order, that is,

$$\beta(\omega) \approx \beta_0^{(0)} + (\omega-\omega_0)\beta_0^{(1)} + \frac{(\omega-\omega_0)^2}{2!}\,\beta_0^{(2)} + \frac{(\omega-\omega_0)^3}{3!}\,\beta_0^{(3)}, \tag{2.14}$$

$$\beta_0^{(i)} = \left.\frac{\mathrm{d}^i\beta(\omega)}{\mathrm{d}\omega^i}\right|_{\omega=\omega_0}, \quad \beta_0 := \beta_0^{(0)}, \quad \Delta\omega = \omega-\omega_0\,, \quad \Delta f = \Delta\omega/(2\pi).$$

We identify the modal phase velocity v_p, the group velocity v_g and the group delay t_g (propagation length $z = L$, group refractive index n_g), which is related to the first-order chromatic dispersion C and its derivatives (e. g., D), Eq. (2.17),

$$v_p = \frac{\omega_0}{\beta_0^{(0)}}\,, \qquad v_g^{-1} = \frac{t_g}{L} = \frac{n_g}{c} = \beta_0^{(1)}, \qquad \frac{1}{L}\frac{\mathrm{d}t_g}{\mathrm{d}\omega} = \beta_0^{(2)}, \qquad \frac{1}{L}\frac{\mathrm{d}^2\,t_g}{\mathrm{d}\omega^2} = \beta_0^{(3)}. \tag{2.15}$$

The length-related group delay time difference $\Delta t_g/L$ of two signals propagating in the same fundamental mode at optical carriers, which differ in λ by $\Delta\lambda$, can be approximately written with the help of the normalized propagation constant $B = \left(\beta^2 - n_2^2k_0^2\right)/\left(n_1^2k_0^2 - n_2^2k_0^2\right) \approx (\beta - n_2k_0)/(n_1k_0 - n_2k_0)$, assuming weak guidance $\Delta \ll 1$,

$$\Delta t_g/L = \left[t_g(\lambda+\Delta\lambda) - t_g(\lambda)\right]/L = C\Delta\lambda = (M+W)\Delta\lambda\,,$$

$$M = M_s = \underbrace{\frac{1}{c}\frac{\mathrm{d}n_{sg}}{\mathrm{d}\lambda}}_{\substack{\text{material}\\ \text{dispersion}}} \quad (s = 1 \text{ or } 2),\; W = -\frac{n_{1g}-n_{2g}}{c\lambda}\,\underbrace{V\frac{\mathrm{d}^2(VB)}{\mathrm{d}V^2}}_{\substack{\text{dispersion}\\ \text{factor}}} \tag{2.16}$$

The first-order material dispersion coefficients in core (M_1) and cladding ($M_2 \approx M_1$) are assumed to be of similar value. The chromatic dispersion is expressed by the first-order coefficient C (unit ps / (km nm)) for a fixed reference wavelength λ_1. We extend Eq. (2.16) by one more term and define a second-order dispersion coefficient D (unit ps / $\left(\mathrm{km\,nm^2}\right)$),

$$\Delta t_g/L = C\,\Delta\lambda + D\,(\Delta\lambda)^2 + \ldots, \qquad C = M + W, \tag{2.17a}$$

$$C = \frac{1}{L}\,\frac{\mathrm{d}\,t_g}{\mathrm{d}\lambda} = -\frac{2\pi c}{\lambda_0^2}\beta_0^{(2)}\,, \tag{2.17b}$$

$$C(\lambda_C) = 0\,, \qquad \lambda_C \text{ first-order dispersion zero wavelength}, \tag{2.17c}$$

$$D = \frac{1}{L}\,\frac{1}{2}\,\frac{\mathrm{d}^2t_g}{\mathrm{d}\lambda^2} = \left(\frac{2\pi c}{\lambda_0^2}\right)^2\beta_0^{(3)} + \frac{4\pi c}{\lambda_0^3}\beta_0^{(2)}\,. \tag{2.17d}$$

When for a certain reference wavelength $\lambda_1 = \lambda_C$ the first-order chromatic dispersion C becomes zero, and the total dispersion is determined by the second-order dispersion coefficient D.

With a slight re-ordering of Eq. (2.17a) we can define a wavelength-dependent chromatic dispersion factor $C_\lambda(\lambda) = C + D\,\Delta\lambda$ which is approximated by a straight line near the reference wavelength λ_1,

$$\Delta t_g/L = C_\lambda(\lambda)\,\Delta\lambda = (C + D\,\Delta\lambda)\;\Delta\lambda = C\,\Delta\lambda + D\,(\Delta\lambda)^2\,, \quad D = \frac{\mathrm{d}C_\lambda(\lambda)}{\mathrm{d}\lambda}\,. \tag{2.17e}$$

Comparing D in Eq. (2.17a), (2.17e) could lead to confusion because of the factor 1/2. In Eq. (2.17a), the dispersion *coefficients* $C \equiv C(\lambda_1)$, $D \equiv D(\lambda_1)$ are constants of the Taylor expansion for the function $t_{gm}(\lambda)$ at a certain reference wavelength $\lambda = \lambda_1$. Therefore, $\mathrm{d}C(\lambda_1)/\,\mathrm{d}\lambda = 0$ holds by definition, and $D(\lambda_1) \neq \mathrm{d}C(\lambda_1)/\,\mathrm{d}\lambda$. On the other hand, the dispersion *function* $C_\lambda(\lambda)$ may be linearly expanded, and its so-called dispersion slope $D = \mathrm{d}C_\lambda(\lambda)/\,\mathrm{d}\lambda|_{\lambda_1}$ is then well defined.

We model the light source by an analytic signal $\underline{a}_s(t)$, which is modulated with a (possibly complex) signal $s\,(t)$,

$$\underline{a}_s\,(t) = A_s\,(t)\,\mathrm{e}^{\,\mathrm{j}\,\omega_0 t}, \qquad A_s\,(t) = s\,(t)\;a\,(t)\,, \tag{2.18}$$

This light source excites a waveguide mode $\Psi_m(\vec{r})$; the mode number m characterizes any set of appropriate mode numbers, e. g., $m \mathrel{\widehat{=}} (\nu, \mu)$ for a fibre. The mode coupling coefficient is c_m, and the normalization $\sum_m |c_m|^2 = 1$ holds. The scalar time-dependent field and its Fourier transform at the waveguide input $z = 0$ read in cylindrical coordinates

$$\Phi_m(t, r, \varphi, 0) = c_m \Psi_m(r, \varphi)\, \underline{a}_s(t), \qquad \breve{\Phi}_m(f, r, \varphi, 0) = c_m \Psi_m(r, \varphi)\, \breve{\underline{a}}_s(f). \tag{2.19}$$

For the waveguide length z the output signal may be calculated by a convolution of the input signal with the causal waveguide impulse response $h_m(t)$ Eq. (2.12). Further, the spectrum can be written as a product of the source spectrum $\breve{\underline{a}}_s(f)$ and the waveguide transfer function $\breve{h}_m(f)$,

$$\begin{aligned} \Phi_m(t, r, \varphi, z) &= c_m \Psi_m(r, \varphi) \int_{-\infty}^{+\infty} h_m(t_1)\, \underline{a}_s(t - t_1)\, \mathrm{d}t_1\,, \\ \breve{\Phi}_m(f, r, \varphi, z) &= c_m \Psi_m(r, \varphi)\, \breve{h}_m(f)\, \breve{\underline{a}}_s(f), \qquad \breve{h}_m(f) = \mathrm{e}^{-\mathrm{j}\,\beta_m(\omega) z}\,. \end{aligned} \tag{2.20}$$

The transverse field dependence of the eigenmode $\Psi_m(r, \varphi)$ does not vary significantly with f because the small frequency dependent changes of the waveguide refractive indices may be usually neglected. Equation (2.20) establishes a linear relation between the modulation $s(t)$ and the response $\Phi_m(t, r, \varphi, z)$.

2.2.2 Propagation in a nonlinear fibre

Propagation in a nonlinear optical fibre can be approximately described by the so-called nonlinear Schrödinger[20] equation (NLSE) as derived in Appendix A, Eq. (A.34) on Page 181. The following quantities are used: Impulse envelope $A(T, z)$ in a retarded time frame $T = T(t, z) := t - \beta_0^{(1)} z = t - z/v_g$ that moves along with the impulse (Eq. (A.26)), fibre nonlinear coefficient γ (Eq. (A.25)), and linear fibre power attenuation coefficient α,

$$\frac{\partial A(T, z)}{\partial z} = \mathrm{j}\,\frac{\beta_0^{(2)}}{2} \frac{\partial^2 A(T, z)}{\partial T^2} - \mathrm{j}\,\gamma\,|A(T, z)|^2\, A(T, z) - \frac{\alpha}{2} A(T, z). \tag{2.21}$$

In general, a simple transfer function as in Eq. (2.12) cannot be specified. In the case of zero linear attenuation $\alpha = 0$, Eq. (2.21), (A.34) resembles the well-known Schrödinger equation of quantum mechanics with a nonlinear (quadratic) potential term $\mathrm{j}\,\gamma\,|A(T, z)|^2\, A(T, z)$. Thus, it is called the *nonlinear Schrödinger equation*[21,22] (NLSE). If during the propagation of a light signal its loss is continuously compensated by gain, then the power loss constant can be set actually to zero, $\alpha = 0$. For including random perturbations by, e. g., ASE noise of optical amplifiers, a random field[23] $-\mathrm{j}\, N_{\mathrm{ASE}}(T, z)$ can be added on the right-hand side of Eq. (2.21), (A.34).

The parameters of a standard single-mode fibre[24] (SSMF) are listed in Table 2.1. For the definition of symbols and their context see also Appendix A.4 on Page 178 ff.

2.2.3 Shannon's channel capacity and spectral efficiency

A quick and intuitive way in understanding the meaning of Shannon's channel capacity formula refers to Eq. (2.9) on Page 17. This maximum possible bit rate R_b for a *real Nyquist signal* with symbol period $T_s = 1/(2B)$, symbol rate $R_s = 1/T_s$, sampling rate $F_s = R_s = 2B$, and bandwidth B is according

[20] Erwin Schrödinger, ⋆ Vienna (Austria) 12.08.1887, † Vienna (Austria) 04.01.1961. Austrian theoretical physicist who contributed to the wave theory of matter and to other fundamentals of quantum mechanics. He shared the 1933 Nobel Prize for Physics with the British physicist P. A. M. Dirac.

[21] See Ref. 17 on Page 6, Sect. 2.3.1 Eq. (2.3.27) Page 43

[22] Boyd, R. W: Nonlinear optics. 3. Ed. San Diego: Academic Press 2008. Section 7.5.2, Eq. (7.5.32)

[23] R.-J. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, B. Goebel: Capacity limits of optical fiber networks. J. Lightw. Technol. 28 (2010) 662–701

[24] R.-J. Essiambre, R. W. Tkach, R. Ryf: Fiber nonlinearity and capacity: Single-mode and multimode fibers. In: Kaminow, I. P.; Li, Tingye; Willner, A. E. (Eds.): Optical Fiber Telecommunications VI B. Systems and Networks, 6th Ed. Elsevier (Imprint: Academic Press), Amsterdam 2013, Chapter 1, pp. 1–43

Parameter	Symbol	SSMF data
Chromatic dispersion	C	$17\,\frac{\text{ps}}{\text{km nm}}$
Dispersion slope	D	$0.07\,\frac{\text{ps}}{\text{km nm}^2}$
Attenuation factor per length	a/L	$0.2\,\frac{\text{dB}}{\text{km}}$
Nonlinear refractive index	n_2^I	$2.5\times10^{-20}\,\frac{\text{m}^2}{\text{W}}$
Effective area	A_{eff}	$80\,\mu\text{m}$
Nonlinear coefficient	γ	$1.27\,\text{W}^{-1}\,\text{km}^{-1}$
Operating wavelength	λ_0	$1.55\,\mu\text{m}$
Operating frequency	f_0	193.41 THz

Table 2.1. Standard single-mode fibre (SSMF) parameters [after Ref. 24 Table 1.1]

to Eq. (2.9) on Page 17 $R_b = B\log_2\left(M^2\right) \approx B\log_2(1+\gamma)$. In more general terms: If a channel has the bandwidth $\mathcal{B} = 2B$ and is able to transport a real Nyquist signal with a certain SNR $\equiv \gamma$ that suffices for "error-free" reception, then the so-called channel capacity C_{real} for *real signals* equals the maximum possible bit rate $C_{\text{real}} = R_b \approx \frac{1}{2}\mathcal{B}\log_2(1+\gamma)$. The spectral efficiency $C'_{\text{real}} = C_{\text{real}}\mathcal{T}$ then denotes how many bit are transmitted per symbol. The symbols repeat with a period (observation time) $\mathcal{T} = 1/\mathcal{B} = 1/(2B) = T_s$, and we find $C'_{\text{real}} = \frac{1}{2}\log_2(1+\gamma) \approx \log_2 M$. — In the following, we first address a linear glass fibre channel, before taking account of the nonlinear properties of a fibre.

Linear Shannon limit

In Shannon's rigorous formulation, a linear communication channel transports a Gaussian distributed noise field (additive white Gaussian noise, AWGN) with average power P_r in the channel bandwidth $\mathcal{B}$. Seen from an information-theoretical point of view the signal field is also a Gaussian-distributed random quantity having the average power P_s. The limiting capacity of a channel is the maximum bit rate that can be transmitted error-free, taking account of noise, available bandwidth, and constrained power[25].

It is remarkable that the capacity can be computed without explicitly considering any specific modulation, coding, or decoding scheme. Likewise, computation of the capacity does not generally tell us which specific modulation, coding, or decoding schemes we should use in order to achieve the capacity. The theory indicates that we must use strong error-correcting codes, and that the decoding complexity and delay must increase exponentially as we approach the limiting capacity.

We define a channel which transmits complex signals $s(t)$ with an in-phase (I, real part) component and a quadrature[26] phase (Q, imaginary part) component. How this mathematical concept is realized will be explained in the IQ-modulator section on Page 28 ff. As we saw in Eq. (2.10) on Page 18, signalling with sinc-shaped Nyquist pulses having complex amplitudes $\Psi(iT_s)$ leads to a symbol rate that equals the signal bandwidth, $R_s = B$. If not stated otherwise, we assume for the following this type of pulse shaping, which leads to a rectangular passband spectrum with a width of B. The channel bandwidth $\mathcal{B} = B = R_s$ be adapted to these transmission requirements.

Limiting channel capacity The theoretical limiting (maximum) channel capacity C (unit bit/s) per polarization for error-free signal transmission, which can be reached only with arbitrarily complicated encoding techniques including the transmission of *complex Nyquist signals*, is given by[27]

$$C = \mathcal{B}\log_2\left(1+\frac{P_s}{P_r}\right) = \mathcal{B}\log_2(1+\gamma), \qquad \gamma \equiv \text{SNR} := \frac{P_s}{P_r}. \tag{2.22}$$

[25] J. M. Kahn, K.-P. Ho: Spectral efficiency limits and modulation / detection techniques for DWDM systems. IEEE J. Sel. Topics Quantum Electron. 10 (2004) 259–272

[26] The points in phase space describe uniquely all possible states of a system (e. g., the momentum p_x of a particle moving in x-direction and its position x). Generally spoken, so-called "quadratures" are quantities that can be used to represent the real (e. g., x) and the imaginary part (e. g., p_x) of a complex quantity. A plot of the quadratures against each other is called a phase diagram. — Here, we plot I vs. Q and name the result a constellation diagram.

[27] See Ref. 8 on Page 2. Shannon's paper from 1948 cites the groundbreaking work of Nyquist (Ref. 11 on Page 3, 1924, Ref. 12 on Page 3, 1928) and Hartley (Ref. 16 on Page 17, 1928).

We cannot go into the details of the derivation of Eq. (2.22). Instead, we refer to Eq. (2.9) on Page 17, where a similar expression was made plausible. It is obvious that for a constant C the channel's SNR and its bandwidth $\mathcal{B}$ can be exchanged: The more elaborate the coding is, the less channel bandwidth $\mathcal{B}$ is required, but the higher the channel's SNR must be. This is also true for the spectral efficiency which is discussed in the following.

We further remark that Eq. (2.22) implies classical noise in one transverse mode and one polarization only. As soon as both classical and quantum noise are involved as is the case with amplified spontaneous emission (ASE) noise of an optical amplifier[28,29,30,31,32,33], the Shannon relation must be properly interpreted. We will return to this problem when treating noise in more detail.

Spectral efficiency If we want to know how many information bits we can transmit per polarization during an observation time $\mathcal{T}$, we have to calculate the so-called spectral efficiency $\mathrm{SE} = C\mathcal{T} = \mathcal{B}\mathcal{T}\log_2(1+\gamma)$ (unit bit/s/Hz; in fact, this "unit" represents a number of bits and is therefore dimensionless). The shortest possible observation time for a *complex symbol* is $\mathcal{T} = 1/\mathcal{B}$ as specified by the sampling theorem Eq. (2.4) on Page 15. Thus, the limiting (maximum) spectral efficiency $C' = C/\mathcal{B}$ for an AWGN channel describes the maximum number of information bits to be transmitted during the minimum observation time $\mathcal{T} = 1/\mathcal{B}$. For a small SNR an approximation can be given, and we find

$$C' = \frac{C}{\mathcal{B}} = \log_2(1+\gamma), \qquad \gamma = 2^{C'} - 1, \tag{2.23}$$

$$C' \approx \frac{1}{\ln 2}\left(\gamma - \frac{1}{2}\gamma^2\right) \quad \text{for } \gamma \ll 1. \tag{2.24}$$

However, if the observation time $\mathcal{T} = 1/B$ becomes longer, i. e., if the actual signal bandwidth $B < \mathcal{B}$ is chosen to be smaller than the channel bandwidth $\mathcal{B}$ so that the signal's information capacity (bit rate) is only $B\log_2(1+\gamma)$, then the practical spectral efficiency C'_{pract} results,

$$C'_{\mathrm{pract}} = \frac{B\log_2(1+\gamma)}{\mathcal{B}} = \frac{B}{\mathcal{B}}\,C'. \tag{2.25}$$

Equations (2.23) and (2.24) show that limiting channel capacity C and spectral efficiency C' tend to zero with the same order as γ. If for $\gamma \ll 1$ (not a very practical case!) the product $\mathcal{B}\times\gamma$ is kept constant, the limiting channel capacity C remains also constant.

Optical amplifiers in a link contribute ASE noise as will be discussed in more detail in Sect. 5.2.3 on Page 128 ff. With a single-pass power gain $\mathcal{G}_s$, an optical bandwidth $B_O = \mathcal{B}$, an amplifier noise figure F, and with the photon energy $w_O = hf_0$ at central frequency f_0, the (extractable) ASE noise output power of such an amplifier is $P_{\mathrm{ASE},x} = N_0\mathcal{B}$ per polarization and mode, where the noise power spectral density is $N_0 = (\mathcal{G}_s - 1)\,F w_O$. This optical noise power can be expressed in terms of minimum-uncertainty quantum fluctuations, characterized by a (non-extractable[34]) minimum quantum noise power $P_{r\,\mathrm{qu}} = w_O\mathcal{B}$ per polarization and mode, Eq. (5.67) on Page 128). The corresponding noise power spectral density is $N_{0\,\mathrm{qu}} = w_O$. In the following, we understand N_0 to be the actual noise power spectral density of the link under consideration.

We define the energy per symbol by dividing the signal power by the symbol rate, $W_s = P_s/R_s = P_s/\mathcal{B}$, and the energy per bit by relating the energy per symbol to the number of bits C' that are

[28] Gordon, J. P.: Quantum effects in communication systems. Proc. Inst. Radio Eng., 50 (1962) 1898–1908

[29] G. Grau: Rauschen und Kohärenz im optischen Spektralbereich. In: W. Kleen, R. Müller (Eds.): Laser. Berlin: Springer-Verlag 1969. Eq. (9.2/21) on Page 476

[30] Helstrom, C. W., Liu, J. W. S., Gordon, P.: Quantum mechanical communication theory. Proc. IEEE 58 (1970) 1578–1598

[31] Helstrom, C. W.: Capacity of the pure-state quantum channel. Proc. IEEE 62 (1974) 140–141

[32] J. R. Pierce: Optical channels: Practical limits with photon counting. IEEE Trans. Commun. COM-26 (1978) 1819–1821

[33] D. O. Caplan: Laser communication transmitter and receiver design. J. Opt. Fiber. Commun. Rep. 4 (2007) 225–362

[34] "Non-extractable power" means that it stands for a quantum uncertainty, and therefore cannot be extracted from the system: You cannot fry eggs with this power. — With the "extractable" ASE noise power, however, you can!

transmitted per symbol, $W_b = W_s/\,C'$. With these definitions, we write the SNR γ and the SNR per bit γ_b in different forms,

$$\gamma \equiv \mathrm{SNR} := \frac{P_s}{P_r} = \frac{P_s}{N_0\mathcal{B}} = \frac{W_s}{N_0}, \qquad W_s = \frac{P_s}{\mathcal{B}}, \tag{2.26}$$

$$\gamma_b \equiv \mathrm{SNR}_b := \frac{W_b}{N_0} = \frac{\gamma}{C'} = \frac{2^{C'}-1}{C'}, \qquad W_b = N_{Pb}\,hf_0 = \frac{W_s}{C'}. \tag{2.27}$$

The energy per bit W_b divided by the energy per photon hf_0 equals the number of photons per bit N_{Pb}. If we had minimum-uncertainty quantum fluctuations only, $N_0 = N_{0\,\mathrm{qu}} = w_O = hf_0$, SNR_b would correspond to the photon number per bit, $\gamma_b = N_{Pb}$. For the limits of large and small spectral efficiencies we find[35] from Eq. (2.27)

$$\gamma_b \equiv \mathrm{SNR}_b = \begin{cases} 2^{C'}/C' & \text{for } C' \gg 1\,, \\ \frac{1+C'\ln 2+\frac{1}{2}\left(C'\ln 2\right)^2-1}{C'} = \ln 2\left(1+\frac{1}{2}C'\ln 2\right) & \text{for } C' \ll 1\,. \end{cases} \tag{2.28}$$

The SNR per bit assumes a minimum value $\gamma_b^{(\min)} = \ln 2 \approx 0.693$, $\gamma_{b\,\mathrm{dB}}^{(\min)} = 10\lg(\ln 2) = -1.58\,\mathrm{dB}$ for a spectral efficiency of $C' \to 0$. This result is plausible and means that a minimum energy per bit is needed to transmit information over an AWGN channel with an ever so small spectral efficiency. The channel capacity $C = \mathcal{B}C'$ would then approach zero if not for an unphysical channel bandwidth $\mathcal{B} \to \infty$. From the first order approximation Eq. (2.28), the spectral efficiency can be written as

$$C' \approx \frac{2}{(\ln 2)^2}\left(\gamma_b - \gamma_b^{(\min)}\right) \approx 4.16 \times (\gamma_b - 0.693) \quad \text{for } C' \ll 1\,. \tag{2.29}$$

If the channel bandwidth $\mathcal{B}$ increases, the channel capacity increases indefinitely according to Eq. (2.23). However, this assumes that the SNR remains constant. This is not true in practice because the noise power spectral density N_0 is essentially frequency-independent. In this case, the limiting channel capacity Eq. (2.22) and the limiting spectral efficiency Eq. (2.23) can be re-written ($\ln(1+x) \approx x$ for $-1 < x \leq +1$),

$$C = \mathcal{B}\log_2\left(1+\frac{P_s}{N_0\mathcal{B}}\right) = \frac{\mathcal{B}}{\ln 2}\ln\left(1+\frac{P_s}{N_0\mathcal{B}}\right), \tag{2.30}$$

$$\lim_{\mathcal{B}\to\infty} C = \frac{1}{\ln 2}\frac{P_s}{N_0} \approx 1.44 \times \frac{P_s}{N_0}, \qquad \lim_{\mathcal{B}\to\infty} C' = \frac{1}{\ln 2}\frac{W_s}{N_0} \approx 1.44 \times \frac{W_s}{N_0}, \tag{2.31}$$

$$C' = \log_2\left(1+\frac{P_s}{N_0\mathcal{B}}\right) = \log_2\left(1+\frac{W_s}{N_0}\right) = \log_2\left(1+\frac{C'W_b}{N_0}\right) = \log_2\left(1+C'\gamma_b\right), \tag{2.32}$$

$$\gamma_b = \frac{2^{C'}-1}{C'}, \quad \text{for quantum limit, } N_{Pb} = \frac{W_b}{N_0} \text{ and } N_0 = hf_0\text{: } N_{Pb} = \gamma_b\,. \tag{2.33}$$

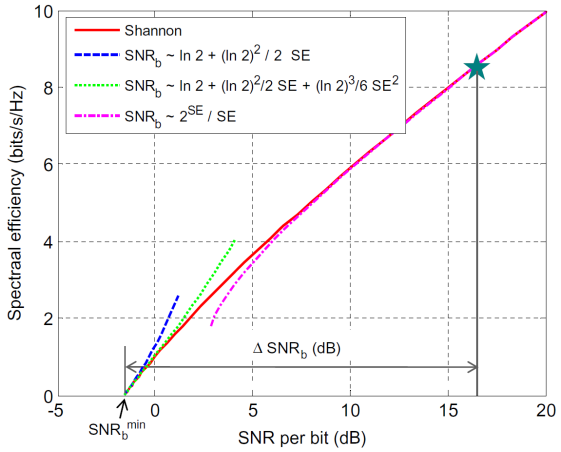
The spectral efficiency C' as a function of the SNR_b along with various approximations of C' are displayed[36,37] in Fig. 2.4(a). This type of graph is especially useful, if different modulation formats are to be compared[38]. Figure 2.4(b) shows that for minimum SNR_b requirements a modulation format like pulse position modulation (PPM) should be preferred (sensitivity-constrained), e. g., for deep-space wireless communication where a few photons per bit $N_{Pb} = \mathrm{SNR}_b$ must suffice. This comes at the cost of a low spectral efficiency. Alternatively, a modulation format with highest spectral efficiency could be chosen, e. g., phase-shift keying with 64 different phases (64PSK), which would be suitable for long-haul communication over fibres with a densely crowded spectrum (capacity-constrained). This comes at the cost of more stringent SNR_b-requirements, i. e., larger transmitting powers. A selection of modulation formats will be explained in more detail in Sect. 2.4 on Page 31.

[35] Expansion of an exponential: $a^x = \mathrm{e}^{x\ln a} \approx 1 + \frac{x\ln a}{1!} + \frac{(x\ln a)^2}{2!} + \ldots$ for $x\ln a \ll 1$.

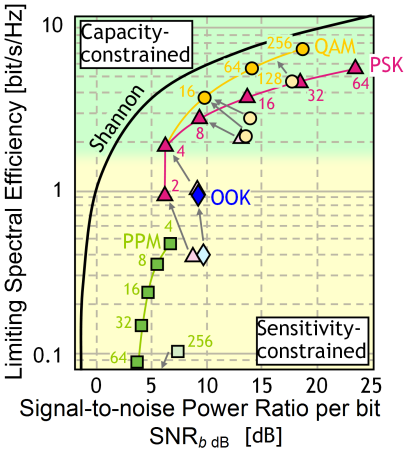
[36] R.-J. Essiambre, R. W. Tkach, Capacity trends and limits of optical communication networks. Proc. IEEE 100 (2012) 1035–1055

[37] See Ref. 24 on Page 20

[38] P. J. Winzer: Modulation and multiplexing in optical communications. Conf. on Lasers and Electro-Optics (CLEO / IQEC 2009), Baltimore (Maryland), USA, May 31–June 05, 2009. Tutorial Paper CTuL3



(a) Limiting spectral efficiency SE $\equiv C'$ as a function of SNR per bit SNR_b [after Fig. 1.4 of Ref. 24 on Page 20]



(b) Spectral efficiency of various modulation formats [after Ref. 38 on Page 23]

Fig. 2.4. Shannon limit for the spectral efficiency $C' \equiv \mathrm{SE}$ in one polarization as a function of the signal-to-noise power ratio per bit $\mathrm{SNR}_{b\,\mathrm{dB}} = 10\lg(\mathrm{SNR}_b)$. For the quantum-limited case, $N_0 = hf_0$ in Eq. (2.33), the SNR_b corresponds to the number of photons per bit. (a) The inset gives a few approximations to $\mathrm{SNR}_b(C')$ of Eq. (2.27). The SNR_b, at which a specifically chosen system operates (marked with ★), is related to the minimum $\mathrm{SNR}_b^{(\mathrm{min})}$ and expressed in dB, $\Delta\,\mathrm{SNR}_{b\,\mathrm{dB}} = \mathrm{SNR}_{b\,\mathrm{dB}} - \mathrm{SNR}_{b\,\mathrm{dB}}^{(\mathrm{min})}$. (b) Trade-off between spectral efficiency and sensitivity (small SNR_b) of various modulation formats limited by AWGN. Modulation formats (bright: theoretical limits; faint: experimental results) for a 7 % overhead code at a pre-FEC BER $= 2 \times 10^{-3}$ (squares: (256,64,32,16,8,4)PPM; triangles: (2,4,8,16,32,64)PSK; circles: (4,16,64,128,256)QAM; 4PSK ≡ QPSK ≙ 4QAM; diamonds: OOK)

Nonlinear Shannon limit

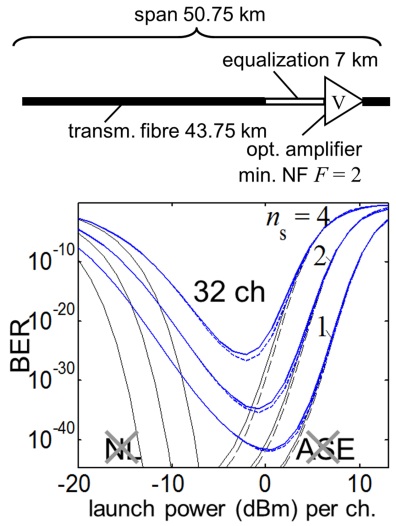
So far we had seen from Eq. (2.32) on Page 23 that for a linear channel the spectral efficiency increases logarithmically with SNR_b. However, this is not true if fibre nonlinearities come into play and wave propagation has to be described by the nonlinear Schrödinger equation (2.21) on Page 20. Because of the usual WDM operation, numerous signals in multiple WDM channels propagate simultaneously on a fixed frequency grid, see Table 1.2 on Page 5.

Consequently, the total power guided in a single-mode fibre increases, and fibre nonlinearities like four-wave mixing (FWM), cross-phase mixing (XPM) and self-phase modulation (SPM) become more and more important (FWM and SPM, see Appendix Page 177); for XPM, the intensity in one WDM channel changes the refractive index and therefore the optical phase in neighbouring WDM channels). The nonlinearities affect the signal itself, but the nonlinear WDM crosstalk adds also more "noise" to neighbouring WDM channels the larger the intensity becomes.

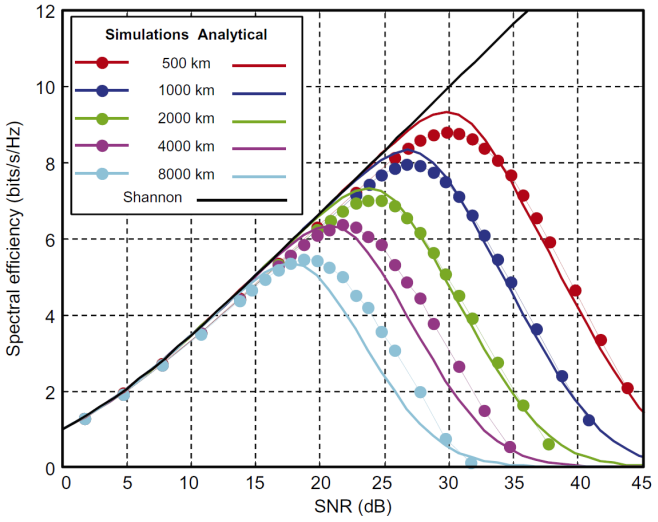
Therefore it is to be expected that the bit error ratio (BER, bit error probability), which first reduces with increasing SNR because of obvious reasons, reaches a minimum and starts increasing for larger WDM channel powers, i. e., for larger SNR. This is illustrated by simulation results[39] displayed in Fig. 2.5(a). A transmission span consists of a SSFM transmission fibre (Table 2.1 on Page 21), a dispersion compensating fibre (equalization) and an ideal optical amplifier with a noise figure $F = 2$, $F_{\mathrm{dB}} = 3\,\mathrm{dB}$. Up to $n_s = 4$ spans are concatenated for a maximum transmission distance of 203 km. A number of 32 WDM channels spaced 100 GHz apart are fed with 40 Gbit/s pseudo-random data in non-return-to-zero (NRZ) on-off keying (OOK) format. Further details are specified in the figure caption.

The linear Shannon limit assumes coding and error correction to be so good that effectively the BER becomes small enough to name the channel "error-free". If with increasing nonlinear noise the BER would deteriorate such that BER ≤ 0.5, it would be just chance how we interpreted the transmitted signal, and error correction cannot help any more. The definition of SNR does not include nonlinear noise. So it

[39] T. Kremp, W. Freude: DWDM transmission optimization in nonlinear optical fibres with a fast split-step wavelet collocation method. Proc. 7th Intern. Conf. on Optoelectronics, Fiber Optics & Photonics (Photonics 2004), Kochi, India, November 9–11, 2004



(a) Span setup and BER simulations for 32 WDM channels and $n_s = 1 \ldots 4$ spans [after Ref. 39 on Page 24]



(b) Limiting spectral efficiency simulations for 5 WDM channels separated by $\Delta f = 100\,\mathrm{GHz}$ and span lengths as specified in the inset [after Fig. 1.10 of Ref. 24 on Page 20]

Fig. 2.5. Simulation of BER as a function of launch power per channel, and simulation of spectral efficiency as a function of SNR $\equiv \gamma$ for nonlinear WDM systems with various transmission distances. (a) BER vs. channel power for 32 WDM channels. Non-return-to-zero (NRZ) data with bit rate 40 Gbit, channel grid spacing 100 GHz, pseudo-random bit sequence (PRBS) with 1 024 bit length per channel, 0.45 mW power per channel, one optical amplifier per span with a theoretically minimum noise figure $F = 2$ (inversion factor $n_{\mathrm{sp}} = 1$). Solid lines (——) include dispersion slope ($\beta_0^{(3)}$), self-steepening and Raman effect, broken lines (– – –), (— —) do not include the aforementioned types of nonlinearities. Solid-line asymptotes (——) without nonlinearities ~~NL~~ and without ~~ASE~~ noise, respectively. The BER numbers are not representative for a practical system, because usual optical amplifier noise figures are in the range $F_{\mathrm{dB}} = 4 \ldots 7\,\mathrm{dB}$. (b) Limiting nonlinear spectral efficiency SE $\equiv C'$ vs. SNR for a symbol rate $R_s = 100\,\mathrm{GBd}$, channel grid spacing 100 GHz, and for various transmission lengths as noted in the inset. If not stated otherwise, the SSMF data of Table 2.1 on Page 21 are assumed. Fibre loss is continuously compensated by Raman gain. Linear Raman amplifier noise has been computed assuming a local gain equal to the local loss with an optimum inversion factor of $n_{\mathrm{sp}} = 1$. Raman excess noise is neglected. Four interfering channels, two on each side of the channel of interest, have been considered, with no guard band in-between. The monotoneously rising solid line (——) represents the linear Shannon limit.

is understandable that with increasing SNR (increasing launch power) the nonlinear Shannon capacity, and consequently the associated spectral efficiency displayed in Fig. 2.5(b), reaches a maximum[40,41]. For larger SNR, the SE starts decreasing.

2.3 Modulation

Modulation[42] denotes the method by which an analogue or digital information signal is imprinted onto an (in our case: optical) carrier wave. The simplest modulation would be to switch the carrier (e. g., as provided by a laser) by turning the laser's power supply on and off. For a semiconductor laser this would be the injection current, and it can be switched[43] such that the light pulses follow each other at a bit rate of 40 Gbit/s. For a number of reasons, more elaborate modulation schemes are frequently used, where the laser operates as a continuous-wave (CW) source, and a modulator external to the laser influences the

[40] A. Mecozzi, R.-J. Essiambre: Nonlinear Shannon limit in pseudo-linear coherent systems. J. Lightw. Technol. 30 (2012) 2011–2024

[41] See Ref. 24 on Page 20

[42] In its most general sense, modulation also includes coding to prevent transmission errors from occurring (line coding, channel coding), or to provide means for correcting already occurred transmission errors (error correcting coding, also forward error correction, FEC).

[43] Fraunhofer Heinrich Hertz Institute, Berlin, May 2013.
`http://www.hhi.fraunhofer.de/fileadmin/Lasers/40Gbit-Laser-2013-05.pdf`

light emission. These concepts are explained in the next sections.

Modulation is a fundamentally nonlinear process where two or more temporal signals interact. Here, we concentrate on the interaction of electromagnetic fields only, but many other interactions, for instance with acoustic waves (photon-phonon interaction) are interesting as well. The signals' spectra can be located at widely different or at rather similar centre frequencies. We talk of modulation, if a baseband signal covering a spectral region of, say, $0 \ldots 100\,\mathrm{GHz}$ interacts with a carrier at a widely different frequency of, say, 193.41 THz (vacuum wavelength $1.55\,\mu\mathrm{m}$). We name the process mixing, if spectra interact that are centred at comparable frequencies, say, at 193.41 THz ($1.55\,\mu\mathrm{m}$) and $2 \times 193.41\,\mathrm{THz} = 386.82\,\mathrm{THz}$ ($0.775\,\mu\mathrm{m}$). A number of such interactions (e. g., four-wave mixing) are mentioned in Appendix A.3.3 on Page 177 ff.

The lowest-order nonlinearity which must be involved for an interaction of electromagnetic fields is a product term of the contributing temporal signals. The physical effects which are employed to perform such an action may differ widely, and they could be based on absorption (as with a photodiode, Eq. (1.1) on Page 2) or on lossless parametric effects (as with a nonlinear fibre, Eq. (2.21)on Page 20, Appendix (A.3.3) on Page 177), but the basic action of modulator and mixer are not different by principle.

In electrical engineering the somewhat misleading terms "multiplicative mixing" and "additive mixing" are used. Multiplicative mixing relies on a physical process which actually multiplies two quantities, $s_1(t)\, s_2(t)$. As an example, a voltage $s_1(t)$ could be applied between source and gate of a field effect transistor (FET), and $s_2(t)$ could control the voltage between source and drain[44]. With additive mixing, we first superimpose the two signals, $s_1(t) + s_2(t)$, and then apply a nonlinear operation (e. g., squaring) to the sum,

$$[s_1(t) + s_2(t)]^2 = s_1^2(t) + s_2^2(t) + 2\overbrace{s_1(t)\, s_2(t)}^{\text{mixing}}. \tag{2.34}$$

Obviously, there is a mixing term, namely the product $s_1(t)\, s_2(t)$.

2.3.1 Analytic signals and phasors

For a better understanding, let us disregard wave propagation and consider two real electrical signals $s_{1,2}$ with real amplitudes $\hat{a}_{1,2}$, phases $\varphi_{1,2}$, and angular frequencies $\omega_{1,2} = 2\pi f_{1,2}$,

$$s_1(t) = \hat{a}_1(t) \cos[\omega_1 t + \varphi_1(t)], \qquad s_2(t) = \hat{a}_2(t) \cos[\omega_2 t + \varphi_2(t)]. \tag{2.35}$$

By inspecting the Fourier transform of any real signal $s(t)$ it is to be seen that its spectrum $\breve{s}(f)$ has the symmetry property

$$\breve{s}(f) = \breve{s}^*(-f) \quad \text{if } s(t) \text{ is real.} \tag{2.36}$$

We now introduce the complex analytic time-dependent amplitude $\underline{a}(t)$ with real part $a(t) = \Re\{\underline{a}(t)\}$, modulus $\hat{a}(t)$ and real phase $\varphi(t)$. The spectrum of $\underline{a}(t)$ is causal in the frequency domain *by definition*, and is related to the two-sided spectrum of $a(t)$,

$$\underline{a}(t) = \hat{a}(t)\, \mathrm{e}^{\mathrm{j}\,\varphi(t)}, \quad \breve{a}(f) = \int_{-\infty}^{+\infty} \underline{a}(t)\, \mathrm{e}^{-\mathrm{j}\,2\pi f t}\, \mathrm{d}t, \quad \breve{a}(f<0) = 0, \quad \breve{a}(f>0) = 2\int_{-\infty}^{+\infty} a(t)\, \mathrm{e}^{-\mathrm{j}\,2\pi f t}\, \mathrm{d}t. \tag{2.37}$$

For a positive time dependency, see Eq. (1.4) on Page 8, the analytic signal $\underline{s}(t)$ and its real part $s_r(t)$ as in Eq. (2.35) read

$$\underline{s}(t) = \underline{a}(t)\, \mathrm{e}^{\mathrm{j}\,\omega_0 t}, \qquad \underline{a}(t) = \hat{a}(t)\, \mathrm{e}^{\mathrm{j}\,\varphi(t)}, \qquad s_r(t) = \Re\{\underline{s}(t)\} = \hat{a}(t) \cos[\omega_0 t + \varphi(t)]. \tag{2.38}$$

[44] S. Preu, S. Kim, P. G. Burke, M. S. Sherwin, A. C. Gossard: Multiplicative mixing and detection of THz signals with a field effect transistor. Conf. on Lasers and Electro-Optics (CLEO'12), San Jose (CA), USA, May 8–11, 2012. Paper CTu2B.7

If the modulus of the analytic amplitude $\underline{a}\,(t) = \hat{a}\,\mathrm{e}^{\mathrm{j}\,\varphi(t)}$ is time-independent, it is called a "phasor"[45] (*German* Zeiger). The time evolution of the phasor $\underline{s}\,(t) = \mathrm{e}^{\mathrm{j}\,\omega_0 t}$ is illustrated in Fig. 2.6.

If the phase depends linearly on time, $\varphi\,(t) = \omega_a t$, it is easy to see that the spectrum $\breve{a}\,(f) = \hat{a}\,\delta\,(f - f_a)$ of $\underline{a}\,(t)$ as well as that of the analytic time signal $\underline{s}\,(t)$ is causal,

$$\underline{s}(t) = \hat{a}\,\mathrm{e}^{\mathrm{j}\,\omega_a t}\,\mathrm{e}^{\mathrm{j}\,\omega_0 t} \quad \circ\!\!-\!\!\bullet \quad \breve{s}(f) = \int_{-\infty}^{+\infty} \underline{s}(t)\,\mathrm{e}^{-\mathrm{j}\,2\pi f t}\,\mathrm{d}t = \hat{a}\,\delta\,(f - (f_0 + f_a))\,. \tag{2.39}$$

An analytic signal $\underline{s}\,(t)$ is generated from its real part $s_r(t) = \Re\{\underline{s}(t)\}$ or from its imaginary part $s_i(t) = \Im\{\underline{s}(t)\}$ by a Hilbert transform, i. e., by applying Cauchy's principal value[46] integral, or by a convolution $(*)$ with $1/\,(\pi t)$,

$$s_i(t) = \mathcal{H}\{s_r(t)\} = \frac{1}{\pi}\mathcal{P}\int_{-\infty}^{+\infty}\frac{s_r(t')}{t-t'}\,\mathrm{d}t' = s_r(t) * \frac{1}{\pi t}, \tag{2.40a}$$

$$s_r(t) = \mathcal{H}^{-1}\{s_i\,(t)\} = -\frac{1}{\pi}\mathcal{P}\int_{-\infty}^{+\infty}\frac{s_i(t')}{t-t'}\,\mathrm{d}t' = -s_i(t) * \frac{1}{\pi t}, \tag{2.40b}$$

$$\mathcal{P}\int_{-\infty}^{+\infty}\frac{f(x)}{x-x_0}\,\mathrm{d}x = \lim_{\epsilon\to 0}\left(\int_{-\infty}^{x_0-\epsilon}\frac{f(x)}{x-x_0}\,\mathrm{d}x + \int_{x_0+\epsilon}^{+\infty}\frac{f(x)}{x-x_0}\,\mathrm{d}x\right) = \lim_{\epsilon\to 0}\int_{-\infty}^{+\infty} f(x)\frac{x-x_0}{(x-x_0)^2+\varepsilon^2}\,\mathrm{d}x. \tag{2.40c}$$

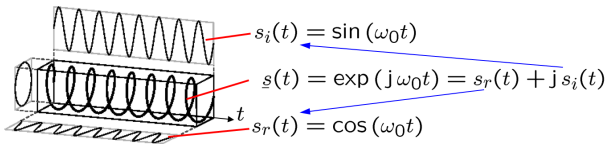


Fig. 2.6. Evolution of an analytic signal $\underline{s}\,(t) = \mathrm{e}^{\mathrm{j}\,\omega_0 t} = s_r\,(t) + \mathrm{j}\,s_i\,(t) = \cos\omega_0 t + \mathrm{j}\sin\omega_0 t$ with time. Projection on the horizontal plane shows the real part $s_r\,(t)$, projection on the vertical plane displays the imaginary part $s_i\,(t)$.

2.3.2 Mixing and modulation

To simplify the setup, we assume in Eq. (2.35) on Page 26 that only the amplitude and phase of $s_2\,(t)$ depend on time, while we choose $\hat{a}_1 = \mathrm{const}_t$ and $\varphi_1 = 0$ for $s_1\,(t)$. The multiplication $s_{12} = s_1 s_2$ (usually understood as a mixing process) results in two signal spectra centred at the difference $f_2 - f_1$ and at the sum frequency $f_1 + f_2$, respectively,

$$\begin{aligned} s_{12}\,(t) &= \frac{\hat{a}_1\hat{a}_2\,(t)}{2}\left\{\cos\left[(\omega_2-\omega_1)\,t + \varphi_2\,(t)\right] + \cos\left[(\omega_1+\omega_2)\,t + \varphi_2\,(t)\right]\right\} \\ &= \frac{\hat{a}_1}{4}\left\{\left[\underline{a}_2\,(t)\,\mathrm{e}^{+\mathrm{j}(\omega_2-\omega_1)t} + \underline{a}_2^*\,(t)\,\mathrm{e}^{-\mathrm{j}(\omega_2-\omega_1)t}\right] + \left[\underline{a}_2\,(t)\,\mathrm{e}^{+\mathrm{j}(\omega_1+\omega_2)t} + \underline{a}_2^*\,(t)\,\mathrm{e}^{-\mathrm{j}(\omega_1+\omega_2)t}\right]\right\}. \end{aligned} \tag{2.41}$$

The Fourier transform of the real function $s_{12}\,(t)$ yields the spectrum $\breve{s}_{12}\,(f) = \breve{s}_{12}^*\,(-f)$,

$$\begin{aligned} \breve{s}_{12}\,(f) = \breve{s}_1\,(f) * \breve{s}_2\,(f) = \frac{\hat{a}_1}{4}\{\ &[\breve{a}_2\,(f - (f_2 - f_1)) + \breve{a}_2^*\,(-f - (f_2 - f_1))] \\ &+ [\breve{a}_2\,(f - (f_1 + f_2)) + \breve{a}_2^*\,(-f - (f_1 + f_2))]\ \}. \end{aligned} \tag{2.42}$$

For the case of modulation and depending on the physical quantity to be modified, we talk of amplitude modulation (AM, $\hat{a}_2\,(t)$), phase modulation (PM, $\varphi_2\,(t)$) or frequency modulation (FM, $\mathrm{d}\varphi_2\,(t)/\,\mathrm{d}t$).

[45] A phasor (*German* Zeiger) is a "vector" in the complex plane. Implicitly we assume that the real part is drawn horizontally and increases to the right, while the imaginary part is drawn vertically and increases in the upward direction. A phasor is displayed as an arrow, pointing from the base point (also named origin, tail, or initial point) to the endpoint (also named tip, head, or final point). The length of the phasor is proportional to its magnitude. The phasor's projection to the horizontal (vertical) axis gives its real (imaginary) part. For an example, see Fig. 2.8(a) on Page 31.

[46] Symbol $\mathcal{P}$ because of *p*rincipal value (Latin *valor principalis*)

Amplitude and phase for each differently polarized wave can be modulated independently, while phase modulation and frequency modulation depend on each other. Note that an amplitude $\hat{a}(t)$ is always a non-negative quantity, $\hat{a}(t) \geq 0$. If the quantity $\hat{a}(t)$, regarded as a multiplicator only, could change its sign, e. g., if $\hat{a}(t) \in \{+1, -1\}$, this would be equivalent to a phase modulation $\varphi(t) = \{0, \pi\}$ for a constant amplitude $\hat{a} = 1$.

Transmission and reception of a complex signal

A physical channel can transmit only physical quantities, i. e., signals that are measurable, e. g., with a voltmeter. The natural choice is then to map these signals to numbers which are real in the mathematical sense. However, at the transmitter, two independent data streams can be regarded as real and imaginary part of a complex signal, and both its real constituents can be transmitted. On reception, real and imaginary parts can be recombined to form a complex number. In this sense a channel is able to transmit also complex data signals.

We start with the real signal $s_r(t)$ from Eq. (2.38). A simple trigonometric manipulation leads to the definition of in-phase signal[47] $I(t)$ and quadrature signal $Q(t)$, which now serve as the representatives of the encoded data instead of amplitude $\hat{a}(t)$ and phase $\varphi(t)$,

$$\begin{aligned} s_r(t) &= \Re\left\{\hat{a}(t)\,\mathrm{e}^{\mathrm{j}\,\varphi(t)}\,\mathrm{e}^{\mathrm{j}\,\omega_0 t}\right\} = \hat{a}(t)\cos\left[\omega_0 t + \varphi(t)\right] = \hat{a}(t)\cos\varphi(t)\cos\omega_0 t - \hat{a}(t)\sin\varphi(t)\sin\omega_0 t \\ &= I(t)\cos(\omega_0 t) - Q(t)\sin(\omega_0 t) \quad \text{for } I(t) = \hat{a}(t)\cos\varphi(t),\ Q(t) = \hat{a}(t)\sin\varphi(t), \qquad (2.43) \\ &\hat{a}(t) = \sqrt{I^2(t) + Q^2(t)}, \qquad \tan\varphi(t) = \frac{Q(t)}{I(t)}. \end{aligned}$$

The naming of $I(t)$ and $Q(t)$ is derived from the fact that the real part $\hat{a}\cos\varphi$ of the complex amplitude $\hat{a}\,\mathrm{e}^{\mathrm{j}\,\varphi}$ is in phase with the carrier phasor[48] $\mathrm{e}^{\mathrm{j}\,\omega_0 t}$ of $\cos(\omega_0 t)$, while the imaginary part $\hat{a}\sin\varphi$ is in phase with the carrier phasor $-\mathrm{j}\,\mathrm{e}^{\mathrm{j}\,\omega_0 t} = \mathrm{e}^{\mathrm{j}(\omega_0 t - \pi/2)}$ of $\sin(\omega_0 t)$ that points at right angles ("is in quadrature") with respect to the phasor of $\cos(\omega_0 t)$, see also Fig. 2.6 on Page 27.

Equation (2.43) provides a simple recipe how to transmit a complex signal: First, take the in-phase component $I(t)$ and let it mix ("multiply it") with a carrier $\cos(\omega_0 t)$. Second, let the quadrature component $Q(t)$ mix with a carrier $\sin(\omega_0 t)$ that lags $\cos(\omega_0 t)$ in phase by 90 °. Third, subtract $Q(t)\sin\omega_0 t$ from $I(t)\cos\omega_0 t$. As a reminder: $I(t)$ and $Q(t)$ are not necessarily non-negative, and therefore this process cannot be named amplitude modulation. The described procedure reproduces the action of a so-called IQ-mixer, which can be used either for modulation or for demodulation.

IQ-mixer

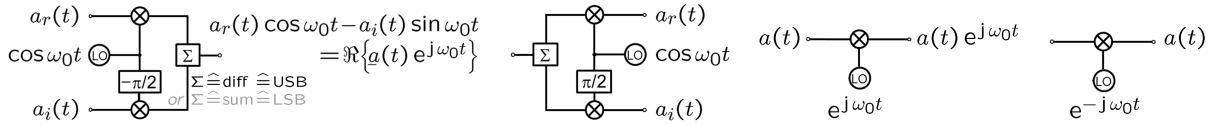
The schematic of an IQ-mixer and of a complex mixer as modulator or demodulator for complex data is shown in Fig. 2.7. The mixers are represented by multiplier symbols $\bigotimes$. Appropriate filters (not drawn) at the mixer outputs select the frequency components of interest. If the mixer symbol in an IQ-demodulator stands for a photodetector, the remarks after Eq. (5.115) on Page 141 have to be observed: In this case and even without filtering, no harmonics of the (optical) carrier appear at the electrical mixer output.

Figure 2.7(a) displays an IQ-modulator for the real part $a_r(t)$ and the imaginary part $a_i(t)$ of a complex band-limited data signal $a(t) = a_r(t) + \mathrm{j}\,a_i(t)$ with spectrum $\breve{a}(|f| > B) = 0$, where $B < f_0$. The local oscillator (LO) supplies the two mixers with orthogonal carriers $+\cos\omega_0 t$ and $+\sin\omega_0 t$. Both modulated carriers are combined (symbol Σ), either by subtraction ($\Sigma \mathrel{\widehat{=}}$ diff $\mathrel{\widehat{=}}$ USB, this is assumed here) or by addition ($\Sigma \mathrel{\widehat{=}}$ sum $\mathrel{\widehat{=}}$ LSB). After subtracting the mixer outputs at node Σ, we have the real part of the complex signal Eq. (2.39) on Page 27 which is then transmitted,

$$s_r(t) = \Re\left\{a(t)\,\mathrm{e}^{\mathrm{j}\,\omega_0 t}\right\} = a_r(t)\cos\omega_0 t - a_i(t)\sin\omega_0 t\,, \qquad a(t) = a_r(t) + \mathrm{j}\,a_i(t)\,. \qquad (2.44)$$

This action is compactly described in Fig. 2.7(c) with a complex mixer and using complex quantities.

[47] Not to be mixed up with an optical intensity or a current that are denoted by the same symbol.
[48] For an explanation of phasors, see Footnote 45 on Page 27



(a) IQ-modulator for encoding real and imaginary data on two orthogonal carriers (Σ ≙ subtract) (b) IQ-demodulator for complex data (Σ ≙ split) (c)Modulation of a complex carrier with complex data (d)Demodulation of complex data

Fig. 2.7. IQ-mixer and complex mixer as modulator or demodulator for complex data. The phase shifter boxes $-\pi/2$ (or $\pi/2$) stand for a phase retardation (or advancement) of the type $\exp[\mathrm{j}(\omega_0 t - \pi/2)]$ (or $\exp[\mathrm{j}(\omega_0 t + \pi/2)]$). Appropriate filters (not drawn) at the mixer outputs select the frequency components of interest. If the mixer symbols in an IQ-demodulator represent photodiodes, no harmonics of the optical carrier appear at the electrical mixer outputs, even without filtering. (a) IQ-modulator for real part $a_r(t)$ and imaginary part $a_i(t)$ of a complex data signal $a(t) = a_r(t) + \mathrm{j}\,a_i(t)$. The local oscillator (LO) supplies the two mixers (real multipliers, symbol ⊗) with orthogonal carriers $+\cos\omega_0 t$ and $+\sin\omega_0 t$. Both modulated carriers are combined (symbol Σ), either by subtraction (Σ ≙ diff ≙ USB, this is assumed here) or by addition (Σ ≙ sum ≙ LSB). If real data $m(t)$ determine a specific analytic signal $\underline{a}(t) = m(t)(\cos\omega_a + \mathrm{j}\sin\omega_a t)$, we can generate single sideband spectra: When subtracting the mixer outputs as assumed in (a), the upper sideband (USB) signal $m_{\mathrm{USB}}(t) = m(t)\cos(\omega_0 + \omega_a)\,t$ results. If the two mixer outputs are added, the lower side band (LSB) is generated, $m_{\mathrm{LSB}}(t) = m(t)\cos(\omega_0 - \omega_a)\,t$. Therefore an IQ-mixer serves also as a single-sideband (SSB) modulator. If $a_r(t)$ and $a_i(t)$ represent independent data, the total spectrum (consisting of USB and LSB) is a superposition of the shifted spectra $\breve{a}_r(f)$ and $\breve{a}_i(f)$. After the superposition at Σ, the real part of a complex signal is transmitted, $\Re\{a(t)\exp(\mathrm{j}\,\omega_0 t)\} = a_r(t)\cos\omega_0 t - a_i(t)\sin\omega_0 t$. (b) IQ-demodulator for recovering a complex data signal $a(t) = a_r(t) + \mathrm{j}\,a_i(t)$ with real part $a_r(t)$ and imaginary part $a_i(t)$, which were modulated on two orthogonal carriers $\cos\omega_0 t$ and $\sin\omega_0 t$, repectively. The incoming signal is split (symbol Σ). The local oscillator (LO) supplies orthogonal carriers $+\cos\omega_0 t$ and $-\sin\omega_0 t$ to the two mixers (real multipliers, symbol ⊗). (c) Complex modulator for encoding a complex data signal $a(t)$ on an analytic carrier $\exp(+\mathrm{j}\,\omega_0 t)$, supplied to the mixer (complex multiplier, symbol ⊗) by a local oscillator (LO). (d) Complex demodulator for recovering a complex data signal $a(t)$, which was modulated on an analytic carrier $\exp(+\mathrm{j}\,\omega_0 t)$. The local oscillator (LO) supplies the complex conjugate carrier $\exp(-\mathrm{j}\,\omega_0 t)$ to the mixer (complex multiplier, symbol ⊗).

Spectrum of IQ-modulator output signal The real IQ-modulator output $s_r(t) = \Re\{a(t)\,\mathrm{e}^{\mathrm{j}\,\omega_0 t}\}$ in Fig. 2.7(a) and Eq. (2.44) has the spectrum

$$\breve{s}_r(f) = \int_{-\infty}^{+\infty} \tfrac{1}{2}\left(a(t)\,\mathrm{e}^{\mathrm{j}\,2\pi f_0 t} + a^*(t)\,\mathrm{e}^{-\mathrm{j}\,2\pi f_0 t}\right)\mathrm{e}^{-\mathrm{j}\,2\pi f t}\,\mathrm{d}t = \tfrac{1}{2}\breve{a}(f - f_0) + \tfrac{1}{2}\breve{a}^*(-(f + f_0))\,. \quad (2.45a)$$

The band-limited baseband spectrum $\breve{a}(|f| > B) = 0$ of the complex modulation signal $a(t)$ in Eq. (2.45a) is shifted to the positive carrier frequency f_0, and in inverted and complex conjugate form also to $-f_0$.

Alternatively, we may use the second form of Eq. (2.44), $s_r(t) = a_r(t)\cos\omega_0 t - a_i(t)\sin\omega_0 t$, and find for the IQ-modulator output spectrum

$$\begin{aligned}\breve{s}_r(f) &= \int_{-\infty}^{+\infty} (a_r(t)\cos\omega_0 t - a_i(t)\sin\omega_0 t)\,\mathrm{e}^{-\mathrm{j}\,2\pi f t}\,\mathrm{d}t = \breve{s}_r^*(-f) \\ &= \tfrac{1}{2}\big(\breve{a}_r(f - f_0) + \mathrm{j}\,\breve{a}_i(f - f_0)\big) + \tfrac{1}{2}\big(\breve{a}_r(f + f_0) - \mathrm{j}\,\breve{a}_i(f + f_0)\big). \end{aligned} \quad (2.45b)$$

By comparing Eq. (2.45a) and (2.45b) for positive and negative frequency shifts, we find

$$\breve{a}(f - f_0) = \breve{a}_r(f - f_0) + \mathrm{j}\,\breve{a}_i(f - f_0) \;\neq\; \breve{a}^*(-f + f_0) = \breve{a}_r(f - f_0) - \mathrm{j}\,\breve{a}_i(f - f_0)\,, \quad (2.45c)$$

$$\breve{a}^*(-(f + f_0)) = \breve{a}_r(f + f_0) - \mathrm{j}\,\breve{a}_i(f + f_0) \;\neq\; \breve{a}(f + f_0) = \breve{a}_r(f + f_0) + \mathrm{j}\,\breve{a}_i(f + f_0)\,. \quad (2.45d)$$

Naturally, we have $\breve{a}(f - f_0) \neq \breve{a}^*(-(f + f_0))$, because $\breve{a}(f - f_0)$ belongs to a *complex* time signal $a(t)$.

The baseband spectra $\breve{a}_{r,i}(f) = \breve{a}_{r,i}^*(-f)$ of the real signals $a_{r,i}(t)$ have a *lowpass* bandwidth $B < f_0$ and comprise correlated positive and negative frequency components in a range $-B < f \leq +B$. After shifting these spectra to the respective carrier frequencies $\pm f_0$, the *passband* spectra of real and imaginary part $\breve{a}_{r,i}(f \mp f_0)$ span a range $-B \pm f_0 < f \leq +B \pm f_0$ *and overlay each other.*

It is obvious that the composite spectra $\breve{a}(f - f_0) = \breve{a}_r(f - f_0) + \mathrm{j}\,\breve{a}_i(f - f_0)$ and $\breve{a}^*(-(f + f_0)) = \breve{a}_r(f + f_0) - \mathrm{j}\,\breve{a}_i(f + f_0)$ likewise span a range of $2B$ centred at $\pm f_0$, but cannot be separated simply in contributions belonging to $\breve{a}_r(f \mp f_0)$ and $\breve{a}_i(f \mp f_0)$. Incoherent square-law detection would not not help. Instead, we must rely on the fact that $\pm\mathrm{j}\,\breve{a}_i(f \pm f_0)$ and $\breve{a}_r(f \mp f_0)$ are orthogonal to each other (all phases are shifted by $\pi/2$), a property which can be exploited with an IQ-demodulator Fig. 2.7(b) that operates with orthogonal LO signals, see Eq. (2.47).

Single-sideband modulation Let a real data signal $m(t)$ modulate the subcarrier $\mathrm{e}^{\mathrm{j}\,\omega_a t}$ having an angular frequency ω_a. The spectral width B_m of $\breve{m}(f)$ is assumed to be limited to $B_m < f_a$. Then the real part $a_r(t)$ and the imaginary part $a_i(t)$ of the analytic signal $\underline{a}(t) = m(t)\,\mathrm{e}^{\mathrm{j}\,\omega_a t}$ are related by a Hilbert transform Eq. (2.40a) on Page 27.

When an IQ-modulator Fig. 2.7(a) is fed with $a_r(t) = m(t)\cos\omega_a t$ and $a_i(t) = m(t)\sin\omega_a t$, a single-sideband (SSB) spectrum is generated: After subtracting the mixer outputs as assumed in Fig. 2.7(a), only the upper sideband (USB) signal $m_{\mathrm{USB}}(t) = m(t)(\cos\omega_a t\cos\omega_0 t - \sin\omega_a t\sin\omega_0 t) = m(t)\cos(\omega_0+\omega_a)t$ appears at the output, because the lower sidebands cancel.

If the two mixer outputs are added, the lower side band (LSB) is generated, because the upper sidebands cancel, $m_{\mathrm{LSB}}(t) = m(t)(\cos\omega_a t\cos\omega_0 t + \sin\omega_a t\sin\omega_0 t) = m(t)\cos(\omega_0-\omega_a)t$. The modulated subcarrier $\underline{a}(t)$ and its (causal) spectrum $\breve{a}(f)$ as well as the generated (non-causal) USB spectrum $\breve{m}_{\mathrm{USB}}(f)$ and LSB spectrum $\breve{m}_{\mathrm{LSB}}(f)$ can be written as

$$\underline{a}(t) = a_r(t) + \mathrm{j}\,a_i(t) = m(t)(\cos\omega_a t + \mathrm{j}\sin\omega_a t) \quad \text{for } m(t) \text{ real}, \tag{2.46a}$$

$$\breve{a}(f) = \breve{m}(f - f_a) \quad \text{causal if spectral width of } \breve{m}(f) \text{ is } B_m < f_a, \tag{2.46b}$$

$$\breve{m}_{\mathrm{USB}}(f) = \tfrac{1}{2}\left[\breve{m}(f-(f_0+f_a)) + \breve{m}(f+(f_0+f_a))\right], \tag{2.46c}$$

$$\breve{m}_{\mathrm{LSB}}(f) = \tfrac{1}{2}\left[\breve{m}(f-(f_0-f_a)) + \breve{m}(f+(f_0-f_a))\right]. \tag{2.46d}$$

If in contrast to the assumption in Eq. (2.46a) the quantities $a_r(t)$ and $a_i(t)$ represent independent data, the associated USB and LSB spectra are also independent and cannot cancel, see Eq. (2.45c).

IQ-demodulator An IQ-demodulator is seen in Fig. 2.7(b). It recovers a complex data signal $a(t) = a_r(t) + \mathrm{j}\,a_i(t)$ with real part $a_r(t)$ and imaginary part $a_i(t)$, which were modulated on two orthogonal carriers $\cos\omega_0 t$ and $\sin\omega_0 t$. The incoming signal is split (symbol Σ). The local oscillator (LO) supplies orthogonal carriers $\cos\omega_0 t$ and $-\sin\omega_0 t$ to the two mixers, the in-phase (I) and quadrature outputs (Q) of which are

$$2I(t) = \quad 2\left[a_r(t)\cos\omega_0 t - a_i(t)\sin\omega_0 t\right]\cos\omega_0 t = a_r(t)(1+\cos 2\omega_0 t) - a_i(t)\sin 2\omega_0 t\,, \tag{2.47a}$$

$$2Q(t) = -2\left[a_r(t)\cos\omega_0 t - a_i(t)\sin\omega_0 t\right]\sin\omega_0 t = a_i(t)(1-\cos 2\omega_0 t) - a_r(t)\sin 2\omega_0 t\,. \tag{2.47b}$$

When filters remove the carrier harmonics at $2f_0$ (or if they are not generated from the beginning, in case the mixers are realized by photodetectors), the receiver recovers the transmitted signals,

$$2I(t) = a_r(t)\,, \qquad 2Q(t) = a_i(t)\,. \tag{2.47c}$$

The schematic Fig. 2.7(d) has the same functionality, but uses a complex mixer and complex quantities for convenience. It is important to note that on reception the complex conjugate of the transmitting carrier serves as a LO, otherwise the quadrature component changes sign, $2Q(t) = -a_i(t)$.

If a different frequency ω_0' was chosen for the receiver's LO, the data spectrum would be located at an intermediate (difference) frequency (IF, *German* Zwischenfrequenz) $\omega_Z = \omega_0 - \omega_0'$. The harmonics $2\omega_0$ would then be replaced by the angular sum frequency $\omega_0 + \omega_0'$.

Homodyne and heterodyne reception

The type of reception as discussed in Fig. 2.7, where a receiver LO has the same frequency as and is (implicitly) phase-locked to the transmitter, is called homodyne[49] reception. The transmitted signal is directly transfered to the baseband. If transmitter and LO frequencies differ, $f_0 - f_0' \neq 0$, we speak of heterodyne[50] reception. Details will be discussed in Sect. 5.4 on Page 140 ff.

[49] From Greek *ὁμός*, same, like, similar, and Greek *δύναμις*, force, power, strength. — Homodyning requires the LO to have the same frequency as the transmitted carrier, and a fixed phase relation with it.

[50] From Greek *ἕτερος*, different, and Greek *δύναμις*, force, power, strength. — Heterodyning is a radio signal processing technique invented in 1901 by Canadian inventor-engineer Reginald Fessenden, in which new frequencies are created by combining or mixing two frequencies. Heterodyning is useful for frequency shifting signals into a new frequency range, and is also

2.4 Modulation formats

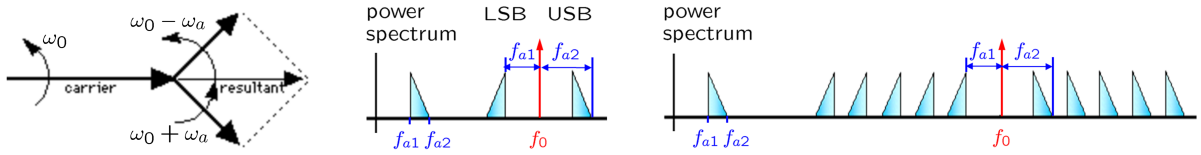
In this section we review a number of important modulation formats, starting with simple analogue amplitude modulation (AM) and ending with advanced digital quadrature amplitude modulation (QAM).

2.4.1 Analogue modulation formats

Analogue modulation formats were the first to be used in early wireline and wireless transmission. We present amplitude, intensity and angle modulation, and inspect (vestigial) single-sideband modulation. Figure 2.11(a) on Page 36 visualizes the temporal signal shapes for some of these modulation formats. The figure refers in addition to analogue polarization-mode modulation (PolM), where the state of polarization of an electromagnetic wave is modulated to carry information. Details on the mathematical description are omitted here.

Amplitude modulation

As the naming suggests, amplitude modulation (AM) modifies the amplitude of an analytic carrier $\mathrm{e}^{\mathrm{j}\,\omega_0 t}$ with a modulation signal $m\,(t)$, which is assumed to be positive real and varies slowly on the scale of



(a) Phasor diagram for a sinusoidal AM. The resultant is in phase with the carrier.

(b) One-sided power spectrum for AM with a baseband spectrum $f_{a1} \ldots f_{a2}$

(c) One-sided power spectrum for IM. Power detection recovers the baseband spectrum $f_{a1} \ldots f_{a2}$. In contrast to (b), the IM-spectrum is infinitely extended.

Fig. 2.8. Analogue amplitude modulation (AM) and intensity modulation (IM) with real modulation signals. (a) Phasors in the complex plane (see Footnote 45 on Page 27, origin located at the base point of the carrier phasor) for sinusoidal AM according to Eq. (2.49a). All phasors rotate counter-clockwise (ccw, in the mathematical positive sense). The carrier phasor, the upper sideband (USB) phasor and the lower sideband (LSB) phasor rotate with angular velocities ω_0, $\omega_0 + \omega_a$ and $\omega_0 - \omega_a$, respectively. If the observer rotates with the carrier, the USB phasor would seemingly rotate ccw with angular velocity ω_a, while the LSB phasor would seemingly rotate clockwise (cw) with the same angular velocity ω_a. (b) One-sided schematic AM power spectrum for a non-sinusoidal baseband modulation spectrum extending from frequency f_{a1} to f_{a2}. Upper and lower sidebands are related by $\mathrm{USB}\,(f) = \mathrm{LSB}^*\,(-f)$, see Eq. (2.48d), (2.48e) (c) One-sided schematic IM spectrum for a non-sinusoidal baseband modulation spectrum extending from frequency f_{a1} to f_{a2}. Because the *power* $\langle s^2_{\mathrm{IM}\,r} \rangle\,(t)$ is modulated in proportion to a modulation signal $1 + p_m\,(t)$, the *amplitude* depends on the square-root of the modulating signal $\sqrt{1 + p_m\,(t)}$, so that the spectrum is infinitely extended.

the carrier period $1/\,f_0$. Its spectrum obeys the symmetry condition $\breve{m}(f) = \breve{m}^*\,(-f)$ as in Eq. (2.36) on Page 26. Modulated analytic carrier $\underline{s}_{\mathrm{AM}}\,(t)$ and its real part $s_{\mathrm{AM}\,r}\,(t)$ ar written as

$$\underline{s}_{\mathrm{AM}}\,(t) = \hat{a}\,m\,(t)\,\mathrm{e}^{\mathrm{j}\,\omega_0 t}, \quad m\,(t) \geq 0 \text{ is real and band-limited, } \breve{m}\,(|f| > B) = 0,\ B < f_0\,, \tag{2.48a}$$

$$s_{\mathrm{AM}\,r}\,(t) = \hat{a}\,m\,(t)\cos\omega_0 t = \tfrac{1}{2}\hat{a}\,m\,(t)\left(\mathrm{e}^{\mathrm{j}\,\omega_0 t} + \mathrm{e}^{-\mathrm{j}\,\omega_0 t}\right). \tag{2.48b}$$

involved in the processes of modulation and demodulation. The two frequencies are combined in a nonlinear signal-processing device such as a vacuum tube, transistor, or diode, usually called a mixer. In the most common application, two signals at frequencies f_1 and f_2 are mixed, creating two new signals, one at the sum $f_1 + f_2$ of the two frequencies, and the other at the difference $f_2 - f_1$. These new frequencies are called heterodynes. Typically only one of the new frequencies is desired, and the other signal is filtered out of the output of the mixer. [Definition cited after `http://en.wikipedia.org/wiki/Heterodyne`]

The spectrum $\breve{s}_{\mathrm{AM}\,r}(f)$ of the real part $s_{\mathrm{AM}\,r}(t)$ of the modulated analytic carrier $\underline{s}_{\mathrm{AM}}(t)$, and the shifted modulation spectra $\breve{m}(f-f_0)$, $\breve{m}(f+f_0)$ of the positive real band-limited modulation signal $m(t)$ are

$$\breve{s}_{\mathrm{AM}\,r}(f) = \tfrac{1}{2}\hat{a}\left[\breve{m}(f-f_0)+\breve{m}(f+f_0)\right] = \breve{s}^*_{\mathrm{AM}\,r}(-f)\,, \tag{2.48c}$$

$$\begin{aligned}\breve{m}(f-f_0) &= \breve{m}(f-f_0)|_{0<f<+f_0} + \breve{m}(f-f_0)|_{f>+f_0}\\ &= \underbrace{\breve{m}^*(-f+f_0)|_{0<f<+f_0}}_{\mathrm{LSB}\,(f-f_0)} + \underbrace{\breve{m}(f-f_0)|_{f>+f_0}}_{\mathrm{USB}\,(f-f_0)} = \breve{m}^*(-f+f_0)\,,\end{aligned} \tag{2.48d}$$

$$\begin{aligned}\breve{m}(f+f_0) &= \breve{m}(f+f_0)|_{f<-f_0} + \breve{m}(f+f_0)|_{0>f>-f_0}\\ &= \underbrace{\breve{m}(f+f_0)|_{f<-f_0}}_{\mathrm{LSB}\,(f+f_0)} + \underbrace{\breve{m}^*(-f-f_0)|_{0>f>-f_0}}_{\mathrm{USB}\,(f+f_0)} = \breve{m}^*(-f-f_0)\,.\end{aligned} \tag{2.48e}$$

Upper sideband (USB) and lower sideband (LSB) are related and carry the same information as can be seen from Eq. (2.48d), (2.48e), $\mathrm{USB}\,(f\mp f_0) = \mathrm{LSB}^*(-f\pm f_0)$.

For definiteness, we now assume a real sinusoidal modulation $m(t) = 1 + m\cos\omega_a t$ with angular frequency $\omega_a = 2\pi f_a$ and a constant modulation index $0 < m < 1$. For the modulated analytic signal $\underline{s}_{\mathrm{AM}}(t)$, its real part $s_{\mathrm{AM}\,r}(t)$ and the one-sided power spectrum $2\left|\breve{s}_{\mathrm{AM}\,r}(f)\right|^2 = 2\Theta_{s\,\mathrm{AM}}(f)$ we find

$$\begin{aligned}\underline{s}_{\mathrm{AM}}(t) &= \hat{a}\,m(t)\,\mathrm{e}^{\mathrm{j}\,\omega_0 t} = \hat{a}\,(1+m\cos\omega_a t)\,\mathrm{e}^{\mathrm{j}\,\omega_0 t}, \qquad 0<m(t)<1,\\ &= \hat{a}\left[1+\tfrac{1}{2}m\left(\mathrm{e}^{\mathrm{j}\,\omega_a t}+\mathrm{e}^{-\mathrm{j}\,\omega_a t}\right)\right]\mathrm{e}^{\mathrm{j}\,\omega_0 t} = \hat{a}\left[\mathrm{e}^{\mathrm{j}\,\omega_0 t}+\tfrac{1}{2}m\left(\mathrm{e}^{\mathrm{j}(\omega_0+\omega_a)t}+\mathrm{e}^{\mathrm{j}(\omega_0-\omega_a)t}\right)\right],\end{aligned} \tag{2.49a}$$

$$\begin{aligned}s_{\mathrm{AM}\,r}(t) &= \hat{a}\,(1+m\cos\omega_a t)\cos\omega_0 t\\ &= \hat{a}\left[\cos\omega_0 t+\tfrac{1}{2}m\left(\cos(\omega_0-\omega_a)\,t+\cos(\omega_0+\omega_a)\,t\right)\right],\end{aligned} \tag{2.49b}$$

$$\begin{aligned}2\Theta_{s\,\mathrm{AM}}(f) &:= 2\left|\breve{s}_{\mathrm{AM}\,r}(f)\right|^2\\ &= \tfrac{1}{2}\hat{a}^2\Big\{\delta(f-f_0)+\tfrac{1}{4}m^2\big[\underbrace{\delta(f-(f_0-f_a))}_{\text{“LSB”}}+\underbrace{\delta(f-(f_0+f_a))}_{\text{“USB”}}\big]\Big\} \quad \text{for } f>0\,.\end{aligned} \tag{2.49c}$$

Figure 2.8(a) displays the phasors of Eq. (2.49a). A schematic (one-sided) power spectrum similar to Eq. (2.49c), but for a non-sinusiodal modulation spectrum extending from frequency f_{a1} to f_{a2}, is to be seen in Fig. 2.8(b). The AM carrier at frequency f_0 contributes a minimum of $\frac{1}{1+2\times(1/4)} = \frac{2}{3}$ of the total spectral power for a maximum modulation index $m = 1$, Eq. (2.49c). This transmitter power could be saved if the carrier is suppressed and re-supplied at the receiver for detection.

Carrier-suppressed double-sideband modulation

With the carrier suppressed, the modulation function is $m(t) = m\cos\omega_a t$ in the case of sinusoidal modulation. This cannot be called AM any more, because $-1 \le m(t) \le +1$ holds as opposed to the requirement Eq. (2.49a). Instead, we talk of carrier-suppressed double-sideband (CS-DSB) modulation. Because only the sidebands remain, the corresponding time function results from the superposition

$$s_{\text{CS-DSB}\,r}(t) = \tfrac{1}{2}\hat{a}\,m\left(\cos(\omega_0-\omega_a)\,t+\cos(\omega_0+\omega_a)\,t\right) = \hat{a}\,m\cos\omega_a t\cos\omega_0 t\,, \tag{2.50}$$

which resembles $s_{\mathrm{AM}\,r}(t)$ of Eq. (2.49b). At the zeros of the modulation function $m(t)$ the phase of the carrier $\hat{a}\cos\omega_0 t$ jumps by π. Such a linear superposition of signals with different frequencies is called a beat signal.

Intensity modulation

Intensity modulation (IM) modifies the *intensity* (or the power) of a carrier, not its amplitude as with AM. The time-dependent power results from an average $\langle\cdot\rangle$ over a few carrier periods. The positive real

modulation signal $m(t) = \sqrt{p_m(t)}$ is assumed to vary slowly on the scale of a carrier period $1/f_0$. The intensity-modulated signal $s_{\mathrm{IM}\,r}(t)$ along with its modulated intensity $\langle s^2_{\mathrm{IM}\,r}\rangle(t)$ then reads

$$s_{\mathrm{IM}\,r}(t) = \hat{a}\sqrt{p_m(t)}\cos\omega_0 t, \quad \langle s^2_{\mathrm{IM}\,r}\rangle(t) = \tfrac{1}{2}\hat{a}^2 p_m(t), \quad \text{slowly varying positive real } p_m(t). \tag{2.51}$$

For a sinusoidal intensity modulation $p(t) = 1 + p_m\cos\omega_a t$ with $\omega_a \ll \omega_0$ and a small modulation index $p_m \ll 1$, the modulated signal $s_{\mathrm{IM}\,r}(t)$ can be expanded in a series,

$$\begin{aligned}
s_{\mathrm{IM}\,r}(t) &= \sqrt{1+p_m\cos(\omega_a t)}\,\hat{a}\cos(\omega_0 t)\\
&\approx \left\{1+\tfrac{p_m}{2}\cos(\omega_a t) - \tfrac{p_m^2}{8}\cos^2(\omega_a t) + \ldots\right\}\hat{a}\cos(\omega_0 t)\\
&\approx \quad \left\{1-\tfrac{p_m^2}{16}+\ldots\right\}\hat{a}\ \cos\ (\omega_0 t)\\
&\quad +\left\{\tfrac{p}{4}+\tfrac{3p_m^3}{128}+\ldots\right\}\hat{a}\left\{\cos\left[(\omega_0 - \ \omega_a)\,t\right] + \cos\left[(\omega_0 + \ \omega_a)\,t\right]\right\}\\
&\quad +\left\{-\tfrac{p_m^2}{32}+\ldots\right\}\hat{a}\left\{\cos\left[(\omega_0 - 2\omega_a)\,t\right] + \cos\left[(\omega_0 + 2\omega_a)\,t\right]\right\}\\
&\quad +\left\{\tfrac{p_m^3}{128}+\ldots\right\}\hat{a}\left\{\cos\left[(\omega_0 - 3\omega_a)\,t\right] + \cos\left[(\omega_0 + 3\omega_a)\,t\right]\right\} + \ldots
\end{aligned} \tag{2.52}$$

A schematic (one-sided) power spectrum $2\langle|\check{s}_{\mathrm{IM}\,r}(f)|^2\rangle$ of a non-sinusiodal modulation spectrum extending from frequency f_{a1} to f_{a2} is to be seen in Fig. 2.8(c). Basically, the spectrum is infinitely extended. If no frequency-dependent time or phase delays modify the partial spectra differently during transmission (due to, e. g., chromatic dispersion in a fibre), the receiver's photodetector current Eq. (1.1) on Page 2 exactly recovers the IM in the photocurrent, $i(t) \sim \langle s^2_{\mathrm{IM}\,r}\rangle(t) = \frac{1}{2}\hat{a}^2(1+p_m\cos\omega_a t)$.

Angle modulation

If the angle of a carrier phasor $\hat{a}\,\mathrm{e}^{\mathrm{j}\,\omega_0 t}$ is changed, we talk of angle modulation. For definiteness, we assume again a real sinusoidal modulation $\eta(t) = \eta\sin\omega_a t$ with angular frequency $\omega_a = 2\pi f_a$ and a constant angle modulation index η,

$$\underline{s}_{\mathrm{PM}}(t) = \hat{a}\,\mathrm{e}^{\mathrm{j}[\omega_0 t+\eta(t)]} = \hat{a}\,\mathrm{e}^{\mathrm{j}(\omega_0 t+\eta\sin\omega_a t)} = \hat{a}\sum_{n=-\infty}^{+\infty}\mathrm{J}_n(\eta)\,\mathrm{e}^{\mathrm{j}[\omega_0+n\omega_a]t}, \quad \mathrm{J}_{-n}(\eta) = (-1)^n\,\mathrm{J}_n(\eta), \tag{2.53a}$$

$$\begin{aligned}
s_{\mathrm{PM}}(t) = \Re\{\underline{s}_{\mathrm{PM}}(t)\} = \hat{a}\big[&\,\mathrm{J}_0(\eta)\cos\omega_0 t - \mathrm{J}_1(\eta)\left[\sin(\omega_0+\omega_a)\,t + \sin(\omega_0-\omega_a)\,t\right]\\
&- \mathrm{J}_2(\eta)\left[\cos(\omega_0+2\omega_a)\,t + \cos(\omega_0-2\omega_a)\,t\right] + \mathrm{J}_3(\eta)\left[\sin(\omega_0+3\omega_a)\,t + \sin(\omega_0-3\omega_a)\,t\right] \pm \ldots\big].
\end{aligned} \tag{2.53b}$$

The exponential can be expanded in terms of Bessel functions[51] $\mathrm{J}_n(\eta)$ of the first kind and order n. Remarkably, the Bessel functions of negative odd order n have the opposite sign of their companions with positive order, $\mathrm{J}_{-n}(\eta) = (-1)^n\,\mathrm{J}_n(\eta)$.

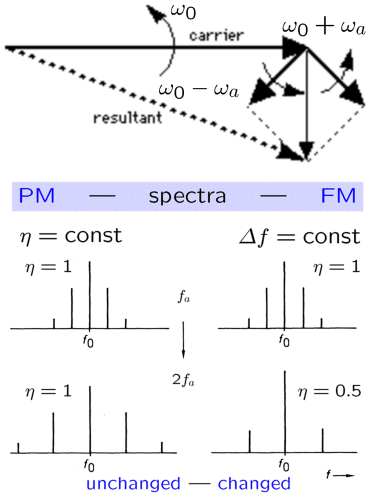
For small-signal angle modulation, where $\eta \ll 1$ holds, the expansion Eq. (2.53a) reduces to three Bessel terms that can be further simplified[52] to resemble the case of AM, Eq. (2.49a) on Page 32 and Fig. 2.8(a),

$$\begin{aligned}
\underline{s}_{\mathrm{PM}}(t) &\approx \hat{a}\left[\mathrm{J}_0(\eta)\,\mathrm{e}^{\mathrm{j}\,\omega_0 t} + \mathrm{J}_1(\eta)\left(\mathrm{e}^{\mathrm{j}(\omega_0+\omega_a)t} - \mathrm{e}^{\mathrm{j}(\omega_0-\omega_a)t}\right)\right]\\
&\approx \hat{a}\left[\mathrm{e}^{\mathrm{j}\,\omega_0 t} + \tfrac{1}{2}\eta\left(\mathrm{e}^{\mathrm{j}(\omega_0+\omega_a)t} - \mathrm{e}^{\mathrm{j}(\omega_0-\omega_a)t}\right)\right] \qquad \text{for } \eta \ll 1.
\end{aligned} \tag{2.54}$$

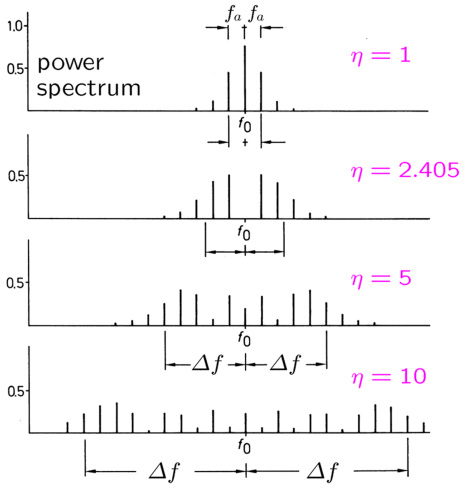
The top of Fig. 2.9(a) displays the associated phasor diagram. Compared to Fig. 2.8(a) on Page 31, the lower sideband phasor $\mathrm{e}^{\mathrm{j}(\omega_0-\omega_a)t}$ is reversed in sign. This corresponds to the fact that for a cos-carrier the $\mathrm{J}_1(\eta)$-terms in Eq. (2.53b) have a sin-dependency for angle modulation, while we see cos-dependencies for the AM sidebands in Eq. (2.49b) on Page 32. Note that for small-signal angle modulation with $\eta \ll 1$ the resultant does not change its length significantly, so that only the angle $\eta(t)$ varies periodically. Naturally,

[51] Abramowitz, M.; Stegun, I. A. (Ed.): Handbook of mathematical functions, 9. Ed. New York: Dover Publications 1970. Chapter 9

[52] See Ref. 51, Eq. (9.1.10), (9.1.12): $\lim_{\eta\to 0}\mathrm{J}_0(\eta) = 1$, and $\lim_{\eta\to 0}\mathrm{J}_1(\eta) = \frac{1}{2}\eta$



(a) Sinusoidal small-signal angle modulation [top] and power spectra for PM and FM [bottom]



(b) Wideband FM for different modulation indices η. The carrier disappears for $\eta = 2.405$. The bandwidth is of order $2\Delta f = 2\eta f_a$.

Fig. 2.9. Angle modulation. Small-signal phasor diagram and one-sided power spectra. The height of the lines represents the area of the associated spectral δ-functions. (a)-[top] Phasors for small-signal sinusoidal angle modulation with $\eta \ll 1$. All phasors rotate counter-clockwise (ccw). The carrier phasor, the upper sideband (USB) phasor and the lower sideband (LSB) phasor rotate with angular velocities ω_0, $\omega_0 + \omega_a$ and $\omega_0 - \omega_a$, respectively. If the observer rotates with the carrier, the USB phasor would seemingly rotate ccw with angular velocity ω_a, while the LSB phasor would seemingly rotate clockwise (cw) with the same angular velocity ω_a. (a)-[bottom] One-sided power spectra for phase modulation (PM) with constant modulation index $\eta = \text{const}$, and for frequency modulation (FM) with constant frequency peak deviation $\Delta f = \eta f_a$. (b) One-sided power spectra for sinusoidal frequency modulation (FM) with different modulation indeces η. The carrier disappears for $\eta = 2.405$. A bandwidth estimate for wideband angle modulation and a fixed modulating signal bandwidth $B = f_{a\,\max}$ is $B_{\text{angle}} = 2\,(\eta + 2)\,B = 2\Delta f + 4B$.

for large-signal angle modulation and taking sufficiently many terms of the expansion Eq. (2.53) into account, the amplitude of the angle-modulated carrier does not change at all.

A sinusoidal angle modulation can be interpreted either as a phase modulation as in Eq. (2.53a), or as a frequency modulation (FM), because the instantaneous frequency is $\mathrm{d}\eta\,(t)/\,\mathrm{d}t = \eta\,\omega_a \cos\omega_a t$. Introducing the frequency peak deviation $\Delta\omega = 2\pi\Delta f = \eta\,\omega_a$, i. e., the maximum deviation of the instantaneous frequency from the carrier frequency f_0, we write the FM signal

$$\underline{s}_{\text{FM}}\,(t) = \hat{a}\,\mathrm{e}^{\,\mathrm{j}\left(\omega_0 + \frac{\mathrm{d}\eta(t)}{\mathrm{d}t}\right)t} = \hat{a}\,\mathrm{e}^{\,\mathrm{j}(\omega_0 + \Delta\omega\cos\omega_a t)t}, \qquad \Delta\omega = 2\pi\Delta f = \eta\,\omega_a. \tag{2.55}$$

The bottom of Fig. 2.9(a) compares the (one-sided) power spectra of sinusoidal PM and FM signals. The length of the vertical lines represents the area of spectral δ-functions. The spectral lines are equidistantly spaced by the modulation frequency f_a. If for $\eta = \text{const}$ the modulation frequency is doubled, $f_a \to 2f_a$, the spectrum retains its shape (but the line separation doubles). This is characteristic for PM. However, if the frequency peak deviation $\Delta\omega = \eta\,\omega_a = \text{const}$ is kept constant while doubling the modulation frequency, $f_a \to 2f_a$, the spectrum changes its shape because $\eta \to \frac{1}{2}\eta$. This is typical for FM.

A one-sided FM power spectrum for a sinusoidal modulation with varying modulation index η is displayed in Fig. 2.9(b). For $\eta = j_{0,1} = 2.405$ the zeroth-order Bessel function has its first zero[53], and the carrier $\mathrm{J}_0\,(j_{0,1}) = 0$ disappears. This fact can be used for determining the associated modulation index $\eta = 2.405$ experimentally. Similarly, from the power ratio of any two lines in the spectrum, e. g., from measuring the ratio $[\mathrm{J}_1\,(\eta)/\,\mathrm{J}_0\,(\eta)]^2$, the modulation index η can be found.

The larger η grows, the wider the significant portion of the (infinitely extended) spectrum becomes. If "significant" means that more than 99 % of the total spectral power is included, then Bessel terms up

[53] See Ref. 51, Table 9.5. The first zero of the zeroth-order Bessel function $\mathrm{J}_0\,(\eta)$ is at $\eta = j_{0,1} = 2.404\,825\,557\,7$.

to the order $|n_{\max}| = \eta + 2$ have to be taken care of[54] if $\eta \leq 50$. For a maximum modulation signal bandwidth $B = f_{a\,\max}$ we then find a significant bandwidth for an angle-modulated signal along with Carson's rule[55] that includes 98 % of the total spectral power

$$B_{\text{angle}}^{(99\,\%)} = 2\,(\eta + 2)\,B \quad \text{for spectral power} > 99\,\%, \text{ modulation bandwidth } B, \text{ and } \eta \leq 50\,, \tag{2.56a}$$

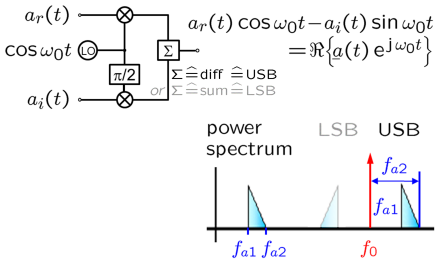
$$B_{\text{angle}}^{(98\,\%)} = 2\,(\eta + 1)\,B \quad \text{for spectral power} > 98\,\%\,. \tag{2.56b}$$

The significant bandwidth of a frequency modulated signal is of the order of double the frequency deviation $2\Delta f = 2\eta B$ as defined in Eq. (2.55). These limits are marked in Fig. 2.9.

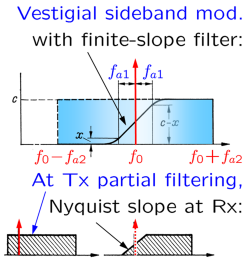
Single-sideband generation and vestigial sideband filtering

For real modulation signals $m\,(t)$ we saw in Eq. (2.48), (2.49) on Page 31 that upper sideband (USB) and lower sideband (LSB) contain identical information, because $\breve{m}\,(f) = \breve{m}^*\,(-f)$ holds and $\text{USB}\,(f) = \text{LSB}^*\,(-f)$, see Eq. (2.36) on Page 26. For improving the practical spectral efficiency C'_{pract} in Eq. (2.25) on Page 22), one of the sidebands and even the carrier could be suppressed. This would be of advantage because the high carrier power could cause nonlinearities in a fibre channel.

Single-sideband generation Single-sideband (SSB) generation can be achieved with an IQ-modulator as described in Eq. (2.46) on Page 30. A graphical illustration is seen in Fig. 2.10(a). An IQ-modulator Fig. 2.10(a)-[top] with LO frequency f_0 as in Fig. 2.7(a) on Page 29 is fed with input quantities $a_r\,(t)$ and $a_i\,(t)$ that are real and imaginary part of a modulated analytic signal $\underline{a}\,(t) = m\,(t)\,(\cos\omega_a t + \mathrm{j}\sin\omega_a t)$, see Eq. (2.46a) on Page 30. If Σ means subtraction ($\Sigma \mathrel{\widehat{=}} \text{diff} \mathrel{\widehat{=}} \text{USB}$), the upper sideband USB is generated.



(a) IQ-modulator for single-sideband (SSB) generation (USB, $\Sigma \mathrel{\widehat{=}} \text{diff} \mathrel{\widehat{=}} \text{subtract}$) [top], and one-sided power spectrum [bottom]



(b) Vestigial sideband (VSB) filtering suppresses redundant spectral parts.

Fig. 2.10. Analogue single-sideband (SSB) generation with an IQ-modulator, and with vestigial sideband (VSB) filtering. (a)-[top] IQ-modulator of Fig. 2.7(a) on Page 29 for SSB generation. The modulator input quantities $a_r\,(t)$ and $a_i\,(t)$ are real and imaginary part of an analytic signal $\underline{a}\,(t) = m\,(t)\,(\cos\omega_a t + \mathrm{j}\sin\omega_a t)$, see Eq. (2.46a) on Page 2.46a. The LO frequency is f_0. (a)-[bottom] The power spectrum $|\breve{a}\,(f)|^2$ extends from f_{a1} to f_{a2}. If Σ means subtraction ($\Sigma \mathrel{\widehat{=}} \text{diff} \mathrel{\widehat{=}} \text{USB}$), the upper sideband USB is generated. If Σ means addition ($\Sigma \mathrel{\widehat{=}} \text{sum} \mathrel{\widehat{=}} \text{LSB}$), the lower sideband LSB results. The carrier is suppressed. For demodulation, a LO at frequency f_0 must be added at the receiver. (b) Vestigial sideband filtering. (b)-[top] If most of the, e. g., LSB is cut off by a transmitter (Tx) filter, the spectral width is reduced as in (b)-[bottom left], but the information is preserved in the USB. As a consequence, the practical spectral efficiency C'_{pract} in Eq. (2.25) on Page 22) increases considerably. (b)-[bottom right] At the receiver, a filter with a so-called Nyquist slope weighs amplitude and phase of the vestigial LSB such that after downconversion with a LO at f_0 the baseband spectrum reproduces the original USB.

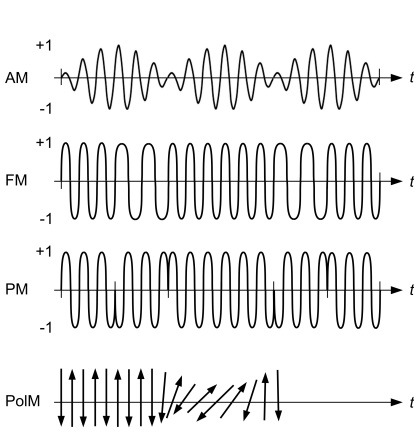
If Σ means addition ($\Sigma \mathrel{\widehat{=}} \text{sum} \mathrel{\widehat{=}} \text{LSB}$), the lower sideband LSB results. The power spectrum $|\breve{a}\,(f)|^2$ from Eq. (2.46b) extends from f_{a1} to f_{a2} and is schematically shown in Fig. 2.10(a)-[bottom], along with the generated one-sided SSB power spectrum $2\,|\breve{m}_{\text{USB}}\,(f)|^2 = |\breve{m}\,(f - (f_0 + f_a))|^2$ as derived from Eq. (2.46c). Remarkably, the carrier frequency f_0 is suppressed.

[54] See Ref. 51, Eq. (9.1.62)
[55] J. R. Carson: Notes on the theory of modulation. Proc. IRE 10 (1922) 57–64

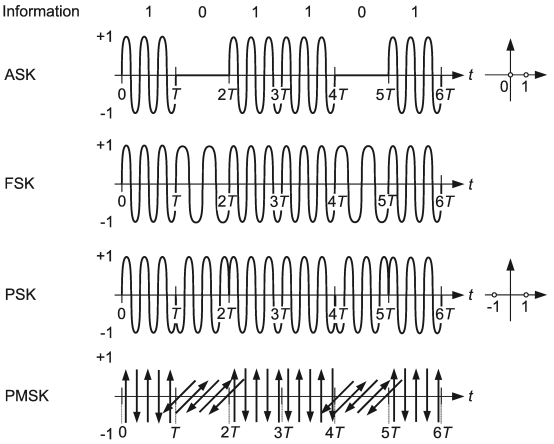
Vestigial sideband filtering Another method for sideband suppression is vestigial sideband (VSB, *German* Restseitenband) filtering. We start with the AM spectrum Fig. 2.8(b) on Page 31. A proper filter could remove any of the sidebands. The problem is that filters with steep slopes have also a strong group delay dispersion which leads to distortion. Therefore a moderately steep filter slope is accepted, so that only a vestige of the unwanted spectra remains (here: LSB and carrier). The method is illustrated in Fig. 2.10(b)-[top]. Most of the unwanted spectral parts are cut off by a transmitter (Tx) filter, Fig. 2.10(b)-[bottom left]. At the receiver (Rx) a filter with a so-called Nyquist slope, see Fig. 2.10(b)-[bottom right], weighs the vestigial lower sideband $\mathrm{LSB}\,(f + f_0) = \mathrm{USB}^*\,(-f + f_0)$ by amplitude and phase such that after mixing with an LO at frequency f_0 the resulting baseband spectrum represents the original USB spectrum.

Analogue modulation formats — Synopsis

A number of temporal signal shapes for analogue modulation like AM, FM and PM are displayed in this synopsis[56] Fig. 2.11(a). The FM and PM examples are drawn for a rectangular modulation function with two states only, and therefore can be also interpreted as binary digital modulation formats. The bottom of Fig. 2.11(a) refers without any previous mathematical description to analogue polarization-mode modulation (PolM) of an electromagnetic wave. If two orthogonal polarizations are transmitted, e. g., in an optical fibre, then four independent data streams can be encoded on the same carrier (IQ-components on two polarizations).



(a) Temporal signal shapes for analogue modulation formats



(b) Temporal signal shapes for digital modulation formats. The quantity T stands for the symbol period.

Fig. 2.11. Signal shapes for various analogue and digital modulation formats. (a) Schematic time dependency of signals with various analogue modulation formats. (AM) sinusoidal amplitude modulation with modulation index $0 < m < 1$. (FM) rectangular 2-frequency modulation (this analogue FM example happens to be identical to (b)-FSK). (PM) rectangular 2-phase modulation (this analogue PM example happens to be identical to (b)-PSK). (PolM) sinusoidal polarization-mode modulation. (b) Schematic time dependency of signals with various digital modulation formats. (ASK) amplitude-shift keying. (FSK) frequency-shift keying. (PSK) phase-shift keying. (PMSK) polarization mode-shift keying. To the right of the ASK and PSK curves the respective constellation diagrams are depicted, using the convention as described in Footnote 26 on Page 21 (Q-component or imaginary part on vertical axis, I-component or real part on horizontal axis). [Modified from Ref. † on the Preface page]

2.4.2 Digital modulation formats

Digital modulation schemes are named similar to the analogue modulation formats. The four basic binary waveform modulations (i. e., the ones with only two levels) are named amplitude-shift keying (ASK, also

[56] Synopsis (pronounced [sɪˈnɒpsɪs]), a brief summary or general survey. Literal meaning "seeing together", from Greek *σύν*, together, and Greek *ὄψις*, the seeing; connected to *ὁράω*, I see, *ὄψομαι*, I shall see, and *ὀφθαλμός*, eye

binary pulse-amplitude modulation, PAM), phase-shift keying (PSK), frequency-shift keying (FSK), and polarization mode-shift keying (PMSK), see Fig. 2.11(b).

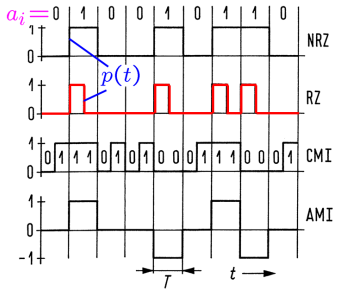
The format ASK can be unipolar or bipolar. In unipolar formats, the sign of the signal does not change when going from mark (logical 1) to space (logical 0). A particular important unipolar ASK format is the case where marks correspond to high signal power and spaces to no signal power (on-off keying, OOK). If the sign of the signal changes during a transition from logical 1 to logical 0 and vice versa, it is common to name the format bipolar. However, as already remarked in Sect. 2.4.1 on Page 31 and in the context of CS-DSB, Sect. 2.4.1 on Page 32, an amplitude is a non-negative quantity. A "bipolar ASK" format is therefore a mixture of ASK and PSK.

The FSK and the PSK modulation formats switch between different carrier frequencies and carrier phases, respectively. Examples of binary FSK and PSK are shown in Fig. 2.11(b). The bottom graph of this figure displays an example of binary PMSK.

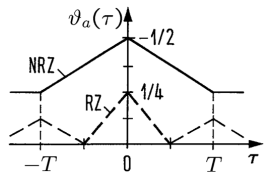
Advanced modulation formats use multilevel coding with M levels. Typically, the amplitude or the phase, or simultaneously both quantities are modulated. This is called M-ary modulation and stands for binary ($M = 2$, 1 bit/symbol), ternary ($M = 3$, on average 1.6 bit/symbol), quaternary ($M = 4$, 2 bit/symbol) etc. modulation formats. According to Eq. (2.7) on Page 16, a symbol with M discrete values can encode $r = \log_2 M$ bit.

ASK modulation

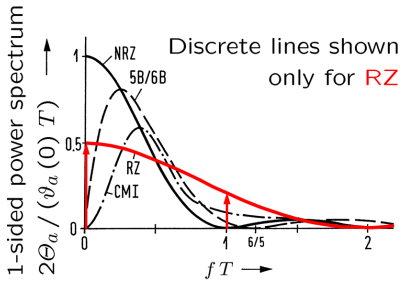
The ASK modulation format is conceptually simple and will therefore be discussed in some length. The format can be encoded in many physical variants and differs in the association of logical 1 and logical 0 to specific pulse shapes $p(t)$ and to transitions between pulses. Here, we assume rectangular pulses which occupy a full or only part of a time slot T. Figure 2.12(a)-[top] shows a bit sequence a_i having a clock period T. Below, various ASK modulation formats are depicted which encode this logical bit sequence physically.



(a) Temporal signal shapes for ASK modulation formats



(b) Autocorrelation functions for NRZ and RZ



(c) Spectra for random bit sequences encoded with different modulation formats

Fig. 2.12. Amplitude-shift keying (ASK) formats, autocorrelation functions of random non-return-to-zero (NRZ) and return-to-zero (RZ) data, and one-sided power spectra for some modulation formats. (a) Binary data $a_i \in \{0, 1\}$, encoded with the formats NRZ, RZ, coded mark inversion (CMI), and alternate mark inversion (AMI, a pseudo-ternary code). Physical pulse shapes $p(t)$ are rect-functions. (b) Autocorrelation functions (ACF) $\vartheta_a(\tau)$ for NRZ and RZ random sequences (c) One-sided normalized power spectra for random data sequences encoded with the formats NRZ, RZ, 5B/6B, and CMI. Discrete lines are only drawn for the RZ format.

If the signal does not return to the level of logical 0 between to neighbouring levels associated with logical 1, we name this format non-return to zero (NRZ). For the return to zero (RZ) format the signal reaches the level of logical 0 in each clock period, for the specific choice of Fig. 2.12(a) at half the clock interval. Time functions $a(t)$ for NRZ and RZ signals are written as

$$\text{NRZ/RZ:} \quad a(t) = \sum_{i=-\infty}^{+\infty} a_i\, p(t - iT) = \sum_{i=-\infty}^{+\infty} a_i\, p(t) * \delta(t - iT), \qquad a_i \in \{0, 1\}\,. \tag{2.57}$$

The binary PCM signal Sect. 6 on Page 16 can be directly transferred this way, but other codings could be more advantageous with respect to available channel bandwidth and noise or other technical peculiarities like clock recovery and error correction. — In the electrical domain, also (pseudo-)ternary codes are used having three signal levels $b_l = 3$ and no direct current (DC) component. — The coded mark inversion (CMI) belongs to the 1B/ 2B code group: One bit (1B) is re-coded into two bit (2B) following the rule (1B $\Rightarrow$ 2B) $0 \Rightarrow 01$ with alternating $1 \Rightarrow 11$, $1 \Rightarrow 00$. — The alternate mark inversion (AMI) code is an example for a (pseudo-)ternary, DC-free line coding. There are three logical states $(0, \pm 1)$ corresponding to, e. g., three voltage levels $0\,\mathrm{V}, \pm 5\,\mathrm{V}$. However, the logical 1 is alternatively represented by the AMI states $+1$ and -1, so that the information per time is identical to NRZ and RZ coding (therefore the naming "pseudo-ternary"). — Much more complex is the 5B/ 6B coding, where each block of 5 bits is treated as a symbol and re-coded according to a look-up table into a block of 6 bits per symbol. For a constant information bit rate R_b, the original symbol rate $R_s = 1/T$ must be increased $R_s \Rightarrow r_c R_s$ by the encoder ratio $r_c = 6/5$. Because the symbol size has been increased from 5 bit to 6 bit, a parity check error correction becomes possible.

In Fig. 2.12(b) the auto-correlation functions $\vartheta_a(\tau) = \overline{a(t+\tau)a(t)}$ (ACF, Table 1.3 on Page 9) for rectangularly shaped binary NRZ and RZ random sequences with equal distribution of logical 1 and 0 are constructed. The maximum $\vartheta_a(0) = \overline{a(t)^2}$ amounts to $1/2$ (NRZ) and $1/4$ (RZ), respectively. This corresponds to the probability $1/2$ (NRZ) and $1/4$ (RZ), respectively, to measure $a(t) = 1$ at any point of time. For the NRZ format and $|\tau| \geq T$, the joint probability for the events $a(t+\tau) = 1$ and $a(t) = 1$ is $1/4$, and therefore $\vartheta_a\,(|\tau| \geq T) = 1/4$ holds.

For the RZ format and $|\tau| = (i - 1/2)\,T$ $(i = 0, \pm 1, \pm 2, \ldots)$, each logical 1 is opposed to a logical 0, and we have $\vartheta_a\,(|(i - 1/2)\,T|) = 0$. With an analogous reasoning as with NRZ, we conclude that the relative extrema $1/8$ of the RZ ACF are reached for $|\tau| = iT$ $(i \neq 0)$. Between the lower and upper corners of the NRZ and RZ ACF, the function $\vartheta_a(\tau)$ changes linearly.

For the ACF $\vartheta_a(\tau)$ and for the associated two-sided power spectrum $\Theta_a\,(f)$ of a binary NRZ random sequence we find, see Table 1.3 on Page 9,

$$\vartheta_{a\,\mathrm{NRZ}}\,(\tau) = \frac{1}{4}\begin{cases} 2 - \dfrac{|\tau|}{T}, & |\tau| \leq T \\ 1\,, & |\tau| \geq T \end{cases}, \qquad \Theta_{a\,\mathrm{NRZ}}\,(f) = \frac{T}{4}\,\mathrm{sinc}^2(fT) + \frac{1}{4}\delta\,(f)\,. \tag{2.58}$$

All other discrete lines that could be expected fall on zeros of the sinc-function $\mathrm{sinc}^2(fT)$ and therefore do not show up. Spectra of NRZ test patterns (which could represent a code transmitting more than one bit per symbol) are nicely derived in an Application Note[57].

With the help of $\vartheta_a\,(\tau)$ in Fig. 2.12(b), the power spectrum for the RZ format with a duty cycle of 50 % as in Fig. 2.12(a) can be calculated[58,59,60]. Because of the periodic part of the ACF, there are discrete spectral lines at frequencies $f = i/T$ $(i = 0, \pm 1, \pm 2, \ldots)$. Figure 2.12(c) displays the one-sided power spectra $2\Theta_a(f)$ for binary NRZ and RZ signals, normalized to the total average power $\vartheta_a(0)$ of the coded signal and to the clock period T, i. e., to the energy per time slot. Discrete lines are shown only for the RZ format.

If the binary data sequence is coded differently, the analytically computed power spectra[61,62] look different. Their continuous part (i. e., without discrete lines) is also displayed in Fig. 2.12(c). Note that after re-coding the clock periods for the same information per second are reduced in comparison to NRZ and RZ transmission, $T_{\mathrm{CMI}} = T/2$ and $T_{\mathrm{5B/6B}} = (5/6)\,T$.

---

[57] "maxim integrated" (`http://www.maximintegrated.com`): Spectral content of NRZ test patterns. Application Note AN3455, `http://pdfserv.maximintegrated.com/en/an/AN3455.pdf`,

[58] E. Hölzler, H. Holzwarth: Pulstechnik, Band I. Grundlagen, 2. Ed. Berlin: Springer-Verlag 1982. General spectra of pseudo-random bit sequences (PRBS) are calculated in Eq. (6.48), (6.49).

[59] Agrawal, G. P.: Lightwave technology. Telecommunication systems. Hoboken (NJ): John Wiley & Sons 2005 (Sect. 2.2)

[60] Ip, E.; Kahn, J. M.: Power spectra of return-to-zero optical signals. J. Lightw. Technol. 24 (2006) 1610–1618

[61] K. W. Cattermole, J. J. O'Reilly (Eds.): Rauschen und Stochastik in der Nachrichtentechnik (Noise and Stochastc Processes in Communications). Weinheim: VCH Verlagsgesellschaft 1988

[62] J. Fluhr, P. Marending, H. Trimmel: Ein Lichtwellenleitersystem für die Übertragung von 8-Mbit/s-Signalen. Siemens Telcom Report 6 (1983), Beiheft "Nachrichtenübertragung mit Licht", 127–132

A line coding of the binary NRZ sequence fits the code's power spectrum to the properties of the channel. The 5B/ 6B and the CMI code have only small contributions at low frequencies, which is favourable if the channel contributes significant noise in this spectral region. From $2\Theta_a(0) = 0$ it is seen that CMI and 5B/6B code have no long sequences of logical 1 (for the AMI code: long $+1$ or -1 sequences), so DC coupled circuits are not required. Finally, long sequences of logical 0 are to be avoided, too. Such sequences can occur for NRZ, RZ, and AMI formats and would prevent clock recovery. The 5B/6B code has a maximum of three[63,64] subsequent 0. For the ternary HDB3 code[65] (high density bipolar code) the longest sequence of zeros is three subsequent 0 (therefore the naming HDB3). Every forth clock cycle holds alternatingly the values ± 1. The minimum probability for a 1 to appear is therefore $1/4$.

Sometimes so-called scramblers randomize the logical 0 and 1 of PCM data following a fixed algorithm, so that a pseudo-random sequence results. This scrambling can be made undone by the receiver.

Clock recovery For proper reception the receiver must recover the clock cycle from the data stream. For rectangular NRZ pulses $a(t)$ Eq. (2.57) with $p(t) = \mathrm{rect}\,(t/T)$ the power spectrum has always a zero at $f = 1/T$ (for the 5B/6B code at $f = (6/5)/T$, for the CMI code at $f = 2/T$), see Fig. 2.12(c). Therefore, a narrowband filter (e. g., a phase locked loop (PLL)[66]) cannot recover the clock frequency at $f = 1/T$. However, RZ codes have pronounced spectral lines at the clock frequency $f = 1/T$. As a disadvantage, double the transmission bandwidth is required compared to the NRZ format.

For clock recovery from NRZ data streams a nonlinear operation is required. The differentiated pulses are rectified, so that two "needles" with a separation of T appear for each pulse. Their spectrum has a strong component at $f = 1/T$. A PLL circuit or a surface acoustic wave (SAW) device acts as a narrowband filter and leads to a data-synchronous clock. For fitting the phase such that each pulse is sampled in its centre, compact self-correcting regenerator circuits are used[67].

ASK modulation formats — Synopsis The various unipolar and bipolar ASK formats are displayed in Fig. 2.13. A word of warning: The established naming "bipolar ASK" is misleading, because an amplitude is positive by definition (like a radius), and no amplitude shift can make it negative. What is meant with bipolar ASK is a combination of unipolar ASK and phase-shift keying (PSK).

Unipolar ASK Figure 2.13(a) displays how a bit sequence (top row) is encoded using various unipolar ASK formats. The following properties can be seen:

NRZ Non-return to zero. Logical 0 is a space (low level of physical signal), logical 1 is a mark (high level of physical signal). A string of consecutive 0 or 1 means no signal change.

RZ Return to zero. Same as NRZ, but marks occupy only a fraction of the bit slot.

Manchester Also phase encoding[68] (PE). Logical 0 is a mark in the first part of the bit slot, and a space in the second one. Logical 1 is a space in the first part of the bit slot, and a mark in the second one. The average signal power is the same for both 0 and 1. The required transmission bandwidth doubles compared to the NRZ format.

Differential Manchester A differential encoding, using the presence or absence of transitions to indicate a logical value. Logical 0 is a level transition in the first part of the bit slot, logical 1 is a level transition in the second part of the bit slot.

[63] A. Stegmeier, H. Trimmel: Ein Lichtwellenleitersystem für die Übertragung von 34-Mbit/s-Signalen. Siemens Telcom Report 6 (1983) Beiheft "Nachrichtenübertragung mit Licht", 133–137

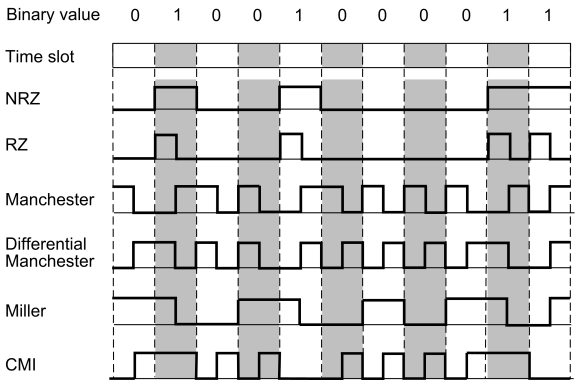
[64] See Reference 61

[65] E. Hölzler, H. Holzwarth: Pulstechnik, Band II. Anwendungen und Systeme, 2. Ed. Berlin: Springer-Verlag 1984. Sect. 8.4.2.2

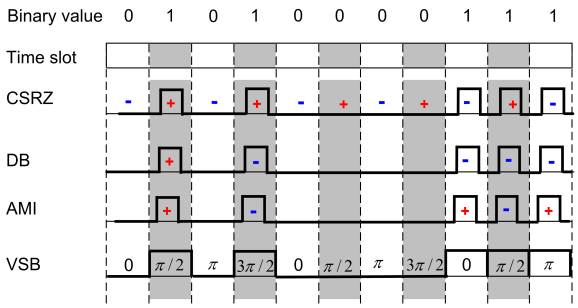
[66] See Reference 65, Sect. 8.2.1.2

[67] C. R. Hogge: A self correcting clock recovery circuit. IEEE J. Lightwave Technol. LT-3 (1985) 1312–1314

[68] The name comes from its development at the University of Manchester, where the coding was used to store data on the magnetic drum of the Manchester Mark 1 computer. — The Manchester Mark 1 was one of the earliest stored-program computers, developed at the Victoria University of Manchester from the Small-Scale Experimental Machine (SSEM) or "Baby" (operational in June 1948). It was also called the Manchester Automatic Digital Machine, or MADM. Work began in August 1948, and the first version was operational by April 1949; a program written to search for Mersenne primes ran error-free for nine hours on the night of 16/17 June 1949 [cited from `http://en.wikipedia.org/wiki/Manchester_Mark_1`]



(a) Temporal signal shapes for unipolar ASK. The signal's polarity is always positive.



(b) Temporal signal shapes for bipolar ASK. The signal's polarity is marked by a negative (−) or a positive sign (+)

Fig. 2.13. Synopsis of various unipolar and bipolar amplitude-shift keying (ASK) formats. (a) Unipolar ASK encoded with the formats NRZ, RZ, Manchester, differential Manchester, and coded mark inversion (CMI). (b) Bipolar ASK encoded with carrier-suppressed return to zero (CSRZ), duobinary (DB), alternate mark inversion (AMI), and vestigial sideband (VSB) filtering with a progrssive phase shift of $\pi/2$. — Amplitudes are always positive, but bipolar "ASK" codes shift the phase as well. [Modified from Fig. 2.11 and 2.12 of Ref. † on the Preface page]

Miller Also delay modulation. Logical 0 means no signal transition, except when 0 is followed by another 0, then there is a transition at the end of the bit slot. Logical 1 is signalled by a transition at the centre of the bit slot from either level.

CMI Coded mark inversion. An NRZ line code, in which logical 0 is encoded as a $0 \to 1$ transition at the centre of the bit slot, and logical 1 remains constantly on the previous level for the entire bit slot. For a sequence of 1 the constant level is inverted for each subsequent bit slot.

Bipolar ASK and VSB Figure 2.13(b) displays how a bit sequence (top row) is encoded using various bipolar ASK formats (ASK with $\{0, \pi\}$-PSK). In addition, vestigial sideband (VSB) encoding shows a combination of ASK with progressive $(\pi/2)$-PSK. The reason for this additional effort is to increase the spectral efficiency C', see Eq. (2.23) on Page 22, and to reduce intersymbol interference (ISI). For example, with VSB, the signals of bit slots adjacent to a central slot interfere destructively, if by dispersion they spill over into the central bit slot.

CSRZ Carrier-suppressed return to zero. This is an RZ format with additional phase modulation. If all even bit slots see a positive signal, then all odd bit slots see a negative signal, i. e., a phase shift by π separates even and odd bit slots.

DB, LP-DB Duobinary, low-pass filtered DB. Logical 0 is a space. Logical 1 is a mark without a phase shift by π if there is an even number of 0 since the last 1, and a mark with a phase shift by π if there is an odd number of 0 since the last 1. Duobinary data encoding is a form of correlative coding in partial response signalling. The modulator drive signal can be produced by adding one-bit-delayed data to the present data bit to give levels 0, 1, and 2. An identical effect can be achieved by applying a low-pass (LP) filter to the ideal binary data signal (LP-DB). The correlated three-level signal can be demodulated into a binary signal by using an optical direct detection receiver.

AMI Alternate mark inversion, also called modified duobinary (bipolar or decode duobinary). Logical 0 is a space. Logical 1 is a mark, where each mark is phase shifted by π compared to the previous mark (even if 0 are between consecutive marks).

VSB, AP Vestigial sideband filtering, also called $\pi/2$ alternating phase change[69]. An optical VSB signal is usually generated from an OOK-NRZ or OOK-RZ signal by an optical filter, the passband of

[69] Schnarrenberger, M., Sotobashi, H., Chujo, W. and Freude, W.: Novel intersymbol interference reduction technique by bit synchronized $\pi/2$ phase shift. Proc. Institute of Electronics, Information and Communication Engineers (IEICE Japan) Spring Conference, Hiroshima, 28.–31.03.2000. `http://www.ipq.kit.edu/staff/freude/ieice2000_pibytwo.pdf`

which is detuned from the carrier, see Fig. 2.10 on Page 35. Logical 0 is a space. Logical 1 is a mark with a progressive phase shift of $\pi/2$ added for each bit slot (even when it contains a 0).

PSK modulation

PSK formatted data streams are generated by modulating the phase η of the carrier $\hat{a}\,e^{j[\omega_0 t+\eta(t)]}$, while its amplitude $\hat{a}$ and frequency f_0 are kept constant, see Fig. 2.9 on Page 34. For binary PSK formats, the phase takes two values, commonly chosen to be $\eta = 0$ and $\eta = \pi$. Because the intensity remains constant, nonlinear effects that depend on intensity are independent from the modulated data stream.

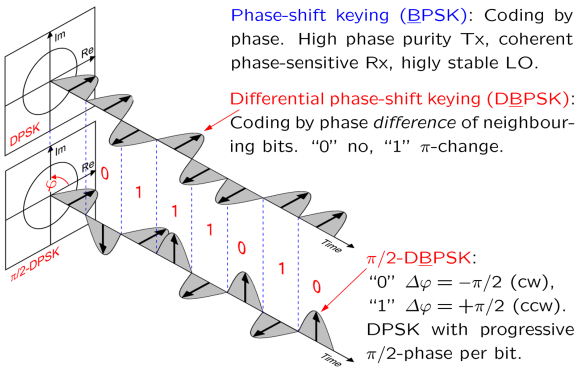


Fig. 2.14. Real part (Re) and imaginary part (Im) of a complex carrier envelope modulated with binary DPSK (DBPSK, upper graph) and binary $\pi/2$-DPSK ($\pi/2$-DBPSK, lower graph). Sinusoidally shaped RZ pulses represent the carrier envelope. The carrier time function itself is not drawn, because it oscillates very rapidly inside each pulse-shaped part of the envelope. For $\pi/2$-DPSK, logical 0 and 1 are represented by a relative phase shift of $-\pi/2$ (clockwise) and $+\pi/2$ (counter clockwise), respectively. The format $\pi/2$-DBPSK resembles minimum-shift frequency keying (MSK). The format $\pi/2$-DBPSK is identical to MSK, if the phase is not switched, but changed continuously during each time slot. [Modified from Ref. 70]

PSK signal transmission so far is only used in backbones, where cost does not matter too much. There are mainly three reasons for this:

- PSK formats require well defined carriers with little phase noise, and phase-sensitive coherent receivers with a laser LO as discussed in Sect. 2.3.2 on Page 30. A photodetector as briefly described in Eq. (1.1) on Page 2 would only be sensitive to the intensity. Unfortunately, such an LO laser adds to cost and complexity, not to speak of the cost of the transmitting laser which must have similarly good properties. So in practice, one tries to avoid such schemes.
- In optical communications, a typical wavelength is $\lambda_0 = 1.55\,\mu\text{m}$ ($f_0 = 193.51\,\text{THz}$, see Table 2.1 on Page 21). Consider a fibre transmission span of $L = 100\,\text{km}$. Following Eq. (2.11) on Page 18, the acquired phase would be $\varphi = -\beta L \approx -k_0 n L$, where the refractive index of the fibre is about $n = 1.5$. A tiny refractive index change by only $\Delta n = 0.77 \times 10^{-11}$ (this happens easily if the temperature changes randomly by fractions of a degree) would cause a phase shift by $\Delta\varphi = -\frac{2\pi}{1.55\,\mu\text{m}} \times 0.77 \times 10^{-11} \times 100\,\text{km} = -\pi$, and thus randomly invert the meaning of space and mark with respect to the phase of the LO.
- Optical signals have a certain polarization. The mixing between the incoming signal and the LO works only if both oscillate in the same state of polarization. However, after hundreds of kilometers of transmission, the state of polarization is usually no longer known.

Despite these difficulties, coherent schemes are gaining ground and will be more and more deployed. However, all of these constraints can be relaxed by using a modified form of PSK, namely differential PSK (DPSK), Fig. 2.14. The scheme does not compare the phase of a transmitter laser and a receiver LO, but rather the phase difference between subsequent time slots. For high data rates, the slot width T and therefore the time difference is small, and the phase even of inexpensive transmitter lasers is stable

enough during such a short time interval of, e. g., $T = 25\,\mathrm{ps}$ at a data rate of 40 Gbit/s. At present, DPSK is globally used in both long-haul backbones and medium-haul networks. Two different DPSK flavours are common:

DPSK Differential phase-shift keying[70]. Information is coded in a phase difference $\Delta\varphi = \{0, \pi\}$ between two neighbouring time slots. If η_k represents the carrier phase for the k-th time slot, the phase difference $\Delta\varphi = \eta_k - \eta_{k-1}$ is $\Delta\varphi = 0$ for encoding a logical 0, and it is $\Delta\varphi = \pi$ for a logical 1.

$\pi/2$-DPSK $\pi/2$ differential phase-shift keying[71,72]. Information is coded in a phase difference $\Delta\varphi = \pm\,\pi/2$ between subsequent time slots. The phase difference is $\Delta\varphi = -\,\pi/2$ for encoding a logical 0, and a phase difference $\Delta\varphi = +\,\pi/2$ encodes a logical 1. The procedure is identical to a progressive $\pi/2$ phase shift on top of the DPSK encoding. The unique advantage[73] of $\pi/2$-DPSK arises from its response to over-filtering, as will be shown in a later section.

For a binary NRZ DPSK signal with a symbol rate $R_s = 1/T$, the total bandwidth B_{DPSK} for signalling is essentially determined by the zero of the NRZ power spectrum at $f_0 \pm R_s$ as in Fig. 2.12(c) on Page 37,

$$B_{\mathrm{DPSK}} \approx 2R_s = \frac{2}{T}\,. \tag{2.59}$$

This is much less than would be expected from distortion-free analogue angle modulation, where Eq. (2.56b) on Page 35 would predict a bandwidth of $2\,(\eta+1)\,B = 2\,(\,\pi/2+1)\,R_s \approx 5\,R_s$ for an average modulation index of $\eta = \langle\,\Delta\varphi/2\rangle = \pi/2$.

FSK modulation and OFDM

Frequency-shift keying (FSK) encodes data by shifting the carrier frequency f_0. Binary data are encoded in two carrier frequencies $f_0 \pm \Delta f$, which are separated by the frequency spacing ("tone" spacing) $2\Delta f$. For estimating the transmission bandwidth with Eq. (2.56b), we set $\eta = \Delta\omega/(2\pi R_s)$ according to Eq. (2.55) on Page 34 and find

$$B_{\mathrm{FSK}} \approx 2\,(\eta+2)\,B = 2\left(\frac{\Delta f}{R_s}+1\right)R_s = 2\,(\Delta f + R_s)\,. \tag{2.60}$$

If $\Delta f \ll R_s$ holds, we speak of narrowband FSK. The case $\Delta f \gg R_s$ is named broadband FSK. Practical implementations of, e. g., a binary FSK modulate the phase $\eta\,(t)$ of the optical carrier $\mathrm{e}^{\mathrm{j}[\omega_0 t+\eta(t)]}$ in a linear fashion according to $\mathrm{d}\eta\,(t)/\,\mathrm{d}t = \pm 2\pi\Delta f$, $\eta\,(t) = \pm\Delta\omega\,t$ with $\Delta\omega = 2\pi\Delta f$. There are two important versions:

CFSK Continuous-phase FSK. For binary CFSK, logical 0 and logical 1 are represented by carrier frequencies $f_0 - \Delta f$ and $f_0 + \Delta f$, respectively. A binary data change $0 \to 1$ is represented by switching the slope of the continuous phase change from $\mathrm{d}\eta\,(t)/\,\mathrm{d}t = -\Delta\omega$ to $\mathrm{d}\eta\,(t)/\,\mathrm{d}t = +\Delta\omega$. The data change $1 \to 0$ requires to switch the phase slope from $\mathrm{d}\eta\,(t)/\,\mathrm{d}t = +\Delta\omega$ to $\mathrm{d}\eta\,(t)/\,\mathrm{d}t = -\Delta\omega$. The "transitions" $0 \to 0$ and $1 \to 1$ leave the phase slope unchanged, and we have $\eta\,(t) = -\Delta\omega\,t$ and $\eta\,(t) = +\Delta\omega\,t$, respectively. The phase function $\eta\,(t)$ is continuous and consists of straight line segments. This saves bandwidth compared to a "hard" switching of independent carriers as in switched FSK.

[70] Wei, X.; Gnauck, A. H.; Gill, D. M.; Liu, X.; Koc, U.-V.; Chandrasekhar, S.; Raybon, G.; Leuthold, J.: Optical $\pi/2$-DPSK and its tolerance to filtering and polarization-mode dispersion. IEEE Photon. Technol. Lett. 15 (2003) 1639–1641

[71] See Ref. 70

[72] H. Grießer, M. Eiselt, B. Teipen, A. Autenrieth, K. Grobe und J.-P. Elbers: Options for Tb/s transmission," ITG Workshop 3.5.1 Karlsruhe (2011)

[73] See Ref. 70

OFDM Orthogonal frequency division multiplexing[74,75] is both, a specialized and a generalized form of CFSK. Generalized in so far, as multiple frequencies can be present at the same time, and specialized because the frequency separation is tied to the symbol duration. If the CFSK symbol shape is rectangular with a duration T, and if the participating carrier frequencies $f_\nu = \nu R_s =$ are chosen to be integer multiples ν of the symbol rate (implying a frequency line separation of $2\Delta f = 1/T$), inter-symbol interference is minimized because the symbols are orthogonal,

$$\frac{1}{T}\int_{T_0-T/2}^{T_0+T/2} \exp(+\mathrm{j}\,2\pi\nu\, t/T)\exp(-\mathrm{j}\,2\pi\nu'\, t/T)\,\mathrm{d}t = \delta_{\nu\nu'} \qquad \text{(arbitrary reference time } T_0\text{)}. \quad (2.61)$$

This means that even if many carrier frequencies with different complex amplitudes are present at the same time, any complex carrier $\exp(+\mathrm{j}\,2\pi\nu\, t/T)$ which is modulated with a rectangular pulse having a complex-valued amplitude $\check{c}_\nu$ can be isolated from any other carrier. At the receiver, an IQ-demodulator as in Fig. 2.7(d) on Page 29 with local oscillator $\exp(-\mathrm{j}\,2\pi\nu'\, t/T)$ mixes with the arriving signal carriers $\check{c}_\nu \exp(+\mathrm{j}\,2\pi\nu\, t/T)$ and integrates over a symbol duration T. Only if the LO frequency $\nu' = \nu$ is chosen properly, the result of this integration will be $\check{c}_\nu$. Received carriers with $\nu \neq \nu'$ do not contribute and are thus ignored. If the complex modulation amplitudes $\check{c}_\nu$ would be real, and if only two alternating carriers are involved, this "OFDM" reduces to CFSK.

SFSK Switched FSK uses independent carriers, which are switched on and off. The phase functions then have discontinuities, which increases the required bandwidth compared to CFSK.

MSK Minimum-shift keying. If the tone spacing equals half the symbol rate, $2\Delta f = R_s/2 = 1/(2T)$, then any data transition changes the phase continuously by $|\Delta\varphi| = \Delta\omega\, T = \pi/2$ during the duration of a time slot T. If further the amplitudes of the two tones $f_0 \pm \Delta f$ are identical, the scheme is identical to the $\pi/2$-DPSK format as depicted in Fig. 2.14: A logical 0 is encoded by a phase change of $\Delta\varphi_0 = -\pi/2$ (the carrier frequency is switched to or remains at $f_0 - \Delta f$), and a logical 1 is encoded by a phase change of $\Delta\varphi_1 = +\pi/2$ (the carrier frequency is switched to or remains at $f_0 + \Delta f$).

FODM Similar to the relationship between CFSK and OFDM, there exists an affinity between MSK and fast orthogonal frequency division multiplexing[76] (FODM) with half the standard OFDM carrier spacing.

Digital modulation formats — Synopsis

The properties[77,78,79,80,81] of various digital modulation formats for transmitting a binary 40 Gbit/s signal (symbol time slot $T = 25$ ps) are summarized in the following. The signal spectra schematics in Fig. 2.15(a) refer to random binary data encoded with various modulation formats. Families of modulation formats are grouped together with the broken blue lines (– – –).

[74] J. Leuthold, W. Freude: Optical OFDM and Nyquist multiplexing. In: Kaminow, I. P.; Li, Tingye; Willner, A. E. (Eds.): Optical Fiber Telecommunications VI B. Systems and Networks, 6th Ed. Elsevier (Imprint: Academic Press), Amsterdam 2013, Chapter 9, pp. 381–432

[75] OFDM transmission is very common, for instance in telephone-line access networks like (V)DSL (short for (*v*ery) high-speed *d*igital *s*ubscriber *l*ine). My 50 Mbit/s (downlink) & 10 Mbit/s (uplink) VDSL connection comprises 4 096 carriers in a spectral region up to 17.664 MHz transmitting 1 . . . 10 bit per carrier, depending on the individual SNR = 25 . . . 55 dB.

[76] J. Zhao, A. D. Ellis: Novel optical fast OFDM with reduced channel spacing equal to half of the symbol rate per carrier. OFC 2010, San Diego. Paper OMR1

[77] Winzer, P. J.: Optical transmitters, receivers, and noise. Wiley Encyclopedia of Telecommunications (2002). `http://www.mrw.interscience.wiley.com/eot/articles/eot404`

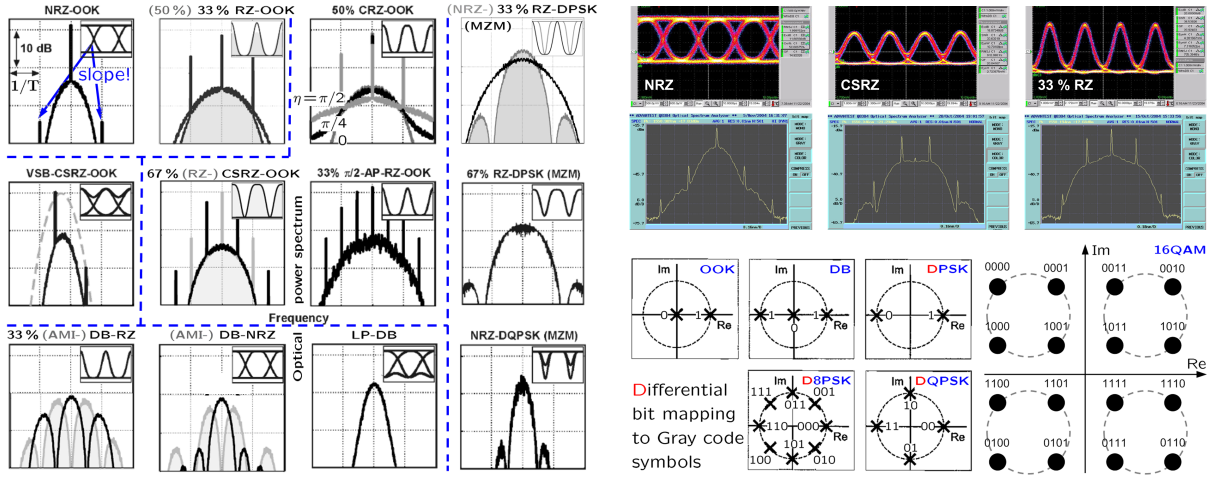
[78] Gnauck, A. H.; Liu, X.; Wei, X.; Gill, D. M.; Burrows, E. C.: Comparison of modulation formats for 42.7-Gb/s single-channel transmission through 1980 km of SSMF. IEEE Photon. Technol. Lett. 16 (2004) 909–911

[79] Gnauck, A. H.: Advanced amplitude- and phase coded formats for 40-Gb/s fiber transmission. Proc. 17th Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS 2004), Puerto Rico, USA, November 7–11, 2004. Paper WR1

[80] P. J. Winzer, R.-J. Essiambre: Advanced optical modulation formats. Proc. IEEE 94 (2006) 952–985

[81] Charlet, G.: Progress in optical modulation formats for high-bit rate WDM transmissions. IEEE J. Sel. Topics Quantum Electron. 12 (2006) 469–483

Variants of spectra are drawn in light grey as well as the associated explanatory texts. Schematics of eye diagrams reveal typical pulse shapes, upper right corner of the subfigures. Figure 2.15(b)-[top] shows measurements[82,83] of eye diagrams for the formats NRZ, CSRZ and RZ with a pulse duty cycle of 33 % along with the associated spectra.



(a) Semi-logarithmic power spectra for random data encoded with various modulation formats

(b) [top] Eye diagrams and spectra [bottom] Constellation diagrams of binary and M-ary modulation formats

Fig. 2.15. Synopsis of properties for various digital modulation formats. (a) Spectra for random binary data, grouped together by the broken blue lines (– – –). The horizontal axis represents frequency (1 div = $R_s = 1/T$, symbol time slot T), the logarithmic vertical axis gives the relative spectral power density in dB (1 div = 10 dB). Spectra in light grey are named in the subfigure texts using the same colour. Eye diagram schematics are shown in the upper right corners of the subfigures. For the NRZ spectrum, unexpected lines are to be seen at a frequency offset $\pm\, 1/T$ from the carrier if we compare to the spectrum in Fig. 2.12(c) at Page 37. These unexpected lines are due to the finite slopes of the actual real-world NRZ pulse, which deviates from an ideal rect-function. [Compiled from Ref. 77, 78, 79, 80, 81] (b)-[top] Measured random signals (so-called "eye diagrams") for NRZ, CSRZ and 33 % RZ formats with measured power spectra [modified from Ref. 82, 83] (b)-[bottom] Constellation diagrams for the binary formats OOK and (D)PSK, and for the M-ary formats DB, (D)QPSK, (D)8PSK, and 16QAM with mappings to so-called Gray code[86] symbols. [Modified from Ref. 84 and 85 (16QAM)]

Finally, Fig. 2.15(b)-[bottom] displays so-called constellation diagrams for binary and M-ary modulation formats[84] showing amplitude and phase of the transmitted signals in a complex plane I-Q plane as was discussed previously in Sect. 2.3.2 on Page 28. Implicitly it is assumed that the horizontal axis represents the real part (Re, the in-phase or I-component) of the electric field and the vertical axis represents the imaginary part (Im, the quadrature or Q-component) of the electric field. Obviously, the QPSK constellation in Fig. 2.15(b)-[bottom] is identical to the constellation of a 4QAM format. The constellation diagram of a more complicated 16QAM format[85] (which in addition uses a Gray code[86]) completes the

[82] Pincemin, E. et al.: Robustness of the OOK modulation formats at 40 Gbit/s in the practical system infrastructure. ECOC (2005). Paper We4.P.112

[83] Gosselin, S.; Joindot, M.: Key drivers and technologies for future optical networks. ECOC (2006). Tutorial We2.2.1, Slide 43

[84] Grobe, K.: 40 Gb/s techniques for metro optical networking. Der Fernmelde-Ingenieur 59 (2005) 1–35. Fig. 19

[85] Kikuchi, N.; Sekine, K.; Sasaki, S.: Proposal of inter-symbol interference (ISI) suppression technique for optical multilevel signal generation. ECOC (2006). Paper Tu4.2.1

[86] Frank Gray, physicist at Bell Laboratories, ⋆Alpine (IN) 13.9.1887, † 23.5.1969. Numerous innovations in television, both mechanical and electronic. Remembered for the invention of the Gray code (reflected binary code) in 1947. The advantage of this code: Consecutive positions of this code differ only by one bit. If a rotary position encoder reads out a number of bits in parallel (encoded in opaque and transparent ring segments with different radii), no simultaneous bit switching is required, which, when not properly done, could lead to glitches. — "In modern digital communications, Gray codes play an important role in error correction. For example, in a digital modulation scheme such as QAM where data are typically transmitted in symbols of 2 bit or more, the signal's constellation diagram is arranged so that the bit patterns conveyed by adjacent constellation points differ by only one bit. By combining this with forward error correction capable of correcting single-bit errors, it is possible for a receiver to correct any transmission errors that cause a constellation point

synopsis. With implicit reference to Fig. 2.15(a), the following list illustrates the application of various modulation formats for transmitting a data stream of 40 Gbit/s:

40 GBd SYMBOL RATE WITH 1 bit / symbol:

OOK On-off keying. NRZ, RZ (duty cycle 33 % and 50 %), chirped RZ[87] (CRZ, duty cycle 50 %), carrier-suppressed RZ (CSRZ, duty cycle 66 %), chirped CSRZ

DB Duobinary. NRZ-DB, chirped NRZ-DB, RZ-DB (duty cycle 33 % and 50 %), CRZ-DB, CSRZ-DB, chirped CSRZ-DB

VSB Vestigial sideband filtering. Sideband and carrier (partially) suppressed

DPSK Differential binary phase shift keying (D(B)PSK), NRZ-DPSK, chirped NRZ-DPSK, RZ-DPSK (duty cycle 33 % and 50 %), CRZ-DPSK, CSRZ-DPSK, chirped CSRZ-DPSK

20 GBd SYMBOL RATE WITH 2 bit / symbol:

DQPSK Differential quaternary[88] PSK (or differential four-PSK, or differential 4QAM). (N)RZ, CSRZ (duty cycle 33 %, 50 % and 66 %)

13.3 GBd SYMBOL RATE WITH 3 bit / symbol:

D8PSK Differential octonary PSK (or differential eight-PSK)

10 GBd SYMBOL RATE WITH 4 bit / symbol:

16QAM Seno-denary quadrature amplitude modulation (or sixteen-QAM)

Pulse-position modulation

Pulse-position modulation (PPM) for optical communication systems was patented[89] as early as 1983. The format aims at optimum sensitivity without putting much weight on spectral efficiency. In this respect it is different from but complements the spectrally more efficient QAM formats. In wired terrestrial long-haul communications PPM is of no interest, but free-space optical (FSO) terrestrial links[90] or links between earth terminals and satellites profit from the inherent sensitivity of the format, see Fig. 2.4(b) on Page 24. Of course[91], for the same bit rate, M-PPM schemes require significantly more bandwidth than spectrally more efficient modulation formats, but due to the high carrier frequency and in contrast to RF wireless systems, there is no intrinsic bandwidth limitation for FSO channels.

Pulse-position modulation is a form of signalling that uses the same transmitter and receiver hardware as with OOK. In M-ary PPM (M-PPM), a number of $r = \log_2 M$ information bits is encoded by the position of an optical pulse within M equidistant time slots of a symbol with duration T, Fig. 2.16. The

to deviate into the area of an adjacent point. This makes the transmission system less susceptible to noise." [Cited after `http://en.wikipedia.org/wiki/Gray_code`]

[87] If a controlled amount of analog phase modulation is applied to a modulation format, the qualifier "chirped" is added. In the case of CRZ, a bit-synchronous periodic chirp spectrally broadens the signal bandwidth. Although this reduces the format's suitability for high spectral efficiency WDM systems, it generally increases its robustness to fiber nonlinearity. [Cited from Ref. 80, Sect. VI.D]

[88] Distributive numbers answer "how many times each?" *Singly* is a distributive number, while *single* is a multiplier. — *Latin* singuli (every one, je einer), bini (every two, je zwei), terni (every three, je drei), quaterni, quini, seni (senary = based on the number six), septeni, octoni (óctonary = based on the number eight), noveni, deni (every ten, denary numbers = decimal numbers), undeni, duodeni, terni deni, quaterni deni, quini deni, seni deni (seno-denary), septeni deni, duodeviceni, undeviceni, viceni (every twenty, je zwanzig)

[89] I. Garrett: United States Patent Number 4,584,720, Apr. 22, 1986. Filed on Aug. 30, 1983
Abstract: An optical communication system using digital pulse position modulation employs a mode locked laser with a mode locking frequency equal to the time slot frequency of the modulation and means dependent on groups of consecutive digits of the data to be transmitted to select pulses from the laser for transmission. In one example, 4-bit groups from the data for transmission select one out of 20 pulses from the laser thus leaving a guard interval of 4 time slot periods between position modulated pulses.

[90] W. Gappmair, S. Hranilovic, E. Leitgeb: Performance of PPM on terrestrial FSO links with turbulence and pointing errors. IEEE Comm. Lett. 14 (2010) 468–470

[91] Until the end of the present "Pulse-position modulation" section, we follow in large parts the text of Ref. 33 on Page 22. Sect. 2.2.1.5, p. 241 ff.

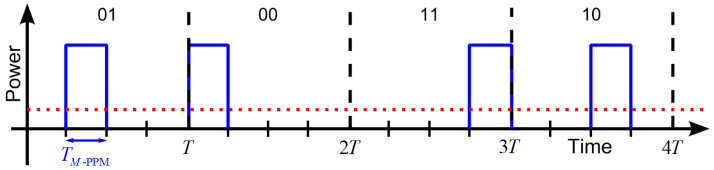


Fig. 2.16. Pulse-position modulation with $M = 4$ different pulse positions inside the frame of a symbol duration T. The signal's duty cycle is $1/M$, so that the peak power in each pulse is larger than the average power (·····) by a factor of M.

resulting waveforms have a low duty cycle $1/M$, and the peak power in each pulse is larger than the average power (dotted red line, ·····) by a factor of M. This makes PPM well suited for average-power limited transmitters with EDFA, but a poor choice for peak-power limited transmitters with SOA.

As can be seen from Fig. 2.12(c) on Page 37 and Fig. 2.15(a) on Page 44, the upper-frequency limit required for NRZ-OOK signalling with a symbol duration T is $1/T$. For identical bit rates $R_{b\,\text{NRZ-OOK}} = R_{b\,M\text{-PPM}} = 1/T_{\text{NRZ-OOK}}$, this leads to a larger M-PPM bandwidth $B_{M\text{-PPM}} = (M/r)\,B_{\text{NRZ-OOK}}$,

$$R_{b\,\text{NRZ-OOK}} = R_{b\,M\text{-PPM}}: \tag{2.62}$$

$$B_{\text{NRZ-OOK}} = \frac{1}{T_{\text{NRZ-OOK}}}, \quad B_{M\text{-PPM}} = \frac{1}{T_{M\text{-PPM}}} = \frac{M/\log_2 M}{T_{\text{NRZ-OOK}}} = \frac{M}{r}\,B_{\text{NRZ-OOK}}, \quad r = \log_2 M.$$

Figure 2.17 shows the spectra of NRZ-OOK, 2PPM, 4PPM and 16PPM signals. The bandwidths are kept identical by having identical pulse widths $T_{M\text{-PPM}} = T_{\text{NRZ-OOK}}$. In this case the M-PPM bit rate becomes smaller, $R_{b\,M\text{-PPM}} = (r/M)\,R_{b\ \text{NRZ-OOK}}$,

$$T_{M\text{-PPM}} = T_{\text{NRZ-OOK}}: \tag{2.63}$$

$$R_{b\,\text{NRZ-OOK}} = \frac{1}{T_{\text{NRZ-OOK}}}, \quad R_{b\,M\text{-PPM}} = \frac{\log_2 M}{T_{M\text{-PPM}}M} = \frac{(\log_2 M)/M}{T_{\text{NRZ-OOK}}} = \frac{r}{M}\,R_{b\,\text{NRZ-OOK}}, \quad r = \log_2 M.$$

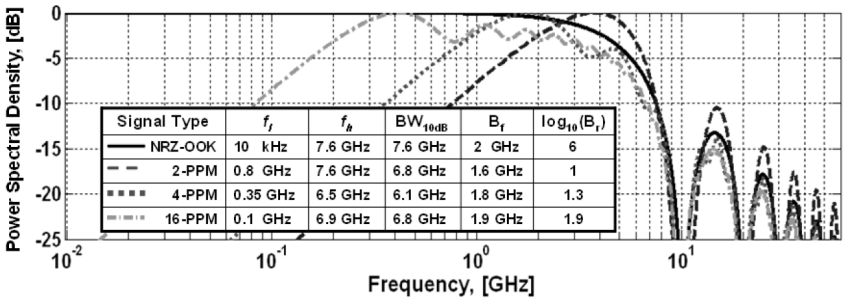


Fig. 2.17. Calculated spectra for square waveforms used in OOK and 2PPM, 4PPM, and 16PPM modulation for a fixed pulse width (OOK *and* M-PPM) of $T_{M\text{-PPM}} = T_{\text{NRZ-OOK}} = 100\,\text{ps}$. Because of this choice the first spectral zero is always at $1/T_{M\text{-PPM}} = 10\,\text{GHz}$, and the bit rates for M-PPM become smaller than for NRZ-OOK, $R_{b\,M\text{-PPM}} = (r/M)\,R_{b\,\text{NRZ-OOK}}$ ($r = \log_2 M$). The M-PPM waveforms have a smaller bandwidth ratio $B_r = f_h/f_l$ of the 10 dB higher and lower limiting frequencies f_h and f_l, respectively, and a significantly smaller fractional (or relative) bandwidth $B_f = B_{10\,\text{dB}}/[(f_h + f_l)/2]$ ($B_{10\,\text{dB}} = f_h - f_l$) than the OOK waveforms. For the OOK spectrum, we set $f_l = 10\,\text{kHz}$, which is a common low-frequency specification for applicable broad-band electronics. For a constant bit rate, the M-PPM spectra are broadened by a factor of M/r, which increases the 10 dB bandwidth $B_{10\,\text{dB}}$, but does not impact B_r. [After Ref. 33 on Page 22. Fig. 9, p. 243]

While electrical bandwidth limitations may determine the maximum slot rate $R_{\text{slot}} = 1/T_{M\text{-PPM}}$ for a single M-PPM channel, commercially available high-speed 10 . . . 40 Gbit/s telecom electronics makes it easy to implement moderately high bit rates. For example, by transmitting 16PPM ($r = 4$) at a slot rate of $R_{\text{slot 16PPM}} = 10\,\text{GSlot/s}$, a bit rate of $R_{b\,16\text{PPM}} = 2.5\,\text{Gbit/s}$ can be delivered with a symbol rate of $R_{s\,16\text{PPM}} = 10\,\text{GBd}\,/\,16 = 625\,\text{MBd}$.

The low duty cycle of M-PPM waveforms can also lead to impairments due to optical nonlinearities, so that the peak transmit power must be limited. Naturally, this does not apply to FSO systems, were the low duty cycle in combination with the high peak power delivered by EDFA during short signal bursts

even helps in bridging longer distances. On the receive side, M-PPM requires two clocks to be recovered, a symbol and a slot clock. Clock acquisition can be challenging for large M since for a given average optical power the received electrical power at the clock frequencies becomes smaller according to $1/M^2$, which may require embedded synchronization bits.

For the same pulse width $T_{M\text{-PPM}} = T_{\text{NRZ-OOK}}$, the M-PPM waveforms compared to NRZ-OOK have a smaller bandwidth ratio B_r, and a significantly smaller fractional (or relative) bandwidth B_f. This is mainly due to the smaller 10 dB bandwidth $B_{10\,\text{dB}}$ which is defined at frequencies f_h and f_l, where the power spectrum is 10 dB down from its maximum,

$$B_r = \frac{f_h}{f_l}\,, \qquad B_f = \frac{B_{10\,\text{dB}}}{(f_h + f_l)/2}\,, \qquad B_{10\,\text{dB}} = f_h - f_l \tag{2.64}$$

For a constant bit rate, the M-PPM spectra are broadened according to Eq. (2.62) on Page 46 by a factor of M/r, which increases the 10 dB bandwidth $B_{10\,\text{dB}}$, but does not impact B_r. Assuming a pseudo-random bit sequence (PRBS) of 10 Gbit/s NRZ-OOK waveforms, B_r extends from a practical lower bound of $f_l = 10\,\text{kHz}$ up to $f_h = 10\,\text{GHz}$ (six decades!). In contrast, the spectra for M-PPM waveforms operating at the same data rate span less than two decades, reducing B_r by over 4 orders of magnitude, despite having more high-frequency content. This relaxes the performance requirements on wide-band electronic amplifiers and drivers. In addition, since the longest string of consecutive logical 1 comprises two logical 1 (from two adjacent PPM symbols), pattern-dependencies in transmit and receive hardware are reduced, making it easier to generate and receive high-quality waveforms.

The M-PPM format also benefits from the sequential nature of the symbol set, which enables a single-chain of drive electronics and associated filters to generate and receive the complete symbol set. This simplifies and improves the decision process, since it is easier to make a fair comparison of the M-samples within a symbol to determine which is the largest.

Finally it should be remarked that M-PPM can be combined with M-ary QAM, PSK, FSK, and additionally with PMSK (see Fig. 2.11(b) on Page 36) for an even higher sensitivity[92].

[92] Ludwig, A.; Schulz, M.-L.; Schindler, P.; Wolf, S.; Koos, C.; Freude, W.; Leuthold, J.: Stacked modulation formats enabling highest sensitivity optical free-space links. Opt. Express 23 (2015) 21942–21957

Chapter 3

Optical transmitters

Among the variety of optical sources, optical fiber communication systems almost always use semiconductor-based light sources such as light-emitting diodes (LED) and laser diodes (LD) because of the advantages such sources have over the others. These advantages include compact size, high efficiency, required wavelength of emission, and the possibility of direct modulation at high speed. However, high data rates and phase-sensitive modulation formats call for external modulators. Besides the references given in the preface, several books covering the topic can be recommended[1,2,3,4], sorted according to complexity.

3.1 Light sources

Laser is an acronym for *l*ight *a*mplification by *s*timulated *e*mission of *r*adiation. Therefore, our first task is to understand what is meant by stimulated (synonym: induced) emission and under what conditions one can achieve amplification of light by stimulated emission. Laser — the device — may be defined as a highly monochromatic, coherent source of optical radiation. In this sense it is analogous to an electronic oscillator, which is a source of electromagnetic waves in the lower frequency range of the electromagnetic spectrum. The acronym "laser" contains the word "amplification", and obviously the optical amplifier and the laser are as closely related as the "transistor amplifier" and the "transistor oscillator". Historically, the advent of lasers preceded that of optical amplifiers, so the chapter on lasers is placed ahead that of optical amplifiers.

A laser consists of an active medium that is capable of providing optical amplification. This medium may be a collection of microsystems like atoms, molecules, or ions in the solid, liquid or gaseous form. Placed around the amplifying medium there is an optical resonator that provides the necessary optical feedback, Fig. 3.1. For an optical amplifier, this feedback is sufficiently suppressed in a certain range of the gain. The optical resonator in its simplest form consists of two plane mirrors aligned suitably to confine the optical energy as light propagates back and forth between the mirrors. Such a structure is called a Fabry-Perot[5] resonator[6,7,8]. It consists of a strip waveguide with height d and width b, and two plane mirrors with power reflection factors $R_{1,2}$ at $z = 0, L$. The active volume amounts to $V = dbL$. The waveguide has a refractive index n and is surrounded by a cladding with index n_2. With semiconductor

[1] Ghatak, A.; Thyagarajan, K.: Introduction to fiber optics. Cambridge: University Press 1998. Chapter 11

[2] Hecht, J.: Understanding fiber optics, 4. Ed. Upper Saddle River: Prentice Hall 2002. Chapter 9

[3] Agrawal, G. P.: Lightwave technology. Vol. 1: Components and devices. Hoboken: John Wiley & Sons 2004. Chapter 5

[4] Iizuka, K.: Elements of photonics, Vol. I and II. New York: John Wiley & Sons 2002. Vol. II Chapter 13 and 14

[5] Charles Fabry, French physicist, ⋆1867, †1945. — Alfred Pérot, French physicist, ⋆1863, †1925

[6] Pérot, A. and Fabry, C.: On the application of interference phenomena to the solution of various problems of spectroscopy and metrology. Astrophys. J. 9 (1899), 87–115. http://dx.doi.org/10.1086/140557
Pérot, A. and Fabry, C.: Théorie et applications d'une novelle méthode de spectroscopie interférentielle. Ann. Chim. Phys. 16 (1899) 115–44.

[7] Born, M.; Wolf, E.: Principles of optics, 6. Ed. Oxford: Pergamon Press 1980

[8] Hecht, E.: Optics. 2nd Ed. Reading: Addison Wesley 1987. See Chapter 9 Sect. 9.6.1 Page 368

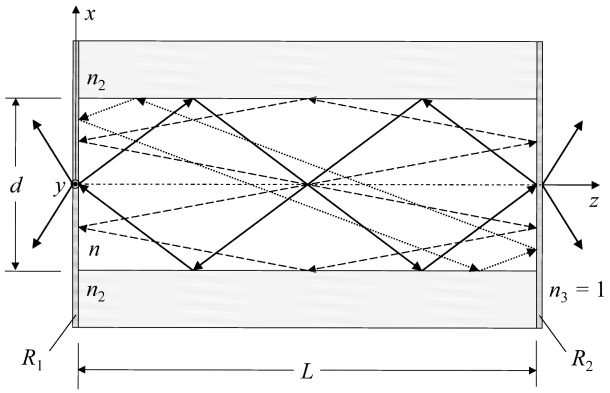


Fig. 3.1. Laser resonator modes. Resonator length L, strip waveguide height d, strip waveguide width b along y-axis, active volume $V = Lbd$, mirrors with power reflection factors $R_{1,2}$

lasers it is common to cleave the crystal at $z = 0, L$ perpendicularly to the z-axis. The power reflection factor for perpendicular incidence at such a cleaved plane semiconductor-air interface can be calculated according to Fresnel's[9] formula[10]

$$R_P = \left(\frac{n - n_3}{n + n_3}\right)^2 . \tag{3.1}$$

For a laser waveguide with refractive index $n \approx 3.6$ (GaAs) and a cleaved waveguide facet at the semiconductor-air interface ($n_3 = 1$) the power reflection factor is $R_P = 32\,\%$. Figure 3.1 shows a few closed ray paths for visualizing possible *modes* in a laser resonator.

Number of modes Inside the homogeneous resonator medium with refractive index n the wave equation (A.2) on Page 175 is solved by monochromatic homogeneous plane waves with complex amplitudes $\widetilde{\Psi}(k_x, k_y, k_z)$, real angular frequency ω and real propagation vector $\vec{k} = k_x\vec{e}_x + k_y\vec{e}_y + k_z\vec{e}_z$ (unit vectors $\vec{e}_{x,y,z}$ in respective directions). If the components of $\vec{k}$ are fixed, a so-called separation condition $|\vec{k}| = n\omega/c$ determines the frequency. A superposition of such waves defines all possible standing or propagating fields.

The propagation constants into the directions of the coordinates $q = \{x, y, z\}$ of Fig. 3.1 are denoted by k_q with $k^2 = \sum_q k_q^2 = (n\omega/c)^2$. Further, the lengths $L_x = d$, $L_y = b$, $L_z = L$ and the integers $m_{x,y,z} = 0, 1, 2, \ldots$ and $l_q = 0, \pm1, \pm2, \ldots$ are introduced for convenience. In addition to the two *transverse* field resonance conditions $2k_{x,y}L_{x,y} = m_{x,y}\times 2\pi$, a third *longitudinal* resonance condition $2k_zL_z = m_z\times 2\pi$ fits the modal phase along the z-axis. Obviously, the possible values of $0 \le k_q \le k$ are discrete and describe standing waves or modes,

$$k_q = m_q\delta k_q\,, \quad \delta k_q = \frac{\pi}{L_q}\,, \qquad q = x, y, z\,, \quad m_q = 0, 1, 2, \ldots \tag{3.2}$$

Figure 3.24 on Page 91 displays a typical spectrum of resonator lines. As an example for computing the number m_q of possible laser resonator modes assume the following: A box-shaped active volume with lengths L_q, an amplification half-power bandwidth Δf_H, a modal frequency spacing δf_q, and no

[9] Augustin-Jean Fresnel (pronounced [ɔgy'stɛ̃ ʒã fʁɛ'nɛl]), French physicist, ⋆ Broglie (France, see Footnote 3 on Page 1) 10.05.1788, † Ville-d'Avray (France) 14.07.1827. Pioneered in optics and did much to establish the wave theory of light advanced by Thomas Young. — Fresnel served as an engineer in various departments of France but lost his post temporarily during the period following Napoleon's return from Elba (1814). About that time he seems to have begun his researches in optics. He studied the aberration of light, created various devices for producing interference fringes, and, by applying mathematical analysis to his work, removed a number of objections to the wave theory.

[10] See Ref. 8 on Page 49. Sect. 4.3, Eq. (4.67)

dispersion of the refractive index n. From Eq. (3.2) we find the estimate for the mode numbers (see also Eq. (3.65) on Page 78)

$$\delta k_q = 2\pi\frac{n}{c}\,\delta f_q = \frac{\pi}{L_q}\,, \quad \delta f_q = \frac{c}{2nL_q}\,, \quad m_{q\,\max} = \max\left(1, \frac{\Delta f_H}{\delta f_q}\right), \quad M_{\rm tot} = 2\prod_{q=x,y,z} m_{q\,\max}\,. \tag{3.3}$$

The max-function guarantees a modal count of at least 1, $m_q \geq 1$. The total number of modes $M_{\rm tot}$ results from multiplying the maximum mode numbers $m_{q\,\max}$ for all three coordinate directions in 2 polarizations: Each longitudinal mode m_z can appear in a number of $m_{x\,\max} \times m_{y\,\max}$ varieties of transverse modes, and in 2 orthogonal polarizations. Usually, semiconductor lasers are transversely single-moded and oscillate in one polarization only, so that $M_{\rm tot} = m_{z\,\max}$.The number of longitudinal resonator modes for a frequency band Δf_H and associated propagation constants $k_z = nk_0$ ($k_0 = \omega/c$) amounts to

$$M_L = m_{z\,\max} = \frac{\Delta f_H}{\delta f_z} = \Delta f_H\,\tau_U, \qquad \tau_U = \frac{2L}{c/n}. \tag{3.4}$$

The quantity τ_U is the round-trip time (German *Umlaufzeit)*. Equation (3.4) is the sampling theorem of Eq. (2.4) on Page 15 in disguise: If we observe an electromagnetic field with a bandwidth Δf_H for a time $\tau_U = 1/\Delta f_H$, then we measure amplitude and phase (or real and imaginary part) of one single longitudinal mode. In contrast to Eq. (3.65) on Page 78, the refractive index n in Eq. (3.4) is assumed to be frequency-independent, i. e., the resonator is dispersion-free.

In the following, we first review the basic emission and absorption processes in a microsystem (an atom, a molecule, or an electron in the conduction band of a semiconductor), and then discuss the conditions for light amplification and laser oscillation in a semiconductor.

3.1.1 Luminescence and laser radiation

The excitation energy W_2 of any microsystem may be released by a transition to a state of lower energy W_1. This transition can be radiative by emission of a photon with energy $hf = W_2 - W_1$ (Planck's constant[11] h, frequency f), or nonradiative. The transition probability depends on the quantum mechanical properties of the microsystem, and on the interaction with an electromagnetic field. The microsystem may also gain energy and make an upward (radiative) transition by absorbing an amount of electromagnetic

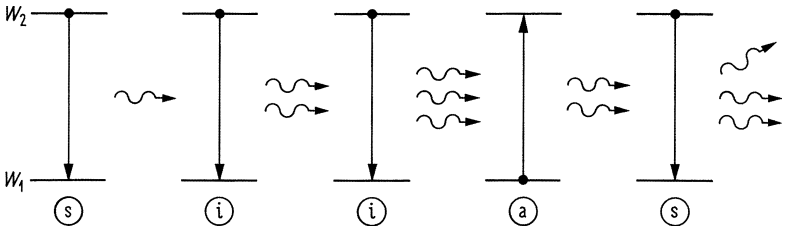


Fig. 3.2. Interaction of a two-level microsystem with electromagnetic radiation, photon energy $hf = W_2 - W_1$. (a) absorption, (s) spontaneous emission, and (i) induced (= stimulated) emission of photons

energy $hf = W_2 - W_1$. The released photon energy is emitted into any mode of the electromagnetic field. For our present purpose the term "mode" may be associated with the resonating modes in the active volume V. Nonradiative transitions transfer the same amount of energy to or from phonons (thermal vibrations of the crystal lattice) or other degrees of freedom of the interacting substances. Referring to Fig. 3.2, there are three types of interactions:

Absorption A microsystem in its ground state W_1 can absorb radiation at a frequency $f = (W_2 - W_1)/h$ and make an upward transition to the higher energy level W_2. This absorption process is therefore *induced* or *stimulated* by an existing electromagnetic field. The absorption rate depends on the electromagnetic energy density, and on the number of microsystems in the ground state, Fig. 3.2(a).

[11] See Footnote 2 on Page 1

Spontaneous emission An excited microsystem in level W_2 can make a downward transition to the ground state W_1 "spontaneously" (apparently without any interaction) by emitting a photon with energy $hf = W_2 - W_1$, Fig. 3.2(s). The spontaneous emission rate depends only on the number of excited microsystems.

The term "spontaneous" needs more explanation. In a *semiclassical theory*, the microsystem is treated as a quantum mechanical system, while the field description is classical. As a result, an excited microsystem would stay in its excited state W_2 for an infinite time period.

Experimentally it is observed that microsystems release their energy at random after a certain average spontaneous lifetime $\tau_{\rm sp}$. In *quantum electrodynamics* the particle (photon) nature of radiation is formulated by a quantization of the electromagnetic field. The outcome is that the electric and the magnetic fields $\vec{E}$ and $\vec{H}$ are connected by an uncertainty relation. Therefore, the simultaneous states $\vec{E}(t,\vec{r}) = 0$ and $\vec{H}(t,\vec{r}) = 0$ for all $t, \vec{r}$ are impossible. However, the quantum electrodynamical vacuum, as defined by the expected values $\overline{\vec{E}(t,\vec{r})} = 0$, $\overline{\vec{H}(t,\vec{r})} = 0$ for all $t, \vec{r}$, *is* allowed. For this case, the average energy density $\overline{\epsilon_0 \vec{E}^2/2 + \mu_0 \vec{H}^2/2}$ is finite such that the total mean energy in each state of the electromagnetic field with a certain polarization amounts to $hf/2$ (zero point energy).

This energy cannot be extracted from the system (it cannot be used to fry eggs), but the fields fluctuating around the expectation zero represent a perturbation for an excited microsystem, and may therefore *induce random transitions* to the ground state. These "spontaneously" emitted photons will be found with equal probability in any possible mode of the electromagnetic field, because all modes possess the same zero point energy $hf/2$. A spontaneously emitted photon modifies an already existing field in a mode by superimposing an additional field with a random phase thus establishing a *noise* signal. Incoherent radiation of this type is called luminescence.

Spontaneous absorption (absorption induced by the zero point field fluctuation) is not an allowed process, because the zero point energy cannot be extracted from the electromagnetic vacuum and therefore not be transferred to a microsystem.

Induced emission A microsystem in an excited level W_2 can also make a downward transition to the ground state W_1 in the presence and induced (synonym: stimulated) by an external radiation of frequency $f = (W_2 - W_1)/h$. As in the case of (induced) absorption, the emission rate depends on the electromagnetic energy density, and on the number of microsystems in the excited state, Fig. 3.2(i). In contrast to spontaneous emission processes (= transitions induced by zero point fluctuations) the emitted radiation is phase coherent with the stimulating radiation. Therefore, the induced radiation adds with the same polarization and phase to the stimulating field and becomes amplified much like by an electronic amplifier.

Lifetime and linewidth

As explained above, spontaneous absorption is impossible. Further, a microsystem in its ground state does not possess the energy to radiate a photon. Therefore, the lifetime of the ground state is infinite. From quantum theoretical considerations the system energy may be determined inside the observation time τ_2 with an uncertainty ΔW_2 ($\hbar = h/(2\pi)$),

$$\Delta W_2 \, \tau_2 \geq \hbar/2 \,, \quad \Delta f_2 \geq 1/(4\pi\tau_2) \,. \tag{3.5}$$

If the excited state energy is not exactly known because of the finite lifetime $\tau_2 \leq \tau_{\rm sp}$ (by induced emission the excited state lifetime may become smaller than $\tau_{\rm sp}$), the spontaneously emitted photon energy hf with expectation $\overline{hf_0} = \overline{W_2 - W_1}$ is uncertain by $\Delta W_2 \geq \frac{1}{2}\hbar/\tau_{\rm sp}$. Therefore, the probability density of the spontaneous emission (luminescence spectrum) has a lineshape $\rho(f)$ with a maximum at $f = f_0$ and a linewidth of $\Delta f \sim 1/\tau_{\rm sp}$.

Laser action

In the presence of external radiation, the stimulated absorption probability per microsystem is the same as the stimulated emission probability. A net emission is possible if the number of excited state microsystems exceeds the number of ground state systems. This does not correspond to the normal population distribution in thermal equilibrium, where the number of ground state microsystems is larger than the number of excited systems, and is therefore called *population inversion*. Such an inversion has to be provided by some kind of a "pump" mechanism, and is a prerequisite for light amplification. For concentrating next to all of the emitted photons in a narrowband spectral range smaller than $\Delta f_2 = 1/(4\pi\tau_{\mathrm{sp}})$, a resonance structure as in Fig. 3.1 provides the necessary means.

Amplification and oscillation The modes of an optical resonator having a lossless transverse guiding structure and partially transmitting mirrors at $z = 0, L$ are characterized at a certain light resonance frequency f_S by a quality factor $Q = f_S/\Delta f_S$, which defines a resonator bandwidth Δf_S and a photon lifetime $\tau_P \sim 1/\Delta f_S$, which is caused mainly by transmission losses of the mirrors. If the resonator volume V contains excited microsystems, the spontaneously emitted photons are collected by all resonator modes. If the maximum of the spontaneous emission line $\rho(f_0)$ is centred at a resonator mode $f_S = f_0$, this special mode collects a larger number N_P of photons than other modes. Because the induced emission is in proportion to N_P, the emission probability increases with N_P. For a population inversion condition the induced absorptions are less than the stimulated emissions, and a net stimulated emission rate results causing a coherent amplification of the light in mode f_S.

With increasing pump rate the gain becomes higher. When the resonator losses are just compensated, the so-called threshold of oscillation is reached. With increasing pump rate, the photon number increases at first exponentially, and so does the probability $1/\tau_2$ of stimulated emissions per second. Each additionally excited microsystem releases its energy practically immediately after an effective lifetime $1/\tau_{2\,\mathrm{eff}} = 1/\tau_2 + 1/\tau_{\mathrm{sp}}$, and the probability $1/\tau_{\mathrm{sp}}$ of spontaneous emissions per second into this mode becomes less and less important, $1/\tau_{2\,\mathrm{eff}} \approx 1/\tau_2$.

Because spontaneous emission is reduced compared to induced emission, the field becomes more coherent. The number of photons N_P in the dominant resonator mode stabilizes at such a high stationary level N_{P0} that practically all microsystems, which are excited additionally by the pump, release their energy immediately by induced (coherent) emission, and the gain gets clamped at the threshold level. The stimulated-emission photons compensate the total resonator losses. A light field develops having a near-sinusoidal time dependence $E(t) = A(t)\cos[\omega_S t + \varphi(t)]$, where amplitude $A(t)$ and phase $\varphi(t)$ vary slowly on the scale of an optical period $1/f_S$ (e. g., for a semiconductor laser $f_S = 193\,\mathrm{THz}$ at $\lambda = 1.55\,\mu\mathrm{m}$, $1/f_S = 5.2\,\mathrm{fs}$). This leads to a very narrow spectral linewidth Δf_S. For a so-called *d*istributed *f*eed*b*ack (DFB) semiconductor laser a typical value is $\Delta f_S = 4 \ldots 40\,\mathrm{MHz}$. Such laser have a narrow-band resonator, where the feedback is established not by endface mirrors, but by a Bragg grating along the whole resonator.

If there was a momentary increase in photon number $N_P > N_{P0}$, the increased stimulated emission would deplete the population inversion, and the gain would decrease, followed by a reduction of the photon number to $N_P < N_{P0}$. This being the case, stimulated emission is below its stationary value, the pump rebuilds the inversion, and a so-called relaxation oscillation of photon number and inversion (gain) is to be expected. This corresponds to an energy exchange between two energy reservoirs, like with inductor and capacitor in a resonant circuit. For semiconductor lasers, relaxation oscillations of the optical intensity occur at microwave frequencies in the order of $f_r = 1 \ldots 30\,\mathrm{GHz}$.

If there were no longitudinal resonator ($R_{1,2} = 0$ in Fig. 3.1), a wave travelling through the active medium would be amplified. Residual mirror reflectivities $R_1, R_2 \neq 0$ could lead to a regenerative oscillation and have to be avoided in this case.

Modulation By changing the pump rate, the population inversion can be modified. If the light source has no optical resonator and therefore emits only spontaneous radiation, the number of excited microsystems cannot decrease faster than their lifetime τ_{sp}, and the maximum light intensity modulation frequency

is in the order of $1/\tau_{\mathrm{sp}}$ if nonradiative transitions are excluded. When the lifetime τ_{sp} is reduced by additional nonradiative processes, $1/\tau_{2\,\mathrm{eff}} = 1/\tau_{\mathrm{nr}} + 1/\tau_{\mathrm{sp}}$, the limiting modulation frequency increases to the order of $1/\tau_{2\,\mathrm{eff}}$, but the radiation efficiency with respect to the pump rate decreases.

In a laser above threshold, the effective lifetime $1/\tau_{2\,\mathrm{eff}} = 1/\tau_2 + 1/\tau_{\mathrm{sp}} \gg 1/\tau_{\mathrm{sp}}$ of the excited microsystems can be made very much smaller than the spontaneous lifetime τ_{sp} by the mechanism of stimulated emission, without paying the prize of a reduced radiation efficiency.

Noise The noise properties of a luminescent device and a laser are completely different. Spontaneous emission represents a noise signal of the electromagnetic field with an expectation of zero and with a non-zero intensity, very much like thermal noise from a resistor (e. g., an incandescent lamp). On the other hand, a laser signal resembles a sinusoidal field perturbed by the noise of spontaneous emissions. The amplitude fluctuations are relatively small because of the nonlinear amplitude control described in Sect. 3.1.1 on Page 53 (gain clamping). The magnitude of the phase (or frequency) fluctuation depends mainly on the resonator bandwidth.

3.1.2 Laser active materials

For a microsystem in thermal equilibrium the occupation probabilities $p(W_i)$ of the various energy levels W_i at any absolute temperature T are given by the Maxwell-Boltzmann statistics (ground state W_1, degeneracy g_i of level W_i, Boltzmann[12] constant $k = 1.380\,658 \times 10^{-23}\,\mathrm{Ws\,/\,K}$)

$$p(W_i) = g_i\,\mathrm{e}^{-W_i/(kT)} \Big/ \sum\nolimits_i g_i\,\mathrm{e}^{-W_i/(kT)}\,. \tag{3.6}$$

Two-level systems

For non-degenerate two-level microsystems as in Fig. 3.2, the population numbers $N_{1,2}$ of a microsystem with energy states $W_{1,2}$ in thermal equilibrium are related by

$$\frac{N_2}{N_1} = \mathrm{e}^{-(W_2-W_1)/(kT)}, \qquad N = N_1 + N_2\,. \tag{3.7}$$

The quantity N is the total number of microsystems. In thermal equilibrium the excited state is less densely populated than the ground state by an exponential factor depending on the difference energy $hf = W_2 - W_1$ with respect to the thermal energy kT. With induced absorption as described in Fig. 3.2(a) the population number N_2 can be increased in proportion to the photon number N_P which is available in a resonator mode of frequency f, and in proportion to the time t. On the other hand, spontaneous emission reduces N_2 in proportion to t, and stimulated emission diminishes N_2 in proportion to N_P and t. Therefore, in the presence of an electromagnetic field of photon energy hf, a dynamic equilibrium will be reached for which the number of spontaneous and induced emissions equals the number of stimulated absorptions. If the photon number N_P is so large that spontaneous emission may be neglected, an dynamic equilibrium state $N_2 = N_1$ (with spontaneous emission: $N_2 \le N_1$) may be reached so that the number of stimulated emissions equals the number of stimulated absorptions. The medium is called transparent in this case. However, with a strict two-level system it is impossible to achieve a gain by population inversion.

Three-level systems

The situation is improved with a three-level system, Fig. 3.3(a). If we pump the system at an absorption frequency $f^{(\mathrm{a})} = (W_3 - W_1)/h$, microsystems can get excited from level W_1 to level W_3, where in the non-degenerate case the occupation numbers are related by $N_3 \le N_1$. The excited microsystems make a downward transition (releasing their energy radiatively or noradiatively) also to level W_2. If

[12] Ludwig Boltzmann, Austrian physicist, ⋆ Wien 20.2.1844, † Duino (Duino-Aurisina, near Trieste) 5.9.1906 (suicide). Professor in Graz, Wien, München, Leipzig

for a given pumping rate the transition rate $1/\tau_{32}$ for $W_3 \to W_2$ exceeds that of the transition rates $1/\tau_{31}$ for $W_3 \to W_1$ and $1/\tau_{21}$ for $W_2 \to W_1$, the microsystems in energy level W_3 deplete (thereby reducing the occupation number of the ground state $N_1 \approx N_3$), and accumulate in energy level W_2, so that a population inversion with $N_2 > N_1$ and an associated net gain for the signal emission frequency $f^{(\mathrm{e})} = (W_2 - W_1)/h$ becomes possible. For reaching the gain threshold a very high pump rate is necessary because the occupation number of the ground state is very high, Eq. (3.7). The pump efficiency cannot be larger than $hf^{(\mathrm{e})}/(hf^{(\mathrm{a})})$,

$$\eta_p \leq \frac{W_2 - W_1}{W_3 - W_1} = \frac{hf^{(\mathrm{e})}}{hf^{(\mathrm{a})}} \,. \tag{3.8}$$

Practically, η_p is much smaller, because not all pump photons excite microsystems with an energy W_3, and not all excited systems end up in level W_2.

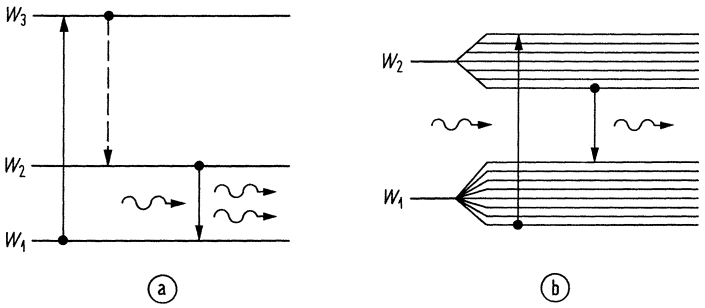


Fig. 3.3. Pump mechanism using energy levels (a) outside (three-level laser system) or (b) inside the energy level group of the laser transition (pseudo-four-level laser system)

Four-level systems and semiconductors

A more efficient pumping scheme can be realized by a four-level system with pump levels $W_{0,3}$ and laser levels $W_{2,1}$ because the final state W_1 for the lasing transition is different from the densely populated ground state W_0. A pseudo-four-level scheme is depicted in Fig. 3.3(b). The lasing levels $W_{2,1}$ split up into closely neighbouring sublevels. According to the equilibrium distribution Eq. (3.7), the occupation probability of the lowest energy level ($\widehat{=} W_0$) is highest, and of the highest energy levels ($\widehat{=} W_3$) lowest, so that absorption from the lowest energy states to the highest ones is a most probable process. On the other hand, emission from a strongly populated level ($\widehat{=} W_2$) to a sparsely populated level ($\widehat{=} W_1$) is very probable. Therefore, the maximum for absorption is found at higher frequencies (shorter wavelengths) than the maximum for luminescence, and effective pumping may be achieved at a slightly shorter wavelength then the lasing emission wavelength. This mechanism can be used in Erbium-doped fibre amplifiers (EDFA), where an absorbed pump power at $\lambda^{(\mathrm{a})} = 1.48\,\mu$m produces an optical gain at the emission signal wavelength $\lambda^{(\mathrm{e})} = 1.53\,\mu$m. The maximum pump efficiency in this case is $\eta_p = 1.48\,\mu\mathrm{m}\,/1.53\,\mu\mathrm{m} = 97\,\%$.

The scheme of Fig. 3.3(b) can be also applied to the case of a semiconductor device. Levels W_2 and W_1 are to be associated with conduction and valence band states, respectively. Pump light with a photon energy $hf^{(\mathrm{a})}$ ($\widehat{=} W_3 - W_0$) is absorbed for producing electron-hole pairs in the appropriate energy levels. This could be also achieved with a forward biased semiconductor pn-diode by injecting electrons and holes into the conduction and valence band, respectively, so that population inversion is reached. For a temperature $T = 0$ the "pump energy" $eU = hf^{(\mathrm{a})}$ given by the forward voltage U (elementary charge $e = 1.602\,177\,33 \times 10^{-19}\,\mathrm{A\,s}$) would define the minimum energetic difference at which electrons and holes could be injected, i. e., the difference of the quasi Fermi[13] levels $W_{Fn} - W_{Fp} = eU$ for electrons in the

[13] Enrico Fermi, Italian physicist, ⋆ Rome 29.9.1901, † Chicago (Illinois) 28.11.1954. Professor in Rome and later in USA. Italian-born American physicist who was one of the chief architects of the nuclear age. He developed the mathematical

conduction band (W_{Fn}) and for holes in the valence band (W_{Fp}), respectively. The energy $hf^{(e)}$ of the emitted photons is therefore smaller than $W_{Fn} - W_{Fp}$, but necessarily larger than the bandgap W_G,

$$W_G < hf^{(e)} < W_{Fn} - W_{Fp} = eU\,, \qquad hf^{(a)} > W_{Fn} - W_{Fp} = eU\,. \tag{3.9}$$

While for a simplified argument we had assumed a fictitious device temperature $T = 0$ (however, for low temperatures, the impurity doping of semiconductors "freezes out"[14], and the device stops functioning), the result Eq. (3.9) holds also true for arbitrary temperatures as we will see in Eq. (3.37) on Page 69.

Such a laser diode could be also used as a photodetector for photon energies $hf^{(a)} > W_G$ without fixing a bias voltage U. Each photo-generated electron-hole pair is separated by the field inside the pn-junction and induces a current with a time integral e in the external circuit. Basically, this unbiased diode represents a solar cell.

3.1.3 Compound semiconductors

In Sect. 3.1.2 we had discussed which properties a material should have for the generation and amplification of light. Here, we specify important properties of the III-V compound semiconductors (Ga,Al)(As,Sb) and (In,Ga)(As,P). The bandgap W_G (and hence the bandgap wavelength λ_G and the refractive index n) depend on the composition. The lattice constant may be chosen to match the lattice constant of a binary substrate semiconductor. Lattice matching is very important for several reasons:

- A close lattice match is necessary in order to grow high-quality crystal layers.
- Excess lattice mismatch between the heterostructure layers results in crystalline imperfections which lead to nonradiative recombination and thus prevent lasing.
- Lattice mismatch causes degradation in devices during operation.

Elemental semiconductors as Si and Ge have a diamond structure, while compound semiconductors as GaAs or InP have a zinc-blende structure. Having no inversion centre, the crystals of the zinc-blende type show a linear electro-optic effect, so these substances my be also used to construct modulators and switches.

Figure 3.4 shows the bandgap W_G and the lattice constant a for two compound material systems. Tables 3.1 and 3.2 summarize the numerical values. With ternary compound crystals (see Table 3.1, Fig. 3.4), active $(Ga_{1-x}Al_x)As$ layers slightly mismatched to a GaAs substrate may be grown for laser diode emission wavelengths $\lambda = 0.69 \ldots 0.87\,\mu$m.

Using quaternary compound crystals $(Ga_{1-x}Al_x)(As_ySb_{1-y})$ lattice-matched to a GaSb-substrate, laser diodes can be fabricated emitting at $\lambda = 1.25 \ldots 1.71\,\mu$m; this material is also well suited for long-wavelength detectors. For photodetectors, indirect semiconductors are applicable, and lattice-matched compound crystals on GaSb may be grown, leading to an absorption energy $W_G = 0.726 \ldots 1.6\,$eV, $\lambda_G = 1.71 \ldots 0.78\,\mu$m. There is an miscibility gap of unknown extent for compound crystals with similar concentrations of As and Sb.

With the material system $(In_{1-x}Ga_x)(As_yP_{1-y})$ laser diodes and photodiodes are grown on lattice-matched InP substrates ($\lambda = 0.92 \ldots 1.65\,\mu$m). With GaAs substrates, emission wavelength in the region $\lambda = 0.87\,\mu$m (GaAs) down to $\lambda = 0.68\,\mu$m ($In_{0.49}Ga_{0.51}P$) become possible. High-quality GaAs substrates are available with relatively large wafer diameters of 3 in (8 cm) and 4 in (10 cm). GaAs-based integrated circuits are a standard technique, and so the construction of optoelectronic integrated circuits with (In,Ga)(As,P), lattice-mismatched to a GaAs substrate, $\lambda = 1.3\,\mu\text{m} \mathrel{\widehat{=}} 0.95\,$eV could mature to become a cost-saving alternative. Much more expensive is the processing of (In,Ga)(As,P) on typically 2 in (5 cm) InP substrates, $\lambda = 1.55\,\mu\text{m} \mathrel{\widehat{=}} 0.8\,$eV.

---

statistics required to clarify a large class of subatomic phenomena, discovered neutron-induced radioactivity, and directed the first controlled chain reaction involving nuclear fission. He was awarded the 1938 Nobel Prize for Physics, and the Enrico Fermi Award of the U. S. Department of Energy is given in his honour.

[14] See Footnote 19 on Page 60

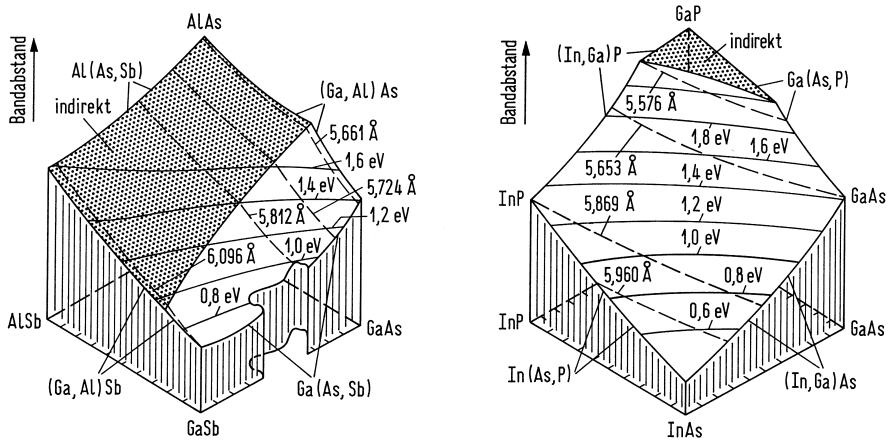


Fig. 3.4. Material systems $(Ga_{1-x}Al_x)(As_ySb_{1-y})$ and $(In_{1-x}Ga_x)(As_yP_{1-y})$, bandgaps and lattice constants. Dotted region: indirect semiconductor (Bandabstand $\widehat{=}$ bandgap)

Semiconductor	$W_G/\,\mathrm{eV}$ $(\lambda_G/\,\mu\mathrm{m})$	n at λ_G	$a/\mathrm{Å}$
GaSb, direct	0.726 (1.708)	3.82	6.096
GaAs, direct	1.424 (0.871)	3.655	5.653
AlSb, indirect	1.58 (0.785)	3.4	6.135
AlAs, indirect	2.163 (0.573)	3.178	5.660
$(Ga_{1-x}Al_x)As$ direct: $x \le 0.3$	$1.424 + 1.247\,x$ $1.424 \ldots 1.798$ $(0.871 \ldots 0.69)$	$3.59 - 0.71\,x +$ $+0.091\,x^2$ (at $\lambda = 0.9\,\mu\mathrm{m}$)	$5.653 + 0.027\,x$
$(Ga_{1-x}Al_x)(As_ySb_{1-y})$ lattice-matched to GaSb direct: $x \le 0.24$ $y = x/1.11$	$0.726 + 0.834\,x +$ $+1.134\,x^2$ $0.726 \ldots 0.991$ $(1.708 \ldots 1.25)$	?	6.096

Table 3.1. Material system $(Ga_{1-x}Al_x)(As_ySb_{1-y})$. W_G bandgap, $\lambda_G = hc/W_G$ bandgap wavelength, n refractive index, a lattice constant

Semiconductor	$W_G/\,\mathrm{eV}$ $(\lambda_G/\,\mu\mathrm{m})$	n at λ_G	$a/\mathrm{Å}$
InAs, direct	0.36 (3.444)	3.52	6.058
InP, direct	1.35 (0.918)	3.45	5.869
GaAs, direct	1.424 (0.871)	3.655	5.653
GaP, indirect	2.261 (0.548)	3.452	5.451
$(In_{0.49}Ga_{0.51})P$, direct lattice-matched to GaAs	1.833 (0.676)	3.451 ?	5.653
$(In_{0.53}Ga_{0.47})As$, direct lattice-matched to InP	0.75 (1.653)	3.61	5.869
$(In_{1-x}Ga_x)(As_yP_{1-y})$ lattice-matched to InP direct: $y \le 1$ $x = y/(2.2091 - 0.06864\,y)$	$1.35 - 0.72\,y +$ $+0.12\,y^2$ $1.35 \ldots 0.75$ $(0.918 \ldots 1.653)$	$3.45 + 0.256\,y -$ $-0.095\,y^2$ $3.45 \ldots 3.61$	5.869

Table 3.2. Material system $(In_{1-x}Ga_x)(As_yP_{1-y})$. W_G bandgap, $\lambda_G = hc/W_G$ bandgap wavelength, n refractive index, a lattice constant

3.1.4 Semiconductor physics

The simplest laser diode structure is a pn-homojunction biased with a forward current I, Fig. 3.5. Spontaneously emitted light leaves the active layer in all possible directions. The field is guided in the active region strip waveguide and reflected from the cleaved end facets at $z = 0, L$, which form a Fabry-Perot resonator. To understand the device properties, we have to recall some semiconductor basics in the following.

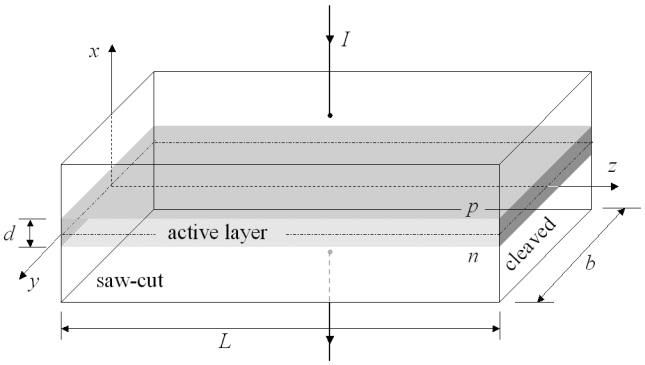


Fig. 3.5. Forward biased semiconductor pn-homojunction acting as a laser diode. Side-walls are saw-cut, the end facets are cleaved. Typical dimensions: $d = 0.1 \ldots 0.2\,\mu$m (active layer), $b = 3 \ldots 6\,\mu$m, $L = 200 \ldots 600\,\mu$m

Energy bands and density of states

Let us consider a linear crystal consisting of a chain of N equal atoms spaced a lattice constant a apart having a length $L = Na$. Each isolated atom contributes a single bound electron with energy levels W_i'. If the N atoms interact, the N electrons do not remain bound to a fixed atom, but belong collectively to the crystal and assume N energy states $W_{i\mu}$ near W_i'. This splitting of energies is in analogy to the coupling of identical resonance circuits to form a bandpass filter. The probability density function $w_{i\mu}(\vec{r})$ to find an electron of energy $W_{i\mu}$ at a position $\vec{r}$ is given by the modulus squared of the quantum mechanical wave function $\Psi_{i\mu}(\vec{r})$ (probability density amplitude, Schrödinger[15] function),

$$\Psi_{i\mu}(x) = \frac{1}{\sqrt{L}}\, u_i(k_\mu, x)\, \mathrm{e}^{\,\mathrm{j}\, k_\mu x}, \qquad u_i(k_\mu, x) = u_i(k_\mu, x + a), \qquad k_\mu = \mu\, \frac{2\pi}{Na}\,, \tag{3.10}$$

$$\int_0^L |\Psi_{i\mu}(x)|^2\, \mathrm{d}x = 1, \qquad W_{i\mu} = W_i(k_\mu), \qquad N \text{ values for } \mu = 0, \pm 1, \pm 2, \ldots, (N-1)/2\,.$$

The functions differ from each other by a parameter $k_\mu = \mu \times 2\pi/L$ having N discrete values. This is analogous to the number of longitudinal modes in a resonator[16], see Eq. (3.2) on Page 50. The higher the electron energy becomes, the less it is influenced by the periodic atomic potentials, and the lattice-periodic function $u_i(k_\mu, x)$ approaches asymptotically one. A free electron of mass m moving in a constant potential W_0 has an energy $W_0 + p_\mu^2/2m$ given by its mechanical momentum p_μ. It may be described by a probability density wave with a de Broglie[17] wavelength $\lambda_\mu = h/p_\mu$, so that

$$k_\mu = \frac{2\pi}{\lambda_\mu} = \frac{2\pi}{h}\, p_\mu = \frac{p_\mu}{\hbar}, \quad \hbar k_\mu = p_\mu, \quad W = W_0 + \frac{p_\mu^2}{2m} = W_0 + \frac{\hbar^2 k_\mu^2}{2m}\,. \tag{3.11}$$

For a free electron, the product $\hbar k_\mu$ denotes the mechanical momentum p_μ, which justifies the plane-wave ansatz Eq. (3.10). For crystal electrons the quantity $\hbar k_\mu$ cannot be interpreted as an electron momentum, but in interactions with photons it represents an invariant together with the photon momentum $p = h/\lambda = \hbar k$ for a field of wavelength λ and propagation constant k. Therefore $\hbar k_\mu$ is called the crystal or pseudo momentum. The number of possible mutual exclusive electron energy states is $2N$ regarding the spin degeneracy. The N energy eigenvalues $W_i(k_\mu)$ represent the bandstructure of the band i, which resulted from the level W_i' of the isolated atom. The function $W_i(k_\mu)$ is periodic in k_μ with a period $2\pi/a$ and

[15] See Footnote 20 on Page 20

[16] As is common in semiconductor physics, the modal index μ takes positive and negative values in contrast to Eq. (3.2), where m_q is non-negative. Therefore, the relation $k_\mu = \mu \times 2\pi/(Na)$ has an additional factor 2 as compared to $k_z = m_z \times \pi/L_z$ in Eq. (3.2).

[17] See Footnote 3 on Page 1

may be therefore restricted to a region $-\pi/a < k_\mu \leq \pi/a$ (first Brillouin zone). The bandstructure Fig. 3.6 has the symmetry property $W_i(k_\mu) = W_i(-k_\mu)$ and possesses extrema at the borders of the Brillouin zone. The topmost band which is fully occupied at $T = 0$ is called valence band (VB), the lowest band

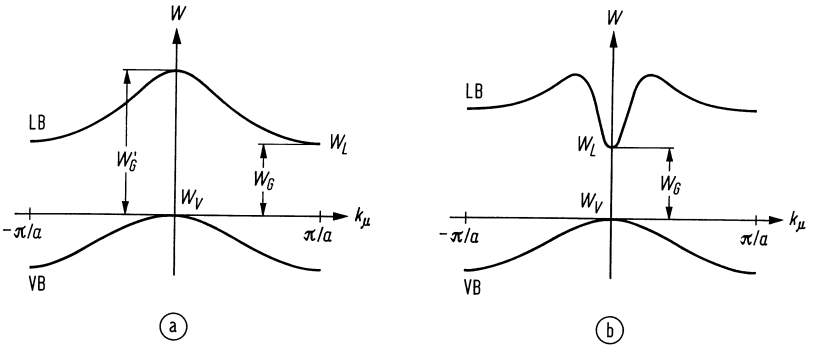


Fig. 3.6. Bandstructures of conduction band (CB, *German* Leitungsband, LB) and valence band (VB) of an (a) indirect semiconductor and of a (b) direct semiconductor. Minimum bandgap energy W_G, bandgap energy W_G' for an indirect semiconductor at $k_\mu = 0$

which is empty for $T = 0$ is called conduction band (CB, *German* Leitungsband, LB). The difference between the lowest energy level of the CB and the highest energy state of the VB is denoted as bandgap energy $W_G = W_C - W_V$. Because of thermal excitation at $T > 0$ the lowest CB states are occupied while the highest VB states are empty. For a transition $W_C \to W_V$ in an indirect semiconductor the crystal momentum changes by an amount of $\hbar k_\mu = \hbar\pi/a$, which cannot be transferred to the emitted photon with momentum $\hbar k = \hbar \times 2\pi/\lambda$ because the lattice constant is much smaller than the wavelength, $a \ll \lambda$. So a phonon as a third interaction particle with sufficient momentum is necessary, but this tree-particle scattering is less probable than a two-particle interaction. Therefore the emission (and the absorption) of photons at energies slightly larger than $hf = W_G$ is a very unlikely process.

The elemental semiconductors Ge and Si are indirect semiconductors and therefore unsuitable for efficient light sources, Fig. 3.6(a). However, they may be successfully used for photodetectors, because the low indirect-semiconductor absorption probability may be effectively increased by an extended interaction length (i. e., by a longer absorption region). If the photon energy becomes larger than W_G', the absorption becomes very likely, because electron transitions with constant momentum near $k_\mu = 0$ are possible. For the bandgap energies the following values hold:

$$W_G = \begin{cases} 0.67\,\text{eV} \mathrel{\widehat{=}} 1.85\,\mu\text{m (Ge)} \\ 1.13\,\text{eV} \mathrel{\widehat{=}} 1.10\,\mu\text{m (Si)} \end{cases} \quad W_G' = \begin{cases} 0.8\,\text{eV} \mathrel{\widehat{=}} 1.55\,\mu\text{m (Ge)} \\ 3.4\,\text{eV} \mathrel{\widehat{=}} 0.36\,\mu\text{m (Si)} \end{cases} \tag{3.12}$$

In direct semiconductors, Fig. 3.6(b), the transitions from the lower CB edge to the upper VB edge and vice versa are possible for a constant crystal momentum $\hbar k_\mu = 0$. These processes are therefore very likely, so that direct semiconductors may be used both for light sources and for detectors.

Consider a direct semiconductor Fig. 3.6(b). In the vicinity of $k_\mu = 0$ the actual bandstructure may be approximated by a parabola,

$$W = W_i(k_\mu) = W_i(0) + \frac{1}{2}\frac{\mathrm{d}^2 W_i}{\mathrm{d}k_\mu^2} k_\mu^2 = W_0 + \frac{p_\mu^2}{2m_{\text{eff}}} = W_0 + \frac{\hbar^2 k_\mu^2}{2m_{\text{eff}}}\,, \tag{3.13}$$

defining an effective mass m_{eff} which should be attributed to a free electron at the same energy. The effective mass is negative for crystal electrons at the upper VB edge. An electron with charge $(-e)$ and effective electron mass $m_n = m_{\text{eff}} < 0$ can be equivalently replaced by a so-called hole (a missing state *not* occupied by an electron) with positive charge $(+e)$ and effective hole mass $m_p = |m_{\text{eff}}| > 0$.

For laser action the population of the CB and VB edges of direct semiconductors near $k_\mu = 0$ is important. If Z stands for the number of electron states with 2 spin directions and a modulus of the

crystal momentum up to $p_\mu = \hbar k_\mu$, we have in analogy to Eq. (3.3) on Page 51

$$Z = \frac{2V_\phi}{h^3}, \qquad V_\phi = VV_p, \qquad V_p = \frac{4\pi}{3}(\hbar k_\mu)^3, \qquad \hbar k_\mu = p_\mu = \sqrt{p_x^2 + p_y^2 + p_z^2}\,. \tag{3.14}$$

In k_μ-space the differential volume in a k_μ-radius interval $k_\mu \ldots k_\mu + \mathrm{d}k_\mu$ is $\mathrm{d}^3\,k_\mu = 4\pi k_\mu^2\,\mathrm{d}\,k_\mu$, and the differential number of states $\mathrm{d}Z$ which corresponds to an energy interval $W \ldots W + \mathrm{d}\,W$ amounts to

$$\mathrm{d}Z = 2V\,\frac{1}{(2\pi\hbar)^3}\,4\pi(\hbar k_\mu)^2\,\mathrm{d}(\hbar k_\mu) = 2V\frac{1}{(2\pi)^3}\,\mathrm{d}^3\,k_\mu = V\rho\,(W)\,\mathrm{d}W\,. \tag{3.15}$$

Using Eq. (3.13) we find for the so-called density of states (DOS) $\rho\,(W)$,

$$\rho(W) = \frac{1}{V}\frac{\mathrm{d}Z}{\mathrm{d}W} = \frac{1}{V}\frac{\mathrm{d}Z}{\mathrm{d}p_\mu}\frac{\mathrm{d}p_\mu}{\mathrm{d}W} = \frac{1}{2\pi^2}\left(\frac{2|m_{\mathrm{eff}}|}{\hbar^2}\right)^{3/2}\sqrt{\pm(W - W_0)}\,. \tag{3.16}$$

For the CB we take the positive sign $\sqrt{+(W - W_0)}$ and $W_0 = W_C$, $|m_{\mathrm{eff}}| = m_n$, and for the VB the negative sign $\sqrt{-(W - W_0)}$, $W_0 = W_V$, $|m_{\mathrm{eff}}| = m_p$. The effective density of states N_B is defined as

$$\frac{\mathrm{d}Z}{\mathrm{d}W}\,kT = V\frac{2}{\sqrt{\pi}}N_B\sqrt{\pm\frac{W - W_0}{kT}}, \qquad N_B = 2\left(\frac{2\pi|m_{\mathrm{eff}}|kT}{h^2}\right)^{3/2}. \tag{3.17}$$

Because $\int_{W_0}^{W_0+kT}\sqrt{\frac{W-W_0}{kT}}\,\frac{\mathrm{d}W}{kT} = \int_0^1\sqrt{W'}\,\mathrm{d}\,W' = \frac{2}{3}$ holds, we find $\frac{1}{V}\int_{W_0}^{W_0+kT}\mathrm{d}Z = \frac{2}{\sqrt{\pi}}\frac{2}{3}N_B \approx 0.752\times N_B$. Therefore the effective density of states N_B specifies approximately the density of states inside an energy interval $W_0 \ldots W_0 \pm kT$ measured from the band edge energy W_0. The effective DOS near the conduction and valence band edges are N_C and N_V, respectively. With the free electron rest mass m_0 and the effective carrier masses for GaAs and InP at $T = 293\,\mathrm{K}$ we find the values specified in Table 3.3. A doped

	m_n/m_0	m_p/m_0	$N_C/\,\mathrm{cm}^{-3}$	$N_V/\,\mathrm{cm}^{-3}$
vacuum	1	–	2.42×10^{19}	–
GaAs	0.067	0.48	4.20×10^{17}	8.05×10^{18}
InP	0.077	0.64	5.17×10^{17}	1.24×10^{19}

Table 3.3. Examples for effective masses and effective DOS ($T = 293$ K)

semiconductor[18] in the saturation range[19] is called degenerately doped, if the dopant concentration is larger than the effective DOS N_B; in this case the Fermi level moves *into* the band.

Filling of electronic states

In a quantum mechanical treatment, particles fall in two categories[20]: fermions and bosons. Particles like photons and phonons are bosons having integer spins $0, \hbar, 2\hbar, \ldots$ Particles such as electrons are fermions with spins $\hbar/2, 3\hbar/2, 5\hbar/2 \ldots$ This subtle difference forces a very important distinction on the occupation statistics. Only one fermion can occupy a quantum state, while any number of bosons can be placed in a particular state. This is the reason why electromagnetic fields can be amplified.

At the absolute temperature $T = 0$ the electrons fill the lowest energy states. At $T > 0$ the distribution which minimizes the free energy of the system is, for fermions, the Fermi-Dirac[21] distribution (Fermi[22]

[18] Singh, J.: Physics of semiconductors and their heterostructures. New York: McGraw-Hill 1993

[19] See Ref. 18 on Page 60, Sect. 8.4.2 Page 270: „In general, there are three regions of interest for doped (extrinsic) semiconductors. At very low temperatures, the electrons (holes) are trapped at the donor (acceptor) levels and the free carrier density goes to zero. This region is called the *freeze-out* range. At higher temperatures, the shallow levels are ionized and there is little change in free carrier density. This region is called the *saturation* range. Finally, at very high temperatures, the intrinsic carrier density exceeds the doping levels and the carrier density ($n \approx p$) increases exponentially as for an *intrinsic* material. The higher the bandgap, the higher the temperature where this regime takes over. However, electronic devices cannot operate in this regime."

[20] See Ref. 18 on Page 60

[21] Paul Adrien Maurice Dirac, physicist, ⋆ Bristol 8.8.1902, † Tallahassee 20.10.1984 (Florida). Nobel prize 1933 (together with E. Schrödinger)

[22] See Footnote 13 on Page 55

function for short),

$$f(W) = \frac{1}{1 + g \exp\left(\frac{W - W_F}{kT}\right)}, \qquad g = \begin{cases} 1 & \text{band states} \\ 1/2 & \text{donor states} \\ 2 & \text{acceptor states} \end{cases}. \tag{3.18}$$

For impurities, a degeneracy factor[23] g has to be taken into account. Figure 3.7 displays the Fermi function for band states. Here W_F is called the chemical potential or Fermi energy, and it represents the energy where the occupation probability $f(W)$ becomes $1/2$ at all temperatures. The transition from a large to a low occupation probability ($0.88 \geq f(W) \geq 0.12$) takes place in a region $4kT$ centred at the Fermi energy W_F (at $T = 293\,\mathrm{K}$ we have $kT = 25\,\mathrm{meV}$ or $\Delta f = 2kT/h = 12.1\,\mathrm{THz}$). In addition to the quantum statistics Eq. (3.18), we also have the classical statistics of Boltzmann[24] which can be derived from Eq. (3.18),

$$\begin{aligned} f(W) &\approx g \exp\left(-\tfrac{W - W_F}{kT}\right) \qquad \text{for } W - W_F > \quad 3kT\,, \\ f(W) &\approx 1 - g \exp\left(\tfrac{W - W_F}{kT}\right) \qquad \text{for } W - W_F < -3kT\,. \end{aligned} \tag{3.19}$$

The residual error is smaller than 5 % ($\mathrm{e}^3 \approx 20$), if the Fermi level has an energetic distance from the band edges W_C, W_V of at least three times the thermal energy kT. Especially for undoped semiconductors the approximation is very good.

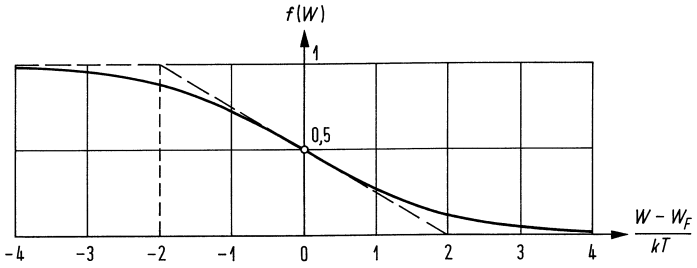


Fig. 3.7. Fermi function for band energy states ($g = 1$)

Impurities and doping

The density n_T of CB electrons and the density p of VB holes may be calculated with the help of the DOS Eq. (3.16), the effective DOS Eq. (3.17) and the Fermi distribution Eq. (3.18),

$$n_T = \int_{W_C}^{\infty} \rho_C(W)\, f(W)\, \mathrm{d}W \qquad p = \int_{-\infty}^{W_V} \rho_V(W) \big[1 - f(W)\big]\, \mathrm{d}W\,. \tag{3.20}$$

[23] A *donor* electron has one of two possible spin orientations. For the occupation of the state with an electron there are two favourite cases (spin $+\hbar/2$, spin $-\hbar/2$) out of three possibilities (electron at donor, i. e., spin $+\hbar/2$, spin $-\hbar/2$; no electron at donor). Therefore we have $f(W_F) = 2/3$ for a donor at $W = W_D = W_F$, Eq. (3.18).
Acceptors bind an electron with a well-defined spin to fill up the outmost shell. For the transition of an electron to an acceptor atom there is one favourite case (fitting spin) out of three possible cases (electron transition to acceptor with fitting spin, no VB electron available with fitting spin, no electron transition to acceptor at all). As a consequence, $f(W_F) = 1/3$ for an acceptor at $W = W_A = W_F$, Eq. (3.18).

[24] See Footnote 12 on Page 54

With the Boltzmann approximations (valid for $n_T \ll N_C$, $p \ll N_V$ only, i. e., for non-degenerate doping) the integrals Eq. (3.20) can be solved,

$$\left.\begin{aligned} n_T &= N_C \exp\left(-\frac{W_C - W_F}{kT}\right) \\ p &= N_V \exp\left(-\frac{W_F - W_V}{kT}\right) \end{aligned}\right\} \quad n_T p = n_i^2 = N_C N_V \exp\left(-\frac{W_G}{kT}\right). \tag{3.21}$$

The intrinsic carrier concentration n_i refers to the electrons n_T (holes p) present in the CB (VB) of a pure undoped semiconductor. It depends on the bandgap W_G as well as the details of the band edge masses, but *not* on the Fermi energy W_F; this is a relation expressing the law of mass action. At $T = 0$ the VB is completely occupied while the CB is empty, and the semiconductor has an extremely high resistance. With increasing temperature some electrons are thermally excited from the VB to the CB so that CB electrons and VB holes are generated in pairs, $n_T = p = n_i$. The presence of intrinsic carriers is detrimental to devices[25] where the current has to be modulated by some means. The Fermi level follows from Eq. (3.21) for charge neutrality $n_T = p$,

$$W_F = \frac{1}{2}(W_C + W_V) + kT \ln\sqrt{\frac{N_V}{N_C}} = \frac{1}{2}(W_C + W_V) + \frac{3}{4}kT \ln\frac{m_p}{m_n}. \tag{3.22}$$

At $T = 0$ the chemical potential W_F of the intrinsic semiconductor is in the centre of the forbidden band, at $T > 0$ the Fermi level shifts into the direction of the faster filling band which owns the smaller effective DOS N_B.

Pure semiconductors would have little use by themselves because of their low conductivity (carrier concentration at room temperature $\sim 10^{11}\,\mathrm{cm}^{-3}$) compared to metals ($\sim 10^{21}\,\mathrm{cm}^{-3}$). By introducing impurities the properties of semiconductors may be tailored to specific needs. When a dopant (impurity) atom is implanted into a crystal, its perfect periodicity is destroyed and additional energy levels for electrons located near the band edges are the outcome. These levels are either near the CB edge (W_D) and can "donate" an electron to the CB (donor), or they are near the VB edge (W_A) where they can accept an electron from the VB (acceptor). The donor (acceptor) concentrations are n_D (n_A), the concentrations of the neutral donor (acceptor) atoms are $n_D^\times$ ($n_A^\times$), and the concentrations of the ionized impurities are n_D^+ (n_A^-). Thus, either a quasi-free CB electron or a quasi-free VB hole is created when the impurity atoms give or take an electron by thermal excitation for $|W_{C,V} - W_{D,A}| < kT$, Fig. 3.8. For a large impurity concentration Fig. 3.8(b),(c) the impurity levels broaden to form impurity bands which may overlap with the CB or the VB, respectively. In this case, the bandgap is decreased to $W_{G_{\mathrm{eff}}}$, and the DOS $\rho(W)$ cannot be approximated by a parabola near the band edges in Eq. (3.16). Equation (3.21) remains valid. The Fermi level W_F can be computed from Eqs. (3.18), (3.20), (3.21) in the case of charge neutrality,

$$n_T + n_A^- = p + n_D^+ \qquad \begin{aligned} n_D &= n_D^\times + n_D^+ \quad \text{donor density,} \\ n_A &= n_A^\times + n_A^- \quad \text{acceptor density.} \end{aligned} \tag{3.23}$$

If for donors $(W_C - W_D)/(kT) \ll 1$ is valid (saturation, practically all donors are ionized), and if $n_D > N_C$, then the Fermi energy W_F is shifted into the CB. An analogue relation holds for acceptors.

For a non-equilibrium condition where a constant perturbation is switched on, the carriers in the CB and VB states need some time to re-arrange. This time is called the intraband relaxation time τ_{CB},

[25] Some intrinsic carrier concentrations at room temperature $T_0 = 293\,\mathrm{K}$: $n_{i\,\mathrm{Si}} = 1.5 \times 10^{10}\,\mathrm{cm}^{-3}$, $n_{i\,\mathrm{Ge}} = 2.4 \times 10^{13}\,\mathrm{cm}^{-3}$, $n_{i\,\mathrm{GaAs}} = 1.8 \times 10^{6}\,\mathrm{cm}^{-3}$ (actually not achievable), $n_{i\,\mathrm{InP}} = 1.2 \times 10^{8}\,\mathrm{cm}^{-3}$. "The fact that n_i^2 is constant at a given temperature is often utilized to produce high resistivity (insulating) materials from impure semiconductors. Consider, for example, impure GaAs with $n_T = 10^{16}\,\mathrm{cm}^{-3}$ and $p = 10^5\,\mathrm{cm}^{-3}$ giving a total free carrier density $n_T + p = 10^{16}\,\mathrm{cm}^{-3}$ and $n_i^2 = n_T p = 10^{21}\,\mathrm{cm}^{-6}$ ($n_i = 3.2 \times 10^{10}\,\mathrm{cm}^{-3}$) at 180 °C. If the p-type carrier concentration is now increased by doping to $3.2 \times 10^{10}\,\mathrm{cm}^{-3}$, the sum concentration becomes $n_T + p \approx 6.4 \times 10^{10}\,\mathrm{cm}^{-3}$ since the $n_T p$ product must remain the same. The Fermi level shifts into the direction of the forbidden-band centre. This greatly reduces the material conductivity. This technique is called compensation. It must be remembered, of course, that the $n_T p$ product is constant only when the system is *in equilibrium*." Sect. 8.1 in Ref. 18 on Page 60

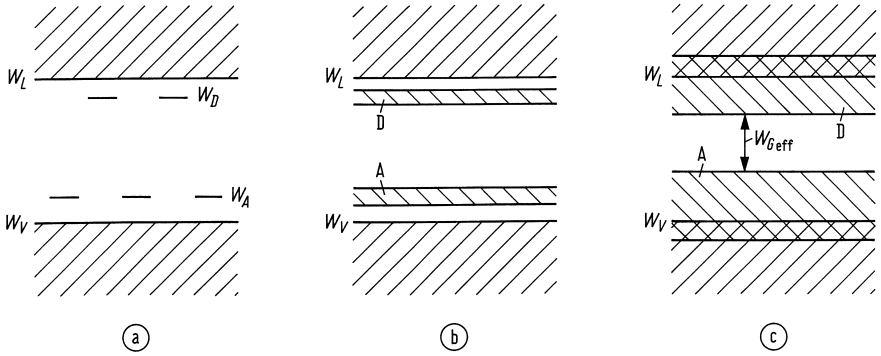


Fig. 3.8. Energy levels of impurities in a semiconductor. (a) isolated donors and acceptors. (b) impurity bands for heavier doping. (c) overlap of impurity bands with CB annd VB for heavy doping ($W_L \mathrel{\widehat{=}} W_C$, Leitungsband $\mathrel{\widehat{=}}$ conduction band)

τ_{VB}. For example: At $t = 0$ we have a constant field strength $\vec{E}$ in the semiconductor. After a transit time τ the electron drift velocity is $\vec{v}_n = -\mu_n \vec{E}$ (mobility μ_n, expectation of electron velocity $\vec{v}_n$). If the relaxation follows the function $\exp(-t/\tau)$, then the time constant τ corresponds to the re-arrangement time of the electrons with respect to the CB states, and $\tau = \tau_{\mathrm{CB}}$ is called the momentum relaxation time. In terms of the intraband relaxation times, the mobilities of electrons and holes are $\mu_n = e\tau_{\mathrm{CB}}/m_n$ and $\mu_p = e\tau_{\mathrm{VB}}/m_p$. For InP we find $\mu_n = 4\,600\,\mathrm{cm}^2\,/\,\mathrm{V}$, $m_n/m_0 = 0.077$, $m_0 = 9.109\,387\,9 \times 10^{-31}\,\mathrm{kg}$, $e = 1.602\,177\,33 \times 10^{-19}\,\mathrm{C}$ from Table 3.3 on Page 60, and $\tau_{\mathrm{CB}} = 0.2\,\mathrm{ps}$ follows. After this time has passed, the occupation probability may be again described by the Fermi distribution $f(W)$ Eq. (3.18). However, because n_T, p assume different values than in the equilibrium case, the Fermi energy W_{Fn} of the electrons in the conduction band (i. e., the energy state with electron occupation probability $f_C(W_{Fn}) = 1/2$) differs from the Fermi level W_{Fp} of valence band holes (i. e., from the energy state with hole occupation probability $f_V(W_{Fp}) = 1/2$). The quantities W_{Fn} and W_{Fp} are therefore denoted as quasi Fermi levels for electrons and holes in the non-equilibrium case. The occupation probabilities for conduction band electrons and valence band holes read now

$$f_C(W) = \frac{1}{1 + \exp\left(\dfrac{W - W_{Fn}}{kT}\right)}\,, \quad f_V(W) = \frac{1}{1 + \exp\left(\dfrac{W - W_{Fp}}{kT}\right)}\,. \tag{3.24}$$

After the intraband relaxation time τ_{LB} the conduction band electrons are in a new dynamic equilibrium as it is the case for the valence band holes when the intraband relaxation time τ_{VB} has passed. However, the conduction band electrons are not in equilibrium with the valence band holes, $W_{Fn} \neq W_{Fp}$. Analogue to Eq. (3.21) we find

$$\left.\begin{aligned} n_T &= N_C \exp\left(-\frac{W_C - W_{Fn}}{kT}\right) \\ p &= N_V \exp\left(-\frac{W_{Fp} - W_V}{kT}\right) \end{aligned}\right\} \quad \begin{aligned} n_T p &= n_i^2 \exp\left(\frac{W_{Fn} - W_{Fp}}{kT}\right) \\ n_i^2 &= N_C N_V \exp\left(-\frac{W_G}{kT}\right). \end{aligned} \tag{3.25}$$

In Sect. 3.1.2 Eq. (3.9) on Page 56 it was made plausible that laser action in a semiconductor laser requires the photon energy to be inside the bounds $W_C - W_V < hf \leq W_{Fn} - W_{Fp}$, so that either the condition $W_{Fn} \geq W_C$ or $W_{Fp} \leq W_V$ must be met. Following Eq. (3.25) the necessary pump can be realized by carrier injection with $n_T p \gg n_i^2$. Then an increased radiative recombination rate leads to an increased emission of spontaneous photons compared to the case of true thermal equilibrium. From Eq. (3.20) we have

$$\frac{\mathrm{d}n_T}{\mathrm{d}W} = \rho_C(W)\, f_C(W), \qquad \frac{\mathrm{d}p}{\mathrm{d}W} = \rho_V(W)\,\big[1 - f_V(W)\big]. \tag{3.26}$$

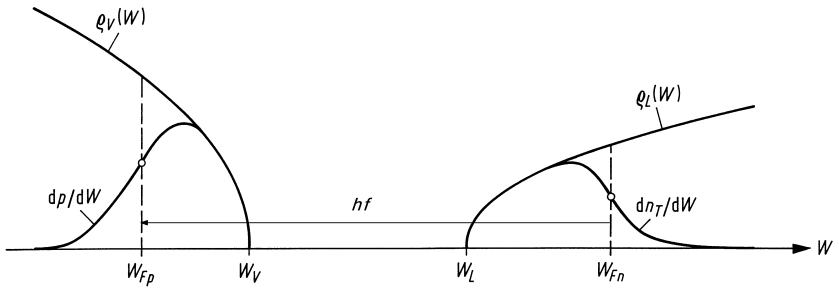


Fig. 3.9. Distribution of valence band holes and conduction band electrons under non-equilibrium conditions $n_T p \gg n_i^2$ for an inverted semiconductor at $T > 0$. The areas under the $\mathrm{d}p/\,\mathrm{d}W$ and $\mathrm{d}n_T/\,\mathrm{d}W$ curves stand for the hole and electron concentrations p and n_T in the valence and in the conduction bands, respectively. The arrow marked by hf indicates the maximum photon energy at which optical amplification is still achieved, see Eq. (3.9) on Page 56. ($\rho_L \mathrel{\widehat{=}} \rho_C$ conduction band DOS)

Figure 3.9 shows the distribution of holes and electrons inside the bands of a population inverted semiconductor laser, and the maximum photon energy $hf = W_{Fn} - W_{Fp}$ at which an optical amplification is still possible. This will be discussed in more detail on Page 67 ff.

To calculate the threshold current of a laser, we need the carrier concentrations n_T, p for shifting the quasi Fermi levels W_{Fn}, W_{Fp} into the bands. We assume a p-doped semiconductor with an equilibrium concentration of n_{T0}, p_0 and $n_{T0}p_0 = n_i^2$. By carrier injection the densities are changed to $n_T = n_{T0} + \Delta n_T$, $p = p_0 + \Delta p$. By substitution into Eq. (3.25) we find

$$W_{Fn} - W_F = kT \ln\left(1 + \frac{\Delta n_T}{n_{T0}}\right), \quad W_F - W_{Fp} = kT \ln\left(1 + \frac{\Delta p}{p_0}\right). \tag{3.27}$$

Further, we assume charge neutrality $\Delta p = \Delta n_T$. By a carrier injection the quasi Fermi level of the minority carriers shifts first (here: W_{Fn}; change of n_T by Δn_T has largest effect because n_{T0} is small). At $\Delta n_T / n_{T0} = 1$ the shift amounts to $W_{Fn} - W_F = 0.7\,kT$. Because $p_0 \gg n_{T0}$ holds in a p-semiconductor the quasi Fermi level for majority carriers starts to shift at much higher injection current levels when $\Delta p = \Delta n_T$ reaches the order of p_0.

Heterojunctions

Heterojunctions are composed of semiconductors with different bandgap energies W_G. They are advantageous for laser diodes and photodetectors. With $(\mathrm{Ga}_{1-x}\mathrm{Al}_x)\mathrm{As}$ of Table 3.1 and $0 \le x \le 0.3$ the bandgap energy W_G can be increased by 374 meV while the refractive index n decreases by nearly 6 %. Lasers are built as 3-layer or 5-layer heterostructures, Fig. 3.10. For the 3-layer heterostructure Fig. 3.10(a) the active layer (the region with induced amplification) has a thickness of $d = 0.1 \ldots 0.2\,\mu$m and consists of p-GaAs. A slight p-doping[26] decreases the electron concentration in the valence band, thereby facilitating a population inversion. The neighbouring layers are formed of (Ga,Al)As having a larger bandgap W_G and a lower refractive index n leading to the following features:

Potential walls exist for carriers n_T, p injected from both sides of the p-GaAs layer. Even from low current densities $J > 0.5\,\mathrm{kA}\,/\,\mathrm{cm}^2$ onwards the carrier concentration n_T inside the active layer is so large that the difference of the quasi Fermi levels exceeds the bandgap, $W_{Fn} - W_{Fp} > W_G$, and laser action starts (for GaAs at about $n_T = 2 \times 10^{18}\,\mathrm{cm}^{-3}$).

Larger W_G in the (Ga,Al)As layers blocks the induced re-absorption in the non-inverted regions.

Smaller n in the (Ga,Al)As layers characterizes the cladding of a slab waveguide where the core is represented by the active layer. Because d is small, a large portion ($\approx 80\,\%$) of the electromagnetic energy propagates inside the cladding (field confinement factor $\Gamma \approx 20\,\%$).

[26] See Footnote 25 on Page 62

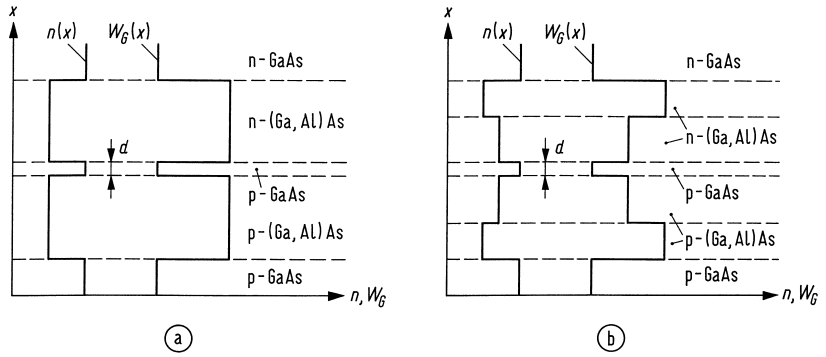


Fig. 3.10. Schematic refractive index dependence n and bandgap W_G as a function of the spatial coordinate x in a (a) 3-layer heterostructure, (b) 5-layer heterostructure

For a 3-layer structure both the carrier and the field confinement are determined by the thickness d of the active layer, and by W_G (i. e., by the refractive index n) of the three central layers. The actual pn-junction is between the p-GaAs active layer and the neighbouring n-(Ga,Al)As layers. Devices with heterojunctions on both sides of the active zone are called double-heterostructures.

With a 5-layer heterostructure Fig. 3.10(b) the carrier confinement and the vertical field confinement in x-direction become independent. The field is guided by two (Ga,Al)As layers on each side of the active zone. Fortunately, the refractive index n in (Ga,Al)As depends only weakly on the doping ($\sim 0.1\,\%$).

Heterojunctions are called "isotype" if the semiconductors have the same conduction type, and "anisotype" if the conduction type differs. The conduction type is specified with small letters n, i, p if the semiconductor has a smaller bandgap than its neighbour, and with capital letters N, I, P if the bandgap is larger. For the structure in Fig. 3.10(a) we see the following junction types from top to bottom: nN, Np, pP and Pp. In the following, we discuss some heterojunction properties in analogy to the ordinary pn-junction.

Band diagram for heterostructures Figure 3.11(a) explains the energy scale. Free electrons are at the vacuum energy level $W = 0$ if they move at a velocity $\vec{v} = 0$ in a region with constant potential $\varphi = 0$; in Fig. 3.11(a) we further assume a potential $\varphi \neq 0$. Electrons leaving the semiconductor with $\vec{v} = 0$ are

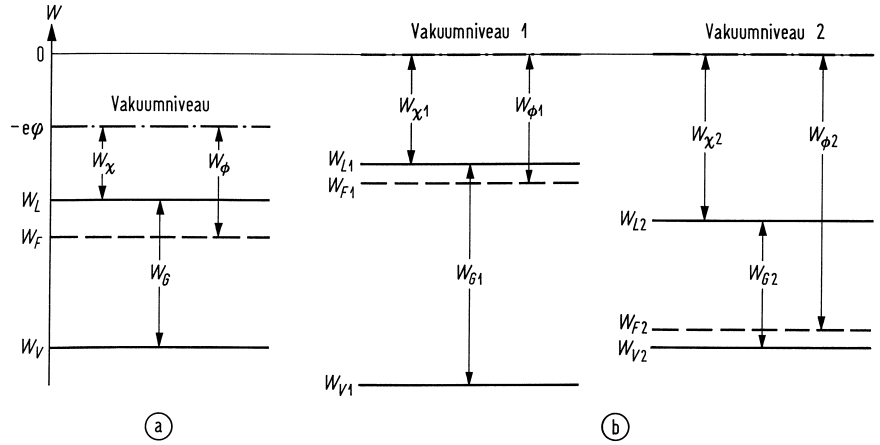


Fig. 3.11. Energy scale for electrons in a semiconductor. (a) Semiconductor at potential $\varphi \neq 0$. (b) Two independent, insulated semiconductors at potential $\varphi = 0$ with different bandgaps. W_χ electron affinity, W_ϕ work function. W_L conduction band edge ($\widehat{=}\, W_C$, Vacuumniveau $\widehat{=}$ vacuum level)

then at the vacuum energy level $W = -e\varphi$. The electron affinity W_χ specifies the energetic distance from the CB edge to the vacuum level. The work function W_ϕ is defined by the energy difference of the Fermi and the vacuum level. All these quantities together with the bandgap $W_G = W_C - W_V$ are positive by definition. W_χ, W_ϕ and W_G fix the energetic distances of the CB edge, the Fermi level and the VB edge in relation to the vacuum level.

In Fig. 3.11(b) the energy-band diagrams of two semiconductors with different bandgaps are displayed. After forming the contact we have a Np-junction as in Fig. 3.10(a). Assuming $\varphi = 0$ for both separated semiconductors we define the quantities

$$\begin{aligned} \Delta W_G &= W_{G2} - W_{G1}\,, \\ \Delta W_C &= W_{C2} - W_{C1} = W_{\chi 1} - W_{\chi 2}\,, \\ \Delta W_V &= W_{V2} - W_{V1} = \Delta W_C - \Delta W_G\,. \end{aligned} \tag{3.28}$$

In the case of Fig. 3.11(b). the relations ΔW_G, $\Delta W_C < 0$ and $\Delta W_V > 0$ hold. For $(Ga_{1-x}Al_x)As$ with $0 \le x \le 0.3$ we find nearly independently of x

$$\frac{\Delta W_C}{\Delta W_G} = 0.65\,, \qquad \frac{\Delta W_V}{\Delta W_G} = -0.35\,. \tag{3.29}$$

For each of the (non-degenerate) semiconductors in Fig. 3.11(b), Eqs. (3.21)–(3.23) are valid. When the contact is formed all states of equal energy are occupied with the same probability in the case of thermal equilibrium. If we fix the potential of semiconductor 2 at $\varphi_2 = 0$, the potential of semiconductor 1 rises until $e\varphi_1 + W_{\phi 1} = W_{\phi 2}$, i. e., the potential is given by the so-called built-in potential $U_D = \varphi_1 = (W_{F1} - W_{F2})/e > 0$ (*German* Diffusionspannung). From Eq. (3.21) we calculate

$$e\varphi_1 = W_{F1} - W_{F2} = W_{C1} - W_{V2} + kT\,\ln\frac{n_{T1}p_2}{N_{C1}N_{V2}} = W_{G1} - \Delta W_V + kT\,\ln\frac{n_{T1}p_2}{N_{C1}N_{V2}}\,. \tag{3.30}$$

Under the assumption of shallow saturated impurities with $n_{T1} = n_D$, $p_2 = n_A$, we find from Eq. (3.21) for a homojunction the diffusion voltage or built-in potential U_D of the pn-junction (U_T is the thermal voltage)

$$\varphi_1 = U_D = U_T\,\ln\frac{n_D n_A}{n_i^2}, \qquad U_T = \frac{kT}{e}, \qquad W_G = -kT\,\ln\frac{n_i^2}{N_C N_V}\,. \tag{3.31}$$

At room temperature $T = 293\,\mathrm{K}$ the thermal voltage is $U_T = 25\,\mathrm{mV}$. Figure 3.12 (not drawn to scale) shows the band-energy diagram of the NpP-heterojunction of Fig. 3.10(a) in the case of thermal equilibrium; we used the semiconductors of Fig. 3.11(b), supplemented by a p-semiconductor with electron affinity $W_{\chi 1}$ and bandgap W_{G1}. The doping of the p-(Ga,Al)As layer was chosen such that the diffusion voltage of the isotype pP-junction is zero, $U_{D\,\mathrm{pP}} = 0$. The band edge energies (and therefore the carrier concentrations) are not continuous but exhibit steps by $|\Delta W_C|, |\Delta W_V|$, see Eq. (3.28). The component of the dielectric displacement vector $\vec{D} = \epsilon_0\epsilon_r\vec{E}$ perpendicularly to the boundary plane is continuous while the refractive index $n = \sqrt{\epsilon_r}$ is discontinuous. Therefore, the normal component of the electric field vector $\vec{E} = -\,\mathrm{grad}\,\varphi$ is also discontinuous. This leads to a kink of the vacuum level at the semiconductor boundary; the slope inside the semiconductor with the larger W_G (smaller n) is larger than inside the semiconductor with the smaller W_G.

Figure 3.13 displays the energy-band diagram from Fig. 3.12 assuming the flat-band case for simplicity. However, because of unavoidable series resistances it is practically impossible to adjust an external forward voltage such that the junction voltage U compensates the diffusion voltage U_D from Fig. 3.12. Far away from the junction the quasi Fermi levels of electrons and holes are practically identical. However, inside the thin p-GaAs layer and inside the diffusion zones we have $W_{Fn} > W_{Fp}$ due to the carrier injection. Because of the longer diffusion length of electrons compared to the diffusion length of holes, $L_n > L_p$, the diffusion zone of the p-semiconductor is larger than inside the n-semiconductor (for GaAs: $L_n/L_p = 5$).

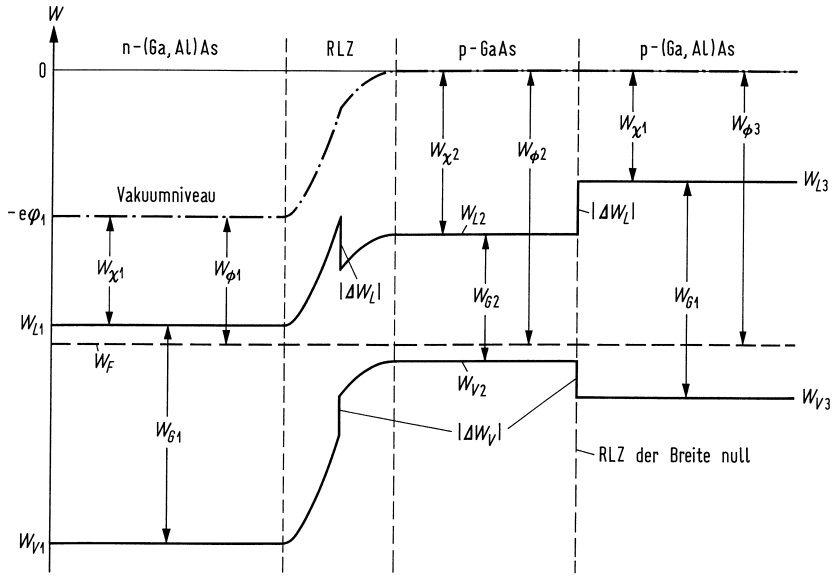


Fig. 3.12. Energy-band diagram of a double-heterostructure with anisotype Np-junction and a special isotype pP-junction with diffusion voltage zero ($W_L \mathrel{\widehat{=}} W_C$, Leitungsband $\mathrel{\widehat{=}}$ conduction band, Vakuumniveau $\mathrel{\widehat{=}}$ vacuum level, Raumladungszone RLZ der Breite null $\mathrel{\widehat{=}}$ space-charge region of zero width)

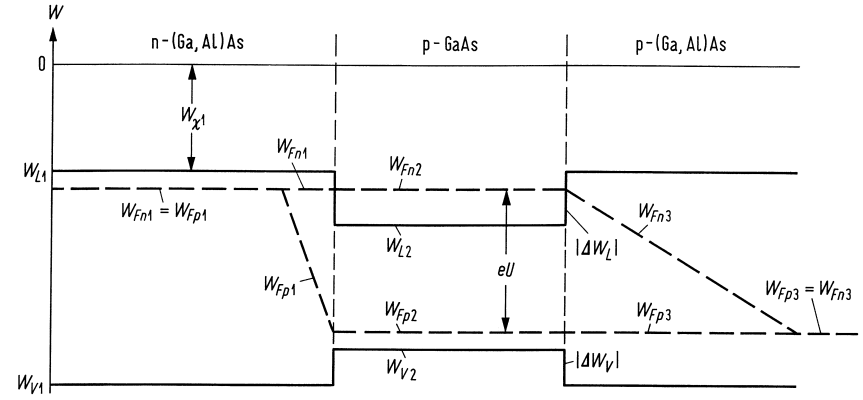


Fig. 3.13. Energy-band diagram of a NpP-heterojunction of Fig. 3.12 with a forward bias voltage. $L_p/L_n \approx 0.2$ for GaAs ($W_L \mathrel{\widehat{=}} W_C$)

Electrons and holes are confined to a potential film[27] (quantum film) inside the p-GaAs layer. In Fig. 3.13 the quasi Fermi level for electrons W_{Fn} was moved into the CB. By appropriate injection (pump) currents, the semiconductor may be population inverted, see Eq. (3.9).

Emission and absorption of light in a semiconductor

General considerations Let us assume a microsystem according to Fig. 3.3(a) with energy levels W_2, W_1 and $W_2 = W_1 + hf$. The electromagnetic field in the active optical volume V is given as an

[27] This quantum film, a thin layer between layers with larger bandgaps, confines electrons in one direction only (the growth direction). Nonetheless, the structure is usually called a potential or quantum "well". However, according to common understanding, a well has a two-dimensional cross-section and confines water (or electrons in our case) in two orthogonal directions.

expansion of orthonormal modes with total number M_{tot}, Eq. (3.3) on Page 51.

A certain mode with frequency f (photon energy hf) contains N_P photons. As a result of a first-order perturbation theory[28] (dubbed "Fermi's Golden Rule") we find the probabilities for the induced ($p_{\mathrm{ind}}^{(\mathrm{e})}$) and spontaneous ($p_{\mathrm{sp}}^{(\mathrm{e})}$) photon emission or absorption of a microsystem, which is initially in the excited state or in the ground state, respectively,

$$p^{(\mathrm{e})} = p_{\mathrm{ind}}^{(\mathrm{e})} + p_{\mathrm{sp}}^{(\mathrm{e})} \sim (N_P + 1)\, t\,, \qquad p^{(\mathrm{a})} \sim N_P t \tag{3.32}$$

The relations are asymptotically valid for $t \to \infty$, because only then the excited state has a well defined energy. Spontaneous transitions which are independent of the photon number N_P are only possible for emission, see the discussion on Page 52.

The transition probability $p = 1$ specifies the lifetime τ of the energy state. The energy is conserved for the transition if $t, \tau \to \infty$ holds, $p^{(\mathrm{e})}, p^{(\mathrm{a})} \sim \delta(W_2 - W_1 - hf)$. If the interaction time t is finite, the δ-function of the Golden Rule changes over to a broadened function $\rho(f)$ so that in this case the energy conservation is no longer strictly satisfied.

Note that according to Eq. (3.5) the energy uncertainty for $\tau_2 = 0.1\,\mathrm{ps}$ is about $\Delta W_2 = 5\,\mathrm{meV} \mathrel{\widehat{=}} 6\,\mathrm{THz}$. Spontaneously emitted photons occupy at random all electromagnetic modes with frequency f in V. With Eq. (3.3) for the total number M_{tot} of modes in V for frequencies $0 \ldots f$ and with the spectral mode density $\varrho_{\mathrm{tot}}(f)\, V = \mathrm{d}M_{\mathrm{tot}}/\,\mathrm{d}f$, the probability of a spontaneous emission, $p_{\mathrm{sp}}^{(\mathrm{e})}$, in any mode of frequency f is $p_{\mathrm{sp}} = \int p_{\mathrm{sp}}^{(\mathrm{e})}\, \mathrm{d}M_{\mathrm{tot}}$.

Because of the finite lifetime the emission is not monochromatic but has a lineshape $\rho(f)$ with a half-maximum bandwidth $\Delta f_H \ll f_0$ centred at f_0. A detailed calculation leads to a Lorentzian,

$$\rho(f) = \frac{2}{\pi \Delta f_H} \frac{1}{1 + \left(\frac{f - f_0}{\Delta f_H/2}\right)^2}\,, \qquad \int_{-\infty}^{+\infty} \rho(f)\,\mathrm{d}f = 1\,, \qquad 2\pi\,\tau_{\mathrm{sp}}\,\Delta f_H = 1\,. \tag{3.33}$$

The probability of an induced emission into a certain electromagnetic mode is by N_P larger than the probability of a spontaneous emission (representing noise) into the same mode. The induced emission of photons from a microsystem in state W_2 happens with the same probability as absorption from the ground state W_1.

To achieve amplification the emission rate must be larger than the absorption rate, i. e., the number N_2 of microsystems in the excited state W_2 must be larger than the number N_1 of microsystems in the ground state W_1.

Purcell[29] suggested more than half a century ago to tailor the spontaneous emission probability of radiating dipoles into a specific mode of frequency f by using a cavity to modify the dipole-field coupling and the density of available photon modes[30,31].

---

[28] See App. I.1 Page 813 in Reference 18 on Page 60

[29] Purcell, E. M.: Spontaneous emission probabilities at radio frequencies. Phys. Rev. 69 (1946) 681

[30] Let a narrow-line cavity with a linewidth of Δf_H centred at a frequency f_0 be described by the line shape function $\rho(f)$ of Eq. (3.33). Such a resonator with a quality factor $Q_R = f_0/\Delta f_H = \omega_0\,\tau_P$ presents exactly one mode having a photon lifetime τ_P, provided that the dipole radiation is spectrally narrower than Δf_H. The number of modes (= 1) per frequency interval $\pi\Delta f_H/2 \approx 1.6 \times \Delta f_H$ is therefore $1/\,(\pi\Delta f_H/2) = 2/(\pi\Delta f_H)$, which represents the equivalent density of cavity modes. Compared with the density of free-space radiating modes $\varrho_{\mathrm{tot}}(f_0)\, V$ in a volume V, the mode density is increased by the so-called Purcell figure of merit[31] F_P,

$$F_P = \frac{\rho(f_0)}{\varrho_{\mathrm{tot}}(f_0)\, V} = \frac{2/(\pi\Delta f_H)}{8\pi V (f_0 n)^2 n_g/c^3} = \frac{Q_R}{4\pi^2 V n^2 n_g (f_0/c)^3}\,, \qquad Q_R = \frac{f_0}{\Delta f_H} = \omega_0\,\tau_P\,. \tag{3.34}$$

Enhancing the spontaneous emission probability (Purcell effect) of a solid-state emitter by making $F_P \gg 1$ would allow in particular the fabrication of high-efficiency light-emitting diodes, see Sect. 3.1.5. This can be achieved with high-Q_R microcavities, and by exploiting the properties of photonic crystals.

[31] Gerard, J.-M.; Gayral, B.: Strong Purcell effect for InAs quantum boxes in three-dimensional solid-state microcavities. J. Lightwave Technol. 17 (1999) 2089–2095. — Here, the Purcell factor definition (see Ref. 29 on Page 68) is larger by 3 which stems from a 1/3 averaging factor accounting for the random polarization of free-space modes with respect to the spontaneously radiating dipole. — Various expressions can be found in the literature for F_P, which differ by a numerical factor as large as ten. Therefore, care must be taken in comparing various experimental outcomes.

Induced and spontaneous transitions

Conduction and valence band states of a semiconductor volume V under non-equilibrium conditions are described by the respective Fermi functions $f_C(W), f_V(W)$ Eq. (3.24), and by the density of states $\rho_C(W), \rho_V(W)$ Eq. (3.16). For the carrier concentrations n_T and p we write Eq. (3.20), substituting $f(W)$ by $f_C(W)$ and $f_V(W)$, respectively. We are interested in a certain electromagnetic mode with photon energy $hf = W_2 - W_1$. The total emission probability into this mode results from the product of the probabilities for

- emission $w^{(\mathrm{e})}$,
- occupation of a CB state $f_C(W_2)$, and
- for the event that the corresponding state in the VB is unoccupied, $[1 - f_V(W_1)]$.

Further, the number of states per energy, i. e., the density of states $\rho_C(W_2)$ and $\rho_V(W_1)$, have to be taken into account, Eqs. (3.15), (3.16). Finally, we have to sum over all possible transitions.

The gain is determined by the difference in induced emission $r_{\mathrm{ind}}^{(\mathrm{eM})}$ and absorption rates $r_{\mathrm{ind}}^{(\mathrm{aM})}$ (unit $\mathrm{m}^{-3}\,\mathrm{s}^{-1}$), i. e., by the net number of photons $r_{\mathrm{ind}}^{(\mathrm{M})}$ *e*mitted or *a*bsorbed per volume and time into a fixed *m*ode with frequency f. The spontaneous emission rate $r_{\mathrm{sp}}^{(\mathrm{eM})}$ into mode f does not depend on the photon number. The quantum mechanical properties of the transition (e. g., the transition matrix element $|\mu_{21}|^2$ specifying the interaction with the electromagnetic field, averaged over all possible spatial orientations of the microsystem) are combined in a quantity K_0 (unit $\mathrm{W}^2\,\mathrm{s}$). For a modulus-$\vec{k}_\mu$ selection rule, i. e., a k_μ-selection rule[32,33], we write (without giving a detailed derivation)

$$\begin{aligned} r_{\mathrm{ind}}^{(\mathrm{M})} &= r_{\mathrm{ind}}^{(\mathrm{eM})} - r_{\mathrm{ind}}^{(\mathrm{aM})} \\ &= \tfrac{1}{2} N_P\, V K_0\, \rho_C(W_0)\, \rho_V(W_0 - hf) \big[f_C(W_0) - f_V(W_0 - hf) \big], \\ r_{\mathrm{sp}}^{(\mathrm{eM})} &= \quad \tfrac{1}{2} V K_0\, \rho_C(W_0)\, \rho_V(W_0 - hf)\, f_C(W_0) \big[1 - f_V(W_0 - hf) \big], \\ W_0 &= W_C + \tfrac{\hbar^2 k_{\mu 0}^2}{2 m_n} = W_C + \tfrac{hf - W_G}{1 + m_n/m_p} \quad \text{for a } k_\mu\text{-selection rule.}^{33} \end{aligned} \tag{3.35}$$

For photon energies exceeding the bandgap energy the DOS product[34] is positive,

$$\rho_C(W_0)\rho_V(W_0 - hf) \sim \left(\frac{hf - W_G}{kT_0} \right) kT_0\,. \tag{3.36}$$

The difference of the Fermi functions in Eq. (3.35) reads

$$\begin{aligned} f_C(W_0) - f_V(W_0 - hf) = {} & \frac{1}{1 + \exp\Big[\Big(W_C + \frac{hf - W_G}{1 + m_n/m_p} - W_{Fn} \Big) \Big/ kT \Big]} \\ & - \frac{1}{1 + \exp\Big[\Big(W_C - hf + \frac{hf - W_G}{1 + m_n/m_p} - W_{Fp} \Big) \Big/ kT \Big]}, \\ f_C(W_0) - f_V(W_0 - hf) > 0 \quad & \text{for} \\ W_C + \frac{hf - W_G}{1 + m_n/m_p} - W_{Fn} < {} & W_C - hf + \frac{hf - W_G}{1 + m_n/m_p} - W_{Fp} \\ \text{or} \quad hf < {} & W_{Fn} - W_{Fp}\,. \end{aligned} \tag{3.37}$$

[32] Adams, M. J.; Landsberg, P. T.: The theory of the injection laser. In: Gooch, C. H. (Ed.): Gallium arsenide lasers. London: Wiley-Interscience 1969. Page 38

[33] On Page 59 the conservation of momentum $\vec{k}_\mu$ for a transition was discussed. The assumption of a $\vec{k}_\mu$-selection rule implies a reasonably pure semiconductor. For an injection laser the impurity concentration has an order of magnitude such that impurity scattering will modify the momentum matrix elements involved in interband transitions, see Ref. 32. The result of such scattering is to effectively relax the strict vectorial $\vec{k}_\mu$-selection rule. Instead, we require the conservation of the *modulus* $|\vec{k}_{\mu\,C,V}| = k_{\mu\,C,V}$, $k_{\mu\,C} = k_{\mu\,V} = k_{\mu 0}$ of the crystal momentum (k_μ-selection rule, not $\vec{k}_\mu$-selection!). Therefore, only the transitions at energies $W_{2,1} = W_{C,V} \pm (\hbar k_{\mu 0})^2 / (2 m_{n,p})$ in Fig. 3.6(b) are allowed, where $W_2 - W_1 = hf$ holds.

[34] As a consequence of the relaxed k_μ-selection rule, the DOS product is linear in frequency. For a strict $\vec{k}_\mu$-selection rule the stimulated $r_{\mathrm{ind}}^{(\mathrm{M})}$ and spontaneous emission rates $r_{\mathrm{sp}}^{(\mathrm{eM})}$ would vary according to $\sqrt{hf - W_G}$, see Ref. 32 on Page 69.

Optical amplification From Eqs. (3.35), (3.37) it follows that for an optical amplification, i. e., for a net induced emission rate $r_{\mathrm{ind}}^{(\mathrm{M})} > 0$ at $T > 0$, at least *one* quasi Fermi level must be inside the CB or the VB, but (accepting a reduced gain) not necessarily both, see Fig. 3.13. However, at $T = 0$ (not very practical, because the electrons and holes are trapped at the donor and acceptor levels, freeze-out[35] range) *both* quasi Fermi levels need be inside the CB and the VB, respectively. The general inversion condition for an amplification of an electromagnetic wave by a semiconductor reads (see Eq. (3.9))

$$W_G < hf \le W_{Fn} - W_{Fp} \quad \text{for} \quad T \ge 0 \quad \text{and} \quad r_{\mathrm{ind}}^{(\mathrm{M})} = r_{\mathrm{ind}}^{(\mathrm{eM})} - r_{\mathrm{ind}}^{(\mathrm{aM})} \ge 0\,. \tag{3.38}$$

This includes the transparency point where $r_{\mathrm{ind}}^{(\mathrm{M})} = r_{\mathrm{ind}}^{(\mathrm{eM})} - r_{\mathrm{ind}}^{(\mathrm{aM})} = 0$. From the definition of $r_{\mathrm{ind}}^{(\mathrm{M})}$ (Eq. (3.35)) and from the gain rate G being defined as the temporal increase of the photon number by stimulated transitions, a relation may be established between both quantities,

$$\left.\begin{aligned} r_{\mathrm{ind}}^{(\mathrm{M})} &= \tfrac{1}{V}\tfrac{\mathrm{d}N_P}{\mathrm{d}t} \\ G &= \tfrac{1}{N_P}\tfrac{\mathrm{d}N_P}{\mathrm{d}t} \end{aligned}\right\} \qquad G = \frac{r_{\mathrm{ind}}^{(\mathrm{M})}}{N_P/V}. \tag{3.39}$$

The spontaneous emission rate into the mode f and its net gain rate are connected by the inversion factor n_{sp}, which is determined by the ratio of the number $N_2 \sim r_{\mathrm{sp}}^{(\mathrm{eM})}$ of excited microsystems to the net number $N_2 - N_1 \sim r_{\mathrm{ind}}^{(\mathrm{eM})} - r_{\mathrm{ind}}^{(\mathrm{aM})}$ of emitting microsystems (total number $N = N_1 + N_2$),

$$\frac{r_{\mathrm{sp}}^{(\mathrm{eM})}}{r_{\mathrm{ind}}^{(\mathrm{M})}/N_P} = \frac{r_{\mathrm{sp}}^{(\mathrm{eM})}}{G/V} = \frac{f_C(W_0)[1 - f_V(W_0 - hf)]}{f_C(W_0) - f_V(W_0 - hf)} = n_{\mathrm{sp}}\,, \tag{3.40}$$

$$n_{\mathrm{sp}} = \frac{1}{1 - \exp\left(\dfrac{hf - (W_{Fn} - W_{Fp})}{kT}\right)} = \frac{N_2}{N_2 - N_1} = \frac{1}{1 - N_1/N_2}\,. \tag{3.41}$$

Maximum gain G is reached for complete inversion $n_{\mathrm{sp}} = 1$, i. e., for $N_1 = 0$ when all microsystems are excited and the VB is empty, $f_V = 0$. For practical operating points we have $n_{\mathrm{sp}} = 1.5 \ldots 2.5$. If the gain rate G is kept fixed for a certain device, the noise caused by the incoherent, spontaneous emission of photons is in proportion to the inversion factor n_{sp}. For low-noise optical amplification the inversion factor should be as closely to 1 as possible.

Figure 3.14 displays the spontaneous and induced emission spectra Eqs. (3.35), (3.37) for fixed quasi Fermi levels $W_{Fn} - W_C = 3kT_0$, $W_V - W_{Fp} = 0.5\,kT_0$, $(W_{Fn} - W_{Fp}) - W_G = 3.5\,kT_0$ and varying ratios T/T_0 (T_0 is a fixed reference temperature). The carrier masses $m_n/m_p = 0.14$ are that of GaAs, Table 3.3. The DOS product and the Fermi functions are

$$f_C(W_0) = \frac{1}{1 + \exp\Big[\ \Big(\underbrace{\frac{1}{1 + m_n/m_p}}_{0.877}\ \underbrace{\frac{hf - W_G}{kT_0}}_{x} - \underbrace{\frac{W_{Fn} - W_C}{kT_0}}_{3}\Big)\Big/\frac{T}{T_0}\Big]}\,, \tag{3.42}$$

$$f_V(W_0 - hf) = \frac{1}{1 + \exp\Big[-\Big(\underbrace{\frac{1}{1 + m_p/m_n}}_{0.123}\ \underbrace{\frac{hf - W_G}{kT_0}}_{x} - \underbrace{\frac{W_V - W_{Fp}}{kT_0}}_{0.5}\Big)\Big/\frac{T}{T_0}\Big]}\,.$$

The photon energy is expressed by the normalized quantity x,

$$x = \frac{hf - W_G}{kT_0}, \qquad x_0 = \frac{(W_{Fn} - W_{Fp}) - W_G}{kT_0} = 3.5\,. \tag{3.43}$$

The maximum spontaneous emission (see Fig. 3.14(a)) at $T \approx 0$ is located at a normalized frequency

[35] See Ref. 19 on Page 60

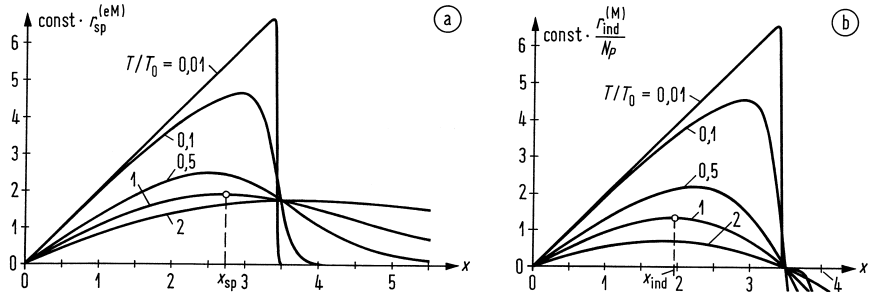


Fig. 3.14. Frequency dependence of spontaneous and induced emission for various temperatures $T/T_0 = 0.01$, 0.1, 0.5, 1, 2 (T_0 reference temperature; $W_{Fn} - W_C$ and $W_V - W_{Fp}$ are kept constant to $3\,kT_0$ and $0.5\,kT_0$, respectively; $m_n/m_p = 0.14$ as in GaAs). Normalized frequency $x = (hf - W_G)/(kT_0)$. (a) Spontaneous emission and (b) induced emission per photon, Eq. (3.42). The multiplicative constant is identical in both diagrams.

$x = x_0$, which corresponds to the difference of the quasi Fermi levels. It shifts for higher temperatures to lower frequencies (down to $x = 2.515$ at $T/T_0 = 0.52$ with the special assumptions of Eq. (3.42)). With a further temperature increase $T/T_0 > 0.52$ the maximum moves continuously to larger frequencies. The spontaneous emission maximum for the reference temperature $T = T_0$ is at $x_{\rm sp} = 2.73$.

The induced net emission rate per photon $r_{\rm ind}^{\rm (M)}/N_P = G/V$ and therefore the optical gain rate G is positive only for $0 < x < x_0$. It is zero at $x = x_0$ and becomes negative for $x > x_0 = 3.5$ because photons with energies $hf > W_{Fn} - W_{Fp}$ are absorbed, Fig. 3.14(b). At $T = 0$ the spectra of spontaneous emission and the gain rate are identical for $x < 3.5$ with an emission maximum at $hf = W_{Fn} - W_{Fp}$. However, with increasing T the maximum gain shifts to lower frequencies. For a fixed temperature the maximum gain is always at a lower frequency than the maximum spontaneous emission. At $T/T_0 = 1$ the maximum emissions are at $x_{\rm sp} = 2.73$ and $x_{\rm ind} = 1.95$, respectively. For GaAs at $T_0 = 293\,\mathrm{K}$ with a gain maximum at $\lambda = 0.842\,\mu\mathrm{m}$ the difference $x_{\rm sp} - x_{\rm ind} = 0.78$ at $T/T_0 = 1$ corresponds to a wavelength shift of $\Delta\lambda = 11.2\,\mathrm{nm}$.

For the chosen model of a k_μ-selection rule the spontaneous and stimulated emission spectra Fig. 3.14 exhibit a non-zero slope at $x = 0$ (this is also true for strict $\vec{k}_\mu$-selection). For the k_μ-selection rule[36] the ratio of the first derivatives $\mathrm{d}\,r_{\rm sp}^{\rm (eM)}/\,\mathrm{d}\,x$ and $\mathrm{d}(r_{\rm ind}^{\rm (M)}/N_P)/\,\mathrm{d}\,x$ amounts to $n_{\rm sp}$. Without k-selection these slopes are zero, and the ratio of the second derivatives would be $n_{\rm sp}$.

Figure 3.15 displays a measured gain curve[37] of an InAlGaAs/InP semiconductor laser. The zero-slope at the bandgap wavelength $\lambda_G = 1.65\,\mu\mathrm{m}$ (see Table 3.2 on Page 57 for the InGaAsP/InP compound),

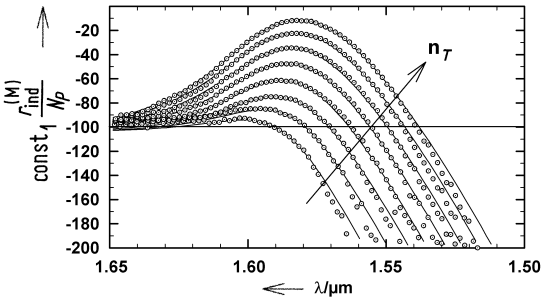


Fig. 3.15. Measured wavelength dependence of induced emission per photon for various carrier densities n_T. The multiplicative constant is different from Fig. 3.14 (after Ref. 37 on Page 71).

[36] See Footnote 33 on Page 69

[37] Wüst, F.: Optischer Gewinn und Alpha-Faktor in InAlGaAs/InP Quantenfilmlasern. PhD Thesis Karlsruhe 1999. — The material system offers a higher characteristic temperature T_0, a larger differential gain $\mathrm{d}G/\,\mathrm{d}n_T$ and a smaller linewidth enhancement factor α. This leads to a faster direct modulation capability than is possible for a InGaAsP/InP laser.

and the shift of $W_{Fn} - W_{Fp}$ to smaller emission wavelengths (larger frequencies) when increasing the carrier concentration n_T are clearly visible.

The ratio of the spontaneous emission rate $r_{\mathrm{sp}}^{(\mathrm{eM})}(f)$ into one mode of frequency f and the total spontaneous emission rate r_{sp} into all possible modes corresponds to the quotient of the probability density $\rho(f) \leq 2/(\pi \Delta f_H)$ that a microsystem with the required energy state is available, and the spectral density of all possible electromagnetic modes $\varrho_{\mathrm{tot}}(f)\, V$ inside the active volume V. The corresponding ratio for the induced emission rate $r_{\mathrm{ind}}^{(\mathrm{eM})}(f)$ into one mode of frequency f is larger by the photon number N_P (Eq. (3.35)),

$$\frac{r_{\mathrm{sp}}^{(\mathrm{eM})}(f)}{r_{\mathrm{sp}}} = \frac{\rho(f)}{\varrho_{\mathrm{tot}}(f)\, V}\,, \qquad \frac{r_{\mathrm{ind}}^{(\mathrm{eM})}(f)}{r_{\mathrm{sp}}} = \frac{r_{\mathrm{ind}}^{(\mathrm{eM})}(f)}{r_{\mathrm{sp}}^{(\mathrm{eM})}(f)}\, \frac{r_{\mathrm{sp}}^{(\mathrm{eM})}(f)}{r_{\mathrm{sp}}} = \frac{N_P}{1}\, \frac{\rho(f)}{\varrho_{\mathrm{tot}}(f)\, V}\,. \tag{3.44}$$

The left-hand side of Eq. (3.44) is an expression similar to Purcell's figure of merit[38] F_P. However, in Eq. (3.44) the relation $\rho(f) \ll \varrho_{\mathrm{tot}}(f)\, V$ holds normally, i. e., the emission linewidth "sees" a large number of modes where to emit. If the quotient $N_P/1$ of the induced and of the spontaneous emission rate into a mode f is reduced by the same factor $\rho(f)/\left(\varrho_{\mathrm{tot}}(f)\, V\right)$, the ratio of the induced emission rate into the mode f and the total emission probability into all modes results, right-hand side of Eq. (3.44).

By integrating the emission rate $r_{\mathrm{sp}}^{(\mathrm{eM})}(f)$ into one mode over all relevant modes the total spontaneous emission rate r_{sp} (unit $\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$) can be calculated ($r_{\mathrm{sp}}^{(\mathrm{eM})} \geq 0$ holds for $f \geq W_G/h$ only),

$$r_{\mathrm{sp}} \equiv \int_{-\infty}^{\infty} \underbrace{r_{\mathrm{sp}}\, \rho(f)}_{r_{\mathrm{sp}}(f)}\, \mathrm{d}f = \int_{-\infty}^{\infty} r_{\mathrm{sp}}^{(\mathrm{eM})}(f)\, \varrho_{\mathrm{tot}}(f) V\, \mathrm{d}f = \int_{W_G/h}^{\infty} r_{\mathrm{sp}}^{(\mathrm{eM})}(f)\, \varrho_{\mathrm{tot}}(f) V\, \mathrm{d}f\,. \tag{3.45}$$

The left-most equation in the chain exploits the identity $1 \equiv \int_{-\infty}^{\infty} \rho(f)\, \mathrm{d}f$, Eq. (3.33).

Radiative and nonradiative transitions For a radiative transition, both an electron and a hole participate. Therefore it is plausible that the total spontaneous emission rate can be written as:

$$r_{\mathrm{sp}} = B n_T p, \quad B = \begin{cases} 1 \times 10^{-10} \ldots 7 \times 10^{-10}\,\mathrm{cm}^3\,\mathrm{s}^{-1} & \text{(Ga,Al)As} \\ 8.6 \times 10^{-11}\,\mathrm{cm}^3\,\mathrm{s}^{-1} & \text{(In,Ga)(As,P)} \end{cases} \tag{3.46}$$

The smaller recombination coefficients B for (Ga,Al)As are valid for band-band transitions, the larger ones for transitions from the CB into non-ionized well localized (small spatial uncertainty Δx) shallow acceptor states which can provide a large difference momentum $\Delta(\hbar k) \geq (\hbar/2)/\Delta x$. For (In,Ga)(As,P) the temperature dependence is $B \sim 1/T^{\kappa}$ with $1 \leq \kappa \leq 1.5$. In thermal equilibrium the carrier concentrations Eq. (3.21) follow, i. e., $r_{\mathrm{sp}} = B n_i^2$. In the non-equilibrium case the spontaneous emission rate is $r_{\mathrm{sp}} = B(n_T p - n_i^2)$, but because normally $n_T p \gg n_i^2$ holds, Eq. (3.46) is a good approximation to the actual case. Figure 3.16(a) shows schematically a radiative recombination. Figure 3.16(b) displays a nonradiative recombination (rate r_{ns}, unit $\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$) via localized impurities in the forbidden band (rate $r_{\ell \mathrm{S}}$; such impurities can help in shortening the lifetime, see the discussion in Sect. 3.1.1 on Page 53). Eventually, Fig. 3.16(c),(d) presents Auger[39] processes (rate: r_{Au}), which are nonradiative. For (In,Ga)(As,P) the process Fig. 3.16(d) is important, while in (Ga,Al)As Auger processes are of no consequence. In summary we have:

$$r_{\mathrm{ns}} = r_{\ell \mathrm{S}} + r_{\mathrm{Au}} \qquad \begin{array}{l} r_{\ell \mathrm{S}} = A n_T \\ r_{\mathrm{Au}} = C n_T p^2 \end{array} \tag{3.47}$$

Measurements in (In,Ga)(As,P) result in coefficients $A = 1/(10\,\mathrm{ns})$ (undoped samples) up to $A = 1/(0.1\,\mathrm{ns})$ ($n_A = 2 \times 10^{18}\,\mathrm{cm}^{-3}$), and in $C = 4 \times 10^{-29}\,\mathrm{cm}^6\,\mathrm{s}^{-1}$ (calculated: $C = 10^{-27} \ldots 10^{-31} \times \mathrm{cm}^6\,\mathrm{s}^{-1}$). The Auger coefficient C increases with temperature; also A increases slightly. The structure

[38] See Eq. (3.34) in Footnote 30 on Page 68

[39] Pierre Victor Auger (pronounced [oˈʒe], *not* [ˈɔːgə(ɹ)]!) ⋆ Paris (France) 14.5.1899, † Paris (France) 24.12.1993, French physicist. He worked in the fields of atomic physics, nuclear physics, and cosmic ray physics.

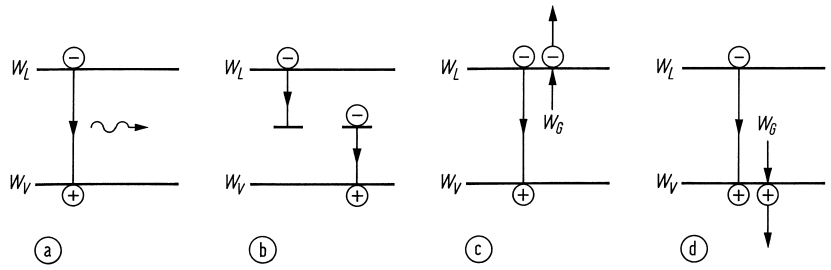


Fig. 3.16. Radiative and nonradiative transitions. (a) Radiative band-band transition. (b) Nonradiative transition via localized states in the forbidden band. (c) (d) Nonradiative Auger recombinations (recombination energy excites an electron in the CB or in the VB)

of $r_{\ell S}$ follows from the proportionality $r_{\ell S} \sim n_T$ because only one carrier type is involved. For Auger recombination we have $r_{Au} \sim n_T p p$ because two carrier types recombine, and additionally a hole (case of Fig. 3.16(d)) must be available to take over the excitation energy[40]. It is useful to define an effective recombination rate

$$r_{\text{eff}} = r_{\text{sp}} + r_{\text{ns}} = r_{\text{sp}} + r_{\ell S} + r_{\text{Au}} \,. \tag{3.48}$$

Inside the recombination zone of a diode (layer height d, cross-section area F) the carrier density changes if the injected carrier rate (injection current density J, elementary charge e) deviates from the recombination rate,

$$\frac{\mathrm{d}n_T}{\mathrm{d}t} = \frac{J}{ed} - r_{\text{eff}}(n_T). \tag{3.49}$$

Strictly speaking, $r_{\text{eff}}(n_T)$ in Eq. (3.49) should be replaced by $r_{\text{eff}}(n_T) - r_{\text{eff}}(n_{T\,\text{equil}})$ for the correct solution at a concentration $n_{T\,\text{equil}}$ for thermal equilibrium $J = 0$. With a step perturbation of the current density from J_0 to $J_0 + J_1$, a perturbation ansatz $n_T(t) = n_{T0} + n_{T1}(t)$ together with a series expansion of $r_{\text{eff}} = r_{\text{eff}\,0} + (\partial r_{\text{eff}}/\partial n_T) n_{T1}$ at n_{T0} results in

$$n_{T1}(t) = \frac{J_1 \tau_{\text{eff}}}{ed} \left(1 - \mathrm{e}^{-t/\tau_{\text{eff}}}\right), \quad \text{with} \quad \tau_{\text{eff}}^{-1} = \frac{\partial r_{\text{eff}}}{\partial n_T} \,. \tag{3.50}$$

In an analogous form the (carrier concentration dependent) lifetimes for the other recombination processes may be defined. With Eqs. (3.46), (3.47) on calculates:

$$\left.\begin{aligned} \tau_{\text{sp}}^{-1} &= \frac{\partial r_{\text{sp}}}{\partial n_T} = B\left(p + n_T \frac{\partial p}{\partial n_T}\right), \\ \tau_{\ell S}^{-1} &= \frac{\partial r_{\ell S}}{\partial n_T} = A, \\ \tau_{\text{Au}}^{-1} &= \frac{\partial r_{\text{Au}}}{\partial n_T} = C\left(p^2 + 2 n_T p \frac{\partial p}{\partial n_T}\right), \end{aligned}\right\} \quad \begin{aligned} \tau_{\text{eff}}^{-1} &= \tau_{\text{sp}}^{-1} + \tau_{\text{ns}}^{-1}, \\ \tau_{\text{ns}}^{-1} &= \tau_{\ell S}^{-1} + \tau_{\text{Au}}^{-1}. \end{aligned} \tag{3.51}$$

The internal quantum efficiency η_{int} of radiative recombination is defined by

$$\frac{1}{\eta_{\text{int}}} = \frac{\tau_{\text{sp}}}{\tau_{\text{eff}}} = 1 + \frac{\tau_{\text{sp}}}{\tau_{\text{ns}}} = 1 + \tau_{\text{sp}} \left(\frac{1}{\tau_{\ell S}} + \frac{1}{\tau_{\text{Au}}}\right). \tag{3.52}$$

The smaller the effective lifetime τ_{eff} is, the faster the spontaneously emitted light can follow, see the discussion in Sect. 3.1.1 on Page 53. However, if nonradiative processes determine the lifetime, the internal quantum efficiency η_{int} deteriorates.

[40] For high electron injection $p \approx n_T$ the recombination rate follows the law $r_{\text{Au}} \sim n_T^3$, i. e., it increases faster than the radiative emission rate Eq. (3.46). Therefore, (In,Ga)(As,P) lasers or LED must not be highly p-doped or operated at high current densities.

For moderate modulation frequencies, for which the modulation period is much smaller than τ_{eff} so that $\left(1 - \mathrm{e}^{-t/\tau_{\text{eff}}}\right) \approx 1$ is a valid approximation, the generated radiation power amplitude $P_1 = (n_{T1}Fd)\ hf/\tau_{\text{sp}}$ follows the instantaneous current amplitude I_1, $P_1 \sim n_{T1} \sim J_1 F = I_1$. The quantity $n_{T1}Fd$ is the number of spontaneously emitted photons,

$$P_1 = \frac{n_{T1}Fd\,hf}{\tau_{\text{sp}}} = \frac{J_1(t)\tau_{\text{eff}}}{ed}\frac{hfFd}{\tau_{\text{sp}}} = \eta_{\text{int}} hf\,\frac{I_1}{e}, \qquad \eta_{\text{int}} = \frac{P_1/(hf)}{I_1/e}\,. \tag{3.53}$$

For high current densities the current dependence of τ_{sp} and η_{int} (see Eq. (3.51)) leads to nonlinear distortions. The internal quantum efficiency η_{int} represents the average number of photons per injected electron.

3.1.5 Light-emitting diode

Output power and modulation properties Light-emitting diodes (LED) operate without end mirrors in a mode where the spontaneous emission rate $r_{\text{sp}}^{(\text{eM})}$ dominates, Eq. (3.35). For communication purposes LED with double-heterostructures are common, Fig. 3.10. The generated light power P is given by Eq. (3.53),

$$P = \frac{n_T Fd\,hf}{\tau_{\text{sp}}} = \eta_{\text{int}} hf\,\frac{I}{e}\,, \qquad \eta_{\text{int}} = \frac{\tau_{\text{eff}}}{\tau_{\text{sp}}} = \frac{P/(hf)}{I/e}\,. \tag{3.54}$$

The mean photon energy hf of the emission line is slightly larger than the bandgap energy W_G. For a flat-diode configuration the power reflection factor at the boundary assuming nearly perpendicular incidence

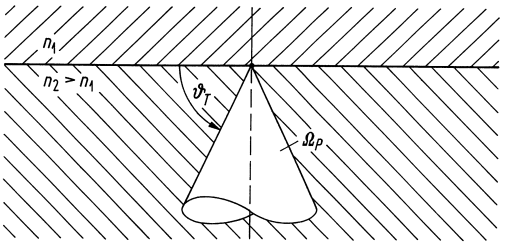


Fig. 3.17. Plane boundary between two media ($n_1, n_2 > n_1$ refractive indices, ϑ_T critical angle of total reflection, R_P power reflection factor). Only the fraction $(1 - R_P)$ of the radiation from the solid angle Ω_P is transmitted into the medium n_1.

is R_P according to Eq. (3.1) on Page 50. The radiation is isotropically emitted into the full solid angle 4π, but only a fraction $(1 - R_P)\,\Omega_P/(4\pi)$ given by the critical solid angle Ω_P for total internal reflection (cone semi-angle $\pi/2 - \vartheta_T$) is transmitted into the medium with lower refractive index $n_1 < n_2$, Fig. 3.17. The optical efficiency η_{opt} describes the amount of usable light. The following numerical values apply for the radiation from a semiconductor into air (or fused silica):

$$\eta_{\text{opt}} = \frac{\Omega_P}{4\pi}\,(1 - R_P) = 1.5\,\%\ (3.5\,\%) \qquad \begin{cases} \Omega_P = 2\pi(1 - \sin\vartheta_T) = 0.27\,\text{sr}\ (0.54\,\text{sr}), \\ \cos\vartheta_T = n_1/n_2 \quad = 73\,^\circ \quad (66\,^\circ), \\ R_P = \left(\dfrac{n_1 - n_2}{n_1 + n_2}\right)^2 \quad = 32\ \% \quad (18\ \%). \end{cases} \tag{3.55}$$

The radiated output power P_a of the LED is

$$P_a = \eta_{\text{opt}} P = \eta_{\text{ext}} hf\,\frac{I}{e}, \quad \rightarrow \quad \eta_{\text{ext}} = \frac{P_a/(hf)}{I/e} = \eta_{\text{opt}}\,\eta_{\text{int}}. \tag{3.56}$$

The external quantum efficiency is denoted as η_{ext}. The voltage drop U over the pn-junction when an injection current I is flowing corresponds to the energetic distance of the quasi Fermi levels (Fig. 3.12), which is in the order of the photon energy, $eU = W_{Fn} - W_{Fp} \approx hf$, Eq. (3.9). With the electrical input power $P_{\text{el}} = UI$ a total power-conversion or wall-plug efficiency η_{tot} may be defined by

$$P_a = \eta_{\text{tot}} P_{\text{el}} = \eta_{\text{ext}} hf \frac{I}{e} \quad \rightarrow \quad \eta_{\text{tot}} = \eta_{\text{ext}} \frac{hf}{eU} \approx \eta_{\text{ext}} = \frac{P_a/(hf)}{P_{\text{el}}/(eU)} \quad \text{for } P_{\text{el}} = UI\,. \tag{3.57}$$

For direct semiconductors we have typically $0.5 \leq \eta_{\text{int}} \leq 0.9$. The larger values are found with moderate p-doping of the active layer, because then the electron number in the CB is reduced and inversion is more simply to achieve, see Footnote 25 on Page 62. The total conversion efficiency for plane surface emitters is in the order $\eta_{\text{tot}} = \eta_{\text{opt}}\,\eta_{\text{int}} = 0.75 \ldots 3.1\,\%$. With operating currents up to 200 mA the emitted power reaches the 10 mW range.

For larger injection currents I the optical power P_a tends to saturate and even to diminish. For (In,Ga)(AsP) this would be true even at a constant junction temperature because of Auger recombination, but actually the increased heat enforces this effect by a thermally reduced B, by an increase of C (not for GaAs), and by the reduced carrier confinement in the double-heterostructure at elevated temperatures (less pronounced in GaAs because of larger barriers than in (In,Ga)(AsP)). The temperature coefficient $c_X = (1/X)(\mathrm{d}X/\,\mathrm{d}T)$ of the power amounts to $c_{Pa} = -1.4 \times 10^{-2}\,\mathrm{K}^{-1}$ for GaAs and $c_{Pa} = -2 \times 10^{-2}\,\mathrm{K}^{-1}$ for (In,Ga)(As,P).

A small signal (perturbation) ansatz $g = g_0 + g_1(\omega)\,\mathrm{e}^{\mathrm{j}\,\omega t}$ for n_T, J in Eq. (3.49) and for P, I, P_a in Eq. (3.54), (3.56) together with the effective lifetime τ_{eff} Eq. (3.50) leads to the spectral relation

$$P_{a1}(\omega) = \eta_{\text{ext}} hf \frac{I_1(\omega)}{e} \frac{1}{1 + \mathrm{j}\,\omega\tau_{\text{eff}}}\,. \tag{3.58}$$

For a constant modulation current amplitude $|I_1(\omega)|$ we find the current-power transfer function

$$\left|\frac{P_{a1}(\omega)}{P_{a1}(0)}\right| = \frac{1}{\sqrt{1 + \omega^2 {\tau_{\text{eff}}}^2}} \qquad \rightarrow \quad \omega_c = \frac{1}{\tau_{\text{eff}}}\,. \tag{3.59}$$

The angular 3-dB cutoff frequency at $|P_{a1}(\omega_c)/P_{a1}(0)| = 1/\sqrt{2}$ is denoted as ω_c. The photodetector current is proportional to the received optical power, $i_1(\omega) \sim P_{a1}(\omega)$. The angular frequency ω_c corresponds to the half-power point of the received signal power $i_1^2(\omega) \sim |P_{a1}(\omega)|^2$ at the angular modulation frequency ω. The signal amplitude increases with the efficiencies and thus decreases with the cutoff frequency, Eqs. (3.58), (3.56), (3.52),

$$i_1(\omega) \sim P_{a1}(0) \sim \eta_{\text{ext}} \sim \eta_{\text{int}} \sim \tau_{\text{eff}} \sim \frac{1}{\omega_c}. \tag{3.60}$$

For a given material the cutoff angular frequency ω_c may be increased only by forcing nonradiative recombination at the cost of efficiency. For high-speed LED $f_c = 1\,\mathrm{GHz}$ is possible. A further decrease of τ_{eff} is counterproductive because eventually the LED junction capacitance C in combination with the source resistance R fixes the time constant $\tau = RC$.

LED spectrum

The detailed distribution of the carriers into states of the CB and VB depends on the Fermi level, i. e., on the impurities in the active layer and on the injection current. With increasing temperature the Fermi function changes, and the spontaneous emission maximum shifts first to lower, then to higher frequencies, Fig. 3.14 on Page 71. The bandgap energy decreases with increasing temperature. Both effects together result in a shift of the emission maximum to lower frequencies (larger wavelengths) by a rate of $0.2\,\mathrm{nm\,K}^{-1}$ for GaAs and $0.4\,\mathrm{nm\,K}^{-1}$ for (In,Ga)(As,P) near $\lambda = 1.3\,\mu\mathrm{m}$.

Figure 3.6(a) on Page 59 shows the bandstructure of a direct semiconductor. The spectral width of the emission is determined mainly by the energetic distribution of the carriers in the bands. The

occupation probabilities are determined by the Fermi function, Fig. 3.7 on Page 61. The quasi Fermi levels $f(W_{Fn,p}) = 1/2$ are usually near the band edge, $W_{Fn,p} \approx W_{C,V}$. The Fermi function changes significantly from $f(W_{Fn,p} - 2kT_0) = 0.88$ to $f(W_{Fn,p} + 2kT_0) = 0.12$ in an energetic interval $4kT_0$ around the Fermi energy. For both GaAs and InP, the curvatures of the CB and VB in Fig. 3.6(b) on Page 59, i. e., the reciprocal effective carrier masses in Eq. (3.13) and in Table 3.3 on Page 60, are significantly different, $m_n \ll m_p$. Compared to the CB, the VB is virtually flat. Therefore, photons are emitted in the spectral range $W_G \leq hf \leq (W_C + 2kT_0) - W_V$. The total spectral width of the emission essentially amounts to $h\,\Delta f_H = 2kT_0$,

$$h\,\Delta f_H = 2\,kT_0 = 50\,\mathrm{meV}, \quad \Delta f_H = 12.1\,\mathrm{THz} \quad \text{at room temperature } T_0 = 293\,\mathrm{K}\ . \tag{3.61}$$

For GaAs we have $\Delta\lambda_H = 30\,\mathrm{nm}$, for (In,Ga)(As,P) $\Delta\lambda_H = 70\,\mathrm{nm}$ at $\lambda = 1.3\,\mu\mathrm{m}$. The quantity Δf_{gain} is also an estimate for the amplification bandwidth of semiconductor laser devices. Basically, it corresponds to the width Δf_H of the lineshape $\rho(f)$, Eq. (3.33) on Page 68.

Devices

The surface emitter (LED) and the edge emitter (ELED) are the two basic device configurations to couple the LED light output into a small-diameter glass fibre.

Surface emitter For the surface emitter Fig. 3.18 the emitting area of the junction is confined by oxide isolation, and the contact is usually $15 \ldots 100\,\mu\mathrm{m}$ in diameter. The active p-GaAs layer is part of a 3-layer heterostructure. The device is known as a Burrus diode[41]. The n-GaAs substrate is thinned

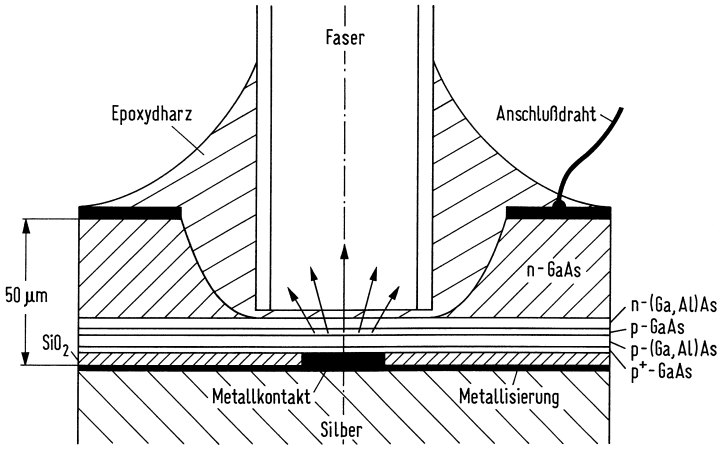


Fig. 3.18. Small-area high-radiance (Ga,Al)As double-heterostructure surface-emitting LED with attached fibre (Burrus diode). Epoxydharz = epoxy resin, Anschlußdraht = bond wire, Metallkontakt = metal contact, Metallisierung = metallization

by etching to reduce the absorption. This is not required in an (In,Ga)(As,P) system because the InP substrate is transparent having a wider bandgap than the active (In,Ga)(As,P) layer. The smaller the junction area F is, the better the heat can be removed, the higher the current density can be chosen, and the brighter the emitted light will be. Depending on the angle γ measured from an axis perpendicular to the emitting surface, the apparent radiating area changes according to $F\cos\gamma$. The radiance L is defined as the differential power $\mathrm{d}P$ radiated from a differential apparent area $\mathrm{d}F\cos\gamma$ into a differential solid angle $\mathrm{d}\Omega$ centred at an angle γ,

$$L = \frac{\mathrm{d}^2 P}{\mathrm{d}F\cos\gamma\,\mathrm{d}\Omega}\,, \qquad \frac{\mathrm{d}P}{\mathrm{d}\Omega} = \mathrm{d}F\cos\gamma\,. \tag{3.62}$$

[41] Burrus, C. A.; Miller, B. I.: Small-area DH Al-Ga-As electroluminescent diode sources for optical fiber transmission lines. Opt. Commun. 4 (1971) 307–309

An emitter with a constant radiance L and a $\cos\gamma$-dependent far-field power distribution P is called a Lambertian source. Its half-power width is $\Delta\gamma_H = 120\,°$. Because of the large emission angle, LED light may be coupled efficiently only to multimode waveguides.

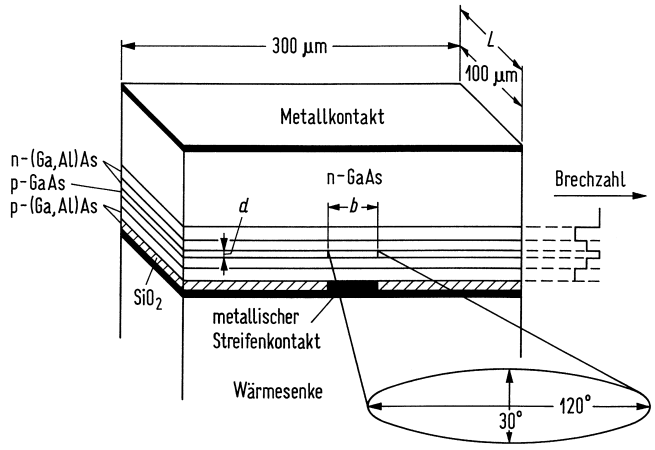


Fig. 3.19. Edge-emitting double-heterostructure LED. L, b, d are length, width and thickness of the active zone. Metallkontakt = metal contact, metallischer Streifenkontakt = metallic contact strip, Wärmesenke = heat sink, Brechzahl = refractive index

Edge emitter The edge emitter is shown in Fig. 3.19. The active layer is very thin, $d = 0.05 \ldots 0.1\,\mu$m. Together with the adjacent layers of reduced refractive index this 5-layer double-heterostructure forms a vertically single-moded strip waveguide. The lateral width of $b = 10 \ldots 50\,\mu$m is effectively defined by the carrier injection from the contact stripe. Because of the vertical waveguiding the far-field half-power width *in vacuum* is less than 120 °, typically $\Delta\gamma_H = 0.52 \mathrel{\widehat{=}} 30\,°$. For a single-mode slab waveguide with a normalized frequency $V = (d/2)k_0 A_N$ and $V \leq V_{1G} = \pi/2$, a vacuum far-field angle $\gamma_N \approx \sin\gamma_N = A_N = (V_{1G}/\pi)(\lambda/d) = \frac{1}{2}\lambda/d$ would be expected. With $\Delta\gamma_H = 2\gamma_N \approx \lambda/d_{\text{eff}}$ and $\lambda = 1.3\,\mu$m an associated effective vertical field extension d_{eff} would be in the order of the vacuum wavelength, $d_{\text{eff}} = \lambda/\Delta\gamma_H = 2.5\,\mu$m. This is larger than the actual significant field extension *in the waveguide* $x_M = 1.8\,d = 0.18\,\mu$m for $h \mathrel{\widehat{=}} d = 0.1\,\mu$m. While immediately at the semiconductor-vacuum interface the transverse field extension is x_M, a few wavelength away from the interface into the vacuum the effective field extension is of the order of λ, because of diffraction.

Laterally the ELED behaves as a Lambertian source. The device radiates perpendicular to the main current flow direction through the cleaved end faces. To increase the efficiency, one endface may be coated to yield a very high reflection factor, while the other end face may be antireflection-coated. Typical device lengths are $L = 100\,\mu$m.

The narrow contact stripe allows high current densities. Therefore, and because of the spatially coherent field in the vertical waveguide, the radiance of edge-emitters is larger (up to $L = 1\,000\,\text{W}\,\text{cm}^{-2}\,\text{sr}^{-1}$) than for surface emitters, and a more efficient power coupling into single-mode fibres becomes possible.

Superluminescent diode With higher current densities and larger lengths up to about $L = 500\,\mu$m induced amplification of the spontaneously emitted light can become important. Such a device is called a superluminescent LED (SLED). Because the effective carrier lifetime τ_{eff} is reduced by stimulated emission, Eq. (3.5), the maximum modulation frequency $f_c = 1/(2\pi\tau_{\text{eff}})$ increases without paying an output power penalty, Eqs. (3.59), (3.60). Simultaneously, the emission linewidth becomes smaller because of the frequency-selective amplification, i. e., the temporal coherence becomes larger. The SLED light may be coupled to external waveguides as well as it this the case for an ELED.

3.1.6 Laser diode

Basic relations

A basic laser diode (LD) that has a rectangular cavity is equivalent to a Fabry-Perot (FP) resonator, Fig. 3.1 on Page 50 and Fig. 3.5 on Page 58, and is thus called a Fabry-Perot laser diode (FP LD). The structure is similar to the one of the edge-emitter Fig. 3.19 having a laser-active volume $V = dbL$ with dimensions $d = 0.1 \ldots 0.2\,\mu$m (vertical, x-direction), $b = 2 \ldots 5\,\mu$m (lateral, y-direction) and $L = 300 \ldots 1\,200\,\mu$m, (longitudinal, z-direction).

Waveguiding properties and resonances The transverse waveguiding mechanism is described by an effective refractive index $n_e < n$ (Eq. (2.13) on Page 18) which is smaller than n in Fig. 3.1. To avoid complicated subscripts we drop the index e for convenience, and implicitly regard the propagation quantities as effective waveguide quantities. This applies, e. g., to the (complex) propagation constant, to the modal loss α_V in Eq. (3.75), and to the longitudinal mode spacing Δf_z.

For plane waves propagating along the z-axis ($k_{x,y} = 0$, $k_z = k$) the longitudinal resonances are given as in Eq. (3.2) and the foregoing text on Page 50 by

$$k \times 2L = k_0 n \times 2L = \omega n \times 2L/c = m_z \times 2\pi\,, \quad m_z = 1, 2, 3, \ldots \tag{3.63}$$

Regarding m_z for the moment as a continuous variable, we differentiate $fn = m_z c/(2L)$ with respect to m_z resulting in

$$\frac{\mathrm{d}(fn)}{\mathrm{d}m_z} = \frac{\mathrm{d}f}{\mathrm{d}m_z}\, n + f\,\frac{\mathrm{d}n}{\mathrm{d}f}\,\frac{\mathrm{d}f}{\mathrm{d}m_z} = \frac{\mathrm{d}f}{\mathrm{d}m_z}\left(n + f\,\frac{\mathrm{d}n}{\mathrm{d}f}\right) = \frac{c}{2L}\,, \qquad n_g = n + f\,\frac{\mathrm{d}n}{\mathrm{d}f}\,. \tag{3.64}$$

Recalling the discrete nature of m_z, the replacements $\mathrm{d}m_z \to 1$ and $\mathrm{d}f \to \Delta f_z$ are appropriate. Introducing the group index n_g, the group velocity v_g (Eq. (2.15) on Page 19) and the photon round-trip time τ_U (Eq. (3.4) on Page 51), this leads to the equidistant longitudinal mode spacing (free spectral range FSR, see also Eq. (3.3) on Page 51),

$$\Delta f_z = \frac{c}{2n_g L} = \frac{v_g}{2L} = \frac{1}{\tau_U}\,. \tag{3.65}$$

The typical comb structure of the spectrum is displayed in Fig. 3.24 on Page 91.

Field confinement factor The field energy is not concentrated in the active volume V alone, because the transversely evanescent field extends into the cladding of the waveguide Fig. 3.1. The extent of the field concentration is given by the field confinement factor Γ of the fundamental TE mode ($\vec{E}$ parallel to y-axis),

$$\Gamma_{\mathrm{TE}} = \int_{-d/2}^{+d/2} |E_y(x)|^2\,\mathrm{d}x \,\Big/ \int_{-\infty}^{+\infty} |E_y(x)|^2\,\mathrm{d}x\,. \tag{3.66}$$

An approximation valid for all V (for the fibre V-parameter, see Eq. (2.13) on Page 18) with a maximum error of 1.5 % is[42]

$$\Gamma_{\mathrm{TE}} = \frac{2V^2}{1 + 2V^2}, \qquad V = \frac{d}{2}\,k_0\sqrt{n_1^2 - n_2^2}. \tag{3.67}$$

The field confinement Γ for the TE mode is slightly larger than the one for the TM mode ($\vec{H}$ parallel to y-axis), $\Gamma_{\mathrm{TE}} > \Gamma_{\mathrm{TM}}$. An example for a laser is $\Gamma_{\mathrm{TE}} = 0.184$ and $\Gamma_{\mathrm{TM}} = 0.145$, for a laser amplifier it is $\Gamma_{\mathrm{TE}} = 0.3$ and $\Gamma_{\mathrm{TM}} = 0.25$. For $d = 0.1 \ldots 0.2\,\mu$m the values $\Gamma = 0.2 \ldots 0.6$ are typical. This and the larger endface reflection factor for TE polarization is the reason why diode lasers usually oscillate in TE polarization.

[42] Botez, D.: Analytical approximation of the radiation confinement factor for the TE_0 mode of a double heterojunction laser. IEEE J. Quantum Electron. QE-14 (1978) 230–232

Emission and absorption rates With respect to Γ, the equations for induced and spontaneous emission as well as the total mode number and density M_{tot} and ϱ_{tot} have to be modified. The oscillating laser mode having a photon number N_P fills effectively a volume V/Γ which is larger than the active volume V. This increases the total mode number M_{tot} and mode density ϱ_{tot}, because the active volume V (not to be mixed up with the normalized frequency parameter V in Eq. (3.67)) has to be replaced by the effective mode volume V/Γ. The number of photons interacting with the active medium is reduced to ΓN_P, and the equivalent energy density for spontaneous and induced transitions $((N_P+1)\,hf\,/V$ (see Eq. (3.32), (3.35)) should be replaced by $\Gamma(N_P+1)\,hf\,/V$. The transverse waveguiding mechanism is described by an effective refractive index $n_e < n$ (Eq. (2.13) on Page 18) which is smaller than n in Fig. 3.1. In summary, we have to substitute in Eqs. (3.39), (3.40)

$$\begin{aligned} r_{\mathrm{ind}}^{(\mathrm{M})} &\longrightarrow r_{\mathrm{ind}\,e}^{(\mathrm{M})} = \Gamma r_{\mathrm{ind}}^{(\mathrm{M})}, \qquad M_{\mathrm{tot}} \longrightarrow M_{\mathrm{tot}\,e} = M_{\mathrm{tot}}/\Gamma\,, \\ r_{\mathrm{sp}}^{(\mathrm{eM})} &\longrightarrow r_{\mathrm{sp}\,e}^{(\mathrm{eM})} = \Gamma r_{\mathrm{sp}}^{(\mathrm{eM})}, \qquad \varrho_{\mathrm{tot}} \longrightarrow \varrho_{\mathrm{tot}\,e} = \varrho_{\mathrm{tot}}/\Gamma\,. \end{aligned} \tag{3.68}$$

The inversion factor n_{sp} Eq. (3.41) remains unchanged because of Eq. (3.40). The same is true for the total spontaneous emission rate r_{sp} Eq. (3.45), and as a consequence also for the effective recombination rate $r_{\mathrm{eff}} = r_{\mathrm{sp}} + r_{\mathrm{ns}} = r_{\mathrm{sp}} + r_{\ell\mathrm{S}} + r_{\mathrm{Au}}$ Eq. (3.48),

$$\begin{aligned} n_{\mathrm{sp}} \longrightarrow n_{\mathrm{sp}\,e} = n_{\mathrm{sp}}\,, \quad & r_{\mathrm{sp}} \longrightarrow r_{\mathrm{sp}\,e} = r_{\mathrm{sp}}\,, \\ & r_{\mathrm{eff}} \longrightarrow r_{\mathrm{eff}\,e} = r_{\mathrm{eff}}\,. \end{aligned} \tag{3.69}$$

Gain and loss Actually, the resonator is longitudinally multimoded, Eq. (3.65). We describe the modes by plane waves with effective propagation properties and a complex (effective) refractive index $\bar{n} = n - \mathrm{j}\,n_i$ with real part n and imaginary part $-n_i$ (dropping the subscript e as discussed on Page 78),

$$\exp(-\mathrm{j}\,\bar{k}z), \qquad \left\{ \begin{array}{l} \bar{k} = k_0\bar{n} = k + \frac{1}{2}\,\mathrm{j}\,(g - \alpha_V), \\ \bar{n} = n - \mathrm{j}\,n_i, \\ k_0 = \omega/c, \end{array} \right\}, \qquad g - \alpha_V = -2\,k_0 n_i. \tag{3.70}$$

The quantities g, α_V are the modal power gain and loss constants corresponding to the net effective gain rate ΓG due to band-band transitions, and a power loss time constant $1/\tau_V$ to be discussed in the following which does not include band-band transitions.

According to Eq. (3.70) the wave experiences a net power gain of $\exp\left[(g-\alpha_V)2L\right]$ for a round-trip of length L between the resonator mirrors with power reflection coefficients $R_{1,2}$. Because of the mixture of length-distributed gain and localized mirror losses one is usually not interested in keeping track of how many times the light goes back and forth for amplification. Instead of using the gain per length it is then a more practical approach to define an equivalent gain per time.

At each partially transparent mirror the localized losses are equivalently described by a power "gain" $R_1R_2 = \exp(-\alpha_{R1}2L)\exp(-\alpha_{R2}2L) = \exp(-\alpha_R 2L)$ distributed over a round-trip through the resonator. The gain rate G (see Eq. (3.39) on Page 70) specifies the number of photons generated per second. The total losses of photons per second $1/\tau_P$ are described by the photon lifetime τ_P. The round-trip time $\tau_U = 2L/v_g$ for a photon can be computed from its group velocity v_g (see Eq. (2.15) on Page 19). Assuming a constant gain rate per round-trip time τ_U, the net increase in photon number per time including all losses is

$$G - \frac{1}{\tau_P} = \frac{1}{N_P}\frac{\mathrm{d}N_P}{\mathrm{d}t}\,, \qquad \frac{N_P(\tau_U)}{N_P(0)} = \exp\left[\left(G - \frac{1}{\tau_P}\right)\tau_U\right], \qquad \tau_U = \frac{2L}{v_g}\,. \tag{3.71}$$

Taking into regard the loss mechanisms discussed above, the following relations hold between the gain rate Eq. (3.71) and the modal power gain Eq. (3.70),

$$\begin{aligned} \exp\big[(G - 1/\tau_P)\tau_U\big] &= \exp\big[(G - 1/\tau_V - 1/\tau_R)\tau_U\big] \\ &= R_1R_2\exp\left[(G-1/\tau_V)\tau_U\right] = R_1R_2\exp\left[(G-1/\tau_V)2L/v_g\right] \\ &= R_1R_2\exp\big[(g-\alpha_V)2L\big] \\ &= \exp\big[(g-\alpha_V-\alpha_R)2L\big] = \exp\big[(g-\alpha_V-\alpha_{R1}-\alpha_{R2})2L\big]. \end{aligned} \tag{3.72}$$

Comparing the various forms in Eq. (3.72) we find

$$\begin{aligned} G &= v_g g, \\ 1/\tau_V &= v_g \alpha_V, \\ 1/\tau_{R1,2} &= v_g \alpha_{R1,2} = -v_g \ln R_{1,2}/(2L), \\ 1/\tau_R &= v_g \alpha_R \ = -v_g \ln(R_1 R_2)/(2L), \\ 1/\tau_P &= v_g(\alpha_V + \alpha_R) = v_g[\alpha_V - \ln(R_1 R_2)/(2L)]. \end{aligned} \tag{3.73}$$

With the same reasoning as above, the net material gain rate $G = v_g g$ must be replaced by the net modal gain rate $\Gamma G = v_g \Gamma g$,.

$$G \longrightarrow G_e = \Gamma G\,, \quad g \longrightarrow g_e = \Gamma g\,. \tag{3.74}$$

The resonant mode is attenuated mainly in the active zone (a background material power loss constant α_V *not* including band-band transitions) and by the adjacent heterolayers (power loss constant $\alpha_{\rm het}$). Additionally, interface scattering and substrate losses could have some influence (power loss constant $\alpha_{\rm add}$). So the material loss has to be replaced by the modal loss,

$$\alpha_V \longrightarrow \alpha_{Ve} = \Gamma\alpha_V + (1-\Gamma)\alpha_{\rm het} + \alpha_{\rm add}\,. \tag{3.75}$$

Gain model The (effective) modal loss constant Eq. (3.75) ($\alpha_V \approx \alpha_{\rm het}$, $\alpha_{\rm add} = 0$, $\alpha_{Ve} \approx \alpha_V$) is typically in the order of $\alpha_V = 20\ldots50\,{\rm cm}^{-1}$. The necessary threshold gain constant $\Gamma g = \alpha_V + \alpha_R$ is in the region $\Gamma g = 25\ldots90\,{\rm cm}^{-1}$. For GaAs and at the spectral gain maximum (see Fig. 3.14(b)) the approximate n_T-dependency is

$$g = g_0 \times (n_T/n_t - 1), \quad g_0 = 330\,{\rm cm}^{-1}, \quad n_t = 1.1\times10^{18}\,{\rm cm}^{-3}\,. \tag{3.76}$$

The carrier concentration for a zero net gain constant $g = 0$ is called the transparency carrier concentration n_t. A typical threshold carrier density amounts to $n_{TS} = 1.2\times10^{18}\ldots1.4\times10^{18}\,{\rm cm}^{-3}$. The refractive index of the active layer is about $n = 3.5$, Table 3.2, the group refractive index is in the region $n_g = 3.75\ldots5$.

The laser oscillates near the frequency f_0 of the maximum spectral gain. However, a larger injected carrier number leads to a nonlinear gain compression, because the energy states near the laser resonance energy hf_0 deplete due to hot carrier effects and spectral hole burning[43]. To fill the depleted states it needs the intraband relaxation time $\tau_{\rm CB}$ (Page 63), and this represents a "bottleneck" for the number of carriers available in energy states near hf_0. Phenomenologically, the nonlinear gain compression is modeled by a photon-number dependent decrease of the gain described by a gain compression factor ε_G. With the differential gain G_d and the transparency concentration n_t the optical gain is

$$G(n_T, N_P) = \frac{G(n_T)}{1+\varepsilon_G\frac{\Gamma N_P}{V}} = G_d\frac{n_T - n_t}{1+\varepsilon_G\frac{\Gamma N_P}{V}}\,. \tag{3.77}$$

[43] Schuster, S.; Haug, H.: Calculation of the gain saturation in cw semiconductor lasers with Boltzmann kinetics for Coulomb and LO phonon scattering. Semicond. Sci. Technol. 10 (1995) 281–289

„Spectral hole burning means the formation of a dip in the carrier distribution function around the laser resonance due to the finite intraband scattering time.

The stimulated emission heats the CB carriers by removing cool particles [between the bandedge and the quasi Fermi energy W_{Fn}], because the energy of the recombining electron-hole pairs is smaller than the average pair energy which is roughly the difference of the quasi Fermi energies $W_{Fn} - W_{Fp}$.

The phenomenological gain saturation coefficient ε_G stems from the finite intraband scattering time which yields an increase of the CB carrier density with increasing pump current. An injected hot electron-hole pair needs a finite time for the scattering into the laser resonance [energy hf] where stimulated recombination may occur. Therefore a higher pump current does not only result in a higher light intensity but also in a more pronounced non-equilibrium carrier distribution function with particle depletion around resonance (spectral hole burning), an increasing average carrier energy (carrier heating) and a growing density due to this kinetic 'bottleneck'.

For smaller pump rates the gain saturation calculated within the microscopic model [treating the carrier distributions in terms of the Boltzmann collision integral] decreases, which indicates that in this regime the kinetic limitation of the pump efficiency due to the carrier scattering into the laser resonance is not yet fully established."

The gain compression factor is in the order of $\varepsilon_G = 2.5 \times 10^{-17} \ldots 3.1 \times 10^{-17}\,\mathrm{cm}^3$. In the case of an empty VB (very high and constant hole concentration $p = n_A$, i. e., $f_V \approx 0$ in the range of interest and complete inversion $n_{\mathrm{sp}} = 1$), the transparency concentration becomes zero, $n_t = 0$, because the slightest electron concentration in the CB already establishes some gain. This can be also seen from Eq. (3.35), (3.37). If we further neglect gain compression, $\varepsilon_G = 0$, a linear gain dependency follows,

$$G(n_T) = G_d\, n_T\,. \tag{3.78}$$

Rate equations

The operating characteristics of a semiconductor laser are well described by a set of rate equations that govern the interaction of photons and electrons inside the active region. A rigorous derivation starts from Maxwell's equations and includes in a semi-classical approach the quantum mechanical calculation of light-matter interaction. If spontaneous emission is to be included rigorously, quantum electrodynamics becomes involved where the optical field is quantized, too.

The rate equation can also be written heuristically by considering the phenomena through which the number N_P of photons and electrons $n_T V$ change with time inside the active volume V. We assume that the valence band is practically emptied of electrons. This is true if — as with GaAs and InP — the curvature of both the CB and VB, i. e., the reciprocal effective carrier masses in Eq. (3.13) and the effective DOS $N_{C,V}$ in Table 3.3 on Page 60, are significantly different, $m_n \ll m_p$ and $N_C \ll N_V$, see the discussion of Eq. (3.61) on Page 76. Therefore the hole concentration in the VB is large and virtually invariant, $\partial p/\partial n_T \approx 0$, so that $\tau_{\mathrm{sp}}^{-1} = \partial r_{\mathrm{sp}}/\partial n_T \approx Bp$ in Eq. (3.51) holds, and the spontaneous recombination rate $r_{\mathrm{sp}} = Bn_T p$ of Eq. (3.46) may be approximated by $r_{\mathrm{sp}} \approx n_T/\tau_{\mathrm{sp}}$. An equivalent procedure approximates the effective recombination rate r_{eff} from Eq. (3.48) by $r_{\mathrm{eff}} \approx n_T/\tau_{\mathrm{eff}}$. The result is

$$\begin{aligned} r_{\mathrm{sp}} &= Bn_T p \approx n_T/\tau_{\mathrm{sp}}\,, \\ r_{\mathrm{eff}} &= r_{\mathrm{sp}} + r_{\ell\mathrm{S}} + r_{\mathrm{Au}} \approx n_T/\tau_{\mathrm{eff}} \\ &\quad \text{if } N_C \ll N_V \text{ and therefore} \end{aligned} \qquad \begin{aligned} \tau_{\mathrm{sp}}^{-1} &= \partial r_{\mathrm{sp}}/\partial n_T \approx Bp\,, \\ \tau_{\ell\mathrm{S}}^{-1} &= \partial r_{\ell\mathrm{S}}/\partial n_T = A\,, \\ \tau_{\mathrm{Au}}^{-1} &= \partial r_{\mathrm{Au}}/\partial n_T \approx Cp^2, \\ \tau_{\mathrm{eff}}^{-1} &= \tau_{\mathrm{sp}}^{-1} + \tau_{\ell\mathrm{S}}^{-1} + \tau_{\mathrm{Au}}^{-1}\,. \end{aligned} \tag{3.79}$$

The lifetimes τ_{sp} and τ_{eff} depend on electron and hole concentrations, if the hole concentration p changes noticeably with the electron concentration n_T. For a longitudinally and laterally single-moded laser, the rate equations take the form[44]

$$\begin{aligned} \underbrace{\frac{\mathrm{d}N_P}{\mathrm{d}t}}_{\substack{\text{change of photon}\\ \text{number per time}}} &= + \underbrace{N_P \Gamma G(n_T, N_P)}_{\substack{\text{stimulatedly generated}\\ \text{photons per time}}} + \underbrace{Q\,\frac{n_T V}{\tau_{\mathrm{eff}}}}_{\substack{\text{spontaneously generated}\\ \text{photons per mode and time}}} - \underbrace{\frac{N_P}{\tau_P}}_{\substack{\text{stimulatedly depleted}\\ \text{photons per time}}}, \\ \underbrace{\frac{\mathrm{d}(n_T V)}{\mathrm{d}t}}_{\substack{\text{change of electron}\\ \text{number per time}}} &= - \underbrace{N_P \Gamma G(n_T, N_P)}_{\substack{\text{stimulatedly depleted}\\ \text{electrons per time}}} - \underbrace{\frac{n_T V}{\tau_{\mathrm{eff}}}}_{\substack{\text{spontaneoulsy depleted}\\ \text{electrons per time}}} + \underbrace{\frac{I}{e}}_{\substack{\text{injected electrons}\\ \text{per time}}}. \end{aligned} \tag{3.80}$$

The first equation (3.80) means in words: The number of photons N_P increases through photons which are generated by stimulated emissions with a net gain rate ΓG, and it increases through photons generated by spontaneous recombinations of electrons at a rate $1/\tau_{\mathrm{eff}}$, where only a fraction Q leads to spontaneous emissions into the mode under consideration. Further, the photon number decreases with a rate $1/\tau_P$ determined by the photon lifetime τ_P from Eq. (3.73).

The second equation (3.80) has to be read as follows: The number of electrons $n_T V$ in the active volume V decreases through electrons which recombine when stimulated by photons existing in the mode

[44] See Sect. 3.5.4 Eq. (3.116) on Page 183 in reference Footnote 47 on Page 89

under consideration, therefore a stimulated increase of the photon number corresponds to an equivalent decrease of the electron count. The number of carriers is further depleted with a rate $1/\tau_{\text{eff}}$ determined by the effective electron lifetime τ_{eff}, which takes into regard radiative (τ_{sp}) and nonradiative recombinations ($\tau_{\ell\text{S}}^{-1}$, τ_{Au}^{-1}) according to Eq. (3.51). Finally, the charge carrier number increases at a rate I/e fixed by the injection current I.

Because of stimulated ($N_P \Gamma G$) and spontaneous recombinations ($Q\, n_T V/\tau_{\text{eff}}$) the rate equations for the photon number N_P and for the CB carrier number $n_T V$ are nonlinear and coupled.

Electrons deplete spontaneously at a rate $1/\tau_{\text{eff}}$. The spontaneous emission factor Q tells how many spontaneous recombinations actually lead to photons being emitted into the mode under consideration. Taking into account the field confinement factor Γ from Eq. (3.66), (3.67), and looking at the translation relations Eq. (3.68), (3.69), the ratio of the spontaneous radiative recombination rate $\Gamma r_{\text{sp}}^{(\text{eM})}$ into one mode and the effective recombination rate r_{eff} defines Q. Using Eq. (3.44), (3.79) one finds[45]

$$Q = \frac{\Gamma r_{\text{sp}}^{(\text{eM})}}{r_{\text{eff}}} = \frac{\Gamma r_{\text{sp}}}{r_{\text{eff}}} \frac{\rho(f)}{\varrho_{\text{tot}}(f) V} = \Gamma \frac{\tau_{\text{eff}}}{\tau_{\text{sp}}} \frac{\rho(f)}{\varrho_{\text{tot}}(f) V}\,. \tag{3.81}$$

Lasing threshold Neglecting spontaneous emission ($Q = 0$) and assuming $N_P G \ll n_T V/\tau_{\text{eff}}$ we define the *lasing threshold* (*German* Schwelle, subscript S) using Eq. (3.80) for the case $\mathrm{d}\,/\,\mathrm{d}t = 0$ with τ_P from Eq. (3.73); above threshold the device starts oscillating as described in on Page 53,

$$\begin{aligned} \Gamma G(n_{TS}, 0) = \Gamma G_S &= \frac{1}{\tau_P} = v_g \Big[\alpha_V - \frac{\ln(R_1 R_2)}{2L}\Big], \\ \frac{I_S}{e} &= \frac{n_{TS} V}{\tau_{\text{eff}}} = r_{\text{eff}} V. \end{aligned} \tag{3.82}$$

It is for the threshold carrier concentration $n_T = n_{TS}$ that the net gain rate ΓG_S just compensates the loss rate, represented by the reciprocal photon lifetime $1/\tau_P$ of Eq. (3.73). Obviously, ΓG_S is larger than the net gain rate $G(n_T, N_P) = G(n_t, N_P) = 0$ where the material becomes transparent. Only above threshold the number of photons generated per time becomes larger than the number of photons annihilated. Excluding other loss mechanisms, the maximum photon lifetime τ_P is determined by the minimum mirror transmission losses for the given configuration. For $Q \neq 0$ the photon number N_P becomes already significant for $\Gamma G < \Gamma G_S = 1/\tau_P$, so that for $\mathrm{d}\,/\,\mathrm{d}t = 0$ the gain rate ΓG is always smaller than the idealized threshold gain ΓG_S as *defined* in Eq. (3.82). From Eq. (3.80) we see that

$$N_P = \frac{(Q/\tau_{\text{eff}}) n_T V}{1/\tau_P - \Gamma G} = \frac{\Gamma r_{\text{sp}}^{(\text{eM})} V}{1/\tau_P - \Gamma G} = \frac{r_{\text{sp}}^{(\text{eM})} V}{G_S - G}\,. \tag{3.83}$$

Threshold current The threshold current density $J_S = I_S/(bL)$ for the 5-layer structure Fig. 3.10(b) becomes minimum for a certain height d of the active layer, because the field confinement factor depends on d, $\Gamma = \Gamma(d)$. If d is small, only a small portion of the field interacts with the amplifying medium and the carrier concentration must be high. If d is large, the field is well confined inside the active layer, but only the region with the maximum field strength interacts efficiently with the population-inverted semiconductor, and again the carrier concentration must be high.

With Eq. (3.82) and g from Eq. (3.76), (3.77) we calculate the threshold current by eliminating n_{TS},

$$J_S = \frac{I_S}{bL} = \frac{e n_t}{\tau_{\text{eff}}} \left[d + \frac{\Gamma \alpha_V + \alpha_R}{g_0} \frac{d}{\Gamma(d)} \right]. \tag{3.84}$$

[45] Usually, the spontaneous emission factor is very small, $Q = 10^{-5} \ldots 10^{-4}$. This is true if the wave is guided by a difference of the real part of the refractive index (index guided laser, Fig. 3.5 on Page 58 and Fig. 3.25(b) on Page 92). However, if the waveguiding is dominated by the gain mechanism itself (gain guided laser, Fig. 3.25(a) on Page 92), the Q-factor is increased to $Q_e = K_e Q$, $K_e = 10 \ldots 20$. More details on these structures are given in Sect. 45 on Page 91 (see also Sect. 3.6.2 Page 191 in Ref. 47 on Page 89).

From the approximation $\Gamma(d) \approx 2V^2/(1+2V^2)$ in Eq. (3.67) and $V \sim d$ we see that $\Gamma \sim d^2$ for d small ($2V^2 \ll 1$), and $\Gamma \to 1$ for d, V large. Equation (3.84) has the structure ($c_{1,2} = \text{const}_d$)

$$J_S = c_1 d + \frac{c_2}{d} \qquad \text{for} \quad d < d_c = \sqrt{10}\,\frac{\lambda}{\sqrt{2}\,\pi}\Big/\sqrt{n^2 - n_2^2} = 0.71 \times \lambda \Big/ \sqrt{n^2 - n_2^2}\,, \tag{3.85}$$

so a minimum J_S is obvious. The term $c_1 d = (e n_t/\tau_{\text{eff}})d$ is the transparency current density for disappearing losses c_2/d. For a specific parameter set of a GaAs-(Ga,Al)As laser (g_0, n_t from Eq. (3.76), $\tau_{\text{eff}} = 1\,\text{ns}$, $\alpha_V + \alpha_R = 50\,\text{cm}^{-1}$, $\lambda = 0.87\,\mu\text{m}$, $n = 3.59$, $n_2 = 3.45$, $d_c = 0.62\,\mu\text{m}$) the minimum threshold current density is $J_S = 2.88\,\text{kA}\,\text{cm}^{-2}$ for an optimum layer thickness of $d = 0.07\,\mu\text{m}$. With $L = 300\,\mu\text{m}$, $b = 5\,\mu\text{m}$ the threshold current is $I_S = 43\,\text{mA}$.

The threshold current depends on temperature. The dependence on a temperature increase ΔT is well described by the empirical function

$$I_S(\Delta T) = I_S(0)\,\mathrm{e}^{\Delta T/T_0}, \quad \frac{1}{T_0} = \frac{1}{I_S(0)}\,\frac{\mathrm{d}I_S(\Delta T)}{\mathrm{d}\Delta T}\bigg|_{\Delta T = 0} = \alpha_{IS}\,. \tag{3.86}$$

The characteristic temperature T_0 has to be found from measurements. The temperature dependence comes from the strongly temperature-dependent distribution of carriers in the CB and VB, see Fig. 3.14(b). It is seen that with increasing T the net gain $r_{\text{ind}}^{(\text{M})}/N_P$ decreases. Further, the carrier confinement by the potential walls of the heterostructure becomes worse with increasing T. Because these walls are lower for the (In,Ga)(As,P) system, and because Auger recombinations become increasingly important with rising temperature, the temperature coefficient α_{IS} for InP is larger than for GaAs (InP: $T_0 = 40 \ldots 80\,\text{K}$. GaAs: $T_0 = 120 \ldots 230\,\text{K}$). Therefore the characteristic $P_a = P_a(I)$ Eq. (3.96) shifts with the temperature-dependent threshold.

Normalized rate equations The rate equations Eq. (3.80) can be normalized with the help of Eq. (3.82),

$$\begin{aligned} \tau_P\,\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{N_P/\tau_P}{I_S/e}\Big) &= \frac{N_P/\tau_P}{I_S/e}\Big[\frac{\Gamma G(n_T, N_P)}{1/\tau_P} - 1\Big] + Q\,\frac{n_T V}{n_{TS} V}\,, \\ \tau_{\text{eff}}(n_{TS})\,\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{n_T V}{n_{TS} V}\Big) &= \frac{I/e}{I_S/e} - \frac{n_T V}{n_{TS} V} - \frac{N_P/\tau_P}{I_S/e}\,\frac{\Gamma G(n_T, N_P)}{\Gamma G_S}\,. \end{aligned} \tag{3.87}$$

We define normalized quantities for the photon and carrier number, for the pump current and for the gain rate,

$$\begin{aligned} N_P^\times &= \frac{N_P/\tau_P}{I_S/e}\,, \qquad N_T^\times(n_T) = \frac{n_T}{n_{TS}}\,, \quad N_t^\times = \frac{n_t}{n_{TS}}\,, \quad \varepsilon_G^\times = \frac{\Gamma}{V}\,\frac{I_S}{e}\,\tau_P\,\varepsilon_G\,, \\ I^\times &= \frac{I}{I_S}\,, \qquad G^\times(N_T^\times, N_P^\times) = \frac{\Gamma G(n_T, N_P)}{1/\tau_P} = \frac{N_T^\times - N_t^\times}{1 - N_t^\times}\,\frac{1}{1 + \varepsilon_G^\times N_P^\times}\,. \end{aligned} \tag{3.88}$$

With the spontaneous emission factor Eq. (3.81), the normalized rate equations are written as

$$\begin{aligned} \tau_P\,\frac{\mathrm{d}N_P^\times}{\mathrm{d}t} &= N_P^\times(G^\times - 1) + Q N_T^\times\,, \\ \tau_{\text{eff}}\,\frac{\mathrm{d}N_T^\times}{\mathrm{d}t} &= I^\times - N_T^\times - N_P^\times G^\times. \end{aligned} \tag{3.89}$$

Characteristic curves For $\mathrm{d}/\mathrm{d}t = 0$ the rate equations (3.89) may be easily solved if spontaneous emission and gain compression is neglected, $Q = 0$ and $\varepsilon_G^\times = 0$. Below and at threshold $I^\times \le 1$ the photon number is zero, $N_P = 0$, and the first line of Eq. (3.89) is fulfilled for any $G^\times$. The carrier number increases with the current, $N_T^\times = I^\times$. Above threshold for $N_P^\times > 0$ the normalized gain is clamped to $G^\times = 1$, and so is the carrier number, $N_T^\times = 1$, as may be seen from Eq. (3.88). Therefore the photon

number increases according to $N_P^\times = I^\times - 1$. The clamped carrier density has the consequence that any residual dependency of $\tau_{\rm sp}$, $\tau_{\rm eff}$ and Q on the carrier concentration becomes essentially unimportant. In summary we have

$$\begin{aligned} &I^\times \le 1\text{:} \quad N_T^\times = I^\times, \quad N_P^\times = 0\,, \qquad\qquad Q = 0\,, \\ &I^\times > 1\text{:} \quad N_T^\times = 1\,, \qquad N_P^\times = I^\times - 1\,, \quad G^\times = 1\,, \end{aligned} \tag{3.90}$$

The normalized light output and the CB carrier density versus the normalized injection current, $N_P^\times = f(I^\times)$ and $N_T^\times = g(I^\times)$ are displayed in Fig. 3.20. If spontaneous emission into the lasing mode is important, $Q \neq 0$, the straight lines are "softened". This is to be seen in Fig. 3.20 where a simplified gain dependence $G^\times = N_T^\times$ according to Eq. (3.78) was assumed.

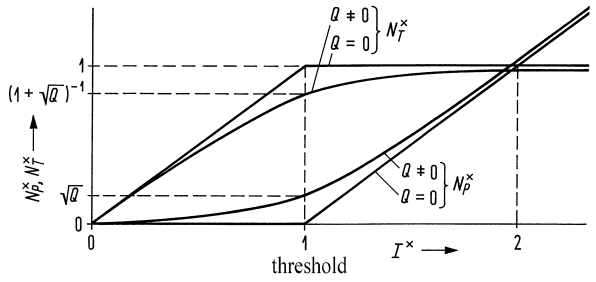


Fig. 3.20. Normalized photon number $N_P^\times$ and normalized CB carrier density $N_T^\times$ as a function of the normalized injection current $I^\times$. For $Q \neq 0$ a simplified gain dependence $G^\times = N_T^\times$ according to Eq. (3.78) is assumed.

Powers and Efficiencies

The totally generated and the total output power, respectively, are denoted as P and P_a. They are given by the total photon energy $N_P hf$ in the resonator volume per lifetime $\tau_{P,R}$ of the photons with respect to total losses (τ_P) or mirror transmittivity (τ_R),[46]

$$P = \frac{N_P hf}{\tau_P}\,, \qquad P_a = \frac{N_P hf}{\tau_R}\,. \tag{3.91}$$

Analogous to Eqs. (3.54), (3.56) one defines an internal and an external quantum efficiency for the laser diode by

$$\eta_{\rm int}^{\rm LD} = \frac{P/(hf)}{I/e}\,, \qquad \eta_{\rm ext}^{\rm LD} = \frac{P_a/(hf)}{I/e}\,, \tag{3.92}$$

From these relations and with Eq. (3.73) we write

$$\frac{\eta_{\rm ext}^{\rm LD}}{\eta_{\rm int}^{\rm LD}} = \frac{\tau_P}{\tau_R} = \frac{1}{1 - \frac{2\alpha_V L}{\ln(R_1 R_2)}}\,. \tag{3.93}$$

Through the slope of the characteristic curve $P_a = P_a(I)$ above threshold one defines a differential quantum efficiency

$$\eta_d = {\rm d}\left(\frac{P_a}{hf}\right) \Big/ {\rm d}\left(\frac{I}{e}\right)\,. \tag{3.94}$$

[46] If an electric heater consumes an energy of 2 kW h within a time of 1 h, then the power consumption of the device is obviously $P = 2\,{\rm kW\,h}/1\,{\rm h} = 2\,{\rm kW}$. The same is true for the laser resonator: We know the energy $N_P hf$ stored inside, and we know the time τ_R after which this energy has been lost through the mirrors (disregarding other loss mechnisms). Therefore, the power output from both mirrors is $P_a = N_P hf/\tau_R$, Eq. (3.91).

For an ideal laser diode with $Q = 0$, $\varepsilon_G^\times = 0$ and with the solution $N_P^\times = I^\times - 1$ Eq. (3.90) we find

$$\left.\begin{aligned} P &= \frac{hf}{e}(I - I_S), \\ P_a &= \frac{hf}{e}\frac{\tau_P}{\tau_R}(I - I_S), \end{aligned}\right\} \qquad \eta_d = \frac{\tau_P}{\tau_R} = \frac{\eta_{\text{ext}}^{\text{LD}}}{\eta_{\text{int}}^{\text{LD}}}\,. \tag{3.95}$$

For an actual laser one approximates the stimulated total power and the output power dependence, respectively, by

$$P = \eta_{\text{ind}}\frac{hf}{e}(I - I_S), \quad P_a = \eta_{\text{ind}}\frac{hf}{e}\frac{\tau_P}{\tau_R}(I - I_S). \tag{3.96}$$

The quantity η_{ind} is the efficiency for induced emission, indicating which percentage of the totally generated photons originated from stimulated emission acts. Because spontaneous emission goes into each of the resonator modes contained inside the linewidth Δf_H of the lineshape $\rho(f)$ Eq. (3.33), η_{ind} is of the order $N_P/(N_P + \varrho_{\text{tot}}(f)V\Delta f_H)$, where $\varrho_{\text{tot}}(f)V\Delta f_H$ is the number of relevant laser modes.From Eq. (3.96) the following interrelations may be derived,

$$\eta_d = \eta_{\text{ind}}\frac{\tau_P}{\tau_R} = \frac{\eta_{\text{ind}}}{1 - \frac{2\alpha_V L}{\ln(R_1 R_2)}} = \eta_{\text{ind}}\frac{\eta_{\text{ext}}^{\text{LD}}}{\eta_{\text{int}}^{\text{LD}}}\,. \tag{3.97}$$

By measuring the differential laser quantum efficiency with $R_1 = R_2 = R$ for varying resonator lengths L, the differential quantum efficiency η_d may be extrapolated for $L = 0$ and the efficiency η_{ind} Eq. (3.96) for induced emission may be determined. Typical differential efficiencies are in the range $\eta_d = 0.5 \ldots 0.8$, efficiencies for induced emission are found to be $\eta_{\text{ind}} = 0.65 \ldots 0.9$. The differential quantum efficiency is small for small reflection factors $R_{1,2}$ or for a small mirror lifetime τ_R Eq. (3.95). The more power is coupled out of the resonator, the higher the intensity modulation sensitivity $\mathrm{d}P_a/\,\mathrm{d}\,I = (hf/e)(\tau_P/\tau_R)$ will be. Typical output powers are in the range $P_a = 1 \ldots 10\,\text{mW}$ for communication lasers with current modulation, and $P_a = 100 \ldots 500\,\text{mW}$ for continuous wave applications.

The current-voltage characteristic of a laser diode is written as (saturation current I_{S0}, threshold current I_S)

$$\begin{aligned} I &= I_{S0}\Big[\exp[\beta(U - R_S I)] - 1\Big] && I \le I_S, \\ U &= W_G/e + R_S I && I \ge I_S. \end{aligned} \tag{3.98}$$

The CB carrier density is assumed to be clamped to the threshold value for $I > I_S$ thereby fixing the energetic difference of the quasi Fermi levels,

$$e(U - R_S I) = W_{Fn} - W_{Fp} \approx W_G\,. \tag{3.99}$$

The series resistance is in the order $R_S = 1 \ldots 10\,\Omega$. The quantity $\beta = 1/(\kappa U_T) = 15 \ldots 30\,\text{V}^{-1}$ is connected to the thermal voltage $U_T = kT_{\text{RT}}/e$ ($U_T = 25\,\text{mV}$ at room temperature $T_{\text{RT}} = 293\,\text{K}$). Typical values for the ideality factor κ are $\kappa = 1.3 \ldots 2.7$.

Small-signal intensity modulation To encode information into the laser beam, the optical output of the laser must be modulated. One of the unique attractions of a laser diode is the possibility of directly modulating the output light power by modulating the injection current. Because of amplitude-phase coupling, Sect. 45 on Page 89, this current modulation leads to a change in the laser mode frequency (chirping). For very high speed communications above bit rates of 10 Gbit/s the chirping of optical pulses can be avoided by employing a continuous wave laser diode and using an external modulator with a much better chirp behaviour, $|\alpha| \le 1$.

This section discusses analytical small-signal approximations of the highly nonlinear rate equations Eq. (3.80), and in the following section we demonstrate some large-signal properties by numerical solutions.

Perturbation ansatz We assume a static operation point above threshold given by the time-independent quantities N_{P0}, n_{T0}, $G_0 = G(n_{T0}, N_{P0})$, τ_P, τ_{eff}, ε_G in Eqs. (3.80), (3.90), and small time-dependent perturbations $N_{P1}(t)$, $n_{T1}(t)$, $I_1(t)$,

$$\begin{aligned} N_P(t) &= N_{P0} + N_{P1}(t), \quad G(t) = G_0 + \frac{\partial G_0}{\partial n_{T0}}\, n_{T1}(t) + \frac{\partial G_0}{\partial N_{P0}}\, N_{P1}(t), \\ n_T(t) &= n_{T0} + n_{T1}(t), \qquad G_0 = G(n_{T0}, N_{P0}), \qquad I(t) = I_0 + I_1(t). \end{aligned} \tag{3.100}$$

The differential gain rate $\partial G_0/\partial n_{T0}$ has typical values of $1.8 \times 10^{-6} \ldots 2.9 \times 10^{-6}\,\text{cm}^3\,\text{s}^{-1}$. Substituting Eq. (3.100) into Eq. (3.80) and neglecting products of perturbation quantities, we solve the linearized rate equations with a Fourier ansatz $X_1(t) = X_1(\omega)\exp(\mathrm{j}\,\omega t)$, where $X_1(\omega)$ is the complex amplitude at the modulation frequency $f = \omega/(2\pi)$,

$$\begin{aligned} N_{P1}(\omega)\Big(\mathrm{j}\,\omega + \frac{1}{\tau_P} - \frac{\Gamma G_0}{1 + \varepsilon_G \frac{\Gamma N_{P0}}{V}}\Big) &= \Big(\frac{Q}{\tau_{\text{eff}}} + \frac{N_{P0}}{V}\,\frac{\partial \Gamma G_0}{\partial n_{T0}}\Big) n_{T1}(\omega) V, \\ n_{T1}(\omega) V \Big(\mathrm{j}\,\omega + \frac{1}{\tau_{\text{eff}}} + \frac{N_{P0}}{V}\,\frac{\partial \Gamma G_0}{\partial n_{T0}}\Big) &= \frac{I_1(\omega)}{e} - \frac{\Gamma G_0}{1 + \varepsilon_G \frac{\Gamma N_{P0}}{V}}\, N_{P1}(\omega)\,. \end{aligned} \tag{3.101}$$

Elimination of $n_{T1}(\omega)$ leads to the modulation transfer function

$$\frac{N_{P1}(\omega)}{I_1(\omega)} = \underbrace{\frac{\left(\frac{N_{P0}}{V}\,\frac{\partial \Gamma G_0}{\partial n_{T0}} + \frac{Q}{\tau_{\text{eff}}}\right)}{\omega_r^2 \tau_P}}_{\approx 1} \frac{\omega_r^2}{(\mathrm{j}\,\omega)^2 + 2\gamma_r(\mathrm{j}\,\omega) + \omega_r^2} \tag{3.102}$$

with the angular relaxation frequency ω_r and the damping constant γ_r,

$$\omega_r^2 \tau_P = \frac{N_{P0}}{V}\,\frac{\partial \Gamma G_0}{\partial n_{T0}} + \underbrace{\frac{\overbrace{\tau_P}^{\ll 1}}{\tau_{\text{eff}}}\Big(\underbrace{\frac{1}{\tau_P} - \Gamma G_0}_{\approx 0}\,\frac{\overbrace{1 - Q}^{\approx 1}}{1 + \varepsilon_G \frac{\Gamma N_{P0}}{V}}\Big)}_{\approx 0} \approx \frac{N_{P0}}{V}\,\frac{\partial \Gamma G_0}{\partial n_{T0}}\,,$$

$$\begin{aligned} 2\gamma_r &= \frac{1}{\tau_{\text{eff}}} + \frac{N_{P0}}{V}\,\frac{\partial \Gamma G_0}{\partial n_{T0}} + \frac{1}{\tau_P}\Big(1 - \overbrace{\Gamma G_0 \tau_P}^{\approx 1}\,\frac{\overbrace{1}^{\approx 1 - \varepsilon_G \frac{\Gamma N_{P0}}{V}}}{1 + \varepsilon_G \frac{\Gamma N_{P0}}{V}}\Big) \\ &\approx \frac{1}{\tau_{\text{eff}}} + \underbrace{\frac{N_{P0}}{V}\,\frac{\partial \Gamma G_0}{\partial n_{T0}} + \Gamma G_0\, \varepsilon_G \frac{\Gamma N_{P0}}{V}}_{= \omega_r^2 K_r}\,. \end{aligned} \tag{3.103}$$

For large photon numbers N_{P0} the relaxation frequency and the damping constant are determined by the differential gain $\partial \Gamma G_0/\partial n_{T0}$. For low photons numbers the effective electron lifetime τ_{eff} controls the damping. Inside the laser two reservoirs exchange their energy: the electromagnetic energy stored in the resonator, and the electronic energy states. An increase in photon number reduces the carrier number, and vice versa.

This relaxation oscillation dies out with a damping constant γ_r, because photons and electrons have finite lifetimes τ_P and τ_{eff}, respectively. Such an energy exchange may be compared to the behaviour of a damped resonance circuit, where the inductor as a magnetic energy store and the capacitor storing the electric field energy are interacting. By a sudden perturbation (e. g., by a injection current step) damped oscillations $\exp(\mathrm{j}\,\omega t)$ are excited. Values of $\mathrm{j}\,\omega$ are defined by the zeros of the denominator in Eq. (3.102):

$$\mathrm{j}\,\omega = \begin{cases} -\gamma_r \pm \mathrm{j}\,\sqrt{\omega_r^2 - \gamma_r^2} & \text{damped oscillation,} \quad \gamma_r/\omega_r < 1\,, \\ -\gamma_r & \text{aperiodic limit,} \quad \gamma_r/\omega_r = 1\,, \\ -\gamma_r \pm \sqrt{\gamma_r^2 - \omega_r^2} & \text{aperiodic behaviour,} \quad \gamma_r/\omega_r > 1\,. \end{cases} \tag{3.104}$$

In the aperiodic region the slower decaying term describes the characteristic time dependence. The angular frequency of the free oscillation is $\sqrt{\omega_r^2 - \gamma_r^2}$. The aperiodic limiting case can be reached for very large photon numbers N_{P0}. For a constant current amplitude $|I_1(\omega)| = I_1 = \text{const}_\omega$ the modulation transfer function Eq. (3.102) shows a maximum at the small-signal resonance angular frequency ω_R,

$$\left|\frac{N_{P1}(\omega)/\tau_P}{I_1/e}\right| \;\rightarrow\; \max: \qquad \omega_R = \sqrt{\omega_r^2 - 2\gamma_r^2}. \tag{3.105}$$

A resonance is only possible for $\gamma_r/\omega_r < 1/\sqrt{2}$. The 3-dB bandwidth (no ripple nor resonance wanted!) is defined by

$$\left|\frac{N_{P1}(\omega_{3\,\mathrm{dB}})}{N_{P1}(0)}\right| = \frac{1}{\sqrt{2}}\,, \quad \omega_{3\,\mathrm{dB}}^2 = (\omega_r^2 - 2\gamma_r^2) + \sqrt{(\omega_r^2 - 2\gamma_r^2)^2 + \omega_r^4}\,. \tag{3.106}$$

Figure 3.21 displays the small-signal current-light transfer function as a function of the normalized current

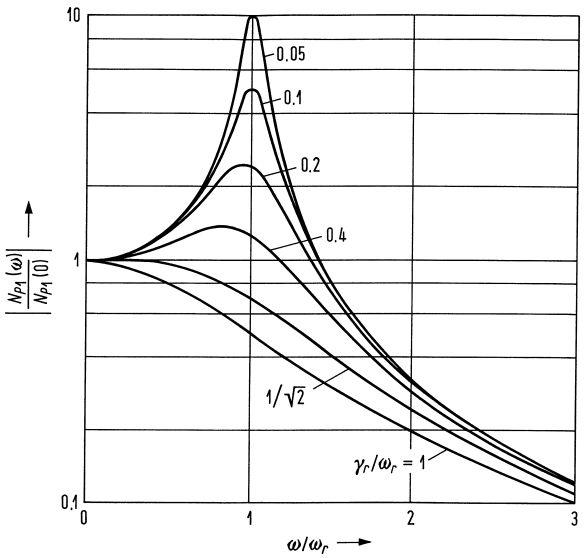


Fig. 3.21. Modulus of current-light modulation transfer function as a function of normalized current modulation frequency for various values of γ_r/ω_r

modulation frequency. The resonance overshoot disappears for $\gamma_r/\omega_r > 1/\sqrt{2}$ (overcritical damping). Because of $\omega_r^2 \sim N_{P0}$, $\gamma_r^2 \sim N_{P0}^2$ the resonance overshoot becomes smaller with increasing operating current. The 3-dB bandwidth becomes larger because ω_r increases. However, γ_r increases faster than ω_r, and for critical damping $\gamma_r/\omega_r = 1/\sqrt{2}$ and large photon numbers the 3-dB modulation bandwidth becomes maximum,

$$\begin{aligned} \omega_{3\,\mathrm{dB}}^{\max} &= \omega_r\,, \qquad \omega_r = \gamma_r\sqrt{2} \approx \frac{\sqrt{2}}{2}\,\omega_r^2\Big(\tau_P + \frac{\varepsilon_G}{\partial G_0/\partial n_{T0}}\Big), \\ \omega_{3\,\mathrm{dB}}^{\max} &= \frac{\sqrt{2}}{K_r}\,, \qquad\qquad K_r = \tau_P + \frac{\varepsilon_G}{\partial G_0/\partial n_{T0}}\,. \end{aligned} \tag{3.107}$$

This is the intrinsic maximum modulation bandwidth. Often it is not possible to operate the laser at this operating point because of temperature limitations or because of the onset of multimode operation. In the region of overcritical damping $\gamma_r/\omega_r > 1/\sqrt{2}$ the modulation bandwidth starts to decrease, Eq. (3.106). The relaxation frequency could be enlarged by decreasing the photon lifetime τ_P (e. g., by reducing the

resonator length L), but this broadens the spectrum, so that only the differential gain $\partial G_0/\partial n_{T0}$ may be manipulated for highest $\omega_{3\,\mathrm{dB}}^{\max}$. This can be achieved by a p-doping of the active layer, see also Page 64. The differential gain $\partial G_0/\partial n_{T0}$ becomes larger by a factor 5 if n_A is increased from $5 \times 10^{16}\,\mathrm{cm}^{-3}$ to $5 \times 10^{18}\,\mathrm{cm}^{-3}$. By using quantum film or quantum wire lasers the differential gain becomes even larger. Relaxation frequencies of 38 GHz could be achieved.

Figure 3.22 shows the electric equivalent circuit if the laser diode. The "internal laser diode" may be regarded as a short circuit because of the voltage clamping, Eq. (3.99). R_S is the series resistance, C_P a parasitic parallel capacitance (in Fig. 3.25(b) parasitic pn-junctions in parallel to the active layer), L_S is the bond wire inductance, typically 1 nH / mm, R_G the generator impedance. The cutoff frequencies defined by $\omega R_S C_P = 1$ and $\omega L_S = R_G$ should not impair the intrinsic modulation capability of the device. For a cutoff frequency of 20 GHz at $R_S = 10\,\Omega$, $R_G = 50\,\Omega$ the values $C_P < 0.8\,\mathrm{pF}$, $L_S < 0.4\,\mathrm{nH}$ should be chosen.

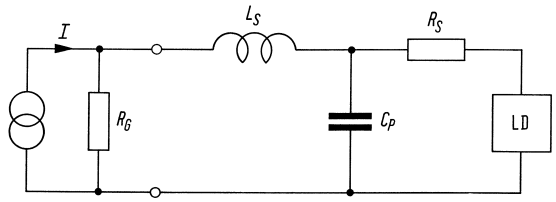


Fig. 3.22. Electric equivalent circuit of the laser diode. The internal laser diode may be regarded as a short circuit.

Large-signal intensity modulation A general analytic solution of the nonlinear rate equations (3.80) is not known, so a specific numerical solution of the simplified normalized rate equations will be discussed, Fig. 3.23. During the delay time t_d the normalized carrier density rises to $N_T^\times = 1$. When the threshold

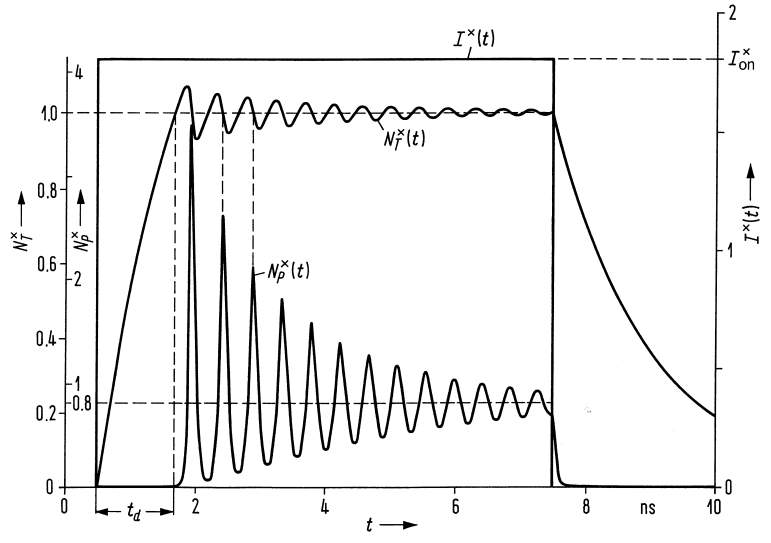


Fig. 3.23. Relaxation oscillation for a current step $I^\times = 1.8$. Parameters are $\tau_P = 2.5\,\mathrm{ps}$, $\tau_{\mathrm{eff}} = 1.5\,\mathrm{ns}$, $Q = 5 \times 10^{-4}$

is reached the photon number $N_P^\times$ starts rising. At first, the carrier density increases further, but when the photon number has become large enough and the stimulated emissions are numerous, the carrier density decreases. As long as $N_T^\times > 1$ holds, the photon number increases further. When the condition $N_T^\times < 1$ is met, the gain rate becomes negative and the photon number decreases rapidly. Therefore the

number of stimulated emissions decreases and the carrier number can be re-established by the injection current.. When the threshold $N_T^\times = 1$ is exceeded, the photon number increases again. A weakly damped relaxation oscillation results for the parameters chosen in Fig. 3.23, which has not yet died out when the current impulse is switched off. A small variation of the carrier density $N_T^\times$ causes very large changes in the photon number $N_P^\times$. When the current is switched on the carrier density starts with a finite slope while the photon number has with a zero slope.

To calculate the delay time t_d when switching from $I_{\text{off}} < I_S$ to $I_{\text{on}} > I_S$ the induced emission term in the second rate equation (3.80) may be neglected. The solution of the resulting equation

$$\frac{\mathrm{d}n_T}{\mathrm{d}t} = \frac{I}{eV} - \frac{n_T}{\tau_{\text{eff}}} \tag{3.108}$$

for the initial condition $I_{\text{off}} < I_S$ for $t < 0$, $I_{\text{on}} > I_S$ for $t > 0$ is

$$n_T(t) = \frac{\tau_{\text{eff}}}{eV}\Big[I_{\text{off}} + (I_{\text{on}} - I_{\text{off}})\left(1 - \mathrm{e}^{-t/\tau_{\text{eff}}}\right)\Big]. \tag{3.109}$$

After the delay time t_d the threshold carrier density is reached,

$$n_T(t_d) = n_{TS} = \frac{\tau_{\text{eff}} I_S}{eV}\,, \tag{3.110}$$

from which the delay time t_d results,

$$t_d = \tau_{\text{eff}} \ln \frac{I_{\text{on}} - I_{\text{off}}}{I_{\text{on}} - I_S}\,, \qquad I_S \sim \frac{1}{\tau_{\text{eff}}}\,, \quad \text{see Eq. (3.84).} \tag{3.111}$$

Thus, the effective carrier lifetime may be measured from the switch-on delay time t_d. For actually modulating the laser, a bias current I_{off} at or slightly above threshold I_S is used.

Amplitude-phase coupling

Because of Eq. (3.40) the gain rate G is a function of the frequency f and (via the quasi Fermi levels) a function of the carrier concentration n_T, Eq. (3.20). This is also true for the modal power gain g and consequently for the imaginary part of the refractive index $-n_i$, Eq. (3.70). The real part n of the complex refractive index $\bar{n}$ depends on f, n_T because of three reasons:

Band filling With the carrier injection the band-band absorption decreases because of the filling of lower CB states, and the absorption energy $hf_1 = W_G + \Delta W_1$ increases (ΔW_1 increasing with n_T). Therefore, $\Delta n < 0$ for $f < f_1$ and $f > f_1$ according to the Kramers-Kronig relations[47], where f is *outside* the region of anomalous dispersion $\mathrm{d}n/\,\mathrm{d}f < 0$. For InP typical data are $\Delta n = -7.7 \times 10^{-21} n_T/\,\mathrm{cm}^{-3}$ at $\lambda = 1.24\,\mu\mathrm{m}$, $\Delta n = -5.6 \times 10^{-21} n_T/\,\mathrm{cm}^{-3}$ at $\lambda = 1.55\,\mu\mathrm{m}$.

Coulomb interaction The interaction of carriers by coulomb forces reduces the bandgap, and the absorption increases, especially in the vicinity of the bandgap energy $hf_2 = W_G$. Therefore, $\Delta n > 0$ for $f < f_2$ and for $f > f_2$ result.

Free carrier The free-carrier absorption causes always $\Delta n < 0$ at optical frequencies.

$$\Delta n = -\frac{e^2 \mu_0 \lambda^2}{8\pi^2 n}\Big(\frac{n_T}{m_n} + \frac{p}{m_p}\Big) = -\frac{4.485 \times 10^{-22}}{n}\Big(\frac{\lambda}{\mu\mathrm{m}}\Big)^2\Big[\frac{n_T/\,\mathrm{cm}^{-3}}{m_n/m_0} + \frac{p/\,\mathrm{cm}^{-3}}{m_p/m_0}\Big]. \tag{3.112}$$

The ratios $m_n/m_0, m_p/m_0$ are given in Table 3.3 on Page 60.

[47] Grau, G.; Freude, W.: Optische Nachrichtentechnik (Optical communications, in German), 3. Ed. Berlin: Springer-Verlag 1991. Since 1997 out of print. Corrected KIT reprint 2005 in electronic form available from W. F. (`w.freude@kit.edu`). Sect. 2.1.1 Page 13 ff., Appendix B Page 371 ff.

At the oscillation frequency of a laser diode the combination of these effects leads to a reduced refractive index n ($\Delta n < 0$) for increasing CB carrier density n_T, and (via the Kramers-Kronig relations[48]) to an increased gain constant g. With the help of Eq. (3.70) one defines a quantity α (α-factor, line broadening factor, Henry factor)

$$\alpha = \frac{\partial n/\partial n_T}{\partial n_i/\partial n_T} = -2k_0 \frac{\partial n/\partial n_T}{\partial(\Gamma g - \alpha_V)/\partial n_T} = -2k_0 \frac{\partial n/\partial n_T}{\partial(\Gamma g)/\partial n_T} > 0. \tag{3.113}$$

The last form of Eq. (3.113) assumes the loss constant α_V to be independent of the carrier density n_T. Typical values for laser diodes are in the range $\alpha = 2 \ldots 8$. Therefore a correlation exists between amplitude and phase of the laser diode oscillator. Spontaneous emissions cause amplitude and phase changes. Because of Eq. (3.113) such an amplitude change gives rise to a secondary phase change, so that a broadening of the emission line is to be expected.

For a stationary laser oscillation the operating point (subscript 0) is given by $G(n_{T0}) = 1/\tau_P$. This is to be seen from Eq. (3.90) and (3.88) where $G^\times = \Gamma G(n_{T0}, N_{P0})\,\tau_P = 1$ may be deduced. The angular optical frequency is ω_0, Eq. (3.63). When changing the carrier density differentially, the gain rate G varies, and the "instantaneous" (on the scale of an optical period $1/f_0$ slowly varying) optical frequency ω deviates from its unperturbed value by a small amount $\mathrm{d}\omega$. This frequency difference $\Delta\omega$ defines the time derivative of the optical phase, $\mathrm{d}\varphi/\,\mathrm{d}t = \overbrace{n_g}\Delta\omega$. From Eq. (3.63) we find

$$\begin{aligned}
\mathrm{d}(\omega n) &= \frac{\partial(\omega n)}{\partial\omega}\,\mathrm{d}\omega + \frac{\partial(\omega n)}{\partial n}\,\mathrm{d}n = \Big(\overbrace{n + \omega\,\frac{\partial n}{\partial\omega}}\Big)\,\mathrm{d}\omega + \omega\,\mathrm{d}n \stackrel{!}{=} 0\,,\\
\mathrm{d}\omega &= -\frac{\omega}{n_g}\,\mathrm{d}n = -\frac{\omega}{n_g}\frac{\partial n}{\partial n_T}\,\mathrm{d}n_T = -\frac{\alpha\omega}{2k_0 n_g}\frac{\partial(\Gamma g)}{\partial n_T}\,\mathrm{d}n_T \approx \Delta\omega = \omega - \omega_0 = \frac{\mathrm{d}\varphi}{\mathrm{d}t}\,,\\
\frac{\mathrm{d}\varphi}{\mathrm{d}t} &= \frac{\alpha}{2} v_g \frac{\partial(\Gamma g)}{\partial n_T}\Delta n_T \approx \frac{\alpha}{2}\frac{\partial(\Gamma G)}{\partial n_T}\Delta n_T\,.
\end{aligned} \tag{3.114}$$

However, with $\mathrm{d}v_g/\,\mathrm{d}n_T, \mathrm{d}n_g/\,\mathrm{d}n_T \neq 0$ the last form of Eq. (3.114) is only approximately valid,

$$\frac{\partial(\Gamma G)}{\partial n_T} = \frac{\partial(\Gamma v_g g)}{\partial n_T} = v_g \frac{\partial(\Gamma g)}{\partial n_T}\left(1 - \frac{\frac{1}{n_g}\frac{\partial n_g}{\partial n_T}}{\frac{1}{\Gamma g}\frac{\partial(\Gamma g)}{\partial n_T}}\right) \approx v_g \frac{\partial(\Gamma g)}{\partial n_T}\,. \tag{3.115}$$

Because of Eqs. (3.76), (3.113) ($\alpha = 2 \ldots 8$) the ratios of the relative change of n_g, g with n_T are smaller than 10^{-2}. Therefore, Eq. (3.114) may be written with reference to Eqs. (3.71), (3.74) and neglecting spontaneous emission as

$$\frac{\mathrm{d}\varphi}{\mathrm{d}t} = \frac{\alpha}{2}\Big(\underbrace{\frac{\partial(\Gamma G)}{\partial n_T}\Delta n_T + \overbrace{\Gamma G(n_{T0})}^{=0}}_{\Gamma G(n_T)} - \frac{1}{\tau_P}\Big) = \frac{\alpha}{2}\Big(\Gamma G - \frac{1}{\tau_P}\Big) = \frac{\alpha}{2}\frac{1}{N_P}\frac{\mathrm{d}N_P}{\mathrm{d}t}\,. \tag{3.116}$$

For the unperturbed stationary oscillation we have $\omega = \omega_0$, $\mathrm{d}\omega = \mathrm{d}\varphi/\,\mathrm{d}t = 0$. A perturbation in the carrier density $\mathrm{d}n_T \neq 0$ leads to a change in optical angular frequency, Eq. (3.114). The phase change by amplitude-phase coupling may be incorporated into the basic equations (3.80). The relation for the photon number is supplemented by an equation for the phase. Neglecting again spontaneously emitted photons we find for the analytic electric field $\underline{E}(t)$ at the laser mirrors

$$\frac{\mathrm{d}N_P}{\mathrm{d}t} = N_P\Big(\Gamma G - \frac{1}{\tau_P}\Big), \quad \frac{\mathrm{d}\varphi}{\mathrm{d}t} = \frac{\alpha}{2}\Big(\Gamma G - \frac{1}{\tau_P}\Big), \quad \underline{E}(t) \sim \sqrt{N_P(t)}\;\mathrm{e}^{\mathrm{j}[\omega_0 t + \varphi(t)]}\,. \tag{3.117}$$

For a measurement of α we derive the oscillation mode frequency dependence from Eq. (3.114),

$$\frac{\mathrm{d}\omega_0}{\mathrm{d}n_T} = -\frac{\omega_0}{n_g}\frac{\partial n}{\partial n_T} \approx \frac{\alpha}{2}\frac{\partial(\Gamma G)}{\partial n_T}\,, \qquad \alpha = 2\,\frac{\mathrm{d}\omega_0}{\mathrm{d}n_T}\Big/\frac{\partial(\Gamma G)}{\partial n_T}\,. \tag{3.118}$$

By a measuring of the (effective) gain rate change and of the shift in angular resonance frequency ω_0 with the carrier density n_T the a-factor may be calculated.

[48] See Reference 47 on Page 89. Sect. 2.1.1 Page 13 ff., Appendix B Page 371 ff.

LD spectrum

When the injection current is below threshold, the laser diode behaves like an LED and the output is mainly due to spontaneous emission and, hence, the spectrum is broad. As the current increases beyond threshold, the longitudinal modes having a larger gain and a smaller resonator loss begin to oscillate,

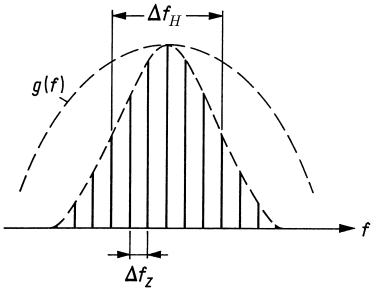


Fig. 3.24. Longitudinal mode spectrum of a gain-guided Fabry-Perot laser diode (see Page 91). Frequency-dependent net gain constant $g(f) \sim r_{\text{ind}}^{(\text{M})}(f)/N_P$ Fig. 3.14(b), longitudinal mode distance Δf_z Eq. (3.65), half-power bandwidth of spectral envelope Δf_H

and the spectrum changes significantly. As the current is further increased, the output spectrum becomes spectrally more concentrated, Fig. 3.24 for a gain-guided Fabry-Perot laser diode (see Footnote 45 on Page 82 and Sect. 45 on Page 91). The longitudinal Fabry-Perot resonator modes are separated by the spectral distance Δf_z, Eq. (3.65) on Page 78. The half-power width Δf_H of the spectral envelope is given by (α_V is the modal or effective loss)

$$\Delta f_H \, P_a = \text{const} \times n_{\text{sp}}(1+\alpha^2) h f v_g^2 (\alpha_V + \alpha_R)\alpha_R \,. \tag{3.119}$$

The quantity Δf_H of the laser oscillator cannot be compared to the spontaneous linewidths Eq. (3.33), (3.61) on Pages 68, 76. It is proportional to the reciprocal of the total output power P_a through both mirrors. Typical values for GaAs gain-guided lasers are $\Delta f_H \, P_a = 3\,000\,\text{GHz}\,\text{mW}$, i. e., with $P_a = 1\,\text{mW}$ and $\Delta f_z = 100\,\text{GHz}$ about 30 modes oscillate simultaneously, while at $P_a = 10\,\text{mW}$ there remain only 3 modes. For index-guided lasers even single-mode oscillation may be achieved.

Devices

As illustrated in Figs. 3.1, 3.5, a simple laser resonator (Fabry-Perot resonator) is a rectangular cavity with six walls, all of which should provide good photon and carrier confinement to reduce the cavity loss. Among the six walls, two are at the longitudinal ends of the cavity ($z = 0, L$) which need to couple light out, and two are the heterojunctions of a 3 or 5-layer heterostructure ($x = \pm d/2$) which achieve both carrier and photon confinement from the energy bandgap and refractive index differences, respectively, Fig. 3.10. To provide the confinement at the two transverse sides ($y = \pm b/2$), two basically different structures have been used, namely *gain-guided* and *index-guided* lasers, Fig. 3.25.

Gain-guided lasers A gain-guided laser, Fig. 3.25(a), has a structure that confines the transverse current flow. There is no physical confinement for photons on the two sides, but the field is concentrated near the z-axis, because $g - \alpha_V = -2k_0 n_i$ (see Eq. (3.70)) has an on-axis maximum and decreases with increasing $|y|$. This profile is due to a lateral decrease of the current density as in Fig. 3.25(a) resulting in a corresponding reduction of g, but it is also possible to increase the lateral loss α_V by moving absorbing regions nearer to the active zone, dashed line. Naturally, the (effective) real part n of the complex refractive index depends on y, too. The high-current region has a lower refractive index causing even an antiguiding effect. For a gain-guided laser the waveguiding by the lateral decrease of n_i dominates. Because gain-guided lasers have no strong transverse photon confinement, they have a relatively large threshold current in the order of $I_S = 100\,\text{mA}$.

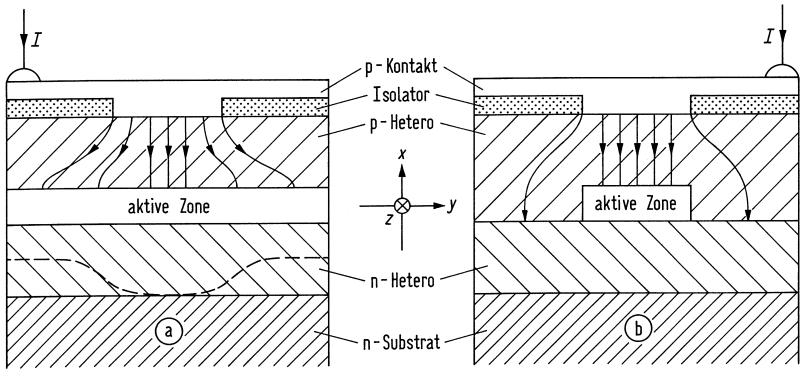


Fig. 3.25. Basic laser diode structures. (a) Gain-guided laser (b) Index-guided laser. The origin of the coordinate system is located in the centre of the active zones (p-Kontakt = p-contact, Isolator = insulator, aktive Zone = active zone).

As mentioned[49] when discussion the rate equations on Page 81 ff., a spontaneous-emission correction factor $K_e = 10 \ldots 20$ has to be introduced for gain-guided lasers. This comes from the non-orthogonality of the gain-guided modes.

Index-guided laser To provide a better photon confinement, index difference must be introduced on the two transverse sides $y = \pm b/2$. Laser diodes that add the index difference are called index-guided lasers, Fig. 3.25(b). Even without a current the strip waveguide cavity is defined, buried into the material with larger bandgap, i. e., lower refractive index. Such a buried-heterostructure laser introduces another two heterojunctions on the lateral sides (total four heterojunctions) to provide both carrier and photon confinement. Because of the excellent confinement, the threshold current can be as low as $I_S = 10\,\mathrm{mA}$.

Vertical cavity surface emitting laser *V*ertical *c*avity *s*urface *e*mitting *l*asers (VCSEL, pronounced ['vɪksəl]) are semiconductor lasers which emit perpendicularly to their pn-junction plane in a manner analogous to that of a surface-emitting LED, and feature circular, low-divergence beams. This new class of lasers emerged during the 1990s[50,51,52]. VCSEL operate in a single longitudinal mode due to an extremely small cavity length $L = \lambda_e/2 \approx 1\,\mu\mathrm{m}$. The mode spacing $\Delta f_z = c/(n_g\lambda_e) = c/\lambda \approx 300\,\mathrm{THz}$ for $\lambda = 1\,\mu\mathrm{m}$ (Eq. (3.65) on Page 78) exceeds the gain bandwidth $\Delta f_H \approx 12\,\mathrm{THz}$ by far (Eq. (3.61) on Page 3.61).

The properties of VCSEL are attractive for many purposes. Traditional edge-emitting diode lasers only partially fulfill these requirements. Such lasers have elliptical, divergent beams which must be optically corrected in order to collimate the beam even over short distances. Furthermore it is often necessary to isolate the laser from back-reflection of the emitted light into the resonator, which can lead to changes in the laser's output characteristics. The divergent, elliptical beam is a result of diffraction at the rectangular emission area of a conventional diode laser resonator. The divergence varies inversely with the size of this emission area. Consequently the beam divergence perpendicular to the junction is significantly greater than parallel to it, see Fig. 3.26(a).

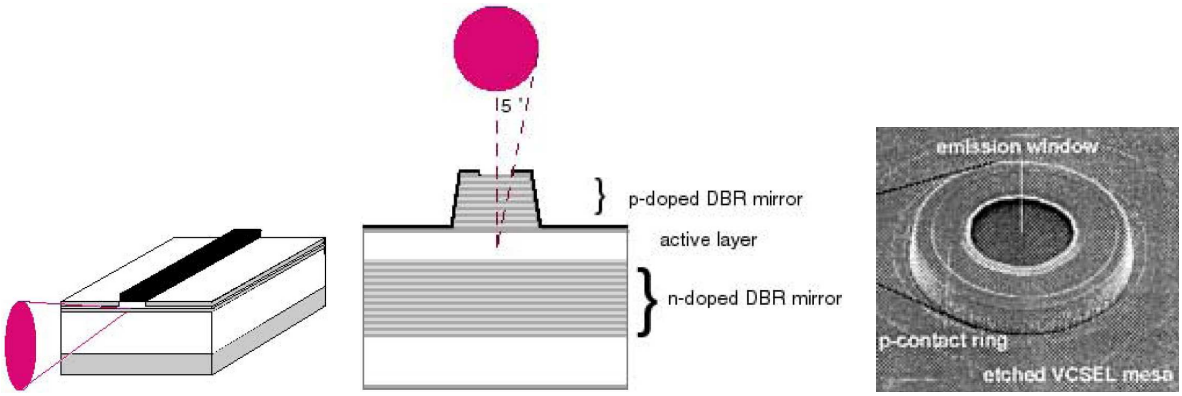
In order to fully control the emitted beam profile it is necessary to define the geometry of the emission area. This is only possible with a vertical resonator, perpendicular to the p-n junction plane, since the active layer is parallel to the emission area, Fig. 3.26(b). To obtain a round, low-divergence beam, a circular output complex may be applied to the entire emitting area. In the case of a vertical resonator the length of the gain medium is defined by the thickness of the pn-junction. In order to achieve laser

[49] See Footnote 45 on Page 82

[50] See Sect. 5.2.4 Page 191–192 in Ref. 3 on Page 49

[51] Li, H.; Iga, K.: Vertical-cavity surface-emitting laser devices. New York: Springer 2001

[52] Laser Components (UK) Ltd. Goldlay House, 114 Parkway, Chelmsford, Essex. CM2 7PR UK. `http://www.lasercomponents.co.uk`



(a) (b) (c)

Fig. 3.26. Edge-emitting and vertically-emitting laser diodes (a) edge-emitting laser diode and far-field radiation characteristic (b) VCSEL layer structure. p-doped DBR mirror: 25 layers $Al_{0.3}Ga_{0.7}As/AlAs$; active zone: 220 nm $Al_{0.3}Ga_{0.7}As$ with 3 $Al_{0.12}Ga_{0.88}As$ quantum films, height about 7 nm each; n-doped DBR mirror: 40 layers $Al_{0.3}Ga_{0.7}As/AlAs$ (c) microscopic image of VCSEL (all after Ref. 52 on Page 92)

operation, this must be compensated by the use of highly efficient distributed Bragg reflectors as resonator mirrors. The lower resonator mirror is made up of 40 alternating layers of $Al_xGa_{1-x}As$ and AlAs, each layer $\lambda_e/4$ thick, which give a power reflection factor in excess of $R_1 = 99.99\,\%$. The output top mirror, with twenty-five $\lambda_e/4$ layers, has a reflectivity of $R_2 = 99.9\,\%$. The high resonator efficiency and small gain medium volume combine to give threshold currents of only a few mA, which means that the low operating currents make VCSEL suitable for any application.

A further advantage of this mirror design is that any light with a different wavelength, reflected back towards the laser from another part of the optical system, cannot re-enter the resonator. VCSEL are effectively isolated against such reflections due to their DBR output coupler structure. The resonator design also means that the laser emission is single-mode. The desired wavelength is achieved by using the correct layer structure.

VCSEL may feature a special mesa structure which ensures that all additional transverse modes are suppressed. The result of this is predominantly single-mode emission with a correspondingly narrow emission linewidth ($< 50\,\text{MHz}$ at $\lambda = 780\,\text{nm}$, for example). This mode will tune within the gain profile as the laser temperature is varied, which allows VCSEL to be temperature-tuned over a range of several nanometers, without mode hopping. The typical tuning rate, about $0.4\,\text{nm}\,/\,\text{mA}$, is significantly higher than in an edge emitter. The vertical structure has a further significant advantage: the laser may be tested before bonding and whilst still on the wafer. Furthermore, VCSEL may be easily incorporated into electronic integrated circuits and are also highly suitable for use in 2-dimensional arrays.

Popular VCSEL are offered[53] with wavelengths in the range $760 \ldots 960\,\text{nm}$. single-mode output powers are typically $0.3 \ldots 0.5\,\text{mW}$, with 5 mW multimode versions also available.

Comparison between edge-emitting lasers and VCSEL The technology of VCSEL is showing great promise as a low cost alternative in new applications such as *v*ery *s*hort *r*each (VSR) transceivers, tunable and high power pump lasers. Consequently, the adoption of VCSEL-based optoelectronics could be the ultimate cost reduction vehicle carriers are searching for[54].

Although advances continue in edge emitter based technology, it seems evident that it just may be a case of diminishing returns. Edge emitter technology was a key enabler of optical communication as we know it today. First used in short reach, single channel applications, the edge emitter evolved to support DWDM and long haul transmission. With the introduction of EDFA, edge emitters adapted to pump laser

[53] See Ref. 52 on Page 92

[54] Hays, T.: A new breed of laser emerges on the optical frontier. Fiber Optic Technology — formerly Fiberoptic Products News (2005). The following section is literally quoted from `http://fiberoptictechnology.net/Scripts`.

applications first at 980 nm and later at 1 480 nm. Edge emitters also filled the vital role of high power continuous wave (CW) lasers for lithium niobate modulators and then morphed once again to emerge in an integrated source modulator combo known as electro-absorptive modulated (EAM) lasers. In spite of these amazing accomplishments, edge emitter innovations now seem to be few and far between. In fact, recent improvements have been accompanied by some level of manufacturing/process heroics that come with a hefty yield-cost impact — not exactly the solution carriers are looking for.

To date, VCSEL have earned a reputation as a superior technology for short reach applications such as fibre channel, Ethernet and intra-systems links (among switches, routers and hubs inside central offices). Within the first two years of commercial availability, VCSEL displaced edge emitters in the local area networks. Consequently, since VCSEL were first adopted in short reach applications, many have prematurely concluded that VCSEL are limited to low power, short wavelength applications. Conversely, a few companies are poised to prove otherwise as recent announcements suggest significant strides in the areas of 1 310 nm source lasers, 1 550 nm tunable lasers and finally high power pump lasers.

VCSEL are grown, processed and tested while still in the wafer form. As such, there is significant economy of scale resulting from the ability to conduct parallel device processing, whereby equipment utilization and yields are maximized and set up times and labor content are minimized. In the case of a VCSEL, the mirrors and active region are sequentially stacked along the x-axis (perpendicularly to the pn-junction, see Fig. 3.5 on Page 58) during epitaxial growth. The VCSEL wafer then goes through etching and metallization steps to form the electrical contacts. At this point the wafer goes to test where individual laser devices are characterized on a pass-fail basis. Finally, the wafer is diced and the lasers are binned for either higher-level assembly (typically $> 95\,\%$) or scrap (typically $5\,\%$).

In a simple Fabry-Perot edge emitter (DFB edge emitters require additional etch and re-growth steps) the growth process also occurs along the x-axis, but only to create the active region, as mirror coatings are later applied along the y-axis (in the junction plane). After epitaxial growth, the wafer goes through the metallization step and is subsequently cleaved along the y-axis, forming a series of wafer strips. The wafer strips are then stacked and mounted into a coating fixture. The x-axis edges of the wafer strips are then coated to form the device mirrors. Now the wafer strips are diced to form discrete laser chips, which are then mounted onto carriers. Finally, the laser devices go in to test where typically more than $50\,\%$ of DFB are scrapped.

It is also important to understand that VCSEL consume less material. In the case of a 3 in wafer, a laser manufacturer can build about 15 000 VCSEL or approximately 4 000 edge emitters. Considering the $2:1$ yield advantage (DFB edge emitter) combined with a $4:1$ wafer throughput edge, the VCSEL cost advantage is obvious.

The main disadvantage comes from the fact that VCSEL are routinely fabricated only in the short-wavelength region $\lambda < 1\,\mu$m which prevents applications in long-haul transmission systems operating near the fibre loss minimum at $\lambda = 1.55\,\mu$m. However, recent progress lead to VCSEL at $1.55\,\mu$m with threshold currents of 0.5 mA. If biased at 5 mA they can be directly modulated with 2.5 Gbit/s for a robust upstream WDM transmission in a low-cost passive optical access network[55]. Recent progress[56] led to transmitting line rates up to 115 Gbit/s over a 4 km long single-mode fibre using a directly modulated $1.55\,\mu$m single-mode VCSEL with discrete multi-tone (DMT) modulation (a variety of orthogonal frequency division multiplexing, OFDM) and direct detection. Also coherent transmission over 5×8 km single-mode fibre was successfully demonstrated[57] with line rates (data rates) of 400 Gbit/s (333 Gbit/s).

[55] Wong, E.; Zhao, X.-x.; Chang-Hasnain, C. J.; Hofmann, W.; Amann, M. C.: Uncooled, optical injection-locked $1.55\,\mu$m VCSELs for upstream transmitters in WDM-PONs. Technical Digest Optical Fiber Communication Conference (OFC'06), Anaheim (CA), USA, 05.–10.03.2006. Postdeadline Paper PDP50

[56] C. Xie, P. Dong, S. Randel, D. Pilori, P. Winzer, S. Spiga, B. Kögel, C. Neumeyr, M.-C. Amann: Single-VCSEL 100-Gb/s short-reach system using discrete multi-tone modulation and direct detection. Technical Digest Optical Fiber Communication Conference (OFC'15), Los Angeles (CA), USA, 22.–26.03.2015. Paper Tu2H.2

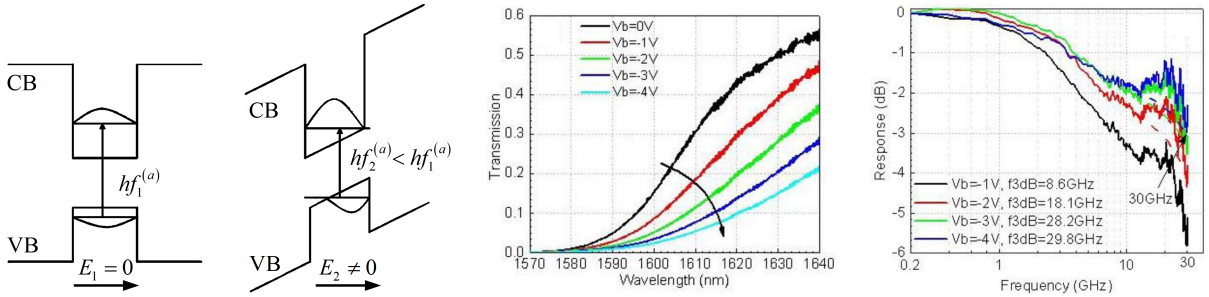
[57] C. Xie, S. Spiga, P. Dong, P. Winzer, M. Bergmann, B. Kögel, C. Neumeyr, M.-C. Amann: 400-Gb/s PDM-4PAM WDM system using a monolithic 2×4 VCSEL array and coherent detection. J. Lightw. Technol. 33 (2015) 670–677

3.2 Modulators

Modulation superimposes a low frequency signal on a high frequency carrier wave. The direct modulation of semiconductor lasers suffers from modulation frequency limitations, chirp, and relaxation oscillation. Therefore it is advantageous to operate the laser as a continuous wave (CW) source, and to modulate the field or the intensity with an external modulator, The external modulators can be broadly classified[58] as electro-absorption type or electro-optic type.

3.2.1 Electro-absorption modulator

Electro-absorption modulators (EAM) exploit the fact that the absorption edge of a semiconductor quantum film[59] can be shifted under the influence of an electric field $\vec{E}$ normal to the film layer. This so-called quantum-confined Stark[60] effect is illustrated in Fig. 3.27.



(a) Energy-band diagram of double-heterostructure quantum film with and without applied electric field $E_{1,2} \sim V_b$ (bias voltage V_b)

(b) Measured transmission spectra of a Ge electro-absorption modulator for various bias voltages V_b

(c) Measured frequency response of a Ge electro-absorption modulator for various bias voltages V_b

Fig. 3.27. Electro-absorption modulator based on the quantum-confined Stark effect. (a) Energy-band diagrams of a quantum film with and without a bias field. The tilt of the band diagram reduces the energy difference between the (schematicall drawn) electronic wave function in the conduction band (CB) and the hole wave function in the valence band (VB) from a photon absorption energy of $hf_1^{(a)}$ to $hf_2^{(a)} < hf_1^{(a)}$. [Modified from Ref. 61, Fig 4] (b) Transmission spectra of an electro-absorption modulator and its (c) frequency response for various bias voltages. [After Ref. 62, Figs. 3(a), 4(a)]

The energy-band diagram of an unbiased and a biased quantum film[61] is shown in Fig. 3.27(a). With a bias field $\vec{E}_2 \neq 0$ the bands tilt, and the energy difference between the electronic wave function in the conduction band (CB) and the hole wave function in the valence band (VB) reduces from a photon absorption energy of $hf_1^{(a)}$ to $hf_2^{(a)} < hf_1^{(a)}$. This is also illustrated in the measured transmission spectra Fig. 3.27(b) of an actual EAM[62]. Modulation bandwidths of 30 GHz and above can be achieved as can be seen from the measurement data in Fig. 3.27(c).

Because in an EAM the absorption is switched, a certain chirp of the optical carrier cannot be avoided. The situation is similar to the case of amplitude-phase coupling for a laser diode. However, in the case of an EAM, the "gain" constant g in Eq. (3.113) on Page 90 is negative and therefore describes loss.

[58] H. Yasaka, Y. Shibata: Semicondurctor-based modulators. In: H. Venghaus, N. Grote (Eds.), Fibre optic communication — Key devices. Heidelberg: Springer-Verlag 2012. Chapter 6

[59] For the naming "quantum film" (instead of "quantum well") see Footnote 27 on Page 67

[60] Johannes Stark, German physicist, ⋆Schickenhof (Germany) 15.4.1874, †Traunstein (Germany) 21.06.1957. Won the 1919 Nobel Prize for Physics for his discovery in 1913 that an electric field would cause splitting of the lines in the spectrum of light emitted by a luminous substance; the phenomenon is called the Stark effect. — Stark became a lecturer at the University of Göttingen in 1900, and from 1917 until he retired in 1922, he was a professor of physics at the University of Greifswald and, later, at the University of Würzburg (all in Germany). A supporter of Adolf Hitler and an anti-semitic racial theorist, Stark was president of the Reich Physical-Technical Institute from 1933 to 1939. In 1947 a denazification court sentenced him to four years in a labour camp.

[61] S. Mokkapati, C. Jagadish: 'III-V compound SC for optoelectronic devices,' materialstoday 12 (2009) 22–32. http://www.sciencedirect.com/science/article/pii/S1369702109701105

[62] Ning-Ning Feng, Dazeng Feng, Shirong Liao, Xin Wang, Po Dong, Hong Liang, Cheng-Chih Kung, Wei Qian, Joan Fong, Roshanak Shafiiha, Ying Luo, Jack Cunningham, Ashok V. Krishnamoorthy, Mehdi Asghari: 30 GHz Ge electro-absorption modulator integrated with 3 μm silicon-on-insulator waveguide. Opt. Express 19 (2011) 7062–7067

3.2.2 Electro-optic modulator

An electro-optic modulator makes use of the electro-optic effect or Pockels[63] effect, see also Eq. (A.10) on Page 177 and the following text. The second-order nonlinear susceptibility $\chi^{(2)}$ is controlled in proportion to a "low" radio-frequency (RF) controlling field E_{RF}, which is in proportion of a controlling voltage V. This enables a change of the refractive index n in a waveguide of geometrical length L propagating ($k_0 = \omega/c = 2\pi/\lambda$) an optical wave E_{opt}. The resulting optical phase ϑ is

$$\vartheta = k_0 n L \sim E_{\mathrm{RF}} L \sim V L, \qquad V_\pi L \text{ for an optical phase shift } \Delta\vartheta = \pi. \tag{3.120}$$

The $V_\pi L$-product is a a quality metric and tells how small the applied voltage $V = V_\pi$ can be to still achieve a phase change $\Delta\vartheta = \pi$ in a waveguide length L. Importantly, the Pockels effect reacts virtually instantaneously to the controlling voltage. In addition, the usual Pockels media are lossless in the wavelength region of interest. Frequently used $\chi^{(2)}$-media are lithium niobate ($LiNbO_3$), gallium arsenide (GaAs), indium phosphide (InP), and $\chi^{(2)}$-nonlinear organic materials[64,65].

Other options are using the plasma dispersion effect[66,67] in depleted pn-junctions, or the injection of carriers into pn-junctions. Both, depletion and injection of charge carriers in an active volume, changes the refractive index in this region. However, the plasma dispersion effect is never lossless (hence cannot realize a pure phase modulation), and it suffers from speed limitations due to the finite carrier lifetime.

Mach-Zehnder modulator

Such a pure phase modulator becomes significantly more versatile when inserted into the arms of a Mach[68]-Zehnder[69] interferometer (MZI) as in Fig. 3.28(a). The transfer function of a MZI modulator (Mach-Zehnder modulator for short, MZM) can be easily calculated when taking into account that the power is split (and combined) evenly between the arms, i. e., the fields have a split (or combine) factor of $1/\sqrt{2}$.

[63] Friedrich Carl Alwin Pockels, German physicist, ⋆ Vicenza (Italy) 18.6.865, † Heidelberg (Germany) 29.8.1913. He obtained a doctorate from the University of Göttingen in 1888, and from 1900 to 1913 he was professor of theoretical physics at the University of Heidelberg. In 1893 he discovered that a static electric field applied to certain birefringent materials causes the refractive index to vary, approximately in proportion to the strength of the field. The coefficient of proportionality is of the order of $10 \times 10^{-10}\,\mathrm{V}^{-1}$ to $10 \times 10^{-12}\,\mathrm{V}^{-1}$. This phenomenon is now called the Pockels effect. — His sister Agnes Pockels (1862–1935) was also a physicist. [Cited from `http://en.wikipedia.org/wiki/Friedrich_Carl_Alwin_Pockels` and `http://de.wikipedia.org/wiki/Friedrich_Pockels`]

[64] Leuthold, J.; Koos, C.; Freude, W.: 'Nonlinear silicon photonics,' Nature Photon. 4 (2010) 535–544

[65] Leuthold, J.; Koos, C.; Freude, W.; Alloatti, L.; Palmer, R.; Korn, D.; Pfeifle, J.; Lauermann, M.; Dinu, R.; Jazbinsek, M.; Waldow, M.; Wahlbrink, T.; Bolten, J.; Fournier, M.; Yu, H.; Wehrli, S.; Fedeli, J. M.; Gunter, P.; Bogaerts, W.: 'Silicon-organic hybrid electro-optical devices', IEEE J. Sel. Topics Quantum Electron. 19 (2013) 3401413

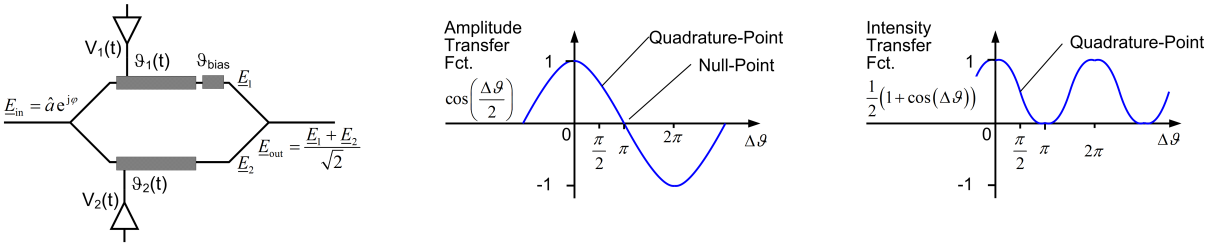
[66] See Ref. 64 on Page 96

[67] G. T. Reed, G. Mashanovich, F. Y. Gardes, and D. J. Thomson: 'Silicon optical modulators,' Nature Photon. 4 (2010) 518–526

[68] Ludwig Mach, German inventor, ⋆ Prague 18.11.1868, † 1951. Eldest son of Ernst Mach. Doctorate in medicine 1895 from the University of Prague. Cooperated with his father in topics of optics and instrument design. Developed together with his father the Mach-Zehnder interferometer in 1892. [Cited after `http://de.wikipedia.org/wiki/Ludwig_Mach`]

Ernst Mach, Austrian physicist and philosopher, ⋆ Chirlitz-Turas (Moravia, Austrian Empire) 18.2.1828, † Haar (Germany) 19.2.1916. Established important principles of optics, mechanics, and wave dynamics and supported the view that all knowledge is a conceptual organization of the data of sensory experience (or observation). — He received his doctorate in physics in 1860 from the University of Vienna and taught mechanics and physics in Vienna until 1864, when he became professor of mathematics at the University of Graz. Mach left Graz to become professor of experimental physics at the Charles University in Prague in 1867, remaining there for the next 28 years. Between 1873 and 1893 he developed optical and photographic techniques for the measurement of sound waves and wave propagation. In 1887 he established the principles of supersonics and the Mach number—the ratio of the velocity of an object to the velocity of sound.

[69] Ludwig (Louis Albert) Zehnder, Swiss physicist, ⋆ Illnau (Switzerland) 4.5.1854, † Oberhofen (Thunersee, Switzerland) 24.3.1949. Studied mechanical engineering in Zürich 1873–1875 and cooperated with Wilhelm Conrad Röntgen. Doctorate in physics from University of Gießen (Germany) in 1887, habilitation in physics from University of Basel (Switzerland) in 1890. Professor in Freiburg (Germany) in 1893, München (Germany) in 1901, and Basel (Switzerland) in 1919–1945. Published the construction of a new interferometer in August 1891 (now the Mach-Zehnder interferometer). Independently, Ludwig Mach had constructed a similar apparatus, which he published seven months later in the same journal. [Cited after `http://de.wikipedia.org/wiki/Ludwig_Zehnder`]



(a) Mach-Zehnder interferometer modulator with voltage-controlled modulator sections in both arms

(b) Amplitude transfer function in dependence of the phase difference $\Delta\vartheta$ in both arms

(c) Intensity transfer function in dependence of the phase difference $\Delta\vartheta$ in both arms

Fig. 3.28. Schematic and characteristics of an electro-optic modulator in form of a Mach-Zehnder interferometer (MZI) with phase modulators in both arms based on the Pockels effect. (a) Schematic of a Mach-Zehnder interferometer with input waveguide splitter, phase modulator sections $\vartheta_{1,2}(t)$, bias phase adjustment ϑ_{bias}, optical electric fields $E_{1,2}(t)$ at the outputs of both arms, and output waveguide combiner. Electrical amplifiers supply the control voltages $V_{1,2}(t)$. (b) Amplitude transfer function of an MZI push-pull modulator as a function of the phase difference $\Delta\vartheta = \vartheta_1 + \vartheta_{\text{bias}} - \vartheta_2$ in both arms. (c) Intensity transfer function of an MZI push-pull modulator as a function of the phase difference $\Delta\vartheta = \vartheta_1 + \vartheta_{\text{bias}} - \vartheta_2$ in both arms [modified from Fig. 2.21(a) and 2.22 of Ref. † on the Preface page]

With this information, the complex amplitude at the interferometer output reads in matrix notation (using also a column and a row matrix)

$$\underline{E}_{\text{out}} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} \mathrm{e}^{\mathrm{j}(\vartheta_1+\vartheta_{\text{bias}})} & 0 \\ 0 & \mathrm{e}^{\mathrm{j}\,\vartheta_2} \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \underline{E}_{\text{in}} \,. \tag{3.121}$$

From Eq. (3.121) the amplitude transfer function $\underline{T}$ follows,

$$\underline{T} = \frac{\underline{E}_{\text{out}}}{\underline{E}_{\text{in}}} = \mathrm{e}^{\mathrm{j}\left(\frac{\vartheta_1+\vartheta_2}{2}+\frac{\vartheta_{\text{bias}}}{2}\right)} \cos\left(\frac{\Delta\vartheta}{2}\right), \qquad \Delta\vartheta = \vartheta_1 - \vartheta_2 + \vartheta_{\text{bias}} \,. \tag{3.122}$$

When varying $\vartheta_{1,2}$, both, the phase and the amplitude of $\underline{T}$ change. However, if $\vartheta_1 = -\vartheta_2$ is chosen, i. e., if $V_1 = -V_2$ holds, the phase factor remains constant (but the sign of $\underline{T}$ could change). This so-called push-pull operation mode is most commonly used. If $\vartheta_1 = \vartheta_2$ is maintained in push-push mode, we have a pure phase modulator.

In the following, we concentrate on push-pull operation, for which the field transfer characteristic is displayed in Fig. 3.28(b), (c). The bias determines the operating point. For an optimum field linearity it has to be chosen at the null-point. The quadrature-point is optimum for intensity linearity, where the intensity transfer characteristic is

$$|\underline{T}|^2 = \left|\frac{\underline{E}_{\text{out}}}{\underline{E}_{\text{in}}}\right|^2 = \cos^2\left(\frac{\Delta\vartheta}{2}\right) = \frac{1}{2}\left(1+\cos\Delta\vartheta\right), \qquad \Delta\vartheta = \vartheta_1 - \vartheta_2 + \vartheta_{\text{bias}} \,. \tag{3.123}$$

The characteristic Eq. (3.123) is displayed in Fig. 3.28(c).

With the MZM described by the field transfer function Eq. (3.122), any point in the complex constellation plane (any amplitude and phase) could be addressed by a proper choice of the modulation voltages $V_1 \sim \vartheta_1$ and $V_2 \sim \vartheta_2$. However, then a sophisticated control of these modulation voltages $V_{1,2}$ is required, and therefore the more tolerant optical IQ-modulator is preferred for this purpose.

Optical IQ-modulator

For optical IQ-modulation we need to realize the scheme as discussed in Sect. 2.3.2 on Page 28 ff. To this end we use two MZM nested in a MZI, Fig. 3.29. The control voltages $V_1(t)$ and $V_2(t)$ represent the in-phase and quadrature signals $I(t)$ and $Q(t)$, respectively, as specified in Eq. (2.43) on Page 28. The combiners in Fig. 2.7 are assumed to perform a *summation* (factors $1/\sqrt{2}$ are omitted), and the phase $\vartheta_{\text{bias}} = \pi/2$ *advances* the local oscillator (LO) signal $\cos(\omega_0 t)$ to be $-\sin(\omega_0 t)$. This is in contrast to

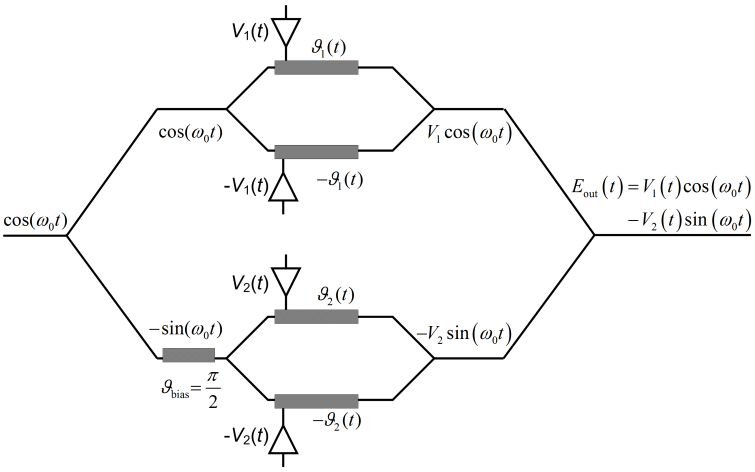


Fig. 3.29. Optical IQ-modulator with two push-pull Mach-Zehnder interferometer modulators nested in a Mach-Zehnder interferometer. The control voltages $V_1(t)$ and $V_2(t)$ represent the in-phase and quadrature signals $I(t)$ and $Q(t)$, respectively, as specified in Eq. (2.43) on Page 28. [Modified from Fig. 2.25 of Ref. † on the Preface page]

Fig. 2.7 on Page 29, where the the symbol Σ stands for a *difference*-forming combiner, while the phase of the LO is *retarded* by $+\pi/2$ to be $+\sin(\omega_0 t)$. The result, however, is the same. The output electric field can be written as

$$\underline{E}_{\text{out}} = (V_1 + \mathrm{j}\, V_2)\, \mathrm{e}^{\,\mathrm{j}\,\omega_0 t}, \qquad E_{\text{out}} = \Re\{\underline{E}_{\text{out}}\} = V_1 \cos(\omega_0 t) - V_2 \sin(\omega_0 t)\,. \tag{3.124}$$

3.3 Implementation of selected modulation formats

In the following, we discuss the implementation of a few selected modulation formats like non-return to zero on-off keying (NRZ-OOK), return to zero on-off keying (RZ-OOK), duobinary (DB) and alternate mark inversion (AMI), and polarization mode shift keying (PMSK).

3.3.1 Non-return to zero on-off keying

Non-return to zero on-off keying (NRZ-OOK) data are generated by either of the following schemes:

- Directly modulating a laser diode by switching the injection current. Directly modulated lasers (DML) can be operated with data rates up 20 Gbit/s. However, significant chirp leads to distortions due to increased chromatic dispersion in the transmitting fiber.
- Externally modulating a CW laser diode with an electro-absorption modulator. This technique is good for data rates up to 40 Gbit/s. Again, a relatively small chirp associated with the absorption change limits the signal quality.
- Externally modulating a CW laser diode with a MZI modulator. This technique is used for data rates up to 40 Gbit/s. Time division multiplexing techniques enable operation up to 160 Gbit/s. If used in push-pull operation mode, there is basically no chirp.

Characteristic for the optical NRZ spectrum are a strong carrier component at the optical carrier frequency, see Fig. 2.12(c) on Page 37 and Fig. 2.13(a) on Page 40. The NRZ spectrum has a spectral width close to $2/T$ (twice the symbol rate). There are spectral zeros at frequency offsets of integer multiples of $\pm\, 1/T$ from the carrier frequency, Eq. (2.58) on Page 38.

3.3.2 Return to zero on-off keying

The idea for generating a RZ format is to encode data with one of the NRZ schemes, and then "carve" an RZ shape into the NRZ data using a second MZM in push-pull mode that avoids chirp. The modulator is biased at a phase ϑ_{bias}, and driven by a sinusoidal voltage in synchronism with the data encoder. The percentage of power left in the optical pulses after carving is indicated by the duty cycle.

Figure 3.30(a) depicts the setup for an RZ duty cycle of 50 %. The total voltage swing over the two arms of the modulator is such that $(\vartheta_1 - \vartheta_2)_{\max} = \pi/2$. The MZM is biased at $\vartheta_{\text{bias}} = \pi/4$. The carver's modulation frequency equals the NRZ symbol rate $1/T$. By reducing the voltage swing applied to the carver MZM and by adjusting the bias, even lower RZ duty cycles can be obtained. However, one typically does not go below 36 %, lest fewer pulse energy results in a jittery signal.

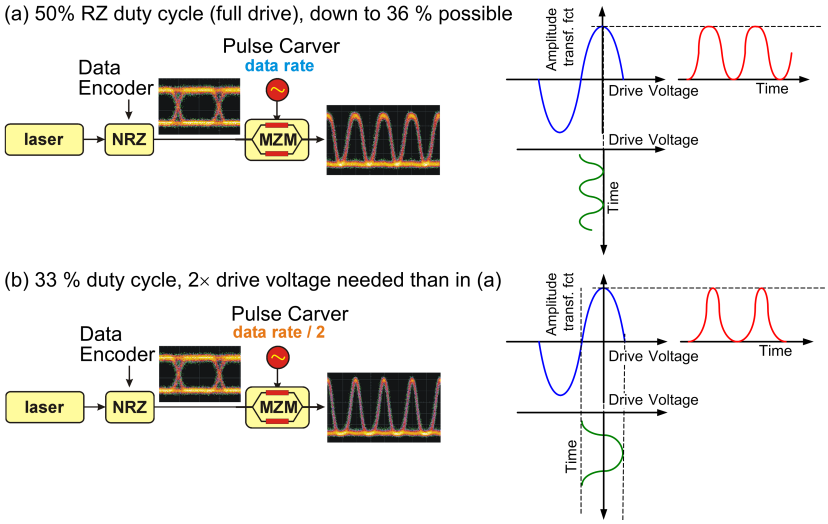


Fig. 3.30. Implementation of RZ-OOK with a pulse "carver" driven in synchronism with the data by a sinusiodal voltage. Eye diagrams of NRZ and resulting RZ data are shown as insets. (a) Pulse carver for RZ pulses with 50 % duty cycle (b) Pulse carver for RZ pulses with 33 % duty cycle. In this case, the drive voltage is twice the drive voltage for a 50 % duty cycle. [Modified from Fig. 2.32 of Ref. † on the Preface page]

A different RZ carver for a 33 % duty cycle is shown in Fig. 3.30(b). Here the modulator voltage swing is doubled such that $(\vartheta_1 - \vartheta_2)_{\max} = \pi$, and the bias is set to $\vartheta_{\text{bias}} = 0$. The carver's modulation frequency $1/(2T)$ equals half the NRZ symbol rate.

The 50 % and 33 % duty cycle schemes provide almost perfect signal quality. Unfortunately, this comes at the price of an additional modulator with its associated driver electronics. A scheme with a single modulator would therefore be preferred. To reduce the number of modulators, one could combine the function of the electrical clock from the pulse carver and the data signal into the electronic circuitry, and then direct these combined electrical signals to a single modulator. This scheme, however, requires electrical circuits and an optical modulator with an exceptionally wideband frequency response.

Pulse carver schemes that produce chirped pulses[70] exist as well, if $\vartheta_1 \neq -\vartheta_2$ is chosen for intentionally inducing a chirp, e. g., for pre-compensating fibre dispersion. A chirped RZ-OOK spectrum (50 % CRZ-OOK) with various phase modulation indices η is displayed in Fig. 2.15(a) on Page 44.

What are the advantages of using RZ-OOK over NRZ-OOK for transmission? The answer is that the RZ format has a receiver sensitivity advantage of about 1 . . . 3 dB. This is a significant improvement, since a 3 dB advantage translates in doubling the transmission distance. The reasons for this improvement are:

- For the same average power, a RZ signal has more power within the pulse centre where sampling takes place. This gives a RZ format a typical 2 dB advantage. This is due to the fact that in a RZ

[70] See Ref. 87 on Page 45

signal all of the energy is confined to within part of a symbol slot. As a drawback, fibre nonlinearities due to high peak powers can lead to degradation.

- Second, RZ signals suffer less from inter-symbol interference (ISI), since leading and trailing edges of the pulse do not easily extend into neighbouring time slots.

One now might wonder if reducing the duty cycle below 33 % will bring an additional advantage (let aside fibre nonlinearities). However, narrower pulses require a larger receiver bandwidth, and this means more noise, so that a 33 % duty cycle is near optimum.

3.3.3 Duobinary and alternate mark inversion

Duobinary (DB) and alternate mark inversion (AMI) signals are bipolar binary signals. As discussed in Sect. 2.4.2 on Page 40 they employ the three-level signalling set $\{-1, 0, +1\}$, where the optical phases of the individual bits additionally depend on the bit pattern: For DB signaling, a phase change occurs whenever there is an odd number of logical 0 between two successive logical 1, whereas for AMI the phase changes for each logical 1 (even for adjacent logical 1), independent of the number of logical 0 in-between. Chirp-free optical DB and AMI signals are obtained when operating the MZI in push-pull mode.

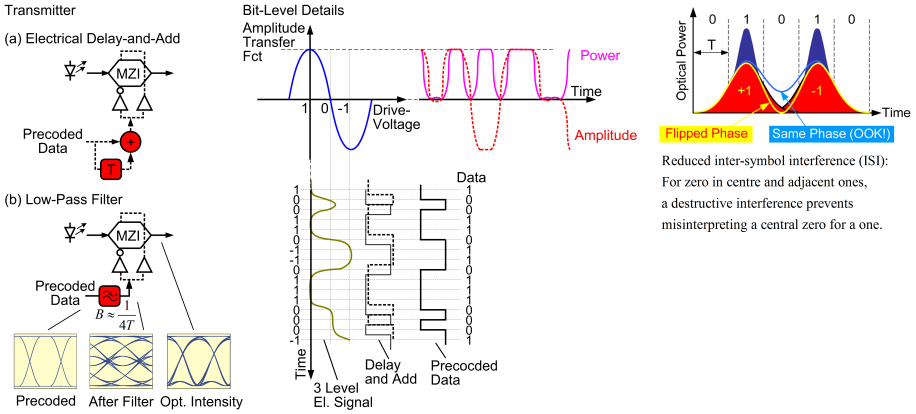


Fig. 3.31. Transmitter for duobinary signals with precoder and (a) delay-and-add circuit, or (b) low-pass filter with bandwidth $B \approx 1/(4T)$ equal to about one quarter of the symbol rate for generating three-level signals. A logical zero enclosed by two logical ones is not perturbed by its neigbours, because the ones interfere destructively in the bit slot of the zero. [Modified from Fig. 2.49 and 2.50 of Ref. † on the Preface page]

The process of generating an optical DB or AMI signal comprises the following:

- Precoding of data by differential encoding. For each occurrence of a logical 0 there is a transition of the electrical level. Logical 1 leaves the previous electrical level unchanged.
- The thus generated precoded drive voltage is added (DB, Fig. 3.31(a)) or subtracted (AMI, not shown) from the 1 bit-delayed replica of itself. Alternatively, a low-pass filter with a bandwidth $B \approx 1/(4T)$ equal to about one quarter of the symbol rate serves the same purpose, Fig. 3.31(b).
- This three-level signal controls the MZI modulator, which is operated in push-pull and biased at $\vartheta = \pi/4$.

Since the required three-level (linear) RF driver electronics are hard to implement in practice, one usually resorts to the DB transmitter version Fig. 3.31(b), where a low-pass filter (bandwidth $B \approx 1/(4T)$ equals a quarter of the symbol rate) processes the precoded data.

Finally, the delay-and-add or the delay-and-subtract action could be also done optically using a 1 bit optical delay interferometer (DI). In this case, a DPSK-encoded optical signal at the DI input converts to an inverted DB signal at the constructive output port of the DI, while AMI formatted data appear at its destructive port. When treating the DPSK receiver, we will come back to this realization.

3.3.4 Polarization mode shift keying

If two orthogonal polarizations (DP) carry independent information, the effective data rate can be doubled when compared to transmission in a single polarization (SP). If not capacity, but sensitivity is of primary importance, polarization mode shift keying (PMSK) in the form of polarization switching (PS) is of interest. Particularly the PS-QPSK format is very noise resilient[71]. Figure 3.32 shows a PS-QPSK

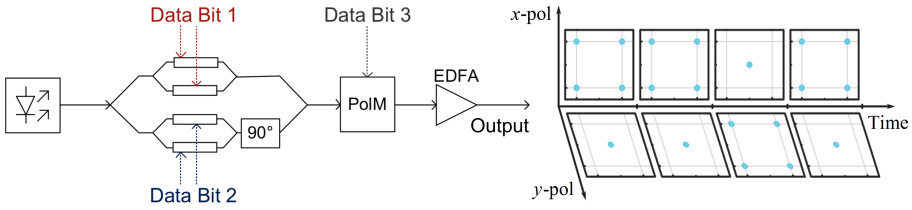


Fig. 3.32. Transmitter for polarization mode shift keying (PMSK), here polarization switching (PS). The example shows QPSK encoding (2 bit), which can appear in either of two polarizations (1 bit) named x-pol and y-pol, respectively. Thus, 3 bit are transmitted in each symbol period. The QPSK constellations illustrate the polarization switching. The central dot represents a zero electric field and indicates that no signal is transmitted in this very polarization. [Modified from Fig. 2.59 of Ref. † on the Preface page]

transmitter along with the constellation diagrams in both polarizations. A single dot in the centre indicates that this polarization has a zero field strength at this very moment. The modulation format worked well[72] for a data rate of 112 Gbit/s (symbol rate 37.3 GBd, 3 bit / symbol.

3.4 Software-defined transmitter

So far, each modulation format required a specific hardware setup. It would be more convenient, if the same hardware could produce a variety of modulation formats. This task is solved by a software-defined transmitter as shown in Fig. 1 on the next but one page. Inside 5 ns and without loosing any data, this transmitter[73] can switch to a different modulation format. The field-programmable gate arrays (FPGA) generate two digital signal streams, which are then converted to analogue voltages by digital-to-analog converters (DAC). The resulting analogue drive voltages control the phase modulator sections of the optical IQ-modulator.

While this approach is most versatile, it would be useful to find configurations without the DAC and their electrical drive circuitry. This would require to translate each and every binary electrical signal directly to an optical constellation. However, this puts much more complexity to the optical modulators.

[71] M. Karlsson, E. Agrell: Which is the most power-efficient modulation format in optical links? Opt. Express 17 (2009) 10814–10819

[72] See Ref. 71 on Page 101

[73] Schmogrow, R.; Hillerkuss, D.; Dreschmann, M.; Huebner, M.; Winter, M.; Meyer, J.; Nebendahl, B.; Koos, C.; Becker, J.; Freude, W.; Leuthold, J.: Real-time software-defined multiformat transmitter generating 64QAM at 28 GBd. IEEE Photon. Technol. Lett. 22 (2010) 1601–1603

IEEE PHOTONICS TECHNOLOGY LETTERS, VOL. 22, NO. 21, NOVEMBER 1, 2010 1601

Real-Time Software-Defined Multiformat Transmitter Generating 64QAM at 28 GBd

R. Schmogrow, D. Hillerkuss, M. Dreschmann, M. Huebner, M. Winter, J. Meyer, B. Nebendahl, C. Koos, J. Becker, W. Freude, and J. Leuthold

***Abstract*—We demonstrate a software-defined real-time optical multiformat transmitter. Here, eight different modulation formats are shown. Data rate and modulation formats are defined through software accessible look-up tables enabling format switching in the nanosecond regime without changing the transmitter hardware. No data are lost during the switching process. SP-64 quadrature amplitude modulation at 28 Gbd has been generated and tested. This allows us to generate a 336-Gb/s real-time pseudorandom bit sequence in a dual polarization setup.**

***Index Terms*—Advanced modulation formats, field-programmable gate array (FPGA), real-time, software defined transmitter.**

I. Introduction

TODAY'S high-performance communication systems rely heavily on optical transmission links. High-speed electronics is crucial to exploit the large bandwidth of optical systems. So far, optical backbone networks were operated mostly with pulse amplitude modulation (PAM) and phase-shift keying (PSK) modulation formats such as differential PSK (DPSK) and differential quadrature PSK (DQPSK) [1]. However, future optical networks will operate with multilevel coded signals such as M -quadrature amplitude modulation (QAM) [2]. Advanced modulation formats promise enhanced spectral efficiency at the cost of more complex transmitters and receivers. There are several ways to implement QAM transmitters, such as discrete electrical digital-to-analog converters (DACs) [3], optical multimodulator schemes [4], all-optical DACs [5], and integrated electrical DACs in the form of arbitrary waveform generators (AWGs). Although the AWG is the most versatile solution, its capability is limited due to finite memory size, and due to the lack of real-time processing. Therefore, a more powerful solution has to be found. Field-programmable gate arrays (FPGAs) are able to handle the required amount of data, and yet offer the flexibility to change their functionality through software. Combined with state-of-the-art high-speed DACs, a software defined transmitter can be implemented. With such a scheme, real-time Nyquist sampling for precompensating dispersion at 10.7 Gb/s was demonstrated [6]. Recently, we have introduced a highly flexible and synchronous transmitter for modulation formats as complex as 16QAM [7], providing online data generation and real-time digital signal processing at the same time.

In this letter, we present the concept and the implementation of a software-defined transmitter that is capable of generating binary PSK (BPSK), QPSK, 4PAM, 6PAM, 8PSK, 16QAM, 32QAM, and 64QAM at symbol rates up to 28 GBd.

II. Experimental Setup

The experimental setup of the software-defined transmitter comprises several electrical and optical components as illustrated in Fig. 1. An external cavity laser source provides the optical carrier to be modulated in nested $LiNbO_3$ complex inphase (I)/quadrature (Q) Mach–Zehnder modulators [(MZMs) electrical bandwidth ≈28 GHz, π-phase shift voltage $V_\pi \approx 2$ V]. The electrical signal is generated by two MICRAM high-speed DACs, the outputs of which are amplified for driving the MZM. Both DACs are supplied with an electrical clock with a maximum frequency of 28 GHz. A variable electrical phase shifter aligns the two DAC outputs in phase with respect to each other. Xilinx Virtex5 FPGAs drive the DACs, each of which providing 24 over-clocked feeding lines operating at up to 7 Gb/s each. The over-clocking did not cause stability issues. The feeding lines are $4:1$ multiplexed by the DAC, resulting in an overall symbol rate of up to 28 GBd with a resolution depth of 6 bits. The electrical clock for the FPGA is generated by frequency dividers located on the DAC board. Real-time computation is performed by the FPGA devices incorporating both, signal generation, or external data accommodation and signal processing. To emulate dual polarization (DP) signals, the modulated signal is split, and a delay of 5.3 ns is applied to one of the paths for decorrelation. A variable optical attenuator equalizes the optical power in the two paths. The orthogonally polarized signals are then combined.

To judge the quality of the transmitter, an Agilent N4391A Optical Modulation Analyzer (OMA) receives, postprocesses, and analyzes the constellations. Further, an Agilent Digital Communications Analyzer (DCA) measures eye diagrams and detects intensities. In order to receive sufficient optical power in the two units, the signal was amplified by an erbium-doped fiber amplifier (EDFA), and then optically bandpass filtered to suppress the amplified spontaneous emissions generated by the

Manuscript received June 29, 2010; revised August 05, 2010; accepted August 27, 2010. Date of publication September 07, 2010; date of current version October 13, 2010. This work was supported by the European Network of Excellence *EuroFOS*, by the Xilinx University Program (XUP), by Micram Microelectronic GmbH, by the Agilent University Relations Program, by the Karlsruhe School of Optics and Photonics (KSOP), and by the German Research Foundation (DFG).

R. Schmogrow, D. Hillerkuss, M. Dreschmann, M. Huebner, M. Winter, J. Meyer, C. Koos, J. Becker, W. Freude, and J. Leuthold are with the Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany (e-mail: rene.schmogrow@kit.edu).

B. Nebendahl is with Agilent Technologies, 71034 Boeblingen, Germany.

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LPT.2010.2073698

1602 IEEE PHOTONICS TECHNOLOGY LETTERS, VOL. 22, NO. 21, NOVEMBER 1, 2010

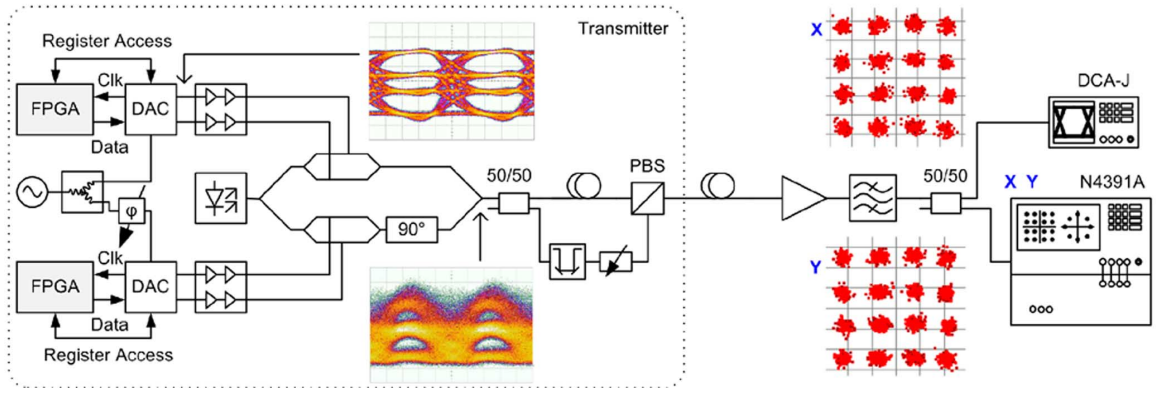


Fig. 1. Experimental setup and transmitter implementation. A semiconductor laser source is I/Q modulated by nested MZMs. Each arm is driven by the amplified output of two DACs that are both clocked at 28 GHz. Two FPGAs are used to drive the DAC and perform real-time signal processing. DAC register access is needed for synchronization. A $2^{15}-1$ PRBS is generated on both FPGAs resulting in a four-level electrical output, which was predistorted to compensate for the nonlinear transfer function of the modulator. The output of the I/Q modulator is polarization multiplexed, amplified, and then split. A DCA with a 50-GHz optical input was used to measure intensity, and the N4391A OMA decoded the DP-16-QAM signal resulting in $\mathbf{X}$ and $\mathbf{Y}$.

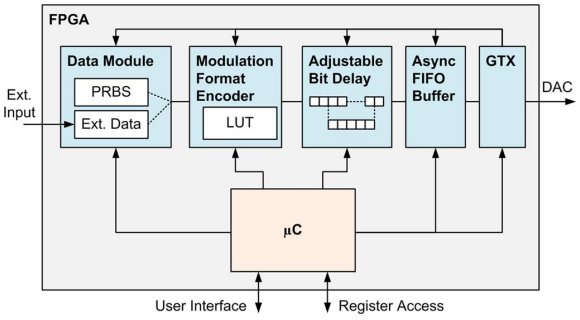


Fig. 2. Design of the on-chip FPGA architecture. All high-speed modules (upper path) are controlled via a central processor (μC). The "Data Module" generates PRBS, or serves as an interface to external data ports. The "Modulation Format Encoder" provides real-time data mapping and can be modified within 5 ns to changing formats. The individually "Adjustable Bit Delay" module is used for synchronization of the "Asynchronous FIFO Buffer" to ensure data consistency. Eventually, data are transmitted through high-speed GTX transceivers at up to 7 Gb/s.

EDFA. The signal is split and fed to the OMA and the DCA. Intensities were measured for each polarization separately. The decoded data served for bit-error ratio (BER) and error vector magnitude (EVM) measurements. We recorded constellation diagrams for eight different modulation formats.

III. FPGA Design

A. Overall FPGA Architecture

The FPGA design can be divided into several blocks as depicted in Fig. 2. Real-time functional blocks are located in the upper path and the data transport is depicted from left to right. The high-speed modules involved in real-time data processing are controlled by a micro-processor located on the FPGA chip. This processor also provides a user interface as well as register access to control the DAC.

B. Binary Data Source

The transmitter can either handle arbitrary externally fed data sequences or generate pseudorandom bit sequences (PRBS) internally. For simplicity, we chose to create gigabit PRBS on-chip. We highly parallelized PRBS generators by creating multiple bits at a single clock period using a parallel XOR gate structure. Additionally, a multiplexing technique taking advantage of well-known PRBS properties was used to create an arbitrary amount of bits per cycle. Several creator polynomials have been made available (2^7-1, $2^{15}-1$, and $2^{31}-1$) resulting from a minor change of the parallel XOR gate structure. The ability to generate PRBS of arbitrary length is an advantage of our setup over an AWG, where bits can be read from a memory of limited length only.

C. Modulation Format Encoder

The characteristic of the transmitter is determined by a software accessible module. A look-up table (LUT)-based structure enables software defined encoding of the data bitstream to arbitrary multilevel signals. Nonlinear mapping (1 Sa/symbol) is applied in order to compensate for the nonlinear transfer function of the I/Q MZM. Switching between different formats is achieved by rewriting the content of the LUT. Therefore, changing the modulation format (e.g., from QPSK to 16QAM) requires only a single clock cycle (<5 ns at 28 GBd). No data at the target data rate are lost at the transmitter side during the switching process. When changing to spectrally less efficient formats (e.g., from 16QAM to QPSK), parts of the input bit stream are excluded. With this capability, the transmitter can adjust the modulation format in real-time, depending on the required throughput and channel quality.

D. FPGA-DAC Interfacing

Successful synchronization between FPGA and DAC modules is crucial for proper operation. All synchronization procedures are controlled by an integrated micro-processor in a system-on-chip approach. Individually adjustable bit delay

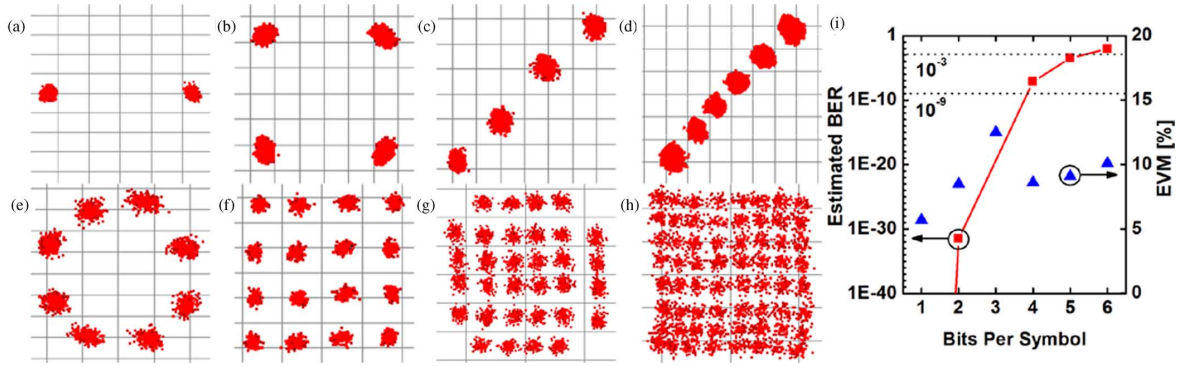


Fig. 3. Received and decoded modulation formats with corresponding constellation diagrams. (a) BPSK at 28 GBd with an EVM of 5.7% exhibited no errors; (b) QPSK at 28 GBd showed no errors with an EVM of 8.5%; (c) 4PAM at 20 GBd; (d) 6PAM at 20 GBd; (e) 8PSK at 28 GBd with an EVM of 12.5%; (f) 16QAM at 28 GBd with an EVM of 8.6% and no errors; (g) 32QAM at 28 GBd with a resulting EVM of 9.1%; (h) 64QAM at 28GBd with an EVM of 10.1%; (i) calculated BER values [8] and measured EVMs for QPSK, 16, 32, and 64QAM formats. FEC limits are indicated for a BER of 10^{-9} and 10^{-3}.

blocks, offering a delay of up to 255 bits, were applied to each of the 24 feeding lines driving the DAC. Because of the 4 : 1 multiplexed DAC inputs, a sampling phase with 90° increment can be selected through DAC-provided registers. Algorithms running on the on-chip processor select the optimum sampling phases as well as appropriate bit delays for all lines. Hence, the synchronization process has been automated completely. A $2^{15} - 1$ PRBS serves as synchronization pattern ensuring an optimal synchronization. Asynchronous first-in first-out (FIFO) buffers guarantee data consistency before they are passed to the multigigabit transceivers (GTX) of the FPGA.

IV. Experimental Results

With an Agilent OMA, we measured the EVM and BER of various formats and symbol rates. For convenience, we chose a $2^{15} - 1$ PRBS as data source. BER calculated according to [8] are denoted with a prefix "$\sim$" as opposed to measured BER. Single-polarization 28-GBd experiments were performed first. For BPSK and QPSK EVMs of 5.7% (BPSK), respectively, 8.1% (QPSK) were found, resulting in a BER below measurement limits. 16QAM exhibited errors below a BER of $\sim 1 \times 10^{-7}$ with an EVM of 8.6%. 8PSK (EVM $= 12.5\%$), 32QAM (EVM $= 9.1\%$/BER $\sim 3.5 \times 10^{-4}$), and 64QAM (EVM $= 10.1\%$/BER $\sim 9.0 \times 10^{-3}$) were also measured at 28 GBd. For DP-16QAM, we found a total BER of 5.3×10^{-5}. We determined an EVM of 10.2% for one, and 10.7% for the other polarization. The remaining formats were recorded at 20 GBd, namely 4PAM and 6PAM.

V. Conclusion and Outlook

The software-defined transmitter can generate eight different modulation formats at symbol rates up to 28 GBd. A total of 48 feeding lines between FPGAs and DACs were synchronized resulting in 336 Gb/s. Dynamic format switching in only 5 ns was performed through the processor-user interface by software adjustable LUTs. No hardware reconfiguration was needed for this purpose. Manual readjustment is only required when changing the symbol rates or when switching to ON–OFF modulation formats. Advanced modulation schemes such as DP-16QAM at 224 Gb/s were successfully tested resulting in BERs well below the forward-error correction (FEC) limit of BER $= 2 \times 10^{-3}$. A variety of different modulation schemes (Fig. 3) including 32QAM and 64QAM were experimentally demonstrated. When using the described transmitter at 64QAM in a dual polarization setup, data rates of up to 336 Gb/s on a single optical carrier can be achieved.

References

[1] A. H. Gnauck, G. Raybon, S. Chandrasekhar, J. Leuthold, C. Doerr, L. Stulz, A. Agarwal, S. Banerjee, D. Grosz, S. Hunsche, A. Kung, A. Marhelyuk, D. Maywar, M. Movassaghi, X. Liu, C. Xu, X. Wei, and D. M. Gill, "2.5 Tb/s (64'42.7 Gb/s) transmission over 40'100 km NZDSF using RZ-DPSK format and all-Raman-amplified spans," in *Proc. OFC 2002*, Anaheim, CA, Paper PDPFC2.

[2] M. Nakazawa, S. Okamoto, T. Omiya, K. Kasai, and M. Yoshida, "256-QAM (64 Gb/s) coherent optical transmission over 160 km with an optical bandwidth of 5.4 GHz," *IEEE Photon. Technol. Lett.*, vol. 22, no. 3, pp. 185–187, Feb. 1, 2010.

[3] A. H. Gnauck, P. J. Winzer, S. Chandrasekhar, X. Liu, B. Zhu, and D. W. Peckham, "10 × 224-Gb/s WDM transmission of 28-Gbaud PDM 16-QAM on a 50-GHz grid over 1 200 km of fiber," in *Proc. OFC 2010*, San Diego, CA, Paper PDPB8.

[4] M. Secondini, E. Forestieri, and F. Cavaliere, "Novel optical modulation scheme for 16-QAM format with quadrant differential encoding," *Photonics in Switching 2009*, 10.1109/PS.2009.5307754.

[5] A. Chiba, T. Sakamoto, T. Kawanishi, K. Higuma, M. Sudo, and J. Ichikawa, "16-level quadrature amplitude modulation by monolithic quad-parallel Mach–Zehnder optical modulator," *Electron. Lett.*, vol. 46, no. 3, pp. 227–228, Feb. 4, 2010.

[6] R. Waegemans, S. Herbst, L. Holbein, P. Watts, P. Bayvel, C. Fürst, and R. I. Killey, "10.7 Gb/s electronic predistortion transmitter using commercial FPGAs and D/A converters implementing real-time DSP for chromatic dispersion and SPM compensation," *Opt. Express*, vol. 17, pp. 8630–8640, 2009.

[7] D. Hillerkuss, R. Schmogrow, M. Huebner, M. Winter, B. Nebendahl, J. Becker, W. Freude, and J. Leuthold, "Software-defined multi-format transmitter with real-time signal processing for up to 160 Gbit/s," in *Proc. Optics & Photonics Congress 2010*, Karlsruhe, Germany, Paper SPTuC4.

[8] R. A. Shafik, M. S. Rahman, and A. H. M. R. Islam, "On the extended relationships among EVM, BER and SNR as performance metrics," in *Proc. 4th Int. Conf. Electrical and Computer Engineering (ICECE 2006)*, Dhaka, Bangladesh, 2006, pp. 408–411.

Chapter 4

Optical amplifiers

The transmission distance of a fibre-optic communications system is limited by fibre loss and dispersion. For long-haul lightwave systems, the loss limitation has traditionally been overcome using optoelectronic repeaters in which the optical signal is first converted into an electric current and then regenerated using a transmitter. Such regenerators become quite complex and expensive for multichannel lightwave systems. An alternative approach makes use of optical amplifiers, which amplify the optical signal directly without requiring its conversion to the electric domain, see Sect. 26 on Page 7 ff.

Optical amplifiers amplify incident light through stimulated emission, the same mechanism as that used by lasers, see Page 67 ff. Indeed, an optical amplifier is nothing but a laser without feedback. Its main ingredient is the optical gain realized when the amplifier is pumped to achieve population inversion[1].

4.1 Semiconductor amplifier

Starting from the gain relations Eq. (3.70)–(3.72) on Page 79 as discussed for the Fabry-Perot laser, we formulate the equations for the gain of an semiconductor optical amplifier (SOA) with residual mirror reflectivities $R_{1,2} \neq 0$. A transition $R_{1,2} \to 0$ leads to a true travelling-wave amplifier. If $R_{1,2} \ll 1$ holds as it is the case for real devices, we talk of a near-travelling-wave amplifier (TWA). However, now the concept of a spatial average of gain and loss must not be applied any more.

Frequently, the symbol G is used for the power gain of an amplifier[2,3]. However, we already associated the character G with the gain rate, see Eq. (3.39) on Page 70. Therefore the calligraphic symbol $\mathcal{G}$ stands for the amplifier power gain in the following text.

Assume an amplifying waveguide region with length L, having a propagation constant β and an effective (modal) refractive index n_e according to Eq. (2.13) on Page 18. For the single-pass power gain $\mathcal{G}_s$ and the phase shift φ along the amplifying region we know from Eqs. (3.70)–(3.74)

$$\mathcal{G}_s = \exp\left[(\Gamma g - \alpha_{Ve})\, L\right], \qquad \varphi = \beta L = k_0 n_e L\,. \tag{4.1}$$

Taking into account the multiple reflections at the mirrors, the amplification factor $\mathcal{G}$ is obtained using the standard theory of a Fabry-Perot interferometer[4],

$$\mathcal{G} = \frac{\mathcal{G}_s(1-R_1)(1-R_2)}{\left(1-\mathcal{G}_s\sqrt{R_1R_2}\,\right)^2 + 4\mathcal{G}_s\sqrt{R_1R_2}\,\sin^2\varphi}\,, \tag{4.2}$$

$$\varphi = \beta L, \qquad \text{resonances Eq. (3.63): } \varphi_z = \omega_z n_e L/c = m_z\pi\,, \quad m_z = 1, 2, 3, \ldots$$

[1] See Ref. 17 on Page 6
[2] See Ref. 17 on Page 6. Sect. 8.2. p. 368 ff.
[3] See Ref. 3 on Page 49, Sect. 5.5. p. 209 ff.
[4] See Ref. 6, 7, 8 on Pages 49 and 49

As is evident from Eq. (4.2), $\mathcal{G}$ peaks whenever the frequency $f = \omega/(2\pi)$ coincides with one of the cavity-resonance frequencies f_z and drops sharply in between them. Because spontaneous emission was disregarded, we find infinite gain for $\mathcal{G}_s\sqrt{R_1R_2} = 1$ at the resonance points $\varphi_z = m_z\pi$ spaced apart by the free spectral range $\Delta f_z = c/(2n_{eg}L)$, see Eq. (3.65) on Page 78 and Fig. 3.24 on Page 91. The usable gain, however, is limited by the existence of spontaneous emission in the region of the lasing threshold, Eq. (3.83) on Page 82. The power gain $\mathcal{G}$ depends on the field confinement factor Γ, Eq. (4.1) and following remarks on Page 105, and on the carrier concentration n_T, Eq. (3.77) on Page 80. Therefore, $\mathcal{G}$ changes with the polarization of the field (the gain for TE polarization is about $5 \ldots 10\,\mathrm{dB}$ larger in bulk amplifiers than for TM polarization), with the injection current I via n_T, Eq. (3.90) on Page 84, and, because of gain saturation, it varies also with the power level of the input signal. Further, because of amplitude-phase coupling, Eq. (3.116) on Page 90, we find a signal-dependent phase shift of the optical output (a chirp of its frequency). The signal power dependency and/or the chirp can be exploited for frequency conversion[5,6,7,8,9].

At resonance and anti-resonance we find the maximum and minimum gain factors

$$\mathcal{G}_{\max} = \frac{\mathcal{G}_s(1-R_1)(1-R_2)}{(1-\mathcal{G}_s\sqrt{R_1R_2})^2}, \qquad \mathcal{G}_{\min} = \frac{\mathcal{G}_s(1-R_1)(1-R_2)}{(1+\mathcal{G}_s\sqrt{R_1R_2})^2}\,. \tag{4.3}$$

From the ripple of the gain curve the single-pass gain $\mathcal{G}_s\sqrt{R_1R_2}$ can be derived[10,11,12],

$$\mathcal{G}_s\sqrt{R_1R_2} = \frac{\sqrt{\mathcal{G}_{\max}/\mathcal{G}_{\min}}-1}{\sqrt{\mathcal{G}_{\max}/\mathcal{G}_{\min}}+1}\,. \tag{4.4}$$

For a 3 dB ripple we have $\mathcal{G}_s\sqrt{R_1R_2} = 0.17$, so for a single-pass gain $10\lg\mathcal{G}_s = 20\,\mathrm{dB}$ the mean mirror reflection factor must be $\sqrt{R_1R_2} < 0.17\times10^{-2}$. Variations in the single-pass gain $\mathcal{G}_s$ have the more effect the larger $\sqrt{R_1R_2}$ is.

4.1.1 Fabry-Perot amplifier

The bandwidth of a Fabry-Perot amplifier (FPA) is determined by the sharpness of the cavity resonance at f_z. The spectral distance between the half-maximum points of the gain $\mathcal{G}$ inside one single Fabry-Perot mode can be computed from Eqs. (4.2), (4.3),

$$B_{\mathcal{G}} = \frac{c}{\pi n_{eg}L}\arcsin\left(\frac{1-\mathcal{G}_s\sqrt{R_1R_2}}{\sqrt{4\mathcal{G}_s\sqrt{R_1R_2}}}\right) = \frac{c}{\pi n_{eg}L}\arcsin\sqrt{\frac{(1-R_1)(1-R_2)}{4\mathcal{G}_{\max}\sqrt{R_1R_2}}}\,. \tag{4.5}$$

Measuring $B_{\mathcal{G}}$ is an alternative method to determine $\mathcal{G}_s\sqrt{R_1R_2}$. Usually, the arcsin-function can be approximated by its argument. For a Fabry-Perot amplifier the maximum power gain $\mathcal{G}_{\max}$ changes with

[5] Nielsen, M. L.; Nord, M.; Petersen, M. N.; Dagens, B.; Labrousse, A.; Brenot, R.; Martin, B.; Squedin, S.; Renaud, M.: 40 Gbit/s standard-mode wavelength conversion in all-active MZI with very fast response. Electron. Lett. 39 (2003) 20th Feb., No. 4

[6] Nielsen, M. L.; Lavigne, B.; Dagens, B.: Polarity-preserving SOA-based wavelength conversion at 40 Gbit/s using band-pass filtering. Electron. Lett. 39 (2003) 4th Sep., No. 18

[7] Leuthold, J.; Ryf, R.; Maywar, D. N.; Cabot, S.; Jaques, J.; Patel, S. S.: Nonblocking all-optical cross connect based on regenerative all-optical wavelength converter in a transparent demonstration over 42 nodes and 16 800 km. IEEE J. Lightw. Technol. 21 (2003) 2863–2870

[8] Leuthold, J.; Marom, D. M.; Cabot, S.; Jaques, J. J.; Ryf, R.; Giles, C. R.: All-optical wavelength conversion using a pulse reformatting optical filter. IEEE J. Lightw. Technol. 22 (2004) 186–192

[9] Leuthold, J.; Möller, L.; Jaques, J.; Cabot, D.; Zhang, L.; Bernasconi, P.; Cappuzzo, M.; Gomez, L.; Laskowski, E.; Chen, E.; Wong-Foy, A.; Griffin, A.: 160 Gbit/s SOA all-optical wavelength converter and assessment of its regenerative properties. Electron. Lett. 40 (2004) 29th Apr., No. 9

[10] Hakki, B. W.; Paoli, T. L.: Gain spectra in GaAs double-heterostructure injection lasers. J. Appl. Phys. 46 (1975) 1299–1306

[11] Guo, Wei-Hua; Huang, Yong-Zhen; Han, Chun-Lin; Yu, Li-Juan: Measurement of gain spectrum for Fabry–Pérot semiconductor lasers by the Fourier transform method with a deconvolution process. IEEE J. Quantum Electron. 39 (2003) 716–721

[12] Fazludeen, R.; Samit Barai; Prasant Kumar Pattnaik; Srinivas, T.; Selvarajan, A.: A novel technique to measure the propagation loss of integrated optical waveguides. IEEE Photon. Technol. Lett. 17 (2005) 360–362

the operating current, so the product of bandwidth $B_\mathcal{G}$ in mode m_z and maximum amplitude gain $\sqrt{\mathcal{G}_{\max}}$ remains constant,

$$B_\mathcal{G}\sqrt{\mathcal{G}_{\max}} = \frac{c}{2\pi n_{eg} L}\sqrt{\frac{(1-R_1)(1-R_2)}{\sqrt{R_1 R_2}}} = \frac{\Delta f_z}{\pi}\sqrt{\frac{(1-R_1)(1-R_2)}{\sqrt{R_1 R_2}}} = \text{const} \ . \tag{4.6}$$

As before, neighbouring modes are spectrally separated by Δf_z, Eq. (3.65) on Page 78. For semiconductor lasers without antireflection coated facets the power reflection factors are $R_1 = R_2 = 0.32$. If the crystal length is $L = 300\,\mu$m and the effective group index $n_{eg} = 3.5$, a bandwidth-gain product of $B_\mathcal{G}\sqrt{\mathcal{G}_{\max}} = 55\,$GHz results. Such a small bandwidth makes Fabry-Perot amplifiers unsuitable for most lightwave system applications.

4.1.2 Travelling-wave amplifier

Antireflection coatings are necessary for travelling-wave amplifiers (TWA). Minimum reflectivities of $\sqrt{R_1 R_2} = 10^{-5}$ are achieved, but considerable technological effort is required. For this reason, alternative techniques help reducing the reflection feedback. In one method[13], the active-region stripe is tilted from the facet normal. If the vertical bars | denote the facets and the dash — or the slash ⁄ represent the stripe, the conventional amplifier is characterized by |—|, while the angled-facet or tilted-strip structure appears like |⁄|. In practice, tilted stripes with antireflection-coated facets have effective reflectivities down to $R_{1,2} \approx 10^{-4}$. The 3 dB bandwidth of typical amplifiers is $\Delta\lambda_H = 70\,$nm near $\lambda = 1.5\,\mu$m. This corresponds to a frequency bandwidth of $\Delta f_H = 9.3\,$THz and compares well with the rough estimate $\Delta f_H = 12.1\,$THz in Eq. (3.61) on Page 76. Gain values of 28 dB with a residual ripple $< 3\,$dB and a polarization dependence of the gain $< 1\,$dB are achievable.

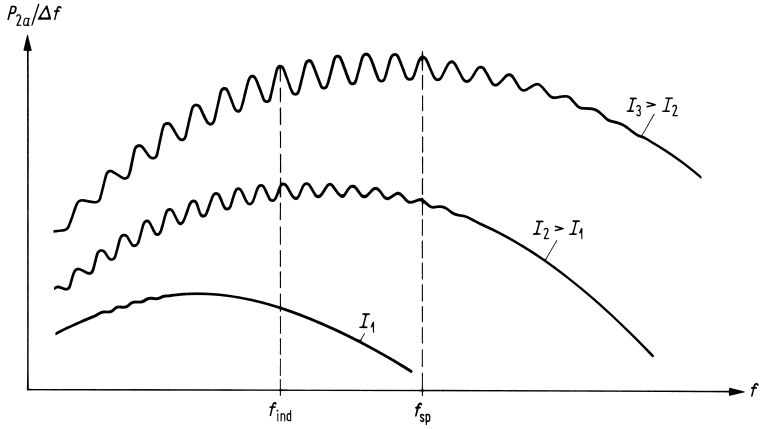


Fig. 4.1. Near-travelling-wave amplifier. Schematic of the spectral output power density $P_{2a}/\Delta F$ of amplified spontaneous emission as transmitted through mirror R_2 for varying injection currents $I_1 < I_2 < I_3$. The frequencies of maximum gain and maximum spontaneous emission are denoted as $f_{\rm ind}$ and $f_{\rm sp}$ for an operating current $I = I_3$.

Figure 4.1 displays the schematic spectral output power density $P_{2a}/\Delta f$ at the exit facet of a near-travelling-wave amplifier. Because of bandfilling at larger injection currents the quasi Fermi levels move deeper into the bands, and the frequencies of maximum spontaneous emission $f_{\rm sp}$ and of maximum induced amplification $f_{\rm ind}$ shift to higher frequencies. This can be also deduced from the diagrams Fig. 3.14 on Page 71, where the point of zero gain $x_0 = (W_{Fn} - W_{Fp} - W_G)/(kT_0)$ depends on the difference $W_{Fn} - W_{Fp}$ of the quasi Fermi levels. As to be seen in Fig. 3.14, the relation $f_{\rm ind} < f_{\rm sp}$ holds.

[13] Zah, C. E.; Osinski, J. S.; Caneau, C.; Menocal, S. G.; Reith, L. A.; Salzman, J.; Shokoohi, F. K.; Lee, T. P.: Electron. Lett. 23 (1987) 990

4.2 Doped fibre amplifier

An alternative to current-pumped semiconductor laser amplifiers are optically pumped Er^{3+}-*d*oped glass *f*ibre *a*mplifiers (EDFA). The optical bandwidth B_O is in the order $B_{\mathrm{OA}} = 4\,\mathrm{THz}$ for an ordinary EDFA. They are commercially available at $\lambda_S = 1.55\,\mu\mathrm{m}$ ($f_S = 193\,\mathrm{THz}$) in the S band (*s*hort-wavelength band, $\lambda \leq 1.528\,\mu\mathrm{m}$), C band (*c*onventional or *c*entral band, $1.528\,\mu\mathrm{m} \leq \lambda \leq 1\,563\,\mu\mathrm{m}$, $\Delta\lambda = 35\,\mathrm{nm}$), and in the L band (*l*ong-wavelength band, $1.563\,\mu\mathrm{m} \leq \lambda \leq 1.606\,\mu\mathrm{m}$, $\Delta\lambda = 43\,\mathrm{nm}$). Combining a C and an S band amplifier[14] the following record results were achieved, Fig. 4.2:

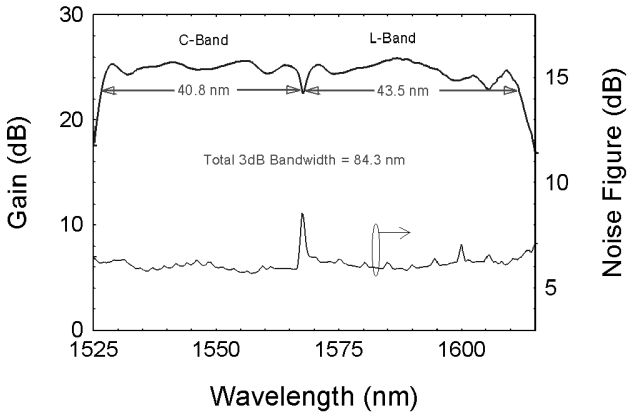


Fig. 4.2. Er-doped optical ultra-broadband amplifier[14] (ECOC'1998)

- With 84.3 nm of bandwidth this ultra wide band amplifier can accommodate 100 WDM (*w*avelength *d*ivision *m*ultiplexing) channels with the proposed ITU (*I*nternational *T*elecommunication *U*nion) standard channel spacing of 100 GHz (0.8 nm @ $\lambda_S = 1.55\,\mu\mathrm{m}$), or 200 WDM channels with 50 GHz (0.4 nm @ $\lambda_S = 1.55\,\mu\mathrm{m}$). There is enough power to support 200 WDM channels. On the other hand, for a given number of WDM channels, this amplifier can be used to allow for a wider channel spacing which may be needed in optical networks with multiple cascaded filters.
- UWB EDFA with a two-section, split band structure. Maximum operating gain is 25 dB and the noise figure is about 6 dB. The output power is very high at 25 dBm, which is required for large numbers of WDM channels. The split bands allow independent optimization of each band for dispersion compensation and span loss variations. The present EDFA, based on field tested erbium-doped silica fiber technology, can be used in Tbit/s capacity DWDM (dense WDM) systems and networks.
- Bandwidth $\Delta\lambda = 84\,\mathrm{nm}$ ($\Delta f = 10\,\mathrm{THz}$ @ $\lambda_S = 1.563\,\mu\mathrm{m}$)
- Noise figure $F = 6\,\mathrm{dB}$, power gain $\mathcal{G}_s = (25 \pm 1.5)\,\mathrm{dB}$, output power $25\,\mathrm{dBm} \mathrel{\widehat{=}} 320\,\mathrm{mW}$
- 100 channels @ 10 Gbit/s @ 400 km, i. e., a length-bandwidth product of $400\,\mathrm{Tbit/s \cdot km}$ (!). The amplifier bandwidth B_{OA} centred at the optical frequency f_S is usually much larger than the signal bandwidth, $B \ll B_{\mathrm{OA}} \ll f_S$. With passive optical filters the optical bandwidth can be varied in the range $100\,\mathrm{GHz} \leq B_O \leq B_{\mathrm{OA}}$.

[14] Sun, Y.; Sulhoff, J. W.; Srivastava, A. K.; Abramov, A.; Strasser, T. A.; Wysocki, P. F.; Pedrazzani, J. R.; Judkins, J. B.; Espindola, R. P.; Wolf, C.; Zyskind, J. L.; Vengsarkar, A. M.; Zhou, J.: A gain-flattened ultra wide band EDFA for high capacity WDM optical communications system. Proc. 24$^{\mathrm{th}}$ Europ. Conf. Opt. Commun. Madrid (ECOC 1998), 20.–24. Sept. 1998. Vol. 1 pp. 53–54 (Lucent Technologies — Bell Laboratories, Holmdel, NJ)

Chapter 5

Optical receivers

The role of an optical receiver is to convert the optical signal back into electrical form, and to recover data transmitted through the lightwave system. Its main component is a photodetector that converts — with a probability η — the received photons to electrons through the photoelectric effect. Because of its speed and sensitivity photoconductive detectors in the form of reverse-biased semiconductor pn-junctions (photodiodes, PD) are commonly used for lightwave systems. Before and after the photodetector there could be additional optical and electronic circuitry, respectively, which will be also discussed in due course. First, we concentrate on the photodiode.

5.1 Pin photodiode

The following sections focus on the pin-photodiode. With *a*valanche *p*hoto*d*iodes (APD) the responsivity can be increased by impact ionization at the cost of additional avalanche noise. However, the combination of optical amplifiers and pin-photodiodes yields a better sensitivity and linearity than the use of an APD.

5.1.1 Basic relations

A reverse-biased pn-junction consists of a region, known as the depletion or space-charge region, that is essentially devoid of free charge carriers and where a large built-in electric field (Eq. (3.30)) opposes flow of electrons from the n-side to the p-side and of holes from the p-side to the n-side. When such a pn-junction is illuminated with light, electron-hole pairs are created through absorption of signal photons with energies $hf_s > W_G$ larger than the bandgap energy W_G of the basic semiconductor material. The probability that a photon generates an electron-hole pair is the quantum efficiency η. Because of the large built-in electric field, electrons and holes generated inside the depletion region accelerate in opposite directions and drift to the regions where they are majority carriers[1,2].

A limiting factor for the bandwidth of pn-photodiodes is the presence of a diffusive component in the photocurrent. Electrons generated in the p-region have to diffuse to the depletion-region boundary before they can drift to the n-side; similarly, holes generated in the n-region must diffuse to the depletion-region boundary. Diffusion is an inherently slow process. Carriers take 1 ns or longer to diffuse over a distance of about 1 μm. In practice, the diffusion contribution depends on the bit rate and becomes negligible by decreasing the widths of the p and n-regions and increasing the depletion-region width so that most of the incident optical power is absorbed inside it. This is the approach adopted for pin-photodiodes, discussed next[3].

A simple way to increase the depletion-region width is to insert a layer of undoped (or lightly doped) semiconductor material between the pn-junction. Since the middle layer consists of (nearly) intrinsic

[1] See Ref. 17 on Page 6, Sect. 4.2.1 p. 141
[2] See Ref. 3 on Page 49, Chapter 7 p. 253 ff.
[3] See Ref. 17 on Page 6, Sect. 4.2.1 p. 143

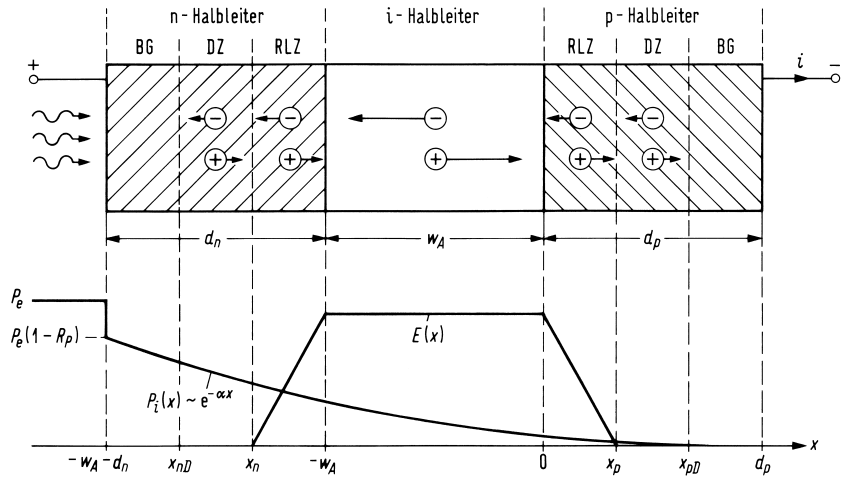


Fig. 5.1. Schematic of a pin-diode. BG contact region (= *B*ahn*g*ebiet), DZ diffusion zone, RLZ space-charge (or depletion) region (= *R*aum*l*adungs*z*one). P_e light power incident from region external of semiconductor, R_P power reflection factor of the semiconductor surface, $P_i(x)$ light power inside the semiconductor, α light power attenuation constant, d_n (d_p) length of n-doped (p-doped) semiconductor, w_A length of intrinsic absorption zone, $E(x)$ x-component of electric bias field. Halbleiter = semiconductor

material, such a structure is referred to as the pin-photodiode[4]. Figure 5.1 shows the device structure (not drawn to scale) together with the electric-field distribution inside it under reverse-bias operation. Because of its intrinsic nature, the middle i-layer offers a high resistance, and most of the voltage drop occurs across it. As a result, a large electric field exists in the i-layer. In essence, the depletion region extends throughout the i-region $-w_A \leq x \leq 0$, and the width w_A of this absorption zone can be controlled by changing the middle-layer thickness. The main difference from the pn-photodiode is that the drift component of the photocurrent dominates over the diffusion component simply because most of the incident power is absorbed inside the i-region (absorption region) of the pin-photodiode.

The external light power P_e is incident perpendicularly to the PD surface and partially reflected with a power reflection factor R_P, see Eq. (3.1) on Page 50. The power inside the semiconductor is attenuated exponentially, see Eq. (1.2) on Page 5. The absorption length corresponding to the reciprocal attenuation constant α for direct semiconductors (e. g., GaAs) is in the order $1/\alpha = 1\,\mu$m, for indirect semiconductors (e. g., Si) it is $1/\alpha = 10 \ldots 20\,\mu$m, see also Fig. 5.4 on Page 114.

A high-speed operation requires small carrier transit times in the depletion region, i. e., small w_A. However, because the depletion-layer capacitance increases with decreasing w_A, an optimum absorption layer width results, and an optimum absorbed power $P_e[1 - \exp(-\alpha w_A)]$ for $R_P = 0$, $d_n = d_p = 0$ has to be found. Thus, high speed (small w_A) means low quantum efficiency and low sensitivity.

To avoid this dilemma, an optical waveguide structure can be used to which the optical signal is edge-coupled, see Fig. 5.7 on Page 119. When the power $P_i(x)$ inside the absorption region changes significantly with x, then the light should be radiated into the direction of the faster moving charge carriers because in this case a larger percentage of the carriers leaves the absorption region faster. In Fig. 5.1 the holes are assumed to drift at a higher velocity than the electrons, $v_p > v_n$. This rule is not necessarily obeyed if discontinuities in the band edges of heterostructures inhibit electrons or holes from leaving the i-region.

Short-circuit photocurrent

We assume a semiconductor without external magnetic field and neglect the magnetic fields associated with flowing currents. The basic equations[5] for semiconductor-device operation can be classified in three

[4] See Ref. 17 on Page 6, Sect. 4.2.2 p. 144

[5] Sze, S. M.: Physics of semiconductor devices. New York: John Wiley & Sons 1985. Chapter 3 p. 70

groups: Continuity equations, current-density or transport equations, and Maxwell's equations. For the *continuity equations* we need the electron and hole concentrations n_T and p, the electron and hole (convection) current densities $\vec{J}_n$ and $\vec{J}_p$, the generation (g_n and g_p) and the recombination rates (r_n and r_p) of electrons and holes,

$$\begin{aligned} \partial p/\partial t + \text{div } \vec{J}_p/e &= g_p - r_p\,, \\ \partial n_T/\partial t - \text{div } \vec{J}_n/e &= g_n - r_n\,. \end{aligned} \tag{5.1}$$

The current-density or *transport equations* consist of the drift component caused by the field and the diffusion component, which in turn is determined by the carrier concentration gradient. Parameters are the drift velocities $\vec{v}_n$ and $\vec{v}_p$, the diffusion constants D_n and D_p, the mobilities μ_n and μ_p for electrons and holes, and the electric field $\vec{E}$,

$$\begin{aligned} \vec{J}_p &= ep\vec{v}_p \quad - eD_p \text{ grad } p, \qquad \vec{v}_p = \mu_p \vec{E}, \\ \vec{J}_n &= -en_T\vec{v}_n + eD_n \text{ grad } n_T, \qquad \vec{v}_n = -\mu_n \vec{E}, \end{aligned} \tag{5.2}$$

It is the Poisson equation which determines important properties of the pn-junction depletion layer. Poisson's equation connects the electric field $\vec{E}$, the total electric charge density ρ, the dielectric displacement $\vec{D}$, and the permittivity ϵ,

$$\text{div } \vec{D} = \rho, \qquad \vec{D} = \epsilon \vec{E}. \tag{5.3}$$

The total current density is source-free,

$$\text{div} \left(\vec{J}_n + \vec{J}_p + \frac{\partial \vec{D}}{\partial t} \right) = 0. \tag{5.4}$$

Equations (5.1)–(5.4) are applied to an i-semiconductor Fig. 5.2 for the one-dimensional case. The electric field due to the externally connected voltage is so high that electrons and holes drift with their saturation velocities v_n and v_p. The recombination rates r_n and r_p are negligibly small because the carrier lifetime is much larger than the drift time in the absorption zone $-w_A \leq x \leq 0$. Diffusion currents can be neglected compared to drift currents. Photogeneration dominates the carrier generation processes, $g_p = g_n = g$.

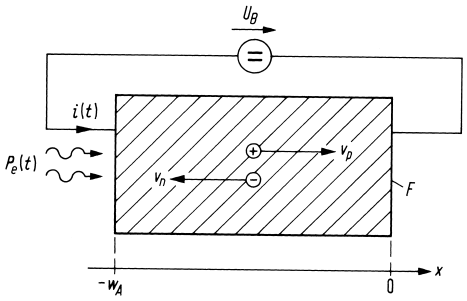


Fig. 5.2. i-layer of a pin-photodiode (one-dimensional case, cross-section area F). Saturation drift velocities $v_n > 0$ and $v_p > 0$ for electrons and holes, incident external optical power $P_e(t)$, total conduction current $i(t)$, open-circuit voltage U_B of a battery with an internal resistance of zero

Introducing convection and conduction currents i instead of current densities J (cross-section area F) we find from Eq. (5.1)–(5.4)

$$\begin{aligned} \frac{1}{v_p}\frac{\partial i_p}{\partial t} + \frac{\partial i_p}{\partial x} &= eFg\,, \qquad i_p = Fepv_p\,, \\ \frac{1}{v_n}\frac{\partial i_n}{\partial t} - \frac{\partial i_n}{\partial x} &= eFg\,, \qquad i_n = Fen_T v_n\,, \end{aligned} \tag{5.5}$$

and

$$\epsilon \frac{\partial E}{\partial x} = e(p - n_T), \qquad \frac{\partial}{\partial x}\left(i_n + i_p + F\epsilon \frac{\partial E}{\partial t}\right) = 0\,. \tag{5.6}$$

The total time-dependent conduction current is

$$i(t) = i_n(x,t) + i_p(x,t) + F\epsilon \frac{\partial E(x,t)}{\partial t}\,. \tag{5.7}$$

Because of

$$\int_{-w_A}^{0} E(x,t)\,\mathrm{d}x = U_B = \text{const} \quad \rightarrow \quad \int_{-w_A}^{0} \frac{\partial E(x,t)}{\partial t}\,\mathrm{d}x = \frac{\mathrm{d}\,U_B}{\mathrm{d}t} = 0\,, \tag{5.8}$$

we find the total conduction current in the external circuit (i. e., the external short-circuit current, $\mathrm{d}\,U_B = 0$) as an average of the sum of the carrier convection currents in the drift region $-w_A \le x \le 0$,

$$i(t) = \frac{1}{w_A} \int_{-w_A}^{0} [\, i_n(x,t) + i_p(x,t)\,]\,\mathrm{d}x\,. \tag{5.9}$$

The carrier transit times τ_n, τ_p are defined by

$$w_A = v_n\tau_n = v_p\tau_p\,. \tag{5.10}$$

The total number of electrons and holes in the semiconductor are

$$N_n(t) = F \int_{-w_A}^{0} n_T(x,t)\,\mathrm{d}x\,, \qquad N_p(t) = F \int_{-w_A}^{0} p(x,t)\,\mathrm{d}x\,. \tag{5.11}$$

From Eqs. (5.9)–(5.11) we calculate with the help of Eq. (5.5) ($i_p = Fepv_p$, $i_n = Fen_Tv_n$) the total conduction current,

$$i(t) = \frac{e}{\tau_n} N_n(t) + \frac{e}{\tau_p} N_p(t). \tag{5.12}$$

If the irradiated n-region is much shorter than the absorption length (i. e., $\alpha\, d_n \to 0$ in Fig. 5.1), the light power in the i-region $P_i(x,t)$ reads

$$P_i(x,t) = P_e(t)\,(1 - R_P)\,\mathrm{e}^{-\alpha(x+w_A)}\,. \tag{5.13}$$

Any light propagation times are neglected. P_e, P_i are classical optical powers averaged over a few optical cycles. The power fraction which is absorbed inside the i-zone represents the quantum efficiency η of the photodiode,

$$\eta = \frac{P_i(-w_A,t) - P_i(0,t)}{P_e(t)} = (1 - R_P)\left(1 - \mathrm{e}^{-\alpha w_A}\right). \tag{5.14}$$

The mean generation rate of electron-hole pairs equals the mean absorption rate of photons, i. e., the absorbed power per quantum energy hf_e. Because photodiodes cannot emit optical power, no second harmonic light frequency is generated as it is common for classical microwave detectors. The generation rate g (unit $\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$) is

$$g(x,t) = -\frac{1}{Fhf_e}\frac{\partial P_i(x,t)}{\partial x} = \frac{\alpha P_i(x,t)}{Fhf_e}\,. \tag{5.15}$$

With Eq. (5.13) (substitute $P_i(x,t)$) and Eq. (5.14) (substitute $(1 - R_P)$) we find

$$eFg(x,t) = \frac{\eta e}{hf_e}\,P_e(t)\,\frac{\alpha\,\mathrm{e}^{-\alpha(x+w_A)}}{1 - \mathrm{e}^{-\alpha w_A}}\,. \tag{5.16}$$

Equivalent electrical circuit

For $\mathrm{d}\,/\,\mathrm{d}t = 0$ with $P_e(t) \equiv P_e = \mathrm{const}_t$ the Eqs. (5.5), (5.16) for the static short-circuit current $i(t) \equiv i = \mathrm{const}_t$ can be easily solved. We integrate Eq. (5.5), (5.16) with the notation $f(x,t) \equiv f(x)$ for space and time dependent functions f in the stationary case. Further, we observe that in Figs. 5.1, 5.2 the minority current injection can be neglected, $i_p(-w_A) = 0$, $i_n(0) = 0$,

$$i = i_p(0) = i_n(-w_A) = \int_{-w_A}^{0} eFg(x)\,\mathrm{d}x = \frac{\eta e}{hf_e}\,P_e\,, \quad i = SP_e\,, \quad S = \frac{\eta e}{hf_e}\,, \quad \frac{S}{\mathrm{A\,/\,W}} = 0.806\,\eta\,\frac{\lambda_e}{\mu\mathrm{m}}\,. \tag{5.17}$$

The quantity S is called photodetector sensitivity (responsivity). The absorbed power is ηP_e and corresponds to a photon absorption rate of $\eta P_e/(hf_e)$. Each absorbed photon generates an electron-hole pair leading to the transport of one elementary charge e through the external circuit. The rate of generated charges i/e equals the photon absorption rate $\eta P_e/(hf_e)$.

The sensitivity of a photodiode increases with the wavelength λ_e simply because more photons are present for the same optical power. Such a linear dependence on λ_e is not expected to continue forever, since eventually the photon energy hf_e becomes smaller than the bandgap energy W_G. The quantum efficiency η then drops to zero[6]. The dependence of η on λ_e enters through the absorption coefficient α, Eq. (5.14) and Fig. 5.4 on Page 114.

For the time-dependent case we substitute Eq. (5.16) in Eq. (5.5). The current $i(t)$ as calculated from Eq. (5.9) or Eq. (5.12) represents the short-circuit current which feeds the equivalent circuit of the electrical embedding network of the photodiode Fig. 5.3. The Fourier transforms of $i(t)$, $i_a(t)$ are denoted

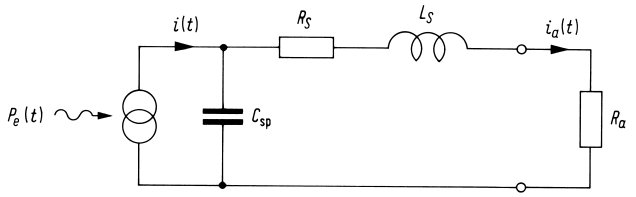


Fig. 5.3. Equivalent electrical circuit of a photodiode. P_e incident external optical power, C_{sp} depletion-layer capacity, R_S series resistance, L_S series inductance, R_a load resistance

as $I(f)$, $I_a(f)$. The transfer function $H_S(f)$ from the photodiode to the load resistance R_a reads

$$\begin{aligned} H_S(f) &= \frac{I_a(f)}{I(f)} = \frac{\omega_r^2}{(\mathrm{j}\,\omega)^2 + 2\gamma_r(\mathrm{j}\,\omega) + \omega_r^2}\,, \\ \omega_r^2 &= \frac{1}{L_S C_{\mathrm{sp}}}\,, \qquad 2\gamma_r = \frac{R_S + R_a}{L_S}\,. \end{aligned} \tag{5.18}$$

The transfer function $H_S(f)$ has the same structure as the small-signal transfer function of the current-modulated laser diode Typical values of the circuit elements are: $C_{\mathrm{sp}} = 0.04 \ldots 0.2\,\mathrm{pF}$, $R_S = 10 \ldots 50\,\Omega$, $L_S = 0.15 \ldots 0.5\,\mathrm{nH}$, $R_a = 50\,\Omega$. The reverse diode voltage depends on the material and on the width if the depletion-layer and is in the order of a few volts. The depletion-layer capacitance C_{sp} can be calculated with the formula for a parallel-plate capacitor, Eq. (5.39). Relative dielectric constants ϵ_r are specified in Sect. 5.1.2. The cross-sectional area F of a photodiode is usually circular having a diameter in the order $7 \ldots 200\,\mu\mathrm{m}$, typically in the range $10 \ldots 80\,\mu\mathrm{m}$.

5.1.2 Materials

Materials commonly used to make photodiodes can be elemental or compound semiconductors. From the elemental semiconductors Ge ($\epsilon_r = 16$) is suitable in the long-wavelength region (direct semiconductor

[6]See Ref. 17 on Page 6. Sect. 4.1 p. 139

for $\lambda < 1.55\,\mu$m, indirect for $\lambda < 1.85\,\mu$m), while Si exhibits excellent properties in the short-wavelength and visible range ($\epsilon_r = 11.7$, direct for $\lambda < 0.36\,\mu$m, indirect for $\lambda < 1.1\,\mu$m).

Compound semiconductors are common in the long-wavelength domain. InP substrates (transparent for $\lambda > 0.92\,\mu$m, $\epsilon_r = 12.35$) with an lattice-matched $In_{0.53}Ga_{0.47}As$ absorption layer are used (in short InGaAs, good for $\lambda < 1.65\,\mu$m, $\epsilon_r = 13.6$).

The bandgap difference $|\Delta W_G| = 0.6\,\text{eV}$ (InP: $W_G = 1.35\,\text{eV}$; InGaAs: $W_G = 0.75\,\text{eV}$) leads to discontinuities for the CB edge of $|\Delta W_L| = 0.2\,\text{eV}$ and for the VB edge of $|\Delta W_V| = 0.4\,\text{eV}$. For an isotype nN-junction between weakly n-doped InGaAs and n-doped InP the built-in voltage of typically $U_D = 0.22\,\text{eV}$ has to be added. Thus, the CB edges of InGaAs and InP on both sides of the contact have nearly the same energy levels. The remaining CB spike is narrow, and electrons may easily penetrate the barrier by tunneling. However, VB holes injected from InGaAs into InP see a VB potential barrier which increased to $|\Delta W_V| + U_D = 0.62\,\text{eV}$.

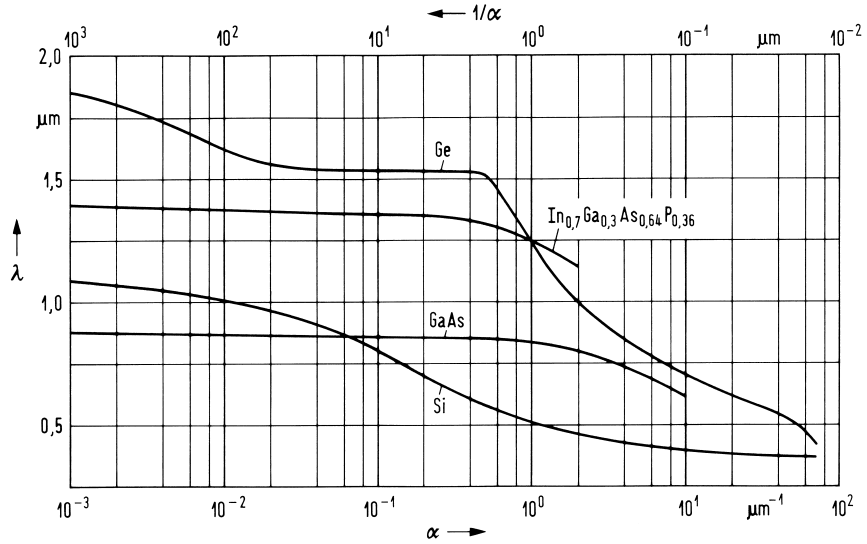


Fig. 5.4. Wavelength dependence of the absorption constant α (penetration depth $1/\alpha$) for several semiconductor materials

Figure 5.4 shows the wavelength dependence of the absorption constant α and of the penetration depth $1/\alpha$ for Ge, Si, GaAs and $(In_{0.7}Ga_{0.3})(As_{0.64}P_{0.36})$; these materials are commonly used to make photodiodes for lightwave systems. Some special values for InGaAs are: $\alpha = 0.68\,\mu\text{m}^{-1}$ at $\lambda = 1.55\,\mu$m, $\alpha = 1.16\,\mu\text{m}^{-1}$ at $\lambda = 1.36\,\mu$m, $\alpha = 2.15\,\mu\text{m}^{-1}$ at $\lambda = 1.06\,\mu$m.

5.1.3 Time and frequency response

The response time of a pin-photodiode is determined by the speed with which it responds to variations of the incident optical power. The absorption layer of a pin-photodiode is displayed in Fig. 5.5. $P_e(t)$ is the incident external light power. The quantum efficiency η was defined in Eq. (5.14). The internal light power dependence $P_i(x,t)$ in the i-region is given by Eq. (5.13).

If we set formally (irrespective of the physical units) $P_e(t) = \delta(t)$, then the short-circuit "current" $i(t)$ computed from Eq. (5.9), (5.12) is denoted as impulse response $h_P(t;\text{pin})$ (unit A / Ws; the subscript P relates the impulse response to an impulse of the optical power). We substitute the generation term eFg from Eq. (5.16) using $P_e(t) = \delta(t)$ into Eq. (5.5), and integrate over the small interval $-\Delta t \le t \le \Delta t$. The currents before the power impulse are zero, and $\int_{-\Delta t}^{+\Delta t} \frac{\partial i_p}{\partial x}\,\mathrm{d}t = \frac{\partial}{\partial x}\int_{-\Delta t}^{+\Delta t} i_p\,\mathrm{d}t \to 0$ for $\Delta t \to 0$ because

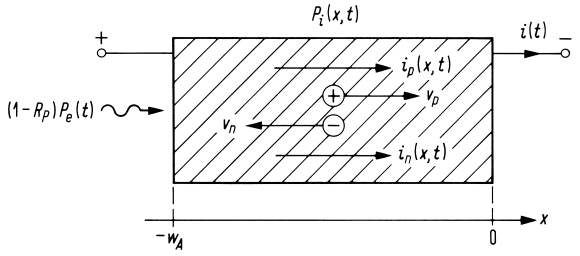


Fig. 5.5. Absorption layer of a pin-photodiode. P_e incicent external light power, R_P power reflection coefficient, $i(t)$ external short-circuit current, P_i internal optical power; i_p, i_n convection currents of electrons and holes; v_p, v_n saturation drift velocities, w_A length of absorption region

$i_p(x,t)$ has no singularity,

$$\frac{1}{v_p}\int_{-\Delta t}^{+\Delta t}\frac{\partial i_p}{\partial t}\,\mathrm{d}t+\int_{-\Delta t}^{+\Delta t}\frac{\partial i_p}{\partial x}\,\mathrm{d}t=\frac{\eta e}{hf_e}\,\frac{\alpha\,\mathrm{e}^{-\alpha(x+w_A)}}{1-\mathrm{e}^{-\alpha w_A}}\int_{-\Delta t}^{+\Delta t}\delta(t)\,\mathrm{d}t\quad\text{for }\Delta t\to 0\,, \tag{5.19}$$

$$\frac{1}{v_p}\Big(i_p(x,+0)-\underbrace{i_p(x,-0)}_{=0}\Big)+\frac{\partial}{\partial x}\underbrace{\int_{-\Delta t}^{+\Delta t}i_p(x,t)\,\mathrm{d}t}_{=0\text{ for }\Delta t\to 0}=\frac{\eta e}{hf_e}\,\frac{\alpha\,\mathrm{e}^{-\alpha(x+w_A)}}{1-\mathrm{e}^{-\alpha w_A}}$$

For $\Delta t\to 0$ we find as an initial condition the convection "currents" at $t=+0$,

$$\frac{1}{v_p}\,i_p(x,+0)=\frac{1}{v_n}\,i_n(x,+0)=\frac{\eta e}{hf_e}\,\frac{\alpha\,\mathrm{e}^{-\alpha(x+w_A)}}{1-\mathrm{e}^{-\alpha w_A}}\,. \tag{5.20}$$

For the δ-excitation $g(x,t>0)=0$ is valid. The homogeneous differential equations (5.5) are solved by arbitrary functions $i_p(x,t)=i_p(x-v_pt)$, $i_n(x,t)=i_n(x+v_nt)$ which fulfill the initial conditions Eq. (5.20). These initial carrier distributions drift to the right and to the left with the saturation velocities v_p and v_n, respectively. With the Heaviside function $H(z)$ and the carrier transit times τ_n,τ_p defined by $w_A=v_p\tau_p=v_n\tau_n$ we calculate

$$\left.\begin{matrix}i_p(x,t)\\ i_n(x,t)\end{matrix}\right\}=\frac{\eta e}{hf_e}\,\frac{\alpha\,\mathrm{e}^{-\alpha w_A}}{1-\mathrm{e}^{-\alpha w_A}}\left\{\begin{matrix}v_p\,\mathrm{e}^{-\alpha(x-v_pt)}\ \times\\ v_n\,\mathrm{e}^{-\alpha(x+v_nt)}\ \times\end{matrix}\right.\\ \times\big[\,H(t)-H(t-\tau_p)\,\big]\big[\,H(x-v_pt+w_A)-H(x)\,\big],\\ \times\big[\,H(t)-H(t-\tau_n)\,\big]\big[\,H(x+w_A)-H(x+v_nt)\,\big]. \tag{5.21}$$

Substituting these "currents" into Eq. (5.9), $i(t)=\frac{1}{w_A}\int_{-w_A}^{0}\big[\,i_n(x,t)+i_p(x,t)\,\big]\,\mathrm{d}x$, we find for the impulse response

$$h_P(t;\mathrm{pin})=\frac{\eta e}{hf_e}\,\frac{1}{1-\mathrm{e}^{-\alpha w_A}}\left\{\frac{1-\mathrm{e}^{-\alpha(w_A-v_pt)}}{\tau_p}\big[\,H(t)-H(t-\tau_p)\,\big]+\right.\\ \left.+\,\frac{\mathrm{e}^{-\alpha v_nt}-\mathrm{e}^{-\alpha w_A}}{\tau_n}\big[\,H(t)-H(t-\tau_n)\,\big]\right\}, \tag{5.22}$$

$$\int_{-\infty}^{+\infty}h_P(t;\mathrm{pin})\,\mathrm{d}t=\frac{\eta e}{hf_e}\,.$$

The total "charge" transported in an external circuit, originating from an optical "power" $P_e(t)=\delta(t)$, is $\eta e/(hf_e)$. The corresponding absorbed light "energy" is η. Thus, $\eta/(hf_e)$ represents the absorbed

photon "number". Each absorbed photon generates an electron-hole pair leading to the transport of an elementary charge e through the external circuit.

The Fourier transform $H_P(f;\mathrm{pin})$ of the impulse response $h_P(t;\mathrm{pin})$ is the transfer function of the pin-photodiode,

$$H_P(f;\mathrm{pin}) = \frac{\eta e}{h f_e}\frac{1}{1-\mathrm{e}^{-\alpha w_A}}\left[\frac{1-\mathrm{e}^{-\mathrm{j}\,\omega\tau_p}}{\mathrm{j}\,\omega\tau_p} + \frac{\mathrm{e}^{-\alpha w_A}-\mathrm{e}^{-\mathrm{j}\,\omega\tau_p}}{\alpha w_A - \mathrm{j}\,\omega\tau_p} + \right.$$
$$\left. + \frac{1-\mathrm{e}^{-\alpha w_A}\,\mathrm{e}^{-\mathrm{j}\,\omega\tau_n}}{\alpha w_A + \mathrm{j}\,\omega\tau_n} - \mathrm{e}^{-\alpha w_A}\,\frac{1-\mathrm{e}^{-\mathrm{j}\,\omega\tau_n}}{\mathrm{j}\,\omega\tau_n}\right]. \tag{5.23}$$

Thus, the external short-circuit current $i(t;\mathrm{pin})$ and its frequency response $I(f;\mathrm{pin})$ for an arbitrary time dependent illumination $P_e(t)$ are

$$\begin{aligned} i(t;\mathrm{pin}) &= \int_{-\infty}^{+\infty} P_e(t')\,h_P(t-t';\mathrm{pin})\,\mathrm{d}t', \\ I(f;\mathrm{pin}) &= \breve{P}_e(f) H_P(f;\mathrm{pin}) \end{aligned} \tag{5.24}$$

$\breve{P}_e(f)$ is the Fourier transform of $P_e(t)$. According to Fig. 5.3 and Eq. (5.18) the transfer function $H_S(f)$ from the photodiode to the load resistance is

$$I_a(f) = \breve{P}_e(f) H_P(f;\mathrm{pin}) H_S(f). \tag{5.25}$$

Strong absorption In the limiting case of strong absorption $\alpha w_A \to \infty$ all the light power is absorbed inside an infinitely thin layer at $x = -w_A$. The quantum efficiency η (Eq. (5.14)) simplifies to

$$\eta = 1 - R_P, \qquad (\alpha w_A \to \infty) \tag{5.26}$$

With an antireflection coating $\eta \to 1$ is achievable. The initial convection currents Eq. (5.20) valid for $-w_A \le x \le 0$ are

$$\frac{1}{v_p}\,i_p(x,+0) = \frac{1}{v_n}\,i_n(x,+0) = \frac{\eta e}{h f_e}\,\frac{\alpha\,\mathrm{e}^{-\alpha(x+w_A)}}{1-\mathrm{e}^{-\alpha w_A}}\,[H(x+w_A) - H(x)]\,. \tag{5.27}$$

With $\lim_{\alpha\to\infty}\left[\alpha\,\mathrm{e}^{-\alpha(x-v_p t+w_A)}\,H(x-v_p t+w_A)\right] = \delta(x-v_p t+w_A)$ we calculate[7] the convection "current" response $i_p(x,t)$ from Eq. (5.21),

$$\begin{aligned} i_p(x,t) &= \frac{\eta e}{h f_e}\,\frac{\alpha\,\mathrm{e}^{-\alpha w_A}}{1-\mathrm{e}^{-\alpha w_A}}\,v_p\,\mathrm{e}^{-\alpha(x-v_p t)}\Big[H(t) - H(t-\tau_p)\Big]\Big[H(x-v_p t+w_A) - H(x)\Big] \\ &= \frac{\eta e}{h f_e}\,\frac{v_p}{1-\mathrm{e}^{-\alpha w_A}}\Big[H(t) - H(t-\tau_p)\Big]\,\mathrm{e}^{-\alpha(x-v_p t+w_A)}\Big[H(x-v_p t+w_A) - H(x)\Big], \end{aligned}$$

[7] The meaning of $\alpha\,\mathrm{e}^{-\alpha(x)}\,H(x)$ may be seen through an integration by parts, $\int uv'\,\mathrm{d}x = uv - \int u'v\,\mathrm{d}x$,

$$\begin{aligned} \lim_{\alpha\to\infty}\alpha\int_{-\infty}^{+\infty}\mathrm{e}^{-\alpha x}\,H(x)\,\Phi(x)\,\mathrm{d}x &= \lim_{\alpha\to\infty}\alpha\int_{0}^{+\infty}\Phi(x)\,\mathrm{e}^{-\alpha x}\,\mathrm{d}x \\ &= \lim_{\alpha\to\infty}\alpha\left[\Phi(x)\frac{-1}{\alpha}\,\mathrm{e}^{-\alpha x}\Big|_0^{+\infty} - \int_0^{+\infty}\Phi'(x)\frac{-1}{\alpha}\,\mathrm{e}^{-\alpha x}\,\mathrm{d}x\right] \\ &= \lim_{\alpha\to\infty}\alpha\left[\Phi(0)\frac{1}{\alpha} - \Phi'(0)\left(\frac{-1}{\alpha}\right)^2 + \int_0^{+\infty}\Phi''(x)\left(\frac{-1}{\alpha}\right)^2\mathrm{e}^{-\alpha x}\,\mathrm{d}x\right] \\ &= \Phi(0) = \int_{-\infty}^{+\infty}\delta(x)\,\Phi(x)\,\mathrm{d}x \quad \rightsquigarrow \quad \delta(x) = \lim_{\alpha\to\infty}\left[\alpha\,\mathrm{e}^{-\alpha x}\,H(x)\right]. \end{aligned} \tag{5.28}$$

Thus, the initial conditions for the convection currents (i. e., the carrier concentrations) in the region $-w_A \le x \le 0$ are in proportion to $\delta(x+w_A)$.

which finally leads to

$$i_p(x,t) = \frac{\eta e}{h f_e}\,\frac{v_p}{1-\mathrm{e}^{-\alpha w_A}}\Big[\,H(t)-H(t-\tau_p)\,\Big] \qquad (5.29)$$
$$\times\Big[\underbrace{\alpha\,\mathrm{e}^{-\alpha(x-v_p t+w_A)}\;H(x-v_p t+w_A)}_{=\delta(x-v_p t+w_A)\ \text{for}\ \alpha\to\infty} - \underbrace{\alpha\,\mathrm{e}^{-\alpha x}\,H(x)\,\mathrm{e}^{\alpha(v_p t-w_A)}}_{=0\ \text{for}\ \alpha\to\infty}\Big]$$
$$= \{\alpha\to\infty\} = \frac{\eta e}{h f_e}\,v_p\Big[\,H(t)-H(t-\tau_p)\,\Big]\,\delta(x-v_p t+w_A).$$

An analogous result can be found for $i_n(x,t)$. With Eq. (5.9) the total external current is $i(t) = \frac{1}{w_A}\times \int_{-w_A}^{0}\big[\,i_n(x,t)+i_p(x,t)\,\big]\,\mathrm{d}x$. For the hole current, $\int_{-w_A}^{0}\delta(x-v_p t+w_A)\,\mathrm{d}x = \int_{-v_p t}^{-v_p t+w_A}\delta(\xi)\,\mathrm{d}\xi = \{t=0\} = \int_{-0}^{+w_A}\delta(\xi)\,\mathrm{d}\xi = 1$ holds. The electron current contribution disappears for $t>0$. From Eqs. (5.22), (5.23) we calculate the external current impulse and frequency responses

$$\left.\begin{aligned} h_P(t;\mathrm{pin}) &= \begin{cases} \dfrac{\eta e}{h f_e}\left(\dfrac{1}{\tau_p}+\dfrac{1}{\tau_n}\right) & (t=0),\\ \dfrac{\eta e}{h f_e \tau_p}\,[\,H(t)-H(t-\tau_p)\,] & (t>0), \end{cases}\\ H_P(f;\mathrm{pin}) &= \frac{\eta e}{h f_e}\;\mathrm{e}^{-\mathrm{j}\,\omega\tau_p/2}\;\frac{\sin(\omega\tau_p/2)}{\omega\tau_p/2}\,, \end{aligned}\right\}\quad (\alpha w_A\to\infty). \qquad (5.30)$$

Weak absorption For the limiting case of weak absorption $\alpha w_A \to 0$ the responses are

$$\left.\begin{aligned} h_P(t;\mathrm{pin}) &= \frac{\eta e}{h f_e}\left\{\frac{1-t/\tau_p}{\tau_p}\,[\,H(t)-H(t-\tau_p)\,] + \right.\\ &\qquad \left. + \frac{1-t/\tau_n}{\tau_n}\,[\,H(t)-H(t-\tau_n)\,]\right\},\\ H_P(f;\mathrm{pin}) &= \frac{\eta e}{\mathrm{j}\,\omega\tau_p h f_e}\left[1-\mathrm{e}^{-\mathrm{j}\,\omega\tau_p/2}\;\frac{\sin(\omega\tau_p/2)}{\omega\tau_p/2}\right] +\\ &\qquad + \frac{\eta e}{\mathrm{j}\,\omega\tau_n h f_e}\left[1-\mathrm{e}^{-\mathrm{j}\,\omega\tau_n/2}\;\frac{\sin(\omega\tau_n/2)}{\omega\tau_n/2}\right]. \end{aligned}\right\}\quad (\alpha w_A\to 0). \qquad (5.31)$$

Figure 5.6 shows the external current impulse responses of a pin-photodiode for the limiting cases of strong ($\alpha w_A\to\infty$) and weak absorption ($\alpha w_A\to 0$) assuming $v_p = v_n = w_A/\tau$.

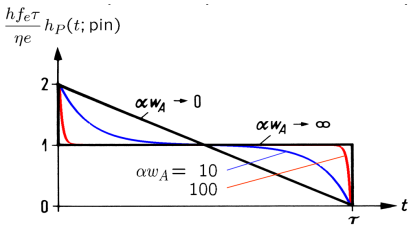


Fig. 5.6. Transit-time limited impulse responses of a pin-photodiode for $v_p = v_n$ ($\tau_p = \tau_n = \tau$) and the cases of strong absorption ($\alpha w_A\to\infty$, in practice $\alpha w_A \geq 1000$) and weak absorption ($\alpha w_A\to 0$, in practice $\alpha w_A \leq 1$). Intermediate-absorption graphs for $\alpha w_A = 100$ and $\alpha w_A = 10$

5.1.4 Cutoff frequency, quantum efficiency and responsivity

The transfer function $H_S(f)$ from the photodiode to the load resistance R_a has the same structure as the small-signal transfer function of the current-modulated laser diode. We are only interested in the short-circuit current $I(f;\mathrm{pin})$ originating from the alternating part of the sinusoidal light power modulation,

$$P_e(t) = P_0 + P_1 \cos(\omega t) \quad (P_0 \geq P_1 = \text{const}_t) \tag{5.32}$$

Again, the limiting cases of strong (Eq. (5.30)) and weak absorption (Eq. (5.31), $\tau_n = \tau_p = \tau$) are regarded,

$$\begin{aligned} \frac{I(f;\text{pin})}{I(0;\text{pin})} &= \mathrm{e}^{-\mathrm{j}\,\omega\tau_p/2}\ \frac{\sin(\omega\tau_p/2)}{\omega\tau_p/2}\,, && (\alpha w_A \to \infty), \\ \frac{I(f;\text{pin})}{I(0;\text{pin})} &= \frac{1}{\mathrm{j}\,\omega\tau/2}\left[1 - \mathrm{e}^{-\mathrm{j}\,\omega\tau/2}\ \frac{\sin(\omega\tau/2)}{\omega\tau/2}\right], && (\alpha w_A \to 0). \end{aligned} \tag{5.33}$$

Transit-time cutoff frequency The transit-time limited 3-dB cutoff frequency is computed from

$$\left|\frac{I(f_{3\,\text{dB}};\text{pin})}{I(0;\text{pin})}\right| = \frac{1}{\sqrt{2}} \tag{5.34}$$

resulting in

$$f_{3\,\text{dB}} = \begin{cases} 0.44/\tau_p & (\alpha w_A \to \infty), \\ 0.55/\tau & (\alpha w_A \to 0,\ \tau_n = \tau_p = \tau). \end{cases} \tag{5.35}$$

Quantum efficiency For weak absorption $\alpha w_A \to 0$ we see from Eq. (5.14) that the quantum efficiency is $\eta = (1 - R_P)\alpha w_A$ (compare $\eta = 1 - R_P$ for $\alpha w_A \to \infty$, Eq. (5.26)), thus the product of quantum efficiency and 3-dB cutoff frequency is

$$\begin{aligned} \eta &= (1 - R_P)\,(1 - \mathrm{e}^{-\alpha w_A}) && (\text{Eq. (5.14)}), \\ \eta f_{3\,\text{dB}} &= 0.55\,(1 - R_P)\,\alpha v\,, && (\alpha w_A \to 0). \end{aligned} \tag{5.36}$$

Assuming $R_P = 0$ the efficiency-bandwidth products depends only on the material parameters α (absorption constant) and v (saturation velocity of carriers). With the InGaAs data of Sect. 5.1.2 at $\lambda = 1.55\,\mu$m ($\alpha = 0.68\,\mu\text{m}^{-1}$, $v = (v_n + v_p)/2 = 56.5\,\mu\text{m}\,/\,\text{ns}$) an efficiency-bandwidth product $\eta f_{3\,\text{dB}} = 21\,\text{GHz}$ results, for $\lambda = 1.36\,\mu$m ($\alpha = 1.16\,\mu\text{m}^{-1}$, $v = 56.5\,\mu\text{m}\,/\,\text{ns}$) it is $\eta f_{3\,\text{dB}} = 36\,\text{GHz}$.

Edge-coupling Better $\eta f_{3\,\text{dB}}$-values can be achieved if the absorption length for photons along a waveguide and the transport distance of charge carriers is decoupled by an edge-coupled photodiode, Fig. 5.7. The active zone is an InGaAs absorbing layer embedded between n-InP and p-InP heterolayers. The structure is operated with reverse bias. The light is coupled into the vertical InP/InGaAs/InP slab waveguide along the z-direction. An InGaAs layer with a height of $w_A = 0.2\,\mu$m results in a field confinement factor of $\Gamma = 0.4$. If InGaAs has an absorption coefficient α, the waveguide shows an effective absorption constant $\Gamma\alpha$ for the fundamental mode propagating into the z-direction. The pin-photodiode has a z-extension of L. The input coupling efficiency from an external source into the waveguide is η_{coupl}. The quantum efficiency of the photodiode is

$$\eta = \eta_{\text{coupl}}\left(1 - \mathrm{e}^{-\alpha\Gamma L}\right). \tag{5.37}$$

With the above values for α, Γ and waveguide lengths $L > 10\,\mu$m the quantum efficiency is $\eta \approx \eta_{\text{coupl}}$. Coupling efficiencies of $\eta_{\text{coupl}} = 0.8$ are feasible. The cutoff frequency for small absorption layers w_A is given by the second line of Eq. (5.35), $f_{3\,\text{dB}} = 0.55/\tau$. Independently of $f_{3\,\text{dB}}$ and of the operating wavelength the quantum efficiency is near $\eta \approx 0.8$. For $w_A = 0.2\,\mu$m the transit-time limited cutoff frequency would be $f_{3\,\text{dB}} = 124\,\text{GHz}$. The bandwidth of such waveguide photodiodes is limited by the $R_a C_{\text{sp}}$ time constant in Fig. 5.3 which can be decreases by controlling the waveguide cross-section area.

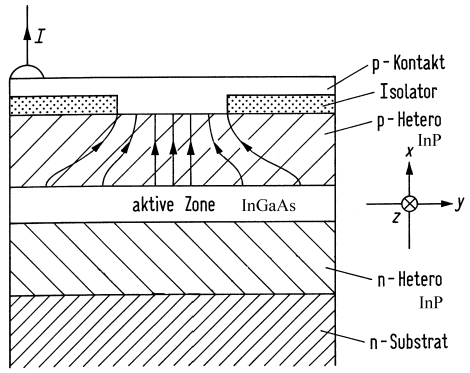


Fig. 5.7. Pin-diode with edge-coupling of light, w_A is the height of the active zone. x-axis: direction of current flow; z-axis: direction of light propagation (Aktive Zone = active zone, p-Kontakt = p-contact, Isolator = insulator, n-substrat = n-substrate).

Diffusion cutoff frequency When absorption inside the diffusion zones must not be neglected (see Fig. 5.1 for a schematic of the pin-photodiode) the cutoff frequency is limited by relatively slow diffusion processes. Carriers having a diffusion constant D move by a mean distance $\Delta x = \sqrt{D\tau}$ inside a time interval τ (random walk). For the case of Fig. 5.1 $\Delta x = x_n - x_{nD} = L_n = \sqrt{D_n \tau_n}$ would be given by the diffusion length L_n, and $\tau = \tau_n$ would equal the minority lifetime τ_n. If the n-doped layer is short (the contact region has a length of zero, the diffusion zone has a length $d_{\text{dif}} < L_n$) we have $\Delta x = d_{\text{dif}}$. The diffusion-limited cutoff frequency is defined by

$$f_{\text{dif}} = \frac{1}{\tau} = \frac{D}{(\Delta x)^2} = \frac{\mu k T_0}{e(\Delta x)^2}\,. \tag{5.38}$$

Assuming a short semiconductor with $\Delta x = 0.2\,\mu\text{m}$ and $\mu_n = 8\,500\,\text{cm}^2\,/\,(\text{V s})$ ($1\,500\,\text{cm}^2\,/\,(\text{V s})$) for GaAs (Si) diffusion-limited cutoff frequencies are computed to be $f_{\text{dif}} = 540\,\text{GHz}$ (94 GHz). For this example the actual cutoff frequency will not be diffusion-limited. However, for indirect semiconductors the minority carrier lifetime increases up to $\tau = 1\,\text{ms}$. Therefore the cutoff frequency can be as small as $f_{\text{dif}} = 1\,\text{kHz}$.

5.1.5 Device structures

The depletion-layer width w and the depletion-layer capacitance C_{sp} for an abrupt pn-junction can be computed if the impurity concentrations n_A, n_D, the intrinsic density n_i, the thermal voltage $U_T = kT/e$, the built-in voltage U_D (*German* Diffusionsspannung), and the reverse voltage U across the depletion region are known. $U > 0$ means that the positive battery contact was connected to the n-doped semiconductor,

$$\begin{aligned} w &= \sqrt{\frac{2\epsilon_0\epsilon_r(U_D+U)}{e}\left(\frac{1}{n_A}+\frac{1}{n_D}\right)}, \\ C_{\text{sp}} &= \frac{\epsilon_0\epsilon_r F}{w}, \qquad\qquad U_D = U_T \ln\frac{n_A n_D}{n_i^2}. \end{aligned} \tag{5.39}$$

In the short-wavelength region most photodiodes are made from Si. The light penetration depth is $1/\alpha = 15\,\mu\text{m}$ at $\lambda = 0.85\,\mu\text{m}$. For the pin-photodiode of Fig. 5.8(a) the light penetrates an anti-reflection coating (SiO_2, Si_3N_4) and a thin p-doped layer ($< 1\,\mu\text{m}$) without significant reflection or absorption. The absorption region is made of n^--doped material (e. g., $n_D = 1.3 \times 10^{14}\,\text{cm}^{-3}$). The propagation direction of the light as discussed on Page 110 coincides with the drift of the faster carriers (electrons

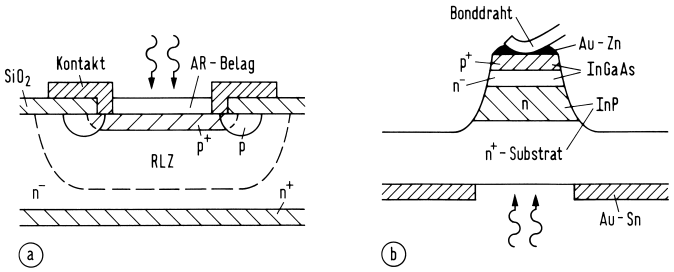


Fig. 5.8. Pin-photodiodes. (a) Planar Si photodiode (b) InGaAs/InP photodiode with mesa structure and illumination through the InP substrate (AR-Belag = anti-reflection coating, RLZ Raumladungszone = depletion region, Kontakt = contact. Bonddraht = bond wire).

in Si). A p-doped guard ring prevents a breakdown at high reverse voltages. The width of the depletion layer for $U = 10\,\mathrm{V}$, $n_D = 1.3 \times 10^{14}\,\mathrm{cm}^{-3} \ll n_A$, $\epsilon_0 = 8.85 \times 10^{-12}\,\mathrm{F\,m}^{-1}$, $\epsilon_r = 11.7$ is about $w = 10\,\mu\mathrm{m}$. With a mean saturation drift velocity $v = (v_n + v_p)/2 = 64\,\mu\mathrm{m}\,/\,\mathrm{ns}$ a transit-time cutoff frequency near $f_{3\,\mathrm{dB}} = 3\,\mathrm{GHz}$ results from Eq. (5.35). With an active photodiode area of $F = (200\,\mu\mathrm{m})^2$ and a resistance of $R_S + R_a = 60\,\Omega$ (see Eq. (5.18)) the RC cutoff frequency is 6.4 GHz. Thus, the photodiode is transit-time limited. For $R_P = 0$ a quantum efficiency of $\eta \approx 0.5$ is to be expected, Eq. (5.14).

The performance of pin-photodiodes can be improved considerably by using a double-heterostructure design. Similar to the case of semiconductor lasers, the middle i-type layer is sandwiched between the p-type and n-type layers of a different semiconductor whose bandgap is chosen such that light is absorbed only in the middle i-layer. Since the bandgap energy of InP is $W_G = 1.35\,\mathrm{eV}$, the material is transparent for light whose wavelength exceeds $\lambda = 0.918\,\mu\mathrm{m}$. By contrast, the bandgap of lattice-matched $(\mathrm{In}_{0.53}\mathrm{Ga}_{0.47})\mathrm{As}$ is $W_G = 0.75\,\mathrm{eV}$ corresponding to a wavelength of $\lambda = 1.653\,\mu\mathrm{m}$. The middle InGaAs layer thus absorbs strongly in the wavelength region $1.3 \ldots 1.6\,\mu\mathrm{m}$. The diffusive component of the detector current is eliminated completely in such a heterostructure photodiode simply because photons are absorbed only inside the depletion region.

For the long-wavelength region, 3-dB cutoff frequencies near 100 GHz were measured. Figure 5.8(b) shows a pin-photodiode with mesa structure. Epitaxial layers of n-InP (buffer layer about $3\,\mu\mathrm{m}$, $n_D = 5 \times 10^{16}\,\mathrm{cm}^{-3}$) and InGaAs ($1.2\,\mu\mathrm{m}$, nominally undoped, $n_D = 3 \times 10^{14}\,\mathrm{cm}^{-3}$) are grown on a n^+-InP-substrate. By diffusion of Zn into the InGaAs layer a $\mathrm{p}^+\mathrm{n}$-junction is formed at a distance of $0.5\,\mu\mathrm{m}$ away from the surface. An etching process forms a mesa to reduce the capacitance. The illumination is through the substrate which is transparent for $\lambda > 0.92\,\mu\mathrm{m}$. By reflection of the light at the p-contact the quantum efficiency is increased to $\eta \approx 0.5$.

Figure 5.9 shows planar pin-photodiodes. In Fig. 5.9(a) the n-type ($n_D \approx 10^{15}\,\mathrm{cm}^{-3}$) InP, the InGaAs, and the (In,Ga)(As,P) ($W_G = 0.95\,\mathrm{eV} \mathrel{\widehat{=}} 1.3\,\mu\mathrm{m}$) layers are nominally undoped. The Zn p-diffusion reaches for about $1\,\mu\mathrm{m}$ into the InGaAs layer. The height of the InGaAs absorption layer is in the range $0.4 \ldots 5\,\mu\mathrm{m}$, depending on the quantum efficiency design and on the tolerated depletion-layer capaci-

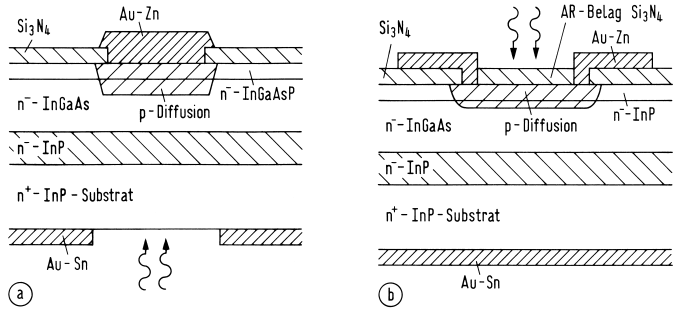


Fig. 5.9. Planar InGaAs/InP pin-photodiodes. (a) Illumination through the substrate (b) illumination from top, AR-Belag = anti-reflection coating

tance. Measured efficiencies are $\eta = 0.3 \ldots 0.6$. Transit time 3-dB frequencies of 20 GHz and RC cutoff frequencies of 80 GHz were achieved.

With similar but top-illuminated planar pin-photodiodes (Fig. 5.9(b)) quantum efficiencies of $\eta = 0.7$ at $\lambda = 1.55\,\mu$m and $f_{3\,\mathrm{dB}} = 25\,\mathrm{GHz}$ are common. The cap layer ($0.5\,\mu$m) consists of n-InP ($n_D = 10^{16}\,\mathrm{cm}^{-3}$) which is moved to a very shallow depth ($0.1\,\mu$m) into the InGaAs layer for increasing the depletion layer over whole of the 2-μm InGaAs layer ($n_D < 5 \times 10^{15}\,\mathrm{cm}^{-3}$).

Instead of the n^+-InP substrate with a buffer layer of n^--InP the diode can be grown on a semi-insulation InP-substrate. In this case the buffer layer is replaced by a n^+-InP contact layer ($n_D = 10^{19}\,\mathrm{cm}^{-3}$) with a lateral contact.

5.2 Noise

Optical receivers convert optical power P_e into electrical current i through a photodiode. The relation $i = SP_e$ in Eq. (5.17) assumes that the current resulting from such a conversion is noise free. However, this is not the case even for a perfect receiver. Two fundamental noise mechanisms, quantum (or shot) noise and thermal noise (or Johnson or Nyquist noise) lead to current fluctuations even when the optical signal has a constant power P_e in the classical sense. The relation $i = SP_e$ still holds if we interpret i as the average current. However, current fluctuations affect the receiver performance. The objective of this section is to briefly review some important noise mechanisms[8].

5.2.1 Noise mechanisms

The classical light power $P_e(t)$ results from an average over a few optical cycles. Fluctuations in $P_e(t)$ are transferred to the photocurrent $i(t)$. The ideal classical signal exhibits a constant amplitude and phase, and no photocurrent fluctuations would be expected. However, quantum mechanics tells that this "ideal" signal (an ideal laser signal) consists of a sequence of independent photons which are Poisson distributed in time. Each photon generates an electron-hole pair with a quantum efficiency η, Eqs. (5.14), (5.36). Thus, the short-circuit photocurrent consists of a stream of statistically independent elementary charges which are also Poisson distributed in time. This type of noise is known as shot[9] noise. Shot noise occurs also in purely electronic circuits if independent electrons cross a biased or unbiased junction at random times. However, this electronic shot noise is independent of any signal photons and is to be observed without any intentional illumination.

The photodiode shot noise as a result of illumination with light is also denoted as quantum noise. The origin of this noise can be attributed neither to the source nor to the detector alone, because the noise shows only when photodiode and light source are interacting during the detection process.

Quantum noise, spontaneous emission noise and shot noise in electronic circuits are unavoidable. In addition to these noise mechanisms there are other noise sources which could be avoided or reduced, for instance thermal noise of a resistor; this noise represents a specific form of spontaneous emission noise and can be reduced by cooling the device.

Photocurrent noise

The photocurrent $i(t)$ (or i for short) fluctuates around its expectation $\overline{i(t)}$ (or $\bar{i}$ for short). The fluctuation is denoted as $\delta i(t)$ (or δi for short),

$$\delta i(t) = i(t) - \overline{i(t)}\,, \qquad \delta i = i - \bar{i}\,. \tag{5.40}$$

[8] See Ref. 17 on Page 6. Sect. 4.4 p. 163

[9] Shot (*pl.* same or shots): A small lead pellet used in quantity in a single charge or cartridge in a shotgun (The Concise Oxford Dictionary. Oxford: Oxford University Press 1990) — Onomatopoetically for the current noise to be heard in a loudspeaker resembling the falling of shot(s) onto a sheet metal.

The autocorrelation function $\vartheta_i(\tau)$ of i is related to the two-sided power spectrum $\Theta_i(f)$ via the Fourier transform,

$$\vartheta_i(\tau) = \overline{i(t+\tau)\, i(t)}\,, \qquad \Theta_i(f) = \int_{-\infty}^{+\infty} \vartheta_i(\tau)\, \mathrm{e}^{-\mathrm{j}\,2\pi f\tau}\, \mathrm{d}\tau\,. \tag{5.41}$$

Because $\vartheta_i(\tau)$ is real, the power spectrum has the property $\Theta_i(f) = \Theta_i^*(-f)$. From the definition, the symmetry relation $\vartheta_i(\tau) = \vartheta_i(-\tau)$ can be seen. As a consequence, the power spectrum is real, $\Theta_i(f) = \Theta_i^*(f)$, and $\vartheta_i(\tau) = \int_0^\infty 2\Theta_i(f)\,\mathrm{d}f$ defines a one-sided real spectrum $2\Theta_i(f)$. The noise variance is obtained by $\sigma_i^2 = \vartheta_i(\tau = 0)$,

$$\sigma_i^2 = \overline{(i-\bar{i}\,)^2} = \overline{\delta i^2} = \int_{-\infty}^{+\infty} \Theta_i(f)\,\mathrm{d}f = \int_0^\infty 2\Theta_i(f)\,\mathrm{d}f\,. \tag{5.42}$$

The spectral density of shot noise is constant and given[10] by $\Theta_i(f) = e\bar{i}$. Usually, the differential fluctuations inside a differential bandwidth $\mathrm{d}f$ centred at the frequency f are of interest,

$$\mathrm{d}\Big(\overline{\delta i^2}\Big) = 2\Theta_i(f)\,\mathrm{d}f = 2e\bar{i}\,\mathrm{d}f, \qquad \overline{|i_{RD}|^2} = 2e\bar{i}\,\mathrm{d}f. \tag{5.43}$$

The complex phasor i_{RD} (effective or root mean square (RMS) value $i_{RD,\mathrm{RMS}} = \big(\,\overline{|i_{RD}|^2}\,\big)^{1/2}$; subscript R for noise, *German* Rauschen) is defined to have the same power $\overline{|i_{RD}|^2}$ per frequency interval[11] $\mathrm{d}f$ as the actual noise process Eq. (5.43). Equation (5.43) expresses a property of the underlying Poisson statistics for the photons: The probability $p_{\,N}(N_P)$ for measuring N_P photons, if the expectation is $\overline{N_P} = N_e$, and the associated second central moment of the process are

$$p_N(N_P) = \frac{\overline{N_P}^{N_P}}{N_P!}\, \mathrm{e}^{-\overline{N_P}}, \quad \overline{\delta N_P^2} = \overline{(N_P - N_e)^2} = \overline{N_P}\,, \qquad \overline{N_P} = N_e\,. \tag{5.45}$$

The expected current $\bar{i} = S\overline{P_e}$ is computed from Eq. (5.17) on Page 113. Classical additional fluctuations from the laser source which has a total output power P_a (Eq. (3.91) on Page 84) are described by the relative intensity noise (RIN),

$$\mathrm{RIN} = \int_0^\infty \mathrm{RIN}(f)\,\mathrm{d}f = \frac{\overline{\delta P_a^2}}{\overline{P_a}^2}\,, \qquad \mathrm{d}\left(\overline{\delta P_a^2}\right) = \overline{P_a}^2\,\mathrm{RIN}(f)\,\mathrm{d}f\,. \tag{5.46}$$

The total differential photocurrent noise fluctuation including the (uncorrelated) received RIN (mean power $\overline{P_a} \sim P_e \sim \bar{i}$) or the noise current i_{RD} in a differential bandwidth $\mathrm{d}f$ are given by

$$\mathrm{d}\Big(\overline{\delta i^2}\Big) = \underbrace{2e\bar{i}\,\mathrm{d}f}_{\substack{\text{shot resp.}\\ \text{quantum}\\ \text{noise}}} + \underbrace{\bar{i}^2\,\mathrm{RIN}(f)\,\mathrm{d}f}_{\text{classical noise}} = \overline{|i_{RD}|^2}. \tag{5.47}$$

The spectral shot noise power density for $\bar{i} = 1\,\mathrm{mA}$ measured at a resistor of $R = 50\,\Omega$ amounts to $\big(2e\bar{i}R\big)_{\mathrm{dB}} = 10\lg\big[2e\bar{i}R\,/\big(1\,\mathrm{mW}\cdot 1\,\mathrm{Hz}^{-1}\big)\big] = -168\,\mathrm{dBm\,Hz^{-1}}$.

[10] Rice, S. O.: Mathematical analysis of random noise. Bell Syst. Techn. J. 23 (1944) 282–332. Eq. (2.6-8)

[11] In Eq. (5.43) the spectral power density $2e\bar{i}$ is frequency-independent. An integration over the frequency range $0 \le f \le B$ results in the total current variance

$$\overline{\delta i^2} = \int_0^B \mathrm{d}\Big(\overline{\delta i^2}\Big) = \int_0^B 2e\bar{i}\,\mathrm{d}f = 2e\bar{i}B = \overline{|i_{RD}|_B^2}\,, \qquad \overline{|i_{RD}|_B^2} = 2e\bar{i}B = \overline{|i_{RD}|^2}\Big|_{\mathrm{d}f\to B}\,. \tag{5.44}$$

However, to avoid an over-complicated notation, we drop the subscript B in the total variance $\overline{|i_{RD}|_B^2}$ and replace $\mathrm{d}f \to B$ wherever it is appropriate, i. e., when the spectral noise current power density $\mathrm{d}\big(\overline{\delta i^2}\big)\big/\,\mathrm{d}f$ does not depend on frequency.

An avalanche photodiode (APD) amplifies the primary photocurrent $i_{\rm pr} = SP_e$ by a current multiplication factor M, the statistical average of which is denoted by M_0. The avalanche multiplication process contributes additional noise, which is described by an excess noise factor F_M,

$$F_M = \frac{\overline{M^2}}{\overline{M}^2} = \frac{\overline{M^2}}{M_0^2} = 1 + \frac{\overline{\delta M^2}}{M_0^2}. \tag{5.48}$$

It is common to approximate F_M by the function

$$F_M = M_0^x, \quad x > 0\,. \tag{5.49}$$

For the APD current $i = Mi_{\rm pr}$ and for the noise current $\mathrm{d}\left(\overline{\delta i^2}\right)$ in a differential bandwidth $\mathrm{d}f$ we find the relations

$$\begin{aligned} \bar{i} &= M_0 \bar{i}_{\rm pr} = M_0 \frac{\eta e}{h f_S} P_e, \\ \mathrm{d}\left(\overline{\delta i^2}\right) &= 2e\bar{i}_{\rm pr} M_0^2 F_M \,\mathrm{d}f + (M_0 \bar{i}_{\rm pr})^2 \,\mathrm{RIN}(f)\,\mathrm{d}f = \overline{|i_{RD}|^2}. \end{aligned} \tag{5.50}$$

For $M_0 = 1$, $F_M = 1$, Eq. (5.50) reduces to the case of the pin photodiode, Eq. (5.47) on Page 122.

Shot noise in semiconductor junctions

A similar relation as for the noise in photodiodes holds for ordinary semiconductor pn-junctions in diodes and transistors, too. The associated current fluctuation is also called shot noise[12], and no illumination with light is needed for this effect. Electrons and holes traverse the junction independently and at random times, depending on the thermal energy a carrier happens to have, and this leads to a Poisson distribution of the arrival times. A pn-junction which carries an average forward or reverse current $\bar{i}$ at a junction

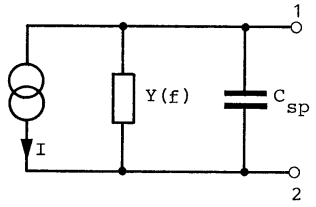


Fig. 5.10. Small-signal equivalent circuit of a pn-junction with shot noise RMS current I in a bandwidth B. Current fluctuation $\overline{|I|^2} = 2e\bar{i}B$, diffusion admittance $Y(f)$, junction capacitance $C_{\rm sp}$

voltage U has a saturation current I_S, and a small-signal conductance G_0 for low frequencies $f \to 0$. With the thermal voltage U_T (*German* Temperaturspannung) we then find the well-known relation

$$i = I_S\left(\mathrm{e}^{U/U_T} - 1\right), \qquad G_0 = \frac{\partial i}{\partial U} = \frac{I_S}{U_T}\,\mathrm{e}^{U/U_T}, \qquad U_T = \frac{kT_0}{e}\,. \tag{5.51}$$

The diffusion admittance of the junction $Y(f) \sim \sqrt{1 + \mathrm{j}\,\omega\tau} \approx 1 + \frac{1}{2}(\mathrm{j}\,\omega\tau) - \frac{1}{8}(\mathrm{j}\,\omega\tau)^2$ (not all carriers cross the junction in a period $1/f$) is given by ($\omega = 2\pi f$, carrier recombination lifetime τ)

$$Y(f) = G_0\sqrt{1 + \mathrm{j}\,\omega\tau} = G(f) + \mathrm{j}\,B_Y(f), \qquad G(f) \approx G_0\left(1 + \frac{1}{8}\omega^2\tau^2\right). \tag{5.52}$$

With Eq. (5.51) we write

$$\mathrm{d}\left(\overline{\delta i^2}\right) = 2e\bar{i}\,\mathrm{d}f + 4eI_S\,\mathrm{d}f + 4kT_0[G(f) - G_0]\,\mathrm{d}f = \begin{cases} 2e\bar{i}\,\mathrm{d}f & \text{for } f \to 0 \\ 4kT_0 G(f)\,\mathrm{d}f & \text{for } \bar{i} = 0 \\ 4kT_0 G_0\,\mathrm{d}f & \text{for } \bar{i} = 0 \text{ and } f \to 0 \end{cases}. \tag{5.53}$$

[12] See Footnote 9 on Page 121

The equivalent circuit of a pn-junction with shot noise RMS current I and fluctuation $\overline{|I|^2} = 2e\bar{i}B$ in a bandwidth B (see Eq. (5.44)[13]), diffusion admittance $Y(f)$ and junction capacitance $C_{\rm sp}$ is depicted in Fig. 5.10.

Thermal noise

In any conductor at a finite temperature T_0, electrons move randomly. Random thermal motion of electrons in a resistor manifests as a fluctuating current even in the absence of an applied voltage. The load resistor R_a of a photodiode (Fig. 5.3) located in the front end of an optical receiver adds such fluctuations to the noise current generated by the photodiode. Figure 5.11 shows a noisy conductance G_Q at temper-

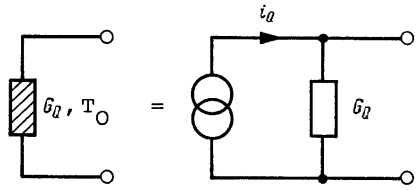


Fig. 5.11. Equivalent circuit of a conductance G_Q with thermal noise

ature T_0 (hatched). In an equivalent circuit, G_Q is replaced by a noiseless conductance with the same value, supplemented by a noise current source with an RMS value $i_{Q,\rm RMS} = \left(\overline{|i_Q|^2}\right)^{1/2}$ in a differential bandwidth $\mathrm{d}f$. This current source represents thermal noise. The available (maximum) differential power $\mathrm{d}P_v$ (*German* verfügbare Wirkleistung) in an impedance-matched load conductance G_Q connected to the open terminals in Fig. 5.11 is $\mathrm{d}P_v = \left(i_{Q,\rm RMS}/2\right)^2/G_Q$. Thermal noise is also called Johnson noise or Nyquist noise after the two scientists who first studied it experimentally and theoretically[14]. The equivalent short-circuit noise current of the noisy conductance G_Q and the available power at temperature T_0 in a bandwidth B are

$$\overline{|i_Q|^2} = 4kT_0G_Q\,\mathrm{d}f, \qquad P_v = \int_0^B \mathrm{d}P_v = \int_0^B \left(\frac{\sqrt{4kT_0G_Q\,\mathrm{d}f}}{2}\right)^2 \frac{1}{G_Q} = kT_0B \quad \text{for } hB \ll kT_0\,. \tag{5.54}$$

The one-sided spectral power density is $2\Theta_T(f) = kT_0$, $(kT_0)_{\rm dB} = 10\lg\left[kT_0\,/\left(1\,\mathrm{mW}\cdot 1\,\mathrm{Hz}^{-1}\right)\right] = -174\,\mathrm{dBm\,Hz}^{-1}$ at room temperature $T_0 = 293\,\mathrm{K}$. To explain Eq. (5.54) on needs to know the average number $\overline{N_P}$ of photons[15] with energy hf per mode in thermal equilibrium. In a bandwidth B these photons provide a power of N_PhfB. Electrical circuits are usually single-moded, and in addition the condition $hf \ll kT_0$ is easily fulfilled for $f < 1\,\mathrm{THz}$ ($hf < 4\,\mathrm{meV}$)[16] at room temperature ($kT_0 = 25\,\mathrm{meV}$).

If the source admittance is complex, i. e., if $Y_Q = G_Q + \mathrm{j}\,B_Q$, then only its real part $G_Q = \Re\{Y_Q\}$ has to be substituted in Eq. (5.54).

5.2.2 Electronic amplifier noise

For characterizing an electronic two-port network as in Fig. 5.12 (four-terminal or fourpole network, *German* Vierpol), we need defining the transducer power gain $\Gamma_{\rm ü}$ (*German* Übertragungsleistungsverstär-

[13] See Footnote 11 on Page 122

[14] See Ref. 17 on Page 6. Sect. 4.4.1 p. 164

[15] If a number $N = N_1 + N_2$ of microsystems (N_1 in state W_1, N_2 in state W_2) is in thermal equilibrium with an electromagnetic radiation mode, the average number of emissions must equal the average number of absorptions, $N_2w^{(\rm eM)} = N_1w^{(\rm aM)}$, i. e., with Eq. (3.32) on Page 68, $N_2(N_P+1) = N_1N_P$. Because in thermal equilibrium we have $N_1/N_2 = \exp[(W_2-W_1)/(kT_0)]$ from Eq. (3.7) on Page 54, we find Planck's formula for the average number of photons per polarization in one transverse and longitudinal mode with frequency f (Bose-Einstein distribution, see Footnote 50 on Page 157),

$$\overline{N_P} = \frac{1}{\exp\left(\frac{hf}{kT_0}\right)-1} \approx \{hf \ll kT_0\} \approx \frac{kT_0}{hf}\,, \qquad hf = W_2 - W_1. \tag{5.55}$$

[16] See Footnote 9 on Page 2

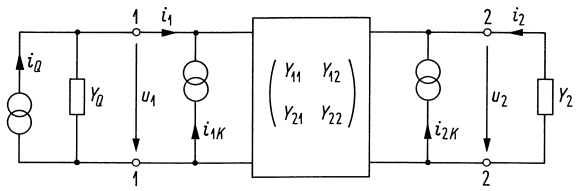


Fig. 5.12. Noisy two-port network (curent sources i_{1K}, i_{2K}) with noisy generator admittance (i_Q, Y_Q) and noise-free load admittance Y_2

kung) as a ratio of the power P_{S2} delivered to the load admittance Y_2 at terminals 2-2, and the available source power P_{Sv1} (*German* verfügbare Wirkleistung). A generator with admittance $Y_Q = G_Q + \mathrm{j}\,B_Q$ would deliver its maximum or available power to a load admittance $Y_Q^* = G_Q - \mathrm{j}\,B_Q$, which is matched to the generator. We further introduce the available power gain (*German* verfügbare Leistungsverstärkung) Γ_v as the ratio of the available power P_{Sv2} at the output terminals 2-2, and the available power of the signal source P_{Sv1},

$$\Gamma_{\ddot{u}}(f) = \frac{P_{S2}}{P_{Sv1}}\,, \qquad \Gamma_v(f) = \frac{P_{Sv2}}{P_{Sv1}}\,. \tag{5.56}$$

Noisy two-port

The equivalent circuit of a noisy electronic amplifier in Fig. 5.12 is driven by a signal source (*German* Quelle) with a deterministic short-circuit current represented by phasor i_S (in parallel to i_Q, not drawn in Fig. 5.12) and admittance Y_Q, the real part $G_Q = \Re\{Y_Q\}$ of which emits thermal noise. This noise is described by a short-circuit current phasor i_Q that fluctuates according to Eq. (5.54), see definition in Eq. (5.43) on Page 122,

$$\overline{|i_Q|^2} = 4kT_0G_Q\,\mathrm{d}f, \quad Y_Q = G_Q + \mathrm{j}\,B_Q, \quad T_0 = 293\,\mathrm{K}\,. \tag{5.57}$$

Noise of the two-port network is described by short-circuit current sources i_{1K}, i_{2K}. The quantities $\overline{|i_{1K}|^2}$, $\overline{|i_{2K}|^2}$ and $\overline{i_{1K}i_{2K}^*}$ are assumed to be known, e. g., by measurement. Because it is is convenient to concentrate all noise sources at the two-port network's input, the output noise source i_{2K} is transformed to the input. Correlations between output and input current noise i_{1K} are taken care of by a two-port noise network[17,18] (noise fourpole, *German* Rauschvierpol) with uncorrelated noise sources i_n, u_n ($\overline{i_n u_n^*} = 0$) and a correlation admittance Y_c, which is electrically not visible outside the two-port noise network because Y_c is connected in parallel to $-Y_c$, see Fig. 5.13. Between terminals 1-1, 2-2 in Fig. 5.12 and Fig. 5.13 the following relations hold:

$$\begin{aligned}
\begin{pmatrix} i_1 \\ i_2 \end{pmatrix} &= \begin{pmatrix} -i_{1K} \\ -i_{2K} \end{pmatrix} + \begin{pmatrix} Y_{11}\ Y_{12} \\ Y_{21}\ Y_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \\
\begin{pmatrix} i_1 \\ i_2 \end{pmatrix} &= \begin{pmatrix} -i_n + Y_{11}u_n - Y_c u_n \\ Y_{21}u_n \end{pmatrix} + \begin{pmatrix} Y_{11}\ Y_{12} \\ Y_{21}\ Y_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}
\end{aligned} \tag{5.58}$$

By comparison and with $\overline{i_n u_n^*} = 0$, the two-port noise network parameters can be calculated,

$$\begin{aligned}
\overline{|i_n|^2} &= \overline{|i_{1K}|^2} - \left|\overline{i_{1K}i_{2K}^*}\right|^2 \big/ \overline{|i_{2K}|^2}, \\
\overline{|u_n|^2} &= \overline{|i_{2K}|^2} \big/ |Y_{21}|^2, \\
Y_c &= Y_{11} - Y_{21}\,\overline{i_{1K}i_{2K}^*} \big/ \overline{|i_{2K}|^2} = G_c + \mathrm{j}\,B_c.
\end{aligned} \tag{5.59}$$

[17] H. Rothe, W. Dahlke: Theory of noisy fourpoles. Proc. IRE 44 (1956) 811–818

[18] Horst Rothe, German microvawe engineer and physicist, ⋆ Hosterwitz, Dresden (Germany) 13.12.1899, † 10.07.1974. Appointed as full professor at Technische Hochschule Karlsruhe on 1.4.1956. Founded the Institut für Hochfrequenztechnik und Hochfrequenzphysik (Institute of High-Frequency Technology and High-Frequency Physics) in 1958. This institute was renamed in Institut für Hochfrequenztechnik und Quantenelektronik (IHQ, High-Frequency and Quantum Electronics Laboratory) in 1971, and again renamed in Institut für Photonik und Quantenelektronik (Institute of Photonics and Quantum Electronics, IPQ) in 2008.

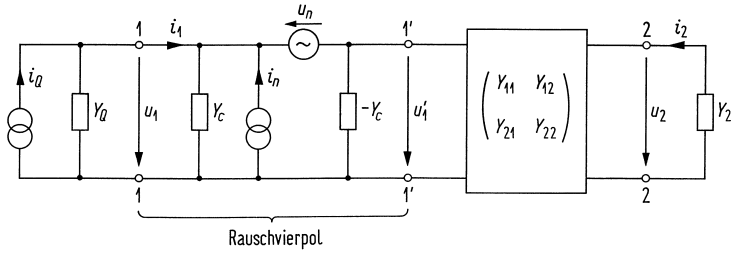


Fig. 5.13. Two-port noise network (*German* Rauschvierpol), representing the noise properties of the two-port network by uncorrelated noise generators i_n, u_n ($\overline{i_n u_n^*} = 0$) together with a correlation admittance Y_c

For convenience, we define a noise resistance R_n and a noise admittance G_n (attention: $R_n \neq 1/G_n$) in a bandwidth B,

$$\overline{|u_n|^2} = 4kT_0R_n\,\mathrm{d}f, \quad \overline{|i_n|^2} = 4kT_0G_n\,\mathrm{d}f. \tag{5.60}$$

Noise figure of electronic amplifiers

As a quality metric for the noisy two-port network, we define a noise figure F (*German* Rauschzahl) by relating signal-to-noise power ratios (SNR) at input and output, or by the ratio of noise powers at the load admittance Y_2 in the case of the actual noisy two-port network, and in the case of a hypothetically noise-free but otherwise identical structure. As before, the signal's short-circuit current source i_S would be in parallel to i_Q, but is not depicted in Fig. 5.12 and 5.13.

From Fig. 5.13 we find the short-circuit currents referred to the (physically not accessible) terminals 1′-1′. The equivalent total short-circuit noise current at this input is

$$i_R = i_Q + i_n + u_n(Y_Q + Y_c). \tag{5.61}$$

All terms are uncorrelated. With the help of Eq. (5.57) and (5.60) we then calculate the noise figure,

$$F = \frac{\mathrm{SNR}_{v1}}{\mathrm{SNR}_2} = \frac{P_{Sv1}}{P_{Rv1}}\,\frac{P_{R2}}{P_{S2}} = \frac{P_{R2}\,/\,\Gamma_{\ddot{u}}}{P_{Rv1}} = \frac{\text{total output noise power in } Y_2 \text{ related to input}}{\text{noise power in } Y_2 \text{ for noise-free TP related to input}} \geq 1,$$

$$F = \frac{\overline{|i_R|^2}}{\overline{|i_Q|^2}} = 1 + \frac{G_n + R_n|Y_Q + Y_c|^2}{G_Q}, \tag{5.62a}$$

$$F = 1 + F_z = 1 + \frac{T_R}{T_0}, \qquad \overline{|i_R|^2} = 4k\,(FT_0)\,G_Q\,\mathrm{d}f = 4k\,(T_0 + T_R)\,G_Q\,\mathrm{d}f. \tag{5.62b}$$

The quantity $F_z = F - 1$ is dubbed excess-noise figure (*German* zusätzliche Rauschzahl), T_R represents the noise temperature of the two-port network (*German* Rauschtemperatur). The signal-to-noise power ratios SNR_{v1} and SNR_2 stand for the ratio of the available signal power and the available noise power at the input, and the ratio of the actual signal and noise powers at the output, respectively.

The noise of an electronic amplifier can be equivalently described by a fictitious increase of the temperature of the source conductance G_Q over the reference temperature T_0 by a factor of F. The noise temperature T_R specifies the equivalent fictitious temperature which must be added to the reference temperature T_0 of the source conductance G_Q.

For a noise-free two-port network $F = 1$ holds ($F_z = 0$, $T_R = 0$). The noise figure F has a relative minimum[19] for the case of noise tuning (*German* Rauschabstimmung) $B_Q = -B_c$ (this allows measurement of B_c). A global minimum for F is found if in addition the source conductance $G_{Q\mathrm{opt}}$ is properly chosen (noise matching, *German* Rauschanpassung). For $G_{Q\mathrm{opt}}, F_{z\mathrm{min}}$ we calculate from Eq. (5.62a)

$$G_{Q\mathrm{opt}} = \sqrt{\frac{G_n}{R_n} + G_c^2}, \qquad F_{z\mathrm{min}} = 2R_n(G_{Q\mathrm{opt}} + G_c). \tag{5.63}$$

[19] See Ref. 17 on Page 125

For negligible output-input correlation $\overline{i_{1K} i_{2K}^*} = 0$ (approximately true for an emitter circuit with a bipolar transistor, BPT, or for a source circuit with a field-effect transistor, FET) we find according to Eq. (5.59) $Y_c = Y_{11}$ (for known Y_c, the quantities G_n, R_n can be deduced by measuring $F_{z\min}$, $G_{Q\text{opt}}$ according to Eq. (5.63)). For a FET we have $\Re\{Y_{11}\} \sim \omega^2$; for a BPT, the quantity $R_{\text{opt}} = 1/G_{Q\text{opt}}$ is smaller than $100\,\Omega$, and for a FET about $1\,\text{k}\Omega$. Minimum noise figures are about $10 \lg F = 1\,\text{dB}$.

Noise figure of an amplifier chain The noise figure of a concatenated arrangement of noisy two-port networks is calculated according to Fig. 5.14. The noise temperature T_{R1} of the first two-port network is defined assuming a source admittance Y_S at temperature T_S. The noise temperature T_{R2} of the second two-port network, however, is defined for a source admittance $Y_{\text{out}}^{(1)}$ at temperature T_S, which is to be seen when looking from the input terminals of the second two-port network to the left,

$$Y_{\text{out}}^{(1)} = Y_{22}^{(1)} - \frac{Y_{12}^{(1)} Y_{21}^{(1)}}{Y_{11}^{(1)} + Y_S}\,.$$

If both amplifiers have no feedback ($Y_{12} = 0$) and identical output short-circuit admittances ($Y_{22} = Y_S$), the two-port networks can exchange position without a change in the *individual* noise figures. However, the noise figure of the amplifier chain changes.

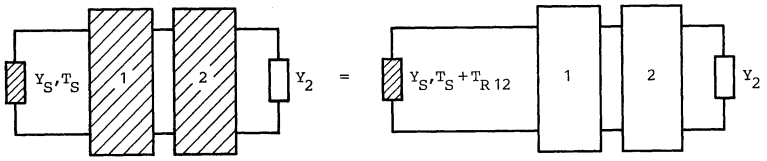


Fig. 5.14. Concatenation of noisy four-port networks

We relate to the transducer power gain $\Gamma_{\ddot{u}}(f)$ and the available power gain Γ_v as defined in Eq. (5.56) on Page 125. For calculating the noise figure of the concatenated two-ports ($\text{TP}_{1,2}$), we employ the definition Eq. (5.62a),

$$F_{12} = \frac{\overbrace{k(T_S + T_{R1})\Gamma_{v1} B}^{\text{available total noise power from TP}_2}\,\Gamma_{\ddot{u}2} \;+\; \overbrace{kT_{R2}\Gamma_{\ddot{u}2} B}^{\text{noise power from TP}_2}}{\underbrace{kT_S\Gamma_{v1}B}_{\substack{\text{noise-free}\\ \text{TP}_1}} \times \underbrace{\Gamma_{\ddot{u}2}}_{\substack{\text{noise-free}\\ \text{TP}_2}}} = 1 + \frac{T_{R1}}{T_S} + \frac{T_{R2}}{\Gamma_{v1} T_S} = 1 + \frac{T_{R12}}{T_S} = 1 + F_{z12} \qquad (5.64)$$

For the noise temperature and the excess-noise figure we find Friis' formulae[20]

$$T_{R\,12} = T_{R\,1} + \frac{T_{R\,2}}{\Gamma_{v\,1}}\,, \qquad F_{z\,12} = F_{z\,1} + \frac{F_{z\,2}}{\Gamma_{v\,1}}\,. \qquad (5.65)$$

Noise measure

The question remains in which sequence amplifiers have to be concatenated to achieve a minimum total noise figure. The following considerations hold for the condition

$$Y_{\text{out}}^{(1)} = Y_{\text{out}}^{(2)} = Y_S\,.$$

[20] Friis, H. T.: Noise figures of radio receivers. Proc. Inst. Radio Engrs. 32 (1944) 419–422. — In 1942, Harald T. Friis, working in Bell Labs in Holmdel NJ, developed the theory of "noise figure" that allows engineers to calculate the signal-to-noise ratio at the output of a complex receiver chain, and thus has a powerful equation named after him.

Harald Friis was born in Naestved Denmark, in 1893. He graduated 1916 in Electrical Engineering from the Polytechnic Institute (founded 1829 by H. C. Oersted, the discoverer of electromagnetics). In 1919 he received a fellowship which enabled him to come to the United States where he studied radio engineering at Columbia University. In 1920, Friis joined a research group headed by at the Western Electric Company and apparently got stuck in the U.S.A. He eventually became a U.S. citizen, which later did not prevent him from being awarded the Valdemar Poulsen Medal of the Danish Academy of Sciences. He held 31 U.S. patents submitted over five decades of research.

As a rule this is true in the microwave region where input and output impedances equal the line impedance $Z_L = 50\,\Omega$. If this assumption is not applicable for a specific problem, then the calculation must be performed considering the different source admittances.

Let us assume that the arrangement Fig. 5.14 leads to the smaller noise figure, i. e., $F_{z12} < F_{z21}$. Further, we consider amplifiers with $\Gamma_{v1} > 1$, $\Gamma_{v2} > 1$ so that $(1 - 1/\Gamma_v) > 0$ is guaranteed. For the concatenation sequence 1-2 we therefore write

$$F_{z12} = F_{z1} + \frac{F_{z2}}{\Gamma_{v1}} < F_{z2} + \frac{F_{z1}}{\Gamma_{v2}} = F_{z\,21} \quad \text{or} \quad \frac{F_{z1}}{1 - 1/\Gamma_{v1}} < \frac{F_{z2}}{1 - 1/\Gamma_{v2}}\,.$$

It is convenient to define a noise measure M (*German* Rauschmaß)[21,22,23] for the two-port network by

$$M_i = \frac{F_{zi}}{1 - 1/\Gamma_{vi}}\,. \tag{5.66}$$

Because we had assumed that the concatenation sequence 1-2 leads to a lower total noise figure than the sequence 2-1, the noise measures of the amplifiers must be related by $M_1 < M_2$. So we find the following rule: A concatenation of amplifiers leads to a minimum total noise figure, if they are arranged in sequence of increasing noise measures.

5.2.3 Optical amplifier noise

For optical amplifiers (OA) in a frequency region $hf \gg kT_0$ where quantum effects become important, the relations of Sect. 5.2.2 must be reconsidered. Especially the uncertainty relation comes now into play which restricts the accuracy of simultaneous measurements of amplitude and phase or real and imaginary part of a field. The minimum input noise equivalent $P_{r\ \mathrm{qu},\,x}$ ascribed to an ideal OA per mode and per polarization (e. g., linear polarization in x-direction) would be such that the uncertainty relation is just fulfilled. Even without any input power, amplified spontaneous emission (ASE) noise power can be extracted[24] from the OA output. This noise is represented by a fictitious OA input noise power $P_{r\,\mathrm{eq},\,x}$, so that a hypothetically noise-free OA with single-pass power gain $\mathcal{G}_s$ had the same ASE noise power $\mathcal{G}_s P_{r\,\mathrm{eq},\,x}$ as the true OA.

Let us observe one mode in one polarization, i. e., we observe for one transverse mode (e. g., a field emitted from a single-mode fibre) inside an optical bandwidth B_O one longitudinal mode during the observation time $1/\,B_O$ by simultaneously measuring both the amplitude and the phase (or the real and imaginary part) of the optical field, see the sampling theorem Eq. (2.4) on Page 15. With the help of the inversion factor n_{sp} from Eq. (3.40) on Page 70, the ASE noise power $P_{\mathrm{ASE},\,x}$ per mode and per polarization, and consequently also the fictitious (not extractable[25]) equivalent input noise power $P_{r\,\mathrm{eq},\,x}$ of a real-world OA can be calculated[26],

$$\begin{aligned} P_{\mathrm{ASE},\,x} &= (\mathcal{G}_s - 1)\, n_{\mathrm{sp}} w_O B_O = \mathcal{G}_s P_{r\,\mathrm{eq},\,x}\,, \\ P_{r\,\mathrm{eq},\,x} &= \frac{\mathcal{G}_s - 1}{\mathcal{G}_s}\, n_{\mathrm{sp}} P_{r\,\mathrm{qu},\,x}\,, \qquad P_{r\,\mathrm{qu},\,x} = w_O B_O, \qquad w_O = h f_e\,. \end{aligned} \tag{5.67}$$

When the OA is effectively removed by making it transparent with the choice $\mathcal{G}_s = 1$, no extractable ASE noise power remains and $P_{r\,\mathrm{eq},\,x} = 0$ results. The quantum fluctuations per mode and per polarization, expressed by the non-extractable minimum quantum noise power $P_{r\,\mathrm{qu},\,x}$ cannot disappear.

Spontaneous emission factor n_{sp} and gain $\mathcal{G}_s$ are linked. If the gain is larger than but close to one, the spontaneous emission factor is very large.

---

[21] Haus, H. A.; Adler, R. B.: Invariants of linear networks, 1956 Inst. Radio Engrs. Convention Record, Part 2, 53 (1956)
[22] Haus, H. A.; Adler, R. B.: Optimum noise performance of linear amplifiers. Proc. Inst. Radio Engrs. 46 (1958) 1517–1533
[23] Haus, H. A.; Adler, R. B.: Circuit theory of linear noisy networks. Technology Press Research Monograph, New York: Wiley 1959
[24] See the remarks on Page 22 and in Footnote 34
[25] See the remarks on Page 22 and in Footnote 34
[26] See Ref. 29 on Page 22. Chapter 9 Eq. (9.2/15)

5.3 Direct receiver

A basic direct optical receiver is displayed in Fig. 5.15. The quantities in this figure are to be interpreted as complex phasors. The receiver consists of an optical front end with a pin photodiode, the current $i\,(t)$ of which feeds an electronic amplifier. Noisy components are hatched. It is advantageous with respect to

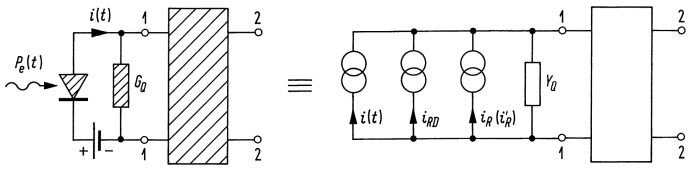


Fig. 5.15. Schematic of an optical receiver with pin-photodiode, source conductance G_Q and amplifier (noisy components are hatched). The phasor i_{RD} specifies the shot noise of the pin-photodiode, and i_R (or i'_R) represents the noise phasor of the source conductance $Y_Q = G_Q + \mathrm{j}\,\omega C_{\mathrm{sp}}$ including the junction capacitance C_{sp} of the photodiode, and of the amplifier without feedback (or of the transimpedance amplifier).

receiver bandwidth and sensitivity to employ a so-called transimpedance amplifier (TIA, *German* Transimpedanzverstärker, TIV). This designation reflects the fact that its complex transfer function between the output voltage at terminals 2-2 and the input current at terminals 1-1 represents an impedance Z_F. First we apply the description of electronic amplifier noise (Sect. 5.2.2 on Page 124 ff.) to an optical receiver, and then a detailed description of the TIA will be given.

Amplifier without feedback We start with the representation of a noisy amplifier as depicted in Fig. 5.13 on Page 126 and complete it with the essentials of the equivalent photodiode circuit displayed in Fig. 5.3 on Page 113. The extended two-port representation is shown in Figure 5.16-[top]. Next we introduce the voltage gain $V < 0$ and the input admittance Y_{1E} of the two-port network and find the equivalent circuit of Fig. 5.16-[bottom],

$$V = -\frac{Y_{21}}{Y_{22} + Y_2}, \quad Y_{1E} = Y_{11} + VY_{12}\,. \tag{5.68}$$

Here, $i_S = SP_e$ represents the phasor of the signal source. The amplifier noise is taken care of by a fictitious temperature increase of the source conductance G_Q according to the amplifier's noise figure F,

$$\overline{|i_R|^2} = F\,\overline{|i_Q|^2} = 4kFT_0G_Q\,\mathrm{d}f. \tag{5.69}$$

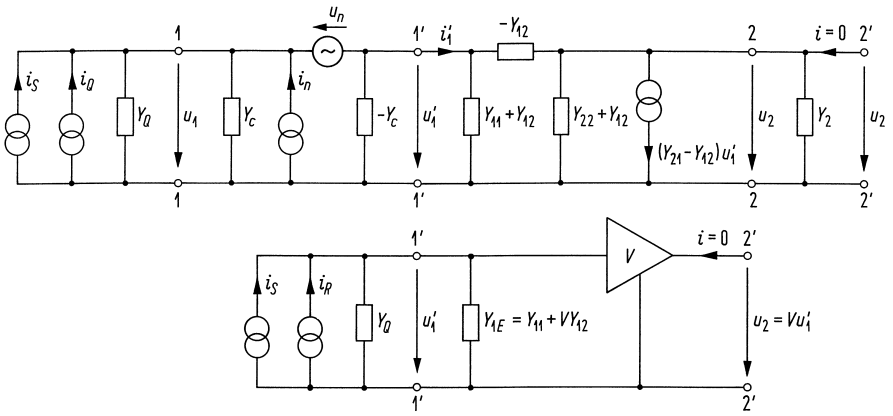


Fig. 5.16. Representation of the two-port network by an input admittance („Eingangsleitwert“) Y_{1E} and a voltage transformer $u_2 = Vu'_1$; the noise sources of four-port and generator admittance are replaced by a noise current i_R.

Transimpedance amplifier We now supplement the amplifier Fig. 5.16 with an ideal voltage amplifier $V_2 > 0$ at terminals $2'$-$2'$ and a negative-feedback with admittance Y_F between the output terminals 3-3 of amplifier V_2 and the input terminals 1-1,

$$Y_F = G_F + \mathrm{j}\, B_F, \quad \overline{|i_F|^2} = 4kT_0 G_F \,\mathrm{d}f. \tag{5.70}$$

With these additions we arrive at the circuit schematic Fig. 5.17(a). The noise current i_F does not influence the voltage u_3, which is impressed at output terminals 3-3 by the ideal voltage amplifier V_2. Because the output impedance of the ideal voltage amplifier V_2 is zero, the feedback admittance Y_F is directly visible between the input terminals 1-1. From Fig. 5.17(a) we find

$$i_1 = Y_F u_1 + Y_c u_1 - i_F - i_n - Y_c u_1' + (Y_{1E} - VV_2Y_F)u_1'. \tag{5.71}$$

In a simplified equivalent circuit without explicit feedback, Fig. 5.17(b), the admittance $-VV_2Y_F$ appears in parallel to Y_{1E}, the feedback admittance Y_F is seen parallel to Y_Q, and the noise current i_F adds to the (uncorrelated) noise current i_R in Fig. 5.16-[bottom]. Using Eq. (5.61) on Page 126 and Eq. (5.68), we write for the equivalent circuit parameters

$$\begin{aligned} Y_Q' &= Y_Q + Y_F, & i_R' &= i_R + i_F = i_Q + i_n + u_n(Y_Q' + Y_c) + i_F, \\ Y_{1E}' &= Y_{1E} - V'Y_F, & V' &= VV_2 < 0. \end{aligned} \tag{5.72}$$

Through the noise current i_F of the feedback conductance G_F, the negative-feedback amplifier exhibits a slightly increased noise current $i_R' > i_R$,

$$\overline{|i_R'|^2} = \overline{|i_Q|^2} + \overline{|i_n|^2} + \overline{|u_n|^2}\left|Y_Q' + Y_c\right|^2 + \overline{|i_F|^2}, \tag{5.73}$$

therefore G_F should be as small as possible within the restrictions set by the limiting R_FC_F bandwidth ($R_F = 1/G_F$) due to a parasitic feedback capacitor C_F parallel to G_F ($B_F = \omega C_F$). However, the real part of the input admittance increases by $-V'G_F > 0$, and therefore the input bandwidth increases dynamically (assuming $B_F \ll G_F$). For the signal output voltage u_3 of this transimpedance amplifier we find from Fig. 5.17(b) and for $V'Y_F \gg Y_Q + Y_F + Y_{1E}$

$$u_3 = i_S \frac{V'}{Y_Q + Y_F + Y_{1E} - V'Y_F} \approx -\frac{i_S}{Y_F} = -i_S Z_F. \tag{5.74}$$

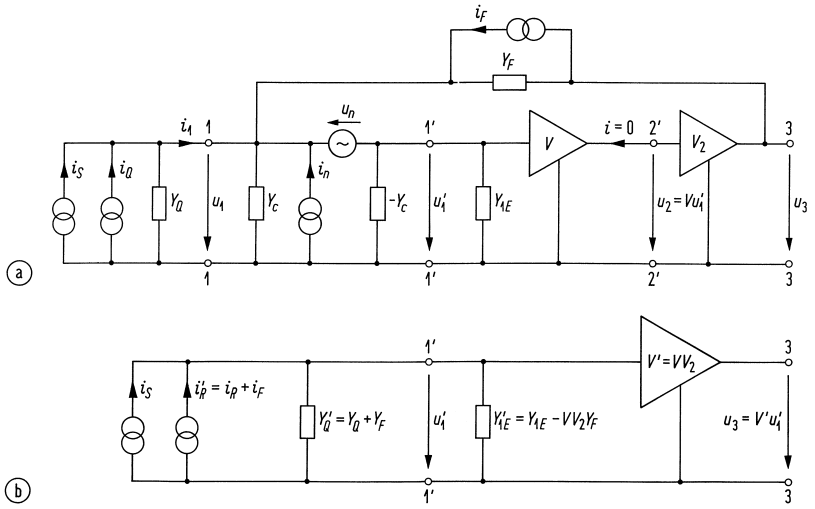


Fig. 5.17. Noisy negative-feedback circuit. (a) Detailed equivalent circuit ($V < 0$, $V_2 > 0$, $V' = VV_2 < 0$) (b) Simplified equivalent circuit (i_S, i_R, Y_Q, Y_{1E} as in Fig. 5.16)

In analogy to Eq. (5.62a) on Page 126, the TIA has a noise figure (Y_F is regarded to be part of the amplifier)

$$F'(f) = \frac{\overline{|i'_R|^2}}{\overline{|i_Q|^2}} = 1 + \frac{G_F + G_n + R_n|Y'_Q(f) + Y_c(f)|^2}{G_Q}. \tag{5.75}$$

The noise figure F' depends on frequency, because both $Y'_Q = G_Q + G_F + \mathrm{j}\,\omega\,(C_{\mathrm{sp}} + C_F)$ (PD junction capacitance C_{sp}, Eq. (5.72) and Eq. (5.57 on Page 125) as well as $Y_c = G_c + \mathrm{j}\,\omega C_c$ (Eq. (5.59) on Page 125) are frequency-dependent with (usually) capacitive imaginary parts.

5.3.1 Direct reception limit

The smallest power which can be received is dictated by receiver noise. We employ a photodiode with transimpedance amplifier. For finding the limiting signal-to-noise power ratio SNR, the equivalent circuit Fig. 5.17 is reduced to its essential parts, Fig. 5.18. For the photodiode noise current i_{RD}, Eq. (5.43) on Page 122) we neglect any RIN. Photodiode noise current i_{RD} and TIA noise current i'_R, Eq. (5.72) on Page 130), stem from different physical processes and are uncorrelated, therefore the noise powers can be added. The signal current $i_S = SP_e$ in Eq. (5.17) on Page 113 is proportional to the received optical

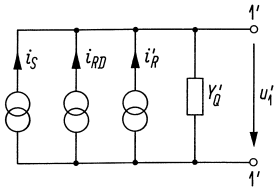


Fig. 5.18. Optical transimpedance receiver input. The quantities i_S, i_{RD} are signal and noise current of the photodetector according to Eq. (5.50) on Page 123; for i'_R, Y'_Q see Eq. (5.72)

signal power P_e. The signal-to-noise power ratio (SNR) γ_{dir} for direct reception is defined as the ratio of the average electrical signal power $P_S \sim i_S^2$ and the electrical noise power P_R in an electrical signal bandwidth B (see Eq. (5.44)[27]),

$$P_R = \int_0^B \mathrm{d}P_R \sim \overline{\delta i^2} \quad \text{with} \quad \mathrm{d}P_R \sim \mathrm{d}\left(\overline{\delta i^2}\right) = \overline{|i_{RD}|^2} + \overline{|i'_R|^2} = 2ei_S\,\mathrm{d}f + 4kF'(f)T_0G_Q\,\mathrm{d}f\,. \tag{5.76}$$

According to Eq. (5.73) and (5.75), and with $Y'_Q = G'_Q + \mathrm{j}\,\omega C'_Q = Y_Q + Y_F = G_Q + G_F + \mathrm{j}\,\omega\,(C_Q + C_F)$ and $Y_c = G_c + \mathrm{j}\,\omega C_c$, the electronic TIA noise $\overline{|i'_R|^2}$ has a part that does not depend on frequency, and a part which increases with f^2,

$$\overline{|i'_R|^2} = 4kT_0\left[G'_Q + G_n + R_n\left(G'_Q + G_c\right)^2 + \omega^2 R_n\left(C'_Q + C_c\right)^2\right]\mathrm{d}f \tag{5.77}$$

The so-called noise corner frequency (*German* Rausch-Eckfrequenz) f_{RE} describes at which frequency the frequency-independent and the frequency-dependent parts contribute alike,

$$\omega_{\mathrm{RE}}^2 = (2\pi f_{\mathrm{RE}})^2 = \frac{G'_Q + G_n + R_n\left(G'_Q + G_c\right)^2}{R_n\left(C'_Q + C_c\right)^2}\,. \tag{5.78}$$

The total short-circuit current noise power at terminals 1'-1' in Fig. 5.18 for a signal bandwidth B is then

$$\overline{\delta i^2} = \int_0^B \mathrm{d}\left(\overline{\delta i^2}\right) = 2ei_SB + 4kT_0\Big[G'_Q + G_n + R_n\left(G'_Q + G_c\right)^2\Big]\Big[1 + \frac{1}{3}\,\frac{B^2}{f_{RE}^2}\Big]B\,. \tag{5.79}$$

[27] See Footnote 11 on Page 122

If $B \ll f_{\mathrm{RE}}$ holds, then the current noise power increases in proportion to the signal bandwidth, $\overline{\delta i^2} \sim B$, while for $B \gg f_{\mathrm{RE}}$ the current noise power increases much stronger, $\overline{\delta i^2} \sim B^3$. To minimize $\overline{\delta i^2}$ for a given signal bandwidth B, the source conductance $G'_Q = G_Q + G_F$ as well as the source capacitance $C'_Q = C_Q + C_F$ should be as small as possible.

For a direct receiver with TIA we find the electrical signal-to-noise power ratio (SNR) γ with the help of Eq. (5.76), (5.79) and (5.75),

$$\gamma = \frac{P_S}{P_R}, \qquad \gamma_{\mathrm{dir}} = \frac{i_S^2}{\overline{|i_{RD}|^2} + \overline{|i'_R|^2}} = \frac{\eta P_e}{2hf_eB} \, \frac{1}{1 + 4kF'T_0 \, G_Q/\left(2eSP_e\right)} \, . \tag{5.80}$$

In most cases of practical interest, the amplifier noise $\overline{|i_R|^2} \gg \overline{|i_{RD}|^2}$ determines the receiver performance (*thermal noise limit*). However, if quantum noise becomes larger than amplifier noise, $\overline{|i_{RD}|^2} \gg \overline{|i_R|^2}$ or $4kF'T_0G_Q/(2eSP_e) \ll 1$, the maximum SNR can be achieved,

$$\gamma_{\mathrm{dir\,qu}} = \frac{\eta P_e}{2hf_eB} \quad (\text{quantum noise limited, } \frac{4kF'T_0G_Q}{2eSP_e} \ll 1). \tag{5.81}$$

This is called the *shot noise* or *quantum noise limit*. In this case, the SNR increases linearly with P_e and depends only on the quantum efficiency η, the signal bandwidth B, and the photon energy hf_e. It is common to choose the symbol duration $T_s = 1/R_s$ (symbol rate R_s, not to be mixed up with the PD series resistance R_S in Fig. 5.3 on Page 113) according to the Nyquist condition $T_s = 1/(2B)$ (sampling theorem). The maximum shot-noise limited SNR is given by the mean number of absorbed photons ηN_e per symbol duration T_s (the absorbed energy in T_s is $\eta P_e T_s = \eta P_e/(2B)$),

$$\gamma_{\mathrm{dir\,qu}} = \eta N_e \, . \tag{5.82}$$

In the shot noise limit, a SNR of 20 dB can be realized for $N_e = 100$ photons per bit (assuming $\eta \approx 1$). By contrast, several 1 000 photons are required to obtain $\gamma_{\mathrm{dir\,qu}} \mathrel{\widehat{=}} 20\,\mathrm{dB}$ when thermal noise dominates the receiver. As a reference, for a 1.55 μm NRZ-OOK shot-noise limited receiver operating at a bit rate of $R_b = R_s = 10\,\mathrm{Gbit/s}$, we receive $N_e = 100$ photons per bit for an average power of $P_e \approx 130\,\mathrm{nW}$.

However, with a simple pin-photodiode receiver the shot noise limit cannot be reached in practice because electronic amplifier noise dominates over shot noise, and a minimum received power of $P_e \approx 1.3\,\mathrm{mW}$ (1 000 photons per bit) required for reaching the shot noise limit would be completely inacceptable.

5.3.2 Signal quality metric for RZ-OOK reception

A physical representation of a direct receiver circuit is shown in Fig. 5.19(a). It corresponds to Fig. 5.17 on Page 130 with the explicit addition of the photodiode circuitry and an equalizer (pulse shaping) network having the complex transfer function $E\,(f)$. The voltage u_A at output A drives a data-recovery section (not drawn) consisting of a decision circuit and a clock-recovery circuit. We assume RZ-OOK modulation. The clock-recovery unit then isolates a spectral component at the clock (= symbol) frequency $1/T$, see Fig. 2.12(c) on Page 37, and helps to synchronize the decision process[28]. Inside each symbol time slot $T_s \equiv T_t \equiv T$ at sampling times t_s determined by the clock-recovery circuit, the decision circuit compares the received signal $u(t_s)$ to a threshold level u_S (*German* Schwelle), and decides whether the signal corresponds to a logical one ($u(t_s) > u_S$) or a logical zero ($u(t_s) < u_S$).

Eye diagram An equalizing filter shapes the received impulse such that the resultant pulse shape $h_A(t)$ has $h_A(t_s) = 1$ for the sampling time $t_s = 0$, and that any interference with impulses $h_A(0 \pm nT_t)$ at neighbouring sampling points $n = 1, 2, \ldots$ disappears, $h_A(\pm nT_t) = 0$ (symbol duration and sampling period T_s as well as the clock period $T_t \equiv T$ are identical). A number of noise-averaged random pulses $\bar{u}_A\,(t)$ is seen in the upper row of Fig. 5.20. The probability of erroneous reception can be judged by superimposing the actual noisy electrical random bit sequences $u_A\,(t)$ inside a symbol duration $T_s \equiv T_t$, lower

[28] See Ref. 17 on Page 6. Sect. 4.3.3 p. 159

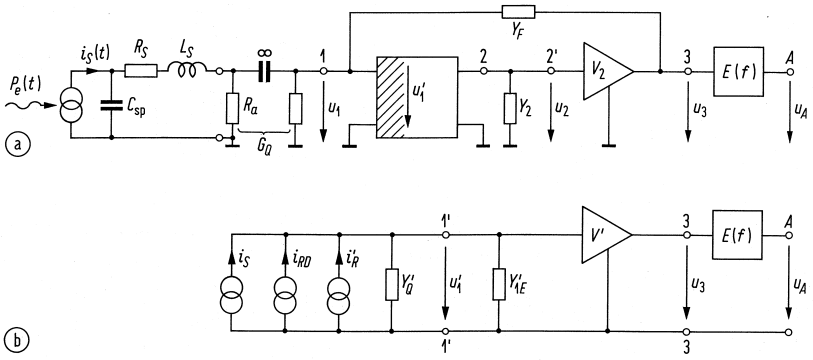


Fig. 5.19. Optical direct receiver circuit with transimpedance amplifier (TIA) and equalizer (pulse shaping) network having a transfer function $u_A/\,u_3 = E\,(f)$. The quantities $i_S = SP_e$ and i_{RD} stand for the phasors of the pin photodetector signal and noise currents, Eq. (5.50) on Page 123. The phasors i_R' and Y_Q' represent the noise current of the TIA and its source admittance, Eq. (5.72) on Page 130. The voltage u_A at output A drives a data-recovery section (not drawn) consisting of a decision circuit and a clock-recovery circuit. (a) Circuit schematic with feedback admittance Y_F connecting terminals 3 and 1 of the TIA. (b) Simplified equivalent circuit, see also Fig. 5.17 on Page 130.

graphs in Fig. 5.20. The resulting shapes resemble an eye and are therefore called eye diagrams. Figure 5.20(a) represents RZ-OOK symbols which are confined to their respective time slots, Fig. 5.20(b) refers to widened pulses that are shaped such that no intersymbol interference occurs at the sampling points t_s, and Fig. 5.20(c) shows strongly widened RZ-OOK pulses with significant inter-symbol interference.

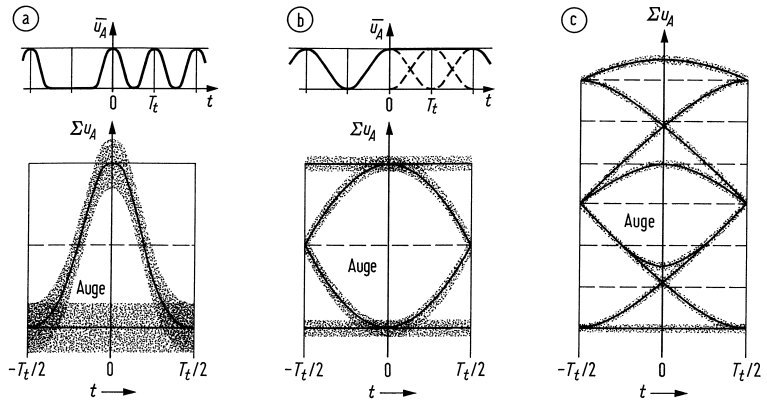


Fig. 5.20. Eye diagrams for RZ pulses, sampling time $t = 0$. Solid lines: No noise. (a) Large noise, no impulse overlap (b) Optimum case: small noise, impulse overlap, but no intersymbol interference at sampling time (c) Low noise, strong impulse overlap, strong intersymbol interference at sampling time. Symbol duration $T \equiv T_s \equiv T_t$, Auge = eye

In the following, the expectations of the voltage $u_A\,(t)$ at sampling time are denoted as $u_{0,1}$ where $u_0 = 0$ is assumed,

$$\begin{aligned} u(t) &= u_{R0}(t) && \text{(0-level received)},\\ u(t) &= u_{R1}(t) + u_1 h_A(t) && \text{(1-level received)}. \end{aligned} \tag{5.83}$$

The best sampling time is found when the signal level difference between noisy 1-level and noisy 0-level is maximum. The optimum decision threshold will be determined in the next sections.

Noise properties The quantities $u_{R0}(t)$, $u_{R1}(t)$ are random voltages. For being definite, and because this assumption holds true if electronic noise dominates, we assume for $u_{R0,1}$ Gaussian probability density functions (PDF) with an expectation zero. The PDF for the decision circuit voltages at the sampling

times are $w_0(u)$ and $w_1(u)$ for a received 0 and 1, respectively,

$$\begin{aligned} w_0(u) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{u^2}{2\sigma_0^2}\right), \qquad \overline{u} = u_0 = 0, \quad \overline{(u-\overline{u})^2} = \overline{u_{R0}^2} = \sigma_0^2, \\ w_1(u) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(u-u_1)^2}{2\sigma_1^2}\right), \quad \overline{u} = u_1, \qquad \overline{(u-\overline{u})^2} = \overline{u_{R1}^2} = \sigma_1^2. \end{aligned} \tag{5.84}$$

Because a logical 1 is transmitted at a higher optical power level than bit 0, its shot noise variance could be slightly larger, $\sigma_1^2 \geq \sigma_0^2$, see Eq. (5.43). Figure 5.21 displays the probability densities $w_0(u)$ and $w_1(u)$.

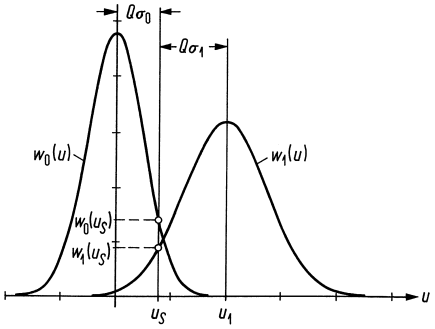


Fig. 5.21. Monomodal probability densities $w_0(u)$, $w_1(u)$ of sampled input voltage of the decision circuit for the received symbols 0 and 1. Standard deviations σ_0, σ_1, expectation of voltage for a received one u_1, specific choice of decision threshold u_S fixed by the bit error parameter Q

Optimum decision threshold The probability of erroneously *d*eciding 1 when 0 is *r*eceived is $p(1\mathrm{d}|0\mathrm{r})$ (and $p(0\mathrm{d}|1\mathrm{r})$ for the opposite case). The probabilities of receiving logical 0 and logical 1 are $p(0\mathrm{r})$ and $p(1\mathrm{r})$, respectively. The optimum decision threshold u_S minimizes the bit error ratio (BER, bit error probability) and requires $\partial\,\mathrm{BER}\,/\partial u_S = 0$,

$$\mathrm{BER} = p(1\mathrm{r})p(0\mathrm{d}|1\mathrm{r}) + p(0\mathrm{r})p(1\mathrm{d}|0\mathrm{r}), \qquad p(0\mathrm{r}) + p(1\mathrm{r}) = 1, \tag{5.85a}$$

$$\mathrm{BER} = p(1\mathrm{r}) \int_{-\infty}^{u_S} w_1(u)\,\mathrm{d}u + p(0\mathrm{r}) \int_{u_S}^{+\infty} w_0(u)\,\mathrm{d}u, \tag{5.85b}$$

$$\frac{\partial\,\mathrm{BER}}{\partial u_S} = p(1\mathrm{r})w_1(u_S) - p(0\mathrm{r})w_0(u_S) \stackrel{!}{=} 0, \tag{5.85c}$$

$$p(1\mathrm{r})w_1(u_S) = p(0\mathrm{r})w_0(u_S). \tag{5.85d}$$

With equal probabilities of logical 0 and 1, $p(1\mathrm{r}) = 1 - p(0\mathrm{r}) = \frac{1}{2}$, the optimum decision threshold for monomodal probability density functions $w_{0,1}(u)$ as in Fig. 5.21 (only one maximum) is to be found at the intersection $w_1(u_S) = w_0(u_S)$, independent of the detailed shapes of the PDF.

Bit error ratio and decision threshold for Gaussian noise Because electronic receiver noise dominates in most practical cases, we now assume Gaussian PDF as in Eq. (5.84) and substitute them in Eq. (5.85b). For performing the integrals, we need the formulae for the error function $\mathrm{erf}(z)$ and the complementary error function $\mathrm{erfc}(z)$,

$$\begin{aligned} &\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2)\,\mathrm{d}t, \qquad \mathrm{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty \exp(-t^2)\,\mathrm{d}t, \\ &\mathrm{erf}(\infty) = 1, \qquad \mathrm{erfc}(\pm z) = 1 \mp \mathrm{erf}(z). \end{aligned} \tag{5.86}$$

An approximation for $z > 2$ with an error $< 3\,\%$ is specified[29] in the first part of Eq. (5.87), while a display of $\lg\{-\lg[\mathrm{erfc}(z)]\}$ vs. $\lg z$ results essentially in a straight line,

$$\mathrm{erfc}(z) = \frac{\exp(-z^2)}{\sqrt{\pi z^2}}\left[1 - \frac{1}{2z^2} + \ldots\right], \quad \lg\{-\lg[\mathrm{erfc}(z)]\} \approx \{z \gg 1\} \approx 2\lg z + 0.434\,. \tag{5.87}$$

For Gaussian density functions

$$w(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(z-A)^2}{2\sigma^2}\right] \tag{5.88}$$

we apply Eqs. (5.86), (5.87) and find

$$\int_{z_1}^{z_2} w(z)\,\mathrm{d}z = \frac{1}{2}\,\mathrm{erf}\left(\frac{z-A}{\sigma\sqrt{2}}\right)\Bigg|_{z_1}^{z_2} = \frac{1}{2}\,\mathrm{erfc}\left(\frac{z-A}{\sigma\sqrt{2}}\right)\Bigg|_{z_2}^{z_1}. \tag{5.89}$$

Referring to Fig. 5.21 and Eq. (5.84) we have

$$\begin{aligned} p(0\mathrm{d}|1\mathrm{r}) &= \int_{-\infty}^{u_S} w_1(u)\,\mathrm{d}u = \frac{1}{2}\,\mathrm{erfc}\left(\frac{u_1 - u_S}{\sigma_1\sqrt{2}}\right), \\ p(1\mathrm{d}|0\mathrm{r}) &= \int_{u_S}^{\infty} w_0(u)\,\mathrm{d}u = \frac{1}{2}\,\mathrm{erfc}\left(\frac{u_S}{\sigma_0\sqrt{2}}\right). \end{aligned} \tag{5.90}$$

The results Eq. (5.90) have to be substituted into Eq. (5.85b). Because of the typical values

$$\begin{aligned} &1 < \sigma_1/\sigma_0 \le \sqrt{2}\,, \qquad \sigma_1 \approx \sigma_0 \quad \text{(electronic noise dominates)}, \\ &p(1\mathrm{r}) = \frac{\sigma_1}{\sigma_0}p(0\mathrm{r}) \gtrapprox p(0\mathrm{r}), \quad p(1\mathrm{r}) \approx p(0\mathrm{r}) \approx 1/2, \end{aligned} \tag{5.91}$$

the relation $p(1\mathrm{r})/\sigma_1 = p(0\mathrm{r})/\sigma_0$ can be replaced approximately by $p(1\mathrm{r}) \approx p(0\mathrm{r}) \approx 1/2$. For $p(1\mathrm{r})/\sigma_1 = p(0\mathrm{r})/\sigma_0$ we find the following simple rule for an optimum decision threshold (see Fig. 5.21, $u_0 = 0$),

$$\begin{aligned} &Q = \frac{u_1 - u_0}{\sigma_0 + \sigma_1} = \frac{u_1 - u_S}{\sigma_1} = \frac{u_S - u_0}{\sigma_0}\,, \\ &p(0\mathrm{d}|1\mathrm{r}) = p(1\mathrm{d}|0\mathrm{r}) = \tfrac{1}{2}\,\mathrm{erfc}(Q/\sqrt{2}\,), \\ &\sigma_0 w_0(u_S) = \sigma_1 w_1(u_S) = \exp(-Q^2/2)/\sqrt{2\pi}\,. \end{aligned} \tag{5.92}$$

Substituting into Eq. (5.85a) (5.85d), the minimum bit error probability reads

$$\mathrm{BER} = \frac{1}{2}\,\mathrm{erfc}\left(\frac{Q}{\sqrt{2}}\right), \tag{5.93}$$

$$\mathrm{BER}_{50\,\%\text{-RZ}} = \frac{1}{2}\,\mathrm{erfc}\left(\frac{\sqrt{\gamma}}{\sqrt{2}}\right) \quad \text{for } Q^2 = \frac{1}{2}\frac{u_1^2/2}{\sigma^2} = \frac{P_S^{(50\,\%\text{-RZ})}}{P_R} = \gamma\,. \tag{5.94}$$

The bit-error parameter Q (signal quality factor) assumes values of $Q = 3.7$, 6, 6.7, 7.3, 7.9 for BER = 10^{-4}, 10^{-9}, 10^{-11}, 10^{-13}, 10^{-15}. For a given Q and known standard deviations σ_0, σ_1 of the noise, the signal voltage u_1 and the optimum threshold can be computed from Eq. (5.92).

For $p(1\mathrm{r})/\sigma_1 \neq p(0\mathrm{r})/\sigma_0$ the choice of u_S according to Eq. (5.92) leads again to a bit error probability Eq. (5.93) (i. e., a BER value independent of $p(1\mathrm{r})$ and $p(0\mathrm{r})$), but this threshold is no more optimum, and the BER is no longer minimum. This drawback is counterbalanced by the advantage that the estimated BER does not depend on the actual bit probabilities $p(1\mathrm{r})$ and $p(0\mathrm{r})$.

For 50 %-RZ signals with average power $P_S^{(50\,\%\text{-RZ})}$ for equally distributed logical 0 and 1, a Gaussian $w_0(u)$ centred at $u_0 = 0$, and Gaussians $w_{0,1}(u)$ having the same variances contributing a noise power $P_R = \sigma_{0,1}^2 = \sigma^2$, we find the minimum BER for the optimum threshold $u_S = u_1/2$ from the electrical signal-to-noise power ratio γ, Eq. (5.94). This relation will be verified for a more general case in the following section, Eq. (5.97) on Page 136.

[29] See Ref. 51 on Page 33. Chapter 7 Eq. (7.1.23) p. 298

Relating signal quality Q and signal-to-noise power ratio SNR For computing the BER, we required the actual shapes of the probability density functions $w_0(u)$, $w_1(u)$. Thus, there is no unique dependency of the SNR defined by moments $\overline{u}$, $\overline{u^2}$ up to the second order, compare Eq. (5.80), and the BER. However, the Gaussian distribution is fully determined by moments up to the second order, and a unique connection between the SNR and the BER can be established, if the noise signals at the decision circuit input are truly Gaussian.

From a simple measurement of the mean $\overline{u}$ and the effective fluctuation $\sqrt{\overline{u^2}}$, the SNR can be determined. This is also important for numerical simulations, where it is practically impossible to simulate 100 erroneous bits out of 100×10^9 bits for BER $= 10^{-9}$. The real-time bit rate achieved by a numerical simulation is about 64 kbit/s for a computer with 3×10^9 floating-point operations per second. This leads to a computing time of 18 days *for only one value* of the optical power P_e.

We assume a signal according to Eq. (5.83) and neglect intersymbol interference of the shaped impulses $h_A(t)$, i. e., the impulses are assumed not to interfere with a signal in neighbouring clock periods. Zeros and ones are equally distributed, R_b is the bit rate, and $T_t = 1/R_b$ the clock period. With Eqs. (5.83), (5.84) the mean electrical power P at the decision circuit is

$$\begin{aligned} P &= \frac{1}{2}\left\{ \frac{1}{T_t}\int_{-T_t/2}^{+T_t/2} \overline{[u_1 h_A(t) + \underbrace{u_{R1}(t)}_{\overline{u_{R1}}=0}\,]^2}\,\mathrm{d}t + \frac{1}{T_t}\int_{-T_t/2}^{+T_t/2} \overline{u_{R0}^2(t)}\,\mathrm{d}t \right\} \\ &= \frac{u_1^2}{2} I(h_A) + \frac{1}{2}(\sigma_0^2 + \sigma_1^2), \\ &\qquad I(h_A) = \frac{1}{T_t}\int_{-T_t/2}^{+T_t/2} h_A^2(t)\,\mathrm{d}t\,. \end{aligned} \tag{5.95}$$

If the SRV is computed from Eq. (5.95), and if we substitute u_1 from Eq. (5.92), the ratio of electrical signal power P_S and electrical noise power P_R is

$$\gamma = \frac{P_S}{P_R} = \frac{u_1^2 I(h_A)/2}{(\sigma_0^2 + \sigma_1^2)/2} = Q^2 \frac{(\sigma_0 + \sigma_1)^2 I(h_A)}{\sigma_0^2 + \sigma_1^2}\,. \tag{5.96}$$

Because we typically have $1 < \sigma_1^2/\sigma_0^2 \le 2$, see Eq. (5.91), and the integral assumes values around $I(h_A) = 1/2$ for usual impulse shapes (e. g., for a raised cosine $h_A(t) = \frac{1}{2}\big[1 + \cos(2\pi t/T_t)\big] = \cos^2(\pi t/T_t)$ we find $I(h_A) = 3/8 \approx 1/2$), we have approximately

$$\gamma = \frac{P_S}{P_R} = (0.97 \ldots 1) \times Q^2, \qquad \rightarrow \qquad \gamma = \frac{P_S}{P_R} = Q^2. \tag{5.97}$$

In summary: If the noise at the decision circuit follows a Gaussian distribution, if the impulse shapes are such that $I(h_A) \approx 1/2$ and if $1 < \sigma_1^2/\sigma_0^2 \le 2$, then the bit-error parameter Q and the BER can be deduced from a measurement of γ. For a special case this was already shown in Eq. (5.94) on Page 135. For BER $= 10^{-9}$ we have from Eq. (5.93) $Q = 6$, and from Eq. (5.97) $\gamma = 36 \mathrel{\widehat{=}} 15.6\,\mathrm{dB}$ follows.

Assuming a signal-independent noise power P_R, (e. g., electronic receiver noise dominates), the signal quality factor Q is proportional to the optical signal power, $Q \sim P_e$. A display of $\lg(-\lg \mathrm{BER})$ vs. $\lg P_e$ then results approximately in a straight line, see Eq. (5.93) and (5.87) on Page 135.

Power penalty Depending on the circumstances, an additional noise source contributing an additional electrical noise power P_{Rz} can be compensated by increasing the electrical signal power from P_S to P_{S+}. Because of Eq. (5.97) we write

$$\gamma = Q^2 = \frac{P_S}{P_R} = \frac{P_{S+}}{P_R + P_{Rz}}\,. \tag{5.98}$$

We define the quantities γ_+, Q_+, BER_+ by

$$\gamma_+ = Q_+^2 = \frac{P_S+}{P_R}\,, \qquad \mathrm{BER}_+ = \frac{1}{2}\,\mathrm{erfc}\left(\frac{Q_+}{\sqrt{2}}\right). \tag{5.99}$$

The SRV which would have been measured for an increased electrical signal power P_{S+} without additional noise sources is denoted as γ_+.

Additive noise If P_{Rz} stands for signal-independent additive electrical noise power (*German* zusätzliche *R*auschleistung), we calculate from Eqs. (5.98), (5.99)

$$\frac{P_{S+}}{P_S} = \left(\frac{Q_+}{Q}\right)^2 = 1 + \frac{P_{Rz}}{P_R}\,. \tag{5.100}$$

Additional additive noise can always be compensated by an increased optical power P_e at the receiver input. For the electrical signal power we have the relation $P_S \sim i_S^2 = S^2 P_e^2$, an we define a power penalty (*German* Leistungs-*B*uße) by

$$p_B = 10\,\lg\left(\frac{P_{e+}}{\overline{P_e}}\right) = 5\,\lg\left(\frac{P_{S+}}{P_S}\right) = 10\,\lg\left(\frac{Q_+}{Q}\right) = 5\,\lg\left(1 + \frac{P_{Rz}}{P_R}\right). \tag{5.101}$$

Multiplicative noise If the additional noise power is always in proportion to the signal power $P_S \sim P_{Rz}$ it can be specified by a fixed residual SNR $\gamma_R = Q_R^2$ (*German* *R*est-Signal-zu-Geräusch-Leistungs-Verhältnis),

$$P_{Rz} = \frac{1}{\gamma_R}\,P_S = \frac{1}{Q_R^2}\,P_S\,, \qquad \gamma_R = Q_R^2 = \frac{P_S}{P_{Rz}}\,. \tag{5.102}$$

For the case of additive noise (Eq. (5.98)) the bit error parameter Q could be increased and the BER decreased arbitrarily by increasing the signal power P_S. For multiplicative noise (Eq. (5.102)) the BER is limited to a minimum BER_R by Q_R. Substituting the additional noise specified in Eq. (5.102) into Eq. (5.98), one calculates

$$\gamma = Q^2 = \frac{P_S}{P_R} = \frac{P_{S+}}{P_R + P_{Rz}} = \frac{P_{S+}}{P_R + P_{S+}/Q_R^2} = \frac{Q_+^2 Q_R^2}{Q_+^2 + Q_R^2}\,. \tag{5.103}$$

With Q_+/Q from Eq. (5.103) the power penalty Eq. (5.101) becomes

$$p_B = 10\,\lg\left(\frac{\overline{P_{e+}}}{\overline{P_e}}\right) = 10\,\lg\left(\frac{Q_+}{Q}\right) = 5\,\lg\left(\frac{Q_R^2}{Q_R^2 - Q^2}\right). \tag{5.104}$$

Only for $Q_R > Q$ (i. e., $P_{Rz} < P_R$) the additional noise can be compensated by an increased signal power. For $p_B \to \infty$ the bit error parameter reaches the limit $Q \to Q_R$. The BER approaches asymptotically the residual or floor error probability (*German* *R*est-Bitfehlerwahrscheinlichkeit)

$$\mathrm{BER}_R = \frac{1}{2}\,\mathrm{erfc}\left(\frac{Q_R}{\sqrt{2}}\right). \tag{5.105}$$

Substituting Q from Eq. (5.103) into Eq. (5.93) leads to

$$\mathrm{BER} = \frac{1}{2}\,\mathrm{erfc}\left(\frac{Q}{\sqrt{2}}\right) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{1}{\sqrt{2}}\,\frac{Q_+ Q_R}{\sqrt{Q_+^2 + Q_R^2}}\right). \tag{5.106}$$

In Fig. 5.22 the bit error probability Eq. (5.106) is displayed as a function of Q and of Q_+. For additive noise only (i. e., $Q_R \to \infty$, see Eq. (5.102)) the BER decreases monotonically with increasing Q (increasing signal power P_S, see Eq. (5.97)). At $Q = Q_0 = 6$ an value of $\mathrm{BER} = 10^{-9}$ is reached. With multiplicative noise as in Eq. (5.102), characterized by a floor error probability (Eq. (5.105)) of $\mathrm{BER}_R = 2.6 \times 10^{-23}$, 5.2×10^{-15}, 2.0×10^{-12} ($Q_R = 9.9$, 7.7, 6.9), a bit error probability of $\mathrm{BER} = 10^{-9}$ requires $Q_{+1} = 7.55$,

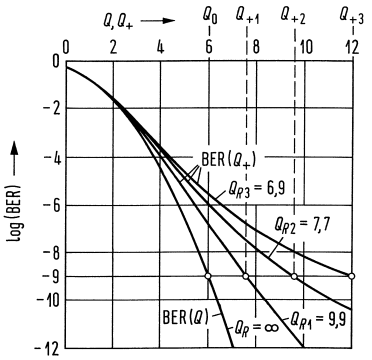


Fig. 5.22. Bit error probability from Eq. (5.106) as a function of the bit error parameters Q (denoted as BER(Q)) and Q_+ (denoted as BER(Q_+)) for various values of the residual bit error parameter Q_R

$Q_{+2} = 9.5$, $Q_{+3} = 12$. After Eq. (5.99), Q_+ measures the increase in signal power, $Q_+ \sim \sqrt{P_{S+}} \sim \overline{P_{e+}}$). The power penalties Eq. (5.104) amount to $p_B = 1\,\mathrm{dB}$, $2\,\mathrm{dB}$, $3\,\mathrm{dB}$.

An example for multiplicative noise Eq. (5.102) are uncertainties (jitter) of the sampling time. If the sampling at $t = 0$ takes place with an effective uncertainty of Δt, the residual bit error parameter is[30]

$$Q_R = \frac{\sqrt{128}}{\pi^2(\Delta t/T_t)^2}\,. \tag{5.107}$$

For $Q_R = 9.9$ ($p_B = 1\,\mathrm{dB}$ at $\mathrm{BER} = 10^{-9}$) the maximum jitter is given by $\Delta t/T_t \leq 0.34$, i. e., $\Delta t = 85\,\mathrm{ps}$ at a bit rate of $R_b = 1/T_t = 4\,\mathrm{Gbit/s}$.

Quantum limit From Eqs. (5.82), (5.97), (5.93) we estimate a shot-noise limited bit error probability

$$\mathrm{BER} = \frac{1}{2}\,\mathrm{erfc}\left(\sqrt{\eta N_e}\Big/\sqrt{2}\right). \tag{5.108}$$

For an optimum receiver with $\eta = 1$ and $\mathrm{BER} = 10^{-9}$ the mean number of photons per 1-bit is $N_e = 36$. As stated after Eqs. (5.82), actual direct receivers are thermal-noise limited.

The BER expression Eq. (5.108) is not truly accurate, since its derivation is based on the Gaussian approximation for the receiver noise statistics. For an ideal detector (no electronic noise, no dark current, quantum efficiency $\eta = 1$), we have $\sigma_0 = 0$, and vanishing quantum noise in the absence of incident optical power, so the decision threshold can be set arbitrarily close to the 0-level signal. Indeed, for such an ideal receiver, 1-bits can be identified without error as long as at least one photon is detected. An erroneous detection occurs only if a 1-bit fails to produce an electron-hole pair. For such a small number of photons and electrons, shot-noise statistics cannot be approximated by a Gaussian distribution, and the exact Poisson statistics Eq. (5.45) must be used[31].

If zero photons arrive for a 0-bit, and $N_e > 0$ is the average number of photons in each 1-bit, the probability that a 0-bit is wrongly taken for a 1-bit is $w(0\mathrm{r}|1\mathrm{d}) = 0$, and the probability of taking wrongly a 1-bit for a 0-bit is $p(0\mathrm{d}|1\mathrm{r}) \neq 0$. For the Poisson distribution the probability of deciding for a 0-bit when actually a 1-bit was received is given by

$$p(0\mathrm{d}|1\mathrm{r}) = p_N(0) = \frac{N_e^0}{0!}\,\mathrm{e}^{-N_e} = \mathrm{e}^{-N_e}\,. \tag{5.109}$$

[30] Shen, T. M.: Power penalty due to decision-time jitter in receivers using avalanche photodiodes. Electron. Lett. 22 (1986) 1043–1045

[31] See Ref. 17 on Page 6. Sect. 4.5.3 p. 175

The probability of a 1-bit be $p(1\mathrm{r}) = 1/2$. The probability of deciding for a 0-bit instead of the received 1-bit is given by Eq. (5.109). The bit error probability according to Eq. (5.85a) is

$$\mathrm{BER} = \overbrace{p(1\mathrm{r})}^{1/2}\overbrace{p(0\mathrm{d}|1\mathrm{r})}^{\exp(-N_e)} + \overbrace{p(0\mathrm{r})}^{1/2}\overbrace{p(1\mathrm{d}|0\mathrm{r})}^{0}, \quad \text{therefore:} \quad \mathrm{BER} = \tfrac{1}{2}\,\mathrm{e}^{-N_e}\,. \tag{5.110}$$

For $\mathrm{BER} = 10^{-9}$ one needs

$$N_e = -\ln\left(2 \times 10^{-9}\right) = 20 \quad \Longrightarrow \quad N_e = 20 \quad \text{for} \quad \mathrm{BER} = 10^{-9} \tag{5.111}$$

photons for a 1-bit. Since this requirement is a direct result of quantum fluctuations associated with the incoming light, it is referred to as the quantum limit. Each 1-bit must contain at least $N_e = 20$ photons to be detected with a $\mathrm{BER} < 10^{-9}$. This requirement can be converted into power by using the relation $P_e = N_e h f_e / T_t$. The receiver sensitivity, defined as $P_{e\,\mathrm{bit}} = (P_e + 0)/2 = P_e/2$ is given by (for the choice of the signal bandwidth $B = 1/(2T_t)$ see text after Eq. (5.81))

$$\begin{aligned} P_e = \frac{N_e h f_e}{T_t}\,, \qquad P_{e\,\mathrm{bit}} &= \frac{\overline{P_e}}{2} = \frac{N_e h f_e}{2T_t} = N_e h f_e B = 2N_{e\,\mathrm{bit}} h f_e B\,, \\ N_{e\,\mathrm{bit}} &= (N_e + 0)/2 = N_e/2\,. \end{aligned} \tag{5.112}$$

Therefore an average number of $N_{e\,\mathrm{bit}} = N_e/2 = 10$ photons per bit (1-bit and 0-bit with equal probability) must be received. This represents the absolute quantum limit for an ideal binary system with direct reception for $\mathrm{BER} = 10^{-9}$. Most receivers operate 20 dB or more above the quantum limit. This is

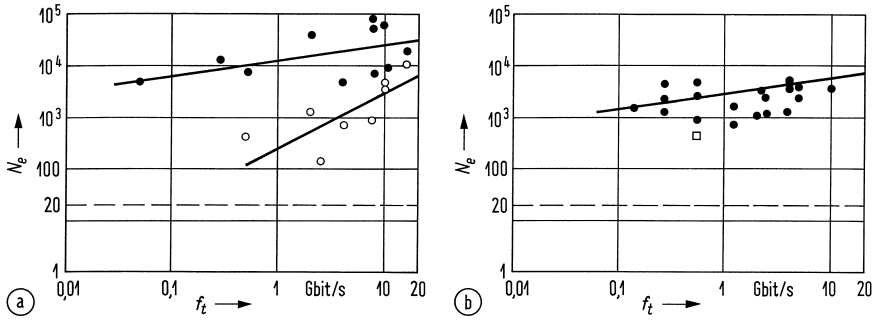


Fig. 5.23. Measured minimum received photon numbers N_e for a 1 bit (η not known, $\mathrm{BER} = 10^{-9}$). Quantum limit $N_e = 20$, $\eta = 1$ (– – –). (a) pin-photodiode $\lambda = 1.3, 1.55\,\mu$m (•), pin-photodiode with optical amplifier (○) (b) avalanche photodiode (APD) $\lambda = 1.55\,\mu$m (•), $\lambda = 0.85\,\mu$m (□)

equivalent to saying that N_e typically exceeds 1 000 photons in practical pin-photodiode receivers. With avalanche photodiodes (APD), optical preamplifiers or with heterodyne reception the quantum limit can be approached more closely as will be discussed in a later section. Figure 5.23 displays some experimental data. With direct reception and for $\mathrm{BER} = 10^{-9}$, typical receiver sensitivities in terms of the number N_e of received photons per bit slot T_t are

$$\begin{aligned} N_e &= 4\,000 \quad \text{pin-photodiode,} \\ N_e &= 150 \quad \text{APD,} \\ N_e &= 152 \quad \text{pin-photodiode and optical preamplifier.} \end{aligned} \tag{5.113}$$

5.4 Coherent receiver

Coherent reception[32,33] allows the full recovery of an optical field $E_s(t) = \hat{E}_s \cos(\omega_s t + \varphi_s)$ with signal bandwidth B, transferred to a lower frequency range, but requires a copolarized[34] optical local oscillator (LO) field $E_O(t) = \hat{E}_O \cos(\omega_O t + \varphi_O)$ as a reference. Depending on the value of $f_Z = |f_s - f_O|$ ($\omega_Z = 2\pi f_Z$, intermediate frequency (IF), *German* Zwischenfrequenz), a coherent receiver is classified as a heterodyne receiver ($f_Z > B$, usually $f_Z > 3B$, see Fig. 5.25 on Page 143), as an intradyne receiver, ($0 < f_Z < B$), or as a homodyne receiver ($f_Z = 0$). The sensitivity of a coherent optical receiver is much better than the sensitivity for direct detection, and shot-noise limited reception becomes possible even with low-power signals.

Figure 5.24(a) displays a schematic of a simple unbalanced coherent receiver with only one photodiode (PD), while Fig. 5.24(b) shows a more elaborate balanced receiver with two photodiodes (PD 1, PD 2). Each PD provides an IF photocurrent $i_Z \cos(\omega_Z t + \varphi_s - \varphi_O)$ at its output. The IF output current of the balanced receiver results from the difference of the individual PD currents, $i(t) = i_2(t) - i_1(t)$. More details will be given in Sect. 5.4.1 on Page 144 ff.

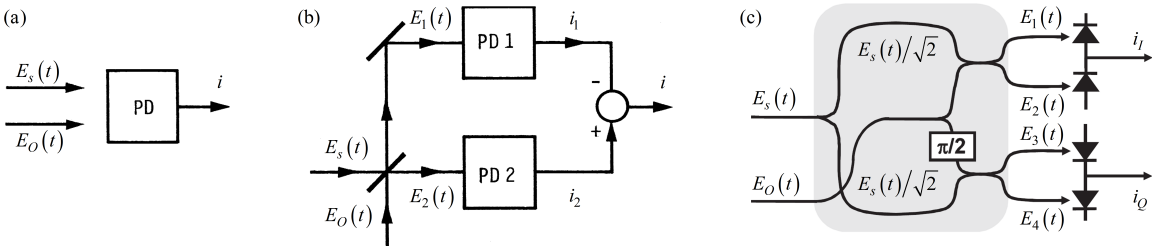


Fig. 5.24. Heterodyne receiver for mixing the superposition of an optical signal field $E_s(t) = \hat{E}_s(t)\cos(\omega_s t + \varphi_s)$ and a copolarized optical local oscillator (LO) field $E_O(t) = \hat{E}_O \cos(\omega_O t + \varphi_O)$ on a photodedetctor (PD), resulting in a photocurrent $i_Z(t)\cos(\omega_Z t - \varphi_O)$ at the intermediate frequency (IF) $f_Z = f_s - f_O = \omega_Z/(2\pi)$. (a) Unbalanced receiver with one PD. (b) Balanced receiver with beam splitter and two photodetectors PD 1 and PD 2. The output IF current $i(t) = i_2(t) - i_1(t)$ represents the difference of the individual photocurrents. [Modified from Ref. 32. Folie 2] (c) Optical hybrid (shaded box) with 2×2 directional couplers (instead of the beam splitter in Subfigure (b)) and an additional $\pi/2$ phase shifter for the LO (which lacks in Subfigure (b)). The circuit is equivalent to the schematic of an IQ-demodulator in Fig. 2.7(b) on Page 29. [After Ref. 33 Fig. 2.6]

For simplifying the calculations, we normalize the electric field strengths (unit $\sqrt{\mathrm{W}}$) of signal $E_s(t)$ and LO $E_O(t)$ such that the associated optical powers as averaged over an optical period are $P_s = \frac{1}{2}\hat{E}_s^2$ and $P_O = \frac{1}{2}\hat{E}_O^2$. We further assume that the IF frequency $f_Z = |f_s - f_O|$ is small ($f_Z < 100\,\mathrm{GHz}$) compared to the optical frequencies ($f_s \approx f_O \approx 200\,\mathrm{THz}$).

Photomixing A photodetector as in Sect. 5.1 on Page 109 ff. delivers electrons at a rate $i(t)/e$ that is determined by the received photon rate $P_e(t)/(hf_s)$, Eq. (5.119). Consider the superposition of a signal and an LO field with frequency $f_O = f_s \pm f_Z$. The envelope of the superimposed (beating) fields is periodic with the difference frequency f_Z,

$$\begin{aligned} E_e(t) &= \hat{E}_s \cos(\omega_s t) + \hat{E}_O \cos(\omega_O t) = \hat{E}_s \cos(\omega_s t) + \hat{E}_O \cos\big[(\omega_s \pm \omega_Z)\, t\big] \\ &= \big[\hat{E}_s + \hat{E}_O \cos(\omega_Z t)\big] \cos(\omega_s t) \mp \hat{E}_O \sin(\omega_Z t) \sin(\omega_s t)\,. \end{aligned} \tag{5.114}$$

[32] G. Grau: Grundlagen der Kohärenten Optischen Nachrichtentechnik. Institut für Hochfrequenztechnik und Quantenelektronik (IHQ), Universität Karlsruhe. Lecture notes WS 1996/1997

[33] K. Kikuchi: Coherent optical communications: Historical perspectives and future directions. In: M. Nakazawa et al. (Eds.): High spectral density optical communication technologies. Optical and Fiber Communications Reports 6, Berlin: Springer-Verlag 2010. Chapter 2

[34] However, polarization diversity receivers can handle arbitrary polarizations with respect to the polarization of the LO. This can be done in hardware, or in software by digital signal processing. An example for the procedure is to be found in: R. Schmogrow, P. C. Schindler, C. Koos, W. Freude, J. Leuthold: Blind polarization demultiplexing with low computational complexity. IEEE Photon. Technol. Lett. 25 (2013) 1230–1233

The rate of arriving photons with energy hf_s depends on the power $P_e(t)$ which results when averaging $\cos^2(\omega_s t)$, $\sin^2(\omega_s t)$ and $\cos(\omega_s t)\sin(\omega_s t) = \frac{1}{2}\sin(2\omega_s t)$ over an optical cycle of $1/f_s$,

$$P_e(t) = \overline{E_e^2(t)} = \tfrac{1}{2}\left(\hat{E}_s^2 + \hat{E}_O^2\right) + \hat{E}_s\hat{E}_O\cos(\omega_Z t)\,, \qquad \omega_Z = 2\pi f_Z = \omega_s - \omega_O\,. \tag{5.115}$$

This scheme is named heterodyne reception. Obviously, the received photon rate $P_e(t)/(hf_s) \sim \cos\omega_Z t$ varies with the intermediate frequency f_Z, and therefore the current rate $i(t)/e \sim \cos\omega_Z t$ reproduces this periodicity. It would be the wrong idea with this type of photodetector to blindly calculate the square of Eq. (5.114), and then worry about the sum frequency $\omega_s + \omega_O$ in the product term $2\hat{E}_s\hat{E}_O\cos(\omega_s t)\cos(\omega_O t)$ — this type of photodetector *cannot emit*, e. g., green light at $\lambda = 0.53\,\mu$m when fed with infrared light having wavelengths of $\lambda_{s,O} = 1.06\,\mu$m.

In preparation of a more detailed description of a heterodyne receiver, the subsequent two paragraphs first discuss the essential properties of a beam splitter and an optical hybrid, and then consider local oscillator noise.

Beam splitter and optical hybrid If a beam splitter is lossless, its scattering matrix is unitary. With a proper choice of the reference planes, the fields at the output of a symmetric, matched beam splitter as in Fig. 5.24(b) are

$$E_1(t) = \frac{1}{\sqrt{2}}\left[E_s(t) - E_O(t)\right], \qquad E_2(t) = \frac{1}{\sqrt{2}}\left[E_s(t) + E_O(t)\right]. \tag{5.116}$$

Such a four-port is mostly used in form of an optical 2×2 directional coupler realized in fibre or in integrated technology.

Figure 5.24(c) displays the schematic of an IQ-demodulator as in Fig. 2.7(b) on Page 29. Optical 2×2 directional couplers are the basic building blocks. Such a circuit (without the photodiode mixers) is called an optical hybrid.

Relative intensity noise and phase noise Both setups Fig. 5.24(a) and (b) have the same limiting sensitivity as long as the LO behaves ideally, i. e., if only the shot (quantum) noise of an (in the classical sense ideally stable) oscillator has to be taken into account. In practice, a laser oscillator with an average output power $\overline{P_O}$ exhibits also classical amplitude noise, so-called relative intensity noise (RIN) with a one-sided power spectrum $\mathrm{RIN}(f)$, which describes power fluctuations due to amplified spontaneous emission (ASE),

$$\mathrm{RIN}_{P_O} = \int_{-\infty}^{+\infty} \mathrm{RIN}\left(f, \overline{P_O}\right) \mathrm{d}f = \frac{\overline{\left(P_O - \overline{P_O}\right)^2}}{\overline{P_O}^2} = \frac{\overline{\delta P_O^2}}{\overline{P_O}^2}\,, \qquad \mathrm{RIN}\left(f, \overline{P_O}\right) = c_{P_O}\frac{\mathrm{RIN}(f)}{\overline{P_O}^3}\,. \tag{5.117}$$

Fortunately, for semiconductor lasers, $\mathrm{RIN}\left(f, \overline{P_O}\right)$ decreases[35] with $c_{P_O}/\overline{P_O}^3$ (c_{P_O} is a constant). In the following, we drop the bar over $\overline{P_O}$ and represent the average just by P_O, if not stated otherwise.

Spontaneous emission is responsible for phase noise, too. This phase noise is characterized by the variance $\sigma_{\varphi_i}^2$ of stationary random phase differences $\varphi_i(t, \tau_0) = \varphi(t + \tau_0) - \varphi(t)$ ($\tau_0 = T$ could be the time between two phase-encoded symbols, i. e., the symbol duration; the random phase $\varphi(t)$ itself does not belong to a stationary process). The variance $\sigma_{\varphi_i}^2 = 2\pi\Delta f_H \tau$ determines the laser linewidth Δf_H which is measured for an observation time τ. Further we saw in Eq. (3.119) on Page 91 how the laser linewidth Δf_H becomes broadened depending on the average output power $P_a = P_O$. For a simplified description, we neglect absorption loss ($\alpha_V = 0$), refer to Eq. (3.73) on Page 80, and substitute the loss due to the finite mirror reflectivity $v_g\alpha_R = 1/\tau_P$ by the reciprocal photon lifetime τ_P.

Combining the informations on $\sigma_{\varphi_i}^2$ and Δf_H, we can establish a relation for the phase noise variance in terms of observation time τ (again, $\tau = T$ could be chosen to be the symbol duration) and various

[35] See Ref. 47 on Page 89. Sect. 6.3.4, p. 303. Typical data: $\mathrm{RIN}_{\mathrm{dB}}(f) = 10\lg\left(\mathrm{RIN}(f)/1\,\mathrm{Hz}^{-1}\right) = -160\ldots-120\,\mathrm{dB\,Hz^{-1}}$

laser parameters,

$$\sigma^2_{\varphi_i} = 2\pi \Delta f_H \tau, \qquad \Delta f_H = \text{const} \times n_{\text{sp}}(1+\alpha^2)\frac{h f_O}{P_O\, \tau_P^2}\,. \tag{5.118}$$

The linewidth Δf_H of the "hot" (oscillating) laser is in proportion to the square of the linewidth $1/\tau_P$ of the "cold" resonator (no oscillation), so a narrow resonator bandwidth decreases Δf_H greatly. The quantities n_{sp} and α are the inversion factor Eq. (3.41) on Page 70 ($n_{\text{sp}} = 1$ for ideally full inversion) and the line broadening factor (Henry factor for amplitude-phase coupling) Eq. (3.113) on Page 90, respectively. From Eq. (5.118) it can be concluded that phase noise influences are minimum for high symbol rates (small $\tau = T$), optimum inversion ($n_{\text{sp}} = 1$), small α-factor, large optical power P_O, and a small "cold" resonator bandwidth.

In the following sections, we first investigate heterodyne reception, and then specialize in homodyne and intradyne receivers.

5.4.1 Heterodyne reception

In a heterodyne[36] receiver the incoming modulated signal light $E_s\,(t)$ is superimposed with light from a copolarized[37] local oscillator $E_O\,(t)$, the power of which is usually much larger than the signal power (but there are exceptions!). We choose unbalanced reception as in Fig. 5.24(a). The photodetector converts the slowly varying optical power $P_e\,(t)$ of the superposition to a current $i\,(t)$ (additive mixing, see Eq. (2.34) on Page 26),

$$P_e\,(t) = \left[\hat{E}_s \cos\left(\omega_s t + \varphi_s\right) + \hat{E}_O \cos\left(\omega_O t + \varphi_O\right)\right]^2, \qquad i = S P_e\,, \qquad S = \frac{\eta e}{h f_O}\,, \tag{5.119}$$

$$P_s\,(t) = \tfrac{1}{2}\hat{E}_s^2\,, \qquad P_O = \tfrac{1}{2}\hat{E}_O^2\,, \qquad P_O \gg P_s\,.$$

When performing the squaring operation in Eq. (5.119), we respect the physical restrictions of the detection process as formulated in Eq. (5.115). The resulting photocurrent (usually amplified with a transimpedance amplifier, Page 130 ff.) comprises an IF component with amplitude i_Z,

$$i\,(t) = S P_e\,(t) \approx S P_O + i_Z \cos\left(\omega_Z t + \varphi_s - \varphi_O\right), \quad i_Z = S \hat{E}_s \hat{E}_O\,, \quad P_s \ll P_O\,. \tag{5.120}$$

Remarkably, the signal amplitude $\hat{E}_s$ in the IF current amplitude i_Z is multiplied by the (large) LO field strength $\hat{E}_O$. By a proper evaluation of the IF current, we can retrieve both, amplitude $\hat{E}_s$ and phase information φ_s of the signal, provided that amplitude $\hat{E}_O$ and phase φ_O of the LO are sufficiently stable.

If $P_O \gg P_s$ holds, the LO contribution dominates the photocurrent shot noise. In addition we have to regard the RIN of the LO according to Eq. (5.47) on Page 122 and Eq. (5.117) on Page 141. However, for the moment we disregard any RIN of the LO, and the photocurrent noise power is then

$$\mathrm{d}\left(\overline{\delta i^2}\right) = 2eSP_O\,\mathrm{d}f\,, \qquad \overline{|i_{RD}|^2} = 2eSP_O \times 2B\,. \tag{5.121}$$

While a large LO power increases the shot noise power, the electronic signal power increases in proportion. As we will see in the next section, this allows shot-noise limited reception even with small optical signals, as opposed to direct reception.

A schematic spectrum is displayed in Fig. 5.25. The baseband signal spectrum $\breve{E}\,(f) = \breve{E}^*\,(-f)$ of the real signal $E\,(t)$ with bandwidth B is transferred to the signal carrier frequency f_s and forms the upper sideband (USB) $\breve{E}\,(f - f_s)$ and the correlated lower sideband (LSB) $\breve{E}^*\,(-f + f_s)$, see Eq. (2.36) on Page 26. The mixing with a LO at frequency f_O shifts the optical spectrum to the IF at $f_Z = f_s - f_O$, where $\breve{E}\,(-f + f_Z) = \breve{E}^*\,(f - f_Z)$ holds. As indicated by the triangular signal spectra in Fig. 5.25, the

[36] Edwin Howard Armstrong, American inventor, ⋆ New York (NY) 18.12.1890, † New York City 1.2.1954. Laid the foundation for much of modern radio and electronic circuitry, including the regenerative and superheterodyne ("superhet") circuits in 1918, and the frequency modulation (FM) system in 1933. After a stint as an instructor at Columbia University, he joined the US Army Signal Corps laboratories in World War I in Paris. Armstrong returned after the war to Columbia University to become assistant to Michael Pupin, the notable physicist and inventor and his revered teacher.— What was called superhet technique at Armstrong's time is now known as heterodyne reception, see also Ref. 39 on Page 145.

[37] See Footnote 34 on Page 140

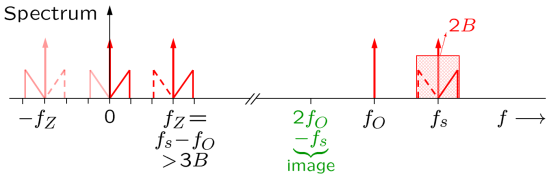


Fig. 5.25. Heterodyne spectra (also homodyne spectra for $f_O = f_s$). A real signal with bandwidth B is modulated on an optical carrier with frequency f_s. Upper (normal position) and lower optical sidebands (inverted position) are complex conjugates and therefore correlated, see Eq. (2.36) on Page 26. This optical signal together with a local oscillator (LO) at frequency f_O illuminates a photodiode and is down-converted to a current at an intermediate-frequency $f_Z = f_s - f_O$. For the IF, the condition $f_Z > B$ must be fulfilled, otherwise the IF-USB at negative frequencies overlaps with the IF-LSB for positive frequencies, and this would lead to distortions. For direct detection of the IF signal, the IF should be chosen according to $f_Z > 3B$. This avoids that in the case of direct IF detection, the baseband signal is perturbed by mixing products of the IF sidebands, which would fall into a frequency range $0 \le f \le 2B$.

USB is in normal position (*German* Gleichlage, larger positive baseband frequencies correspond to larger IF components), while the LSB is in an inverted position (*German* Kehrlage, larger positive baseband frequencies correspond to smaller IF components).

The condition $f_Z > B$ must be fulfilled, otherwise the IF-USB at negative frequencies overlaps with the IF-LSB for positive frequencies, and this would lead to distortions. For direct detection of the IF signal, an IF $f_Z > 3B$ should be chosen. This avoids interference with mixing products of the IF sidebands, which would fall into a frequency range $0 \le f \le 2B$, see Fig. C.2(d) on Page C.2 of Appendix C. The bandwidth of the electronic IF amplifier after the photodetector must be $2B$ to cover both USB and LSB.

Heterodyne reception limit

For calculating the SNR in the IF range, we first have to average the squared electrical IF signal over an intermediate frequency cycle, which results in the average electrical signal power $P_S = i_Z^2/2$. The electrical noise power P_R is determined by photodiode shot noise $\overline{|i_{RD}|^2} = 2eSP_O \times 2B$, Eq. (5.47) on Page 122, caused by the strong LO (where we neglect for the moment any RIN), and by electrical noise $\overline{|i'_R|^2}$ from the (transimpedance) amplifier. Analogous to Eq. (5.80) on Page 132 we find the IF SNR

$$\gamma = \frac{P_S}{P_R}, \quad \gamma_{\text{IF-het}} = \frac{i_Z^2/2}{\overline{|i_{RD}|^2} + \overline{|i'_R|^2}} = \frac{\frac{1}{2}S^2\hat{E}_s^2\,2P_O}{(2eSP_O + 4kF'T_0G_Q)\,2B} = \frac{\eta P_s}{2hf_OB}\,\frac{1}{1 + +\frac{4kF'T_0G_Q}{2eSP_O}}\,. \quad (5.122)$$

If the LO power is chosen large enough, we actually realize shot (quantum) noise limited reception, even with small received optical signal powers P_s,

$$\gamma_{\text{IF-het qu}} = \frac{\eta P_s}{2hf_OB} = \gamma_{\text{dir qu}} \quad \text{(quantum noise limited, } \frac{4kF'T_0G_Q}{2eSP_O} \ll 1\text{).} \quad (5.123)$$

This SNR is the same as for direct reception (Eq. (5.81) on Page 132), $\text{SNR}_{\text{IF-het qu}} = \text{SNR}_{\text{dir qu}}$, albeit both SNR are not really comparable — we relate the IF-band SNR of heterodyne reception to the baseband SNR of direct reception. For a fair comparison we have to transfer the IF band electronically to the baseband. This demodulation can be accomplished either coherently (by mixing with an electrical local oscillator at frequency f_Z), or incoherently (with a rectifier). These subtleties will be discussed in the context of intradyne reception, Sect. 5.4.3 on Page 146 ff.

If the $\text{SNR}_{\text{IF-het qu}}$ is large enough, and if the phase information $\varphi_s - \varphi_O$ need not be recovered, incoherent demodulation is sufficient. The demodulation process adds the signal components in USB and LSB coherently (by amplitude, because they are correlated), while the associated USB and LSB noise components add incoherently (by power, because the noise is uncorrelated). As a result, the electrical signal power quadruples, while the noise power in the baseband (BB) frequency range $0 \ldots B$ doubles. Therefore, the shot-noise limited BB SNR only doubles as compared to the IF-band SNR,

$$\gamma_{\text{BB-het qu}} = \frac{\eta P_s}{hf_OB} = 2\gamma_{\text{dir qu}} \quad \text{(quantum noise limited, } \frac{4kF'T_0G_Q}{2eSP_O} \ll 1\text{).} \quad (5.124)$$

Compared to the quantum noise limited SNR $\gamma_{\text{dir qu}}$ for direct reception, Eq. (5.80) on Page 132 and Eq. (5.121) on Page 142, the baseband SNR for heterodyne reception is twice as large. It corresponds to two times the mean number ηN_s of received signal photons per symbol duration $T_s = 1/(2B)$,

$$\gamma_{\text{BB-het qu}} = 2\eta N_s \,. \tag{5.125}$$

Influence of amplitude and phase noise of the LO

Any real-world LO shows relative intensity noise (RIN) and phase noise, so its field amplitude $\hat{E}_O(t)$, its power $P_O(t)$, and its phase $\varphi_O(t)$ fluctuate around their expectations, Eq. (5.117) and (5.118) on Page 141.

Amplitude noise When taking RIN into account, the SNR relation Eq. (5.122) on Page 143 has to be modified. As in Eq. (5.47) on Page 122 and with the help of Eq. (5.117) on Page 141, the photodiode noise power is

$$\overline{|i_{RD}|^2} \;\longrightarrow\; 2eSP_O\,2B + \big(SP_O\big)^2 \int_{f_Z-B}^{f_Z+B} \text{RIN}\,(f, P_O) \approx 2eSP_O\,2B + \big(SP_O\big)^2 \frac{c_{P_O}}{P_O^3}\,\text{RIN}\,(f_Z)\;2B.$$

The resulting SNR in the IF band including the influence of RIN then becomes

$$\begin{aligned}\gamma_{\text{IF-het}} &= \frac{\frac{1}{2}S^2\hat{E}_s^2\,2P_O}{\big(2eSP_O + \big(SP_O\big)^2\frac{c_{P_O}}{P_O^3}\,\text{RIN}\,(f_Z) + 4kF'T_0G_Q\big)2B} \\ &= \frac{\eta P_s}{2hf_OB}\,\frac{1}{1+\eta c_{P_O}\,\text{RIN}\,(f_Z)\,/(2hf_OP_O^2) + 4kF'T_0G_Q/(2eSP_O)}.\end{aligned} \tag{5.126}$$

If shot-noise limited reception should be achieved, the LO power must be chosen large enough to render RIN and electronic amplifier noise unimportant. If in the case of significant LO RIN an unbalanced receiver would be used, the LO laser should have $20\ldots25\,\text{dB}$ more power[38] than the incoming signal.

Phase noise If the actual LO phase fluctuates, the constant phase in Eq. (5.120) on Page 142 has to be replaced by a random phase $\varphi_O \to \varphi_O(t)$. The IF current is inevitably influenced, and a measurement of the signal phase φ_s becomes inaccurate to a certain degree. The only countermeasure is to reduce the LO phase noise variance $\sigma^2_{\varphi_i}$, see Eq. (5.118) on Page 142.

Balanced heterodyne reception

We now investigate a balanced heterodyne receiver according to Fig. 5.24(b) on Page 140 and consider RIN and phase noise from the LO. Its field amplitude $\hat{E}_O(t)$, its power $P_O(t)$, and its phase $\varphi_O(t)$ fluctuate, Eq. (5.117) and (5.118). With Eq. (5.116) and in analogy to Eq. (5.120) we find the photocurrents

$$\begin{aligned} i_{1,2}(t) = SE_{1,2}^2(t) &\approx \frac{1}{2}SP_O(t) \mp \frac{1}{2}i_Z(t)\cos\left[\omega_Z t + \varphi_s - \varphi_O(t)\right], \quad i_Z(t) = S\hat{E}_s\hat{E}_O(t)\;, \quad P_s \ll P_O \\ i(t) = i_2(t) - i_1(t) &= i_Z(t)\cos\left[\omega_Z t + \varphi_s - \varphi_O(t)\right]. \end{aligned} \tag{5.127}$$

The IF current amplitude $i_Z(t)$ is identical to the one of an unbalanced heterodyne receiver, Eq. (5.120) on Page 142 and Eq. (5.117) on Page 141. However, the balanced heterodyne receiver eliminates the term $SP_O(t)$ and thus also the local oscillator's RIN, which is detected in both photodetectors alike: The fully correlated RIN current fluctuations in i_1 and i_2 cancel in the difference current i.

Still, even with a balanced receiver, the IF current amplitude $i_Z(t)$ is slightly perturbed by the fluctuating $\hat{E}_O(t)$, but much less than through the LO RIN when using an unbalanced receiver. The LO

[38] Infinera Corporation White Paper: Coherent DWDM technologies. Document Number WP-CT-10-2012 http://www.infinera.com/pdfs/whitepapers/Infinera_Coherent_Tech.pdf

phase noise $\varphi_O(t)$, however, directly influences the photocurrent phase even for a balanced receiver. Thus the LO quality $\sigma^2_{\varphi_i}$, i. e., the linewidth Δf_H and the observation time τ in terms of the symbol duration T determine the reception quality for phase sensitive modulation formats.

If $P_O \gg P_s$ holds, the photocurrent shot noise is dominated by the LO. The shot noise fluctuations δi_1 and δi_2 of the photodetector currents are statistically independent, and we find according to Eq. (5.42) and (5.47) on Page 122 for the differential current fluctuations $\mathrm{d}\left(\overline{\delta i^2}\right)$ and for the equivalent photodiode shot noise $\overline{|i_{RD}|^2_{\mathrm{eq}}}$ of the balanced receiver output current $i = i_2 - i_1$

$$\mathrm{d}\left(\overline{\delta i^2}\right) = \mathrm{d}\left(\overline{(\delta i_2 - \delta i_1)^2}\right) = \mathrm{d}\left(\overline{\delta i_1^2}\right) + \mathrm{d}\left(\overline{\delta i_2^2}\right) = 2 \times 2e\left(\tfrac{1}{2} S P_O\right) \mathrm{d}f, \quad \overline{|i_{RD}|^2_{\mathrm{eq}}} = 2eSP_O \times 2B\,. \tag{5.128}$$

The equivalent photodiode shot noise $\overline{|i_{RD}|^2_{\mathrm{eq}}}$ for the balanced heterodyne receiver is determined by LO shot noise only, even if RIN of the LO is taken into account.

Quantum noise limited sensitivity The shot (quantum) noise limited sensitivity for both, balanced and unbalanced heterodyne receivers is identical, as can be seen by comparing Eqs. (5.120) and (5.127), and by inspecting Eqs. (5.122), (5.123) and (5.124).

The following considerations assume an ideal LO and unbalanced receivers, but the considerations for a noisy LO in heterodyne reception with a balanced receiver can be easily transferred to the special cases of homodyne and intradyne reception.

5.4.2 Homodyne reception

To further improve the receiver sensitivity, homodyne[39] reception can be chosen. In the same setup as with unbalanced heterodyne reception Fig. 5.24 on Page 140, the receiver's copolarized[40] LO must then have the same frequency as the optical carrier so that $f_Z = f_s - f_O = 0$. This implies that the phases of carrier and LO have to be aligned properly by an optical phase-locked loop (PLL). An LO power much larger than the signal power $P_O \gg P_s$ guarantees that mixing products of the signal sidebands, which would fall into a frequency range $0 \le f \le 2B$, see the discussion in Fig. 5.25 on Page 143 and also Fig. C.2(d) on Page C.2 of Appendix C remain below the shot noise power level. The photodetector current Eq. (5.120) then becomes

$$i = S P_e(t) = S P_O + i_Z \cos(\varphi_s - \varphi_O)\,, \quad i_Z = S\hat{E}_s\hat{E}_O\,, \quad f_Z = 0\,, \quad P_s(t) \ll P_O\,. \tag{5.129}$$

The signal-dependent part $i_Z \cos(\varphi_s - \varphi_O)$ of the photocurrent i is maximum if $\varphi_s - \varphi_O = 0$ is chosen, i. e., we receive only the in-phase component with respect to the LO phase, and the quadrature component $\sin(\varphi_s - \varphi_O)$ cannot be detected. This renders homodyne reception more sensitive than heterodyne reception, if we are interested in the signal amplitude $\hat{E}_s$ only. However, when employing an optical hybrid and IQ-demodulation as in Fig. 5.24 on Page 140, both the in-phase and the quadrature component of the signal can be retrieved.

[39] In a first homodyne experiment, R. A. Fessenden demonstrated in 1901 a "direct-conversion heterodyne receiver" or beat receiver as a method of making continuous wave radiotelegraphy signals audible. Fessenden's receiver did not see much application because of its local oscillator's stability problem. While complex isochronous electromechanical oscillators existed, a stable yet inexpensive local oscillator would not be available until the invention of the triode vacuum tube oscillator. In a 1905 patent, Fessenden stated the frequency stability of his local oscillator was one part per thousand. — Reginald Aubrey Fessenden, Canadian-American radio pioneer, ⋆ Milton (Quebec, Canada) 6.10.1866, † Hamilton (Bermuda) 22.7.1932. Broadcast the first program of music and voice ever transmitted over long distances. Working as a tester at the Thomas Edison Machine Works, he met Thomas Edison and in 1887 became chief chemist of the Edison Laboratory at Orange (NJ). In 1890 he became chief electrician at the Westinghouse works at Pittsfield, Mass., and in 1892 turned to an academic career, as professor of electrical engineering first at Purdue University, West Lafayette, Ind., then at the Western University of Pennsylvania (now the University of Pittsburgh), where he worked on the problem of wireless communication. [Cited in parts from `http://en.wikipedia.org/wiki/Heterodyne`]. See also Ref. 36 on Page 142.

[40] See Footnote 34 on Page 140

Phase diversity If both the in-phase component $i_I = i_Z \cos(\varphi_s - \varphi_O)$ and the quadrature component $i_Q = i_Z \sin(\varphi_s - \varphi_O)$ are measured, a phase-diversity reception scheme can be designed by $i_Z = \sqrt{I^2 + Q^2}$. This is the usual practice with so-called lock-in amplifiers (operating up to a few 100 MHz), were weak signals buried in heavy noise are to be detected. The actual phase difference $\varphi_s - \varphi_O$ becomes irrelevant, see also Eq. (2.43) on Page 28 and Eq. (5.143) on Page 150.

Homodyne processing could be also applied to the heterodyne IF signal Eq. (5.120) on Page 142. In fact, if phase information has to be retrieved, the final LO must be always phase-locked. The only exception is a differential encoding of the phase[41], where phase transitions between the present and the previous symbols are evaluated (self-coherent receiver)[42,43]. In this case, no LO is required.

Homodyne reception limit

The limiting sensitivity for homodyne reception is derived in analogy to the heterodyne case, only that we have to observe $f_O = f_s$ which leads to an IF frequency $f_Z = 0$. Further, the optical band is directly transferred to the baseband, see Fig. 5.25 on Page 143. The relevant receiver bandwidth corresponds to the signal bandwidth B, but the average electrical signal power is $P_S = i_Z^2$ (not $P_S = i_Z^2/2$ as before). Compared to $\gamma_{\text{BB-het qu}}$ of Eq. (5.124), the SNR doubles,

$$\gamma_{\text{hom qu}} = \frac{\eta P_s}{\frac{1}{2} h f_O B} = 2\gamma_{\text{BB-het qu}} = 4\gamma_{\text{dir qu}} \quad \text{(quantum noise limited, } \frac{4kF'T_0G_Q}{2eSP_O} \ll 1\text{)}. \tag{5.130}$$

Compared to the quantum noise limited SNR for direct reception, Eq. (5.80) on Page 132, the SNR for homodyne reception is four times as large. It corresponds to four times the mean number ηN_s of received signal photons with energy $hf_s \approx hf_O$ per symbol duration $T_s = 1/(2B)$,

$$\gamma_{\text{hom qu}} = 4\eta N_s \,. \tag{5.131}$$

5.4.3 Intradyne reception

Optical intradyne reception[44] takes an intermediate position between heterodyne and homodyne reception, Fig. 5.26. Compared to a heterodyne receiver, the IF of an intradyne receiver is larger than zero, $f_Z > 0$, but smaller than the signal bandwidth, $0 < f_Z < B$. As a result, it requires only a lower bandwidth for the electronic circuits, and frequency offset as well as phase shift could be compensated more easily by digital signal processing. However, this comes at the prize that a direct detection of the down-converted signal is no longer possible, because negative and positive frequency components overlap, second row of Fig. 5.26. Instead, we *must* employ coherent and phase-locked IQ-demodulation to recover the optical signal's amplitude and phase. This means that in this case we need electrical homodyne detection for establishing a reference phase. Apart from that and generally spoken, intradyne reception embraces both heterodyne and homodyne techniques.

To see this, we start with a detailed discussion of the intradyne setup Fig. 5.27. For simplicity's sake we omit the balanced photoreceivers. As shown in Sect. 38 on Page 144, this leads to identical results if local oscillator RIN is disregarded for the simplified setup.

The incoming optical signal field, assumed to be copolarized[45] with the LO, is written either compactly, or in terms of the real ($\hat{E}_{s,r}$) and the imaginary part ($\hat{E}_{s,i}$) of a complex modulation function as in

[41] See DPSK encoding, Fig. 2.14 on Page 41

[42] J. Li: Optical delay interferometers and their application for self-coherent detection. PhD Thesis, Karlsruhe Institute of Technology, 2012

[43] Li, J.; Billah, M. R.; Schindler, P. C.; Lauermann, M.; Schuele, S.; Hengsbach, S.; Hollenbach, U.; Mohr, J.; Koos, C.; Freude, W.; Leuthold, J.: Four-in-one interferometer for coherent and self-coherent detection. Opt. Express 21 (2013) 13293–13304

[44] F. Derr: Coherent optical QPSK intradyne system: Concept and digital receiver realization. J. Lightwave Technol. 10 (1992) 1290–1296

[45] See Footnote 34 on Page 140

COHERENT OPTICAL TRANSMISSION SYSTEMS (IF = INTERMEDIATE FREQUENCY; B = BANDWIDTH OF BASEBAND SIGNAL)

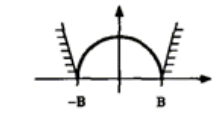
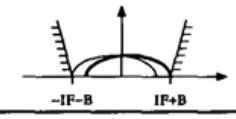
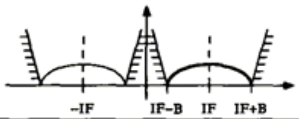
system	IF spectrum	IF
homodyne		IF = 0 Optical PLL
intradyne		IF < B
heterodyne		IF > B

Fig. 5.26. Intradyne IF spectra (IF $\widehat{=}\ f_Z$) compared to heterodyne and homodyne spectra for a baseband spectral width B. [After Ref. 44 Table I]

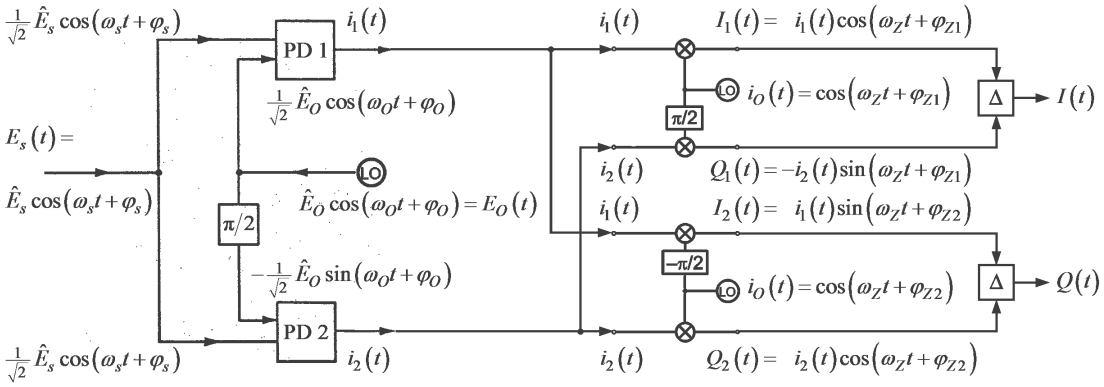


Fig. 5.27. Optical intradyne receiver with electrical IQ-demodulator (simplified optical hybrid as in Fig. 5.24(c) on Page 140, but without balanced photoreceivers). The incoming optical signal field $E_s(t) = \hat{E}_s\cos(\omega_s+\varphi_Z)$ and the in-phase and quadrature component of an optical local oscillator field $E_O(t) = \hat{E}_O\cos(\omega_O+\varphi_Z)$ are superimposed on photodetectors PD_1 and PD_2, which both have the sensitivity $S = \eta e/(hf_s)$. The respective photocurrents comprise an IF component at angular frequency $\omega_Z = \omega_s - \omega_O$ and are in proportion to $i_1(t) \sim i_Z\cos(\omega_Z t+\varphi_s-\varphi_O)$ and $i_2(t) \sim i_Z\sin(\omega_Z t+\varphi_s-\varphi_O)$ with IF amplitude $i_Z = S\hat{E}_s\hat{E}_O/2$. Due to the specific PD properties, see the discussion in Sect. 5.4 on Page 140, the angular sum frequency $\omega_s+\omega_Z$ does not exist. Down-conversion of the IF signal to the baseband can be done either incoherently, or coherently with with another pair of electrical IQ-demodulators as in Fig. 2.7(b) on Page 29. Both electrical IQ-demodulators are different in so far, as $i_1(t)$ in the upper IQ-demodulator and $i_2(t)$ in the lower IQ-demodulator are mixed with $\cos(\omega_Z t+\varphi_Z)$, while $i_2(t)$ in the upper IQ-demodulator is mixed with $-\sin(\omega_Z t+\varphi_Z)$ and $i_1(t)$ in the lower IQ-demodulator is mixed with $\sin(\omega_Z t+\varphi_Z)$. The difference outputs (Δ) of the upper and lower ID-demodulator provide the in-phase (I) and the quadrature (Q) components of the transmitted data. — Another arrangement of connecting $i_{1,2}(t)$ to the electrical IQ-demodulators would be possible, too: Both inputs of the upper IQ-demodulator could be connected to $i_1(t)$, and both inputs of the lower IQ-demodulator could be connected to $i_2(t)$, but then the 4 outputs $I_1(t)$, $Q_1(t)$, $I_2(t)$ and $Q_2(t)$ must be cross-combined according to $I(t) = I_1(t)+I_2(t)$ and $Q(t) = Q_1(t)+Q_2(t)$. The phases of the local IF oscillators must be chosen to be $\varphi_{Z1} = \varphi_{Z2} = \varphi_Z = -\varphi_O$.

Eq. (2.44) on Page 28,

$$E_s(t) = \hat{E}_s\cos(\omega_s+\varphi_s) = \hat{E}_{s,r}\cos\omega_s t - \hat{E}_{s,i}\sin\omega_s t = \Re\left\{\left(\hat{E}_{s,r}+\mathrm{j}\,\hat{E}_{s,i}\right)\mathrm{e}^{\mathrm{j}\,\omega_s t}\right\}, \tag{5.132a}$$

$$\hat{E}_{s,r} = \hat{E}_s\cos\varphi_s, \qquad \hat{E}_{s,i} = \hat{E}_s\sin\varphi_s\,. \tag{5.132b}$$

The optical signal field is split, superimposed with an optical local oscillator $E_O(t) = \hat{E}_O\cos(\omega_O+\varphi_Z)$ ($E_O(t) = -\hat{E}_O\sin(\omega_O+\varphi_Z)$) and detected by photodetectors PD 1 (PD 2). The PD have a sensitivity

$S = \eta e/(hf_s)$, see Eq. (5.17) on Page 113. The respective photocurrents have an IF component at angular frequency $\omega_Z = \omega_s - \omega_O$ and are in proportion to $i_1(t) \sim i_Z \cos(\omega_Z + \varphi_s - \varphi_O)$ ($i_2(t) \sim i_Z \sin(\omega_Z + \varphi_s - \varphi_O)$) with IF amplitude $i_Z = S\hat{E}_O\hat{E}_s$. The LO power $P_O = \frac{1}{2}\hat{E}_O^2$ is assumed to be much larger than the signal power $P_s = \frac{1}{2}\hat{E}_s^2 \ll P_O$, so only the LO contributes shot noise, leading to noise currents $n_{1,2}(t)$. The noise powers $\overline{n_{1,2}^2}$ are specified for the IF band in a bandwidth $2B$. We find for the total PD currents

$$i_{1\,\mathrm{tot}}(t) = \tfrac{1}{4}S\hat{E}_O^2 + \tfrac{1}{2}i_Z\cos(\omega_Z t + \varphi_s - \varphi_O) + n_1(t), \qquad i_1(t) = i_{1\,\mathrm{tot}}(t) - \tfrac{1}{4}S\hat{E}_O^2, \tag{5.133a}$$

$$i_{2\,\mathrm{tot}}(t) = \tfrac{1}{4}S\hat{E}_O^2 + \tfrac{1}{2}i_Z\sin(\omega_Z t + \varphi_s - \varphi_O) + n_2(t), \qquad i_2(t) = i_{2\,\mathrm{tot}}(t) - \tfrac{1}{4}S\hat{E}_O^2, \tag{5.133b}$$

$$i_Z = S\hat{E}_O\hat{E}_s, \quad \overline{n_{1,2}^2} = 2e\left(\tfrac{1}{4}S\hat{E}_O^2\right) \times 2B, \quad P_O = \tfrac{1}{2}\hat{E}_O^2, \quad P_s = \tfrac{1}{2}\hat{E}_s^2, \quad P_s \ll P_O. \tag{5.133c}$$

Due to the specific photodiode properties, see the discussion in Sect. 5.4 on Page 140, the angular sum frequency $\omega_s + \omega_O$ does not exist.

As discussed in Sect. 5.4.1 on Page 144, a real-world implementation would employ balanced photodetectors, so that the DC part is removed from the resulting photocurrents $i_{1,2}(t)$. With real and imaginary signal parts given in Eq. (5.132b) we then write

$$i_1(t) = \tfrac{1}{2}S\hat{E}_O\left[\hat{E}_{s,r}\cos(\omega_Z t - \varphi_O) - \hat{E}_{s,i}\sin(\omega_Z t - \varphi_O)\right] + n_1(t), \quad \hat{E}_{s,r} = \hat{E}_s\cos\varphi_s, \tag{5.134a}$$

$$i_2(t) = \tfrac{1}{2}S\hat{E}_O\left[\hat{E}_{s,i}\cos(\omega_Z t - \varphi_O) + \hat{E}_{s,r}\sin(\omega_Z t - \varphi_O)\right] + n_2(t), \quad \hat{E}_{s,i} = \hat{E}_s\sin\varphi_s. \tag{5.134b}$$

By comparing with Eq. (5.132) on Page 147 we see that $i_1(t)$ and $i_2(t)$ represent the down-converted real part and the imaginary part of the complex transmitted signal, respectivley.

For discriminating between optical and electrical reception, we use in the following the word "detection" if electrical reception is meant.

Incoherent detection With incoherent detection, using an electrical square-law detector and subsequent filtering, we measure the average powers $\overline{i_{1,2}^2(t)}$ in a signal bandwidth B,

$$\overline{i_1^2(t)} = \overline{\left(\tfrac{1}{2}S\hat{E}_O\hat{E}_s\right)^2 \tfrac{1}{2}\left[1 + \cos(2\omega_Z t + 2\varphi_s - 2\varphi_O)\right]} + \overline{n_1^2(t)}, \quad \overline{n_1^2} = 2e\left(\tfrac{1}{4}S\hat{E}_O^2\right) \times B, \tag{5.135a}$$

$$\overline{i_2^2(t)} = \overline{\left(\tfrac{1}{2}S\hat{E}_O\hat{E}_s\right)^2 \tfrac{1}{2}\left[1 - \cos(2\omega_Z t + 2\varphi_s - 2\varphi_O)\right]} + \overline{n_2^2(t)}, \quad \overline{n_2^2} = 2e\left(\tfrac{1}{4}S\hat{E}_O^2\right) \times B. \tag{5.135b}$$

However, this averaging is done to remove the spectrum centred at the harmonic angular frequency $2\omega_Z$, while the baseband signal spectrum must remain untouched. For heterodyne reception this is only possible, if the IF is larger than the signal bandwidth, $f_Z > B$, such excluding intradyne reception with a low IF, $f_Z < B$.

For optical *heterodyne reception* and high IF, specifically for $f_Z > 3B$ as discussed in Sect. 5.4.1 and Fig. 5.25 on Page 143, we then calculate the shot noise limited SNR in the baseband width B,

$$\gamma_{\text{BB-het qu},\,1,2} = \frac{\frac{1}{4}S^2\hat{E}_O^2\hat{E}_s^2\frac{1}{2}}{2e\left(\frac{1}{4}S\hat{E}_O^2\right) \times B} = \frac{\eta P_s/2}{hf_O B} \qquad \text{for} \quad P_s = \tfrac{1}{2}\hat{E}_s^2. \tag{5.136}$$

Phases are not evaluated. Due to the power splitter in the front end of the receiver Fig. 5.27 on Page 147, the input signal power for each PD is halved, $P_s \to P_s/2$. Obviously, one PD would be enough. Then the power splitter can be dropped, and comparing to Eq. (5.123) on Page 143 we see that the SNR are identical, $\gamma_{\text{BB-het qu},\,1,2}(P_s/2 \to P_s) = \gamma_{\text{BB-het qu}}$.

With optical *homodyne reception*, $\omega_Z = 0$ holds, but the LO phase φ_O must be in a fixed relation to the average signal phase $\overline{\varphi_s}$, choosing for instance $\varphi_O = 0$ if $\overline{\varphi_s} = 0$. Locking of the LO to the average phase to the optical signal requires an optical phase-locked loop (OPPL). Phase-locking of optical signals

is principally possible and has been done previously. However, because of the high carrier frequency, it is much more complicated than an electrical PLL. For optical homodyne reception, Eq. (5.134) becomes

$$i_1(t) = \tfrac{1}{2} S \hat{E}_O \hat{E}_{s,r} + n_1(t)\,, \qquad \hat{E}_{s,r} = \hat{E}_s \cos\varphi_s\,, \tag{5.137a}$$

$$i_2(t) = \tfrac{1}{2} S \hat{E}_O \hat{E}_{s,i} + n_2(t)\,, \qquad \hat{E}_{s,i} = \hat{E}_s \sin\varphi_s\,. \tag{5.137b}$$

Obviously, for the down-converted signal the photocurrents i_1 and i_2 recover the I and the Q-component (or real and imaginary part) of the optical signal, respectively.

With two electrical square-law detectors and disregarding the noise contribution ($n_1 = n_2 = 0$), we receive the modulated signal power irrespective of the actual modulated signal phase φ_s,

$$P_s(t) = \frac{i_1^2(t) + i_2^2(t)}{S^2 P_O}\,,\ \varphi_s(t) = \arctan\frac{i_2(t)}{i_1(t)} \quad \text{for } P_s = \tfrac{1}{2}\hat{E}_s^2\,,\ P_O = \tfrac{1}{2}\hat{E}_O^2\,. \tag{5.138}$$

The shot noise limited SNR at any of the outputs i_1 or i_2 with optimally adjusted phases $\varphi_{s,1} = 0$ or $\varphi_{s,2} = \pi/2$ (assuming $\varphi_O = 0$) and the noise powers $\overline{n_{1,2}^2}$ of Eq. (5.135) in the baseband width B is

$$\gamma_{\text{hom qu, 1,2}} = \frac{\frac{1}{4} S^2 \hat{E}_O^2 \hat{E}_s^2}{2e\left(\frac{1}{4} S \hat{E}_O^2\right) \times B} = \frac{\eta P_s/2}{\frac{1}{2} h f_O B} = 2\gamma_{\text{BB-het qu, 1,2}} \qquad \text{for} \quad P_s = \tfrac{1}{2}\hat{E}_s^2\,. \tag{5.139}$$

As before, one PD would be enough. Then the power splitter can be dropped, and comparing to Eq. (5.130) on Page 146 we see that the SNR are identical, $\gamma_{\text{hom qu, 1,2}}(P_s/2 \to P_s) = \frac{1}{2}\gamma_{\text{hom qu}}$. However, if both current outputs were added, the electrical signal power was $\overline{(i_1 + i_2)^2}$. At best, this could contribute double the maximum power for one output, because with $\varphi_s = \pi/4$ a compromise phase has to be found. But now two PD are involved, so the noise power doubles, and therefore the electrical baseband SNR for the combined output currents would be

$$\gamma_{\text{hom qu, 1+2}} = \frac{\frac{1}{4} S^2 \hat{E}_O^2 \hat{E}_s^2 \times 2}{2 \times 2e\left(\frac{1}{4} S \hat{E}_O^2\right) \times B} = \frac{\eta P_s}{h f_O B} = \tfrac{1}{2}\gamma_{\text{hom qu, 1,2}} \qquad \text{for} \quad P_s = \tfrac{1}{2}\hat{E}_s^2\,. \tag{5.140}$$

The SNR in Eq. (5.140) is worse than that of Eq. (5.139) by a factor of two, because both PD are used, and in that sense must be compared directly to Eq. (5.130) on Page 146, $\gamma_{\text{hom qu, 1+2}} = \frac{1}{2}\gamma_{\text{hom qu}}$.

Equations (5.135) reveal that in contrast to optical homodyne reception, an optical *heterodyne* receiver with electrical square-law detection can evaluate the magnitude of the received optical field only, while its phase goes unnoticed: There is no reference phase available. If both the optical amplitude $\hat{E}_s$ and the phase φ_s (or real part $\hat{E}_{s,r} = \hat{E}_s \cos\varphi_s$ and imaginary part $\hat{E}_{s,i} = \hat{E}_s \sin\varphi_s$) are to be received, coherent detection must be employed.

Coherent detection With an optical intradyne receiver according to Fig. 5.27 on Page 147 having a low IF $f_Z < B$, harmonic filtering of the output currents proves to be impossible. Instead, we employ coherent detection with an electrical IQ-demodulator as in Fig. 2.7(b) on Page 29, but without the splitter Σ, upper right in Fig. 5.27. The $i_{1,2}$-terminals of the optical intradyne receiver output are connected to the appropriate input terminals of the electrical IQ-demodulator, where the currents $i_{1,2}$ in Eq. (5.134) on Page 148 are mixed (multiplied) with the in-phase and quadrature component of a local electrical oscillator at an IF f_Z. The in-phase and quadrature signals at the output of IQ-mixer 1 are

$$\begin{aligned} I_1(t) &= i_1(t)\cos(\omega_Z t + \varphi_{Z\,1}) \\ &= \tfrac{1}{4} i_Z \cos(\varphi_s - \varphi_O - \varphi_{Z\,1}) + \tfrac{1}{4} i_Z \cos(2\omega_Z t + \varphi_s - \varphi_O + \varphi_{Z\,1}) + n_1 \cos(\omega_Z t + \varphi_{Z\,1})\,, \end{aligned} \tag{5.141a}$$

$$\begin{aligned} Q_1(t) &= -i_2(t)\sin(\omega_Z t + \varphi_{Z\,1}) \\ &= -\tfrac{1}{4} i_Z \cos(\varphi_s - \varphi_O - \varphi_{Z\,1}) + \tfrac{1}{4} i_Z \cos(2\omega_Z t + \varphi_s - \varphi_O + \varphi_{Z\,1}) - n_2 \sin(\omega_Z t + \varphi_{Z\,1})\,. \end{aligned} \tag{5.141b}$$

A similar procedure applies for a second electrical IQ-demodulator having the same LO frequency f_Z, but a different LO phase setting, lower right in Fig. 5.27. The in-phase and quadrature signals at the output of IQ-mixer 2 are

$$\begin{aligned} I_2(t) &= i_1(t)\sin(\omega_Z t+\varphi_{Z\,2}) \\ &= -\tfrac{1}{4}i_Z\sin(\varphi_s-\varphi_O-\varphi_{Z\,2})+\tfrac{1}{4}i_Z\sin(2\omega_Z t+\varphi_s-\varphi_O+\varphi_{Z\,2})+n_1\sin(\omega_Z t+\varphi_{Z\,2})\,, \quad (5.142a)\\ Q_2(t) &= i_2(t)\cos(\omega_Z t+\varphi_{Z\,2}) \\ &= \tfrac{1}{4}i_Z\sin(\varphi_s-\varphi_O-\varphi_{Z\,2})+\tfrac{1}{4}i_Z\sin(2\omega_Z t+\varphi_s-\varphi_O+\varphi_{Z\,2})+n_2\cos(\omega_Z t+\varphi_{Z\,2})\,. \quad (5.142b)\end{aligned}$$

In both IQ-mixers the electrical LO has to be phase-locked to the average phase $\overline{\varphi_s}$ of the IF data. For IQ-mixer 1 (2) we choose $\varphi_{Z\,1}+\varphi_O=0$ ($\varphi_{Z\,2}+\varphi_O=\pi$), form the difference signals $I=I_1-Q_1$ ($Q=I_2-Q_2$), respectively, thereby eliminate the harmonic spectrum centered at $2f_Z$, and recover the transmitted in-phase and quadrature data $I(t)$ and $Q(t)$ within a signal bandwidth B,

$$I=I_1-Q_1=\tfrac{1}{2}i_Z\cos\varphi_s+n_1\cos(\omega_Z t+\varphi_{Z\,1})+n_2\sin(\omega_Z t+\varphi_{Z\,1}) \quad \text{for } \varphi_{Z\,1}+\varphi_O=0\,, \quad (5.143a)$$

$$Q=I_2-Q_2=\tfrac{1}{2}i_Z\sin\varphi_s+n_1\sin(\omega_Z t+\varphi_{Z\,2})-n_2\cos(\omega_Z t+\varphi_{Z\,2}) \quad \text{for } \varphi_{Z\,2}+\varphi_O=\pi\,, \quad (5.143b)$$

$$i_Z=S\hat{E}_O\hat{E}_s, \quad \overline{n_{1,2}^2}=2e\left(\tfrac{1}{4}S\hat{E}_O^2\right)\times B\,, \quad P_O=\tfrac{1}{2}\hat{E}_O^2, \quad P_s=\tfrac{1}{2}\hat{E}_s^2, \quad P_s\ll P_O\,. \quad (5.143c)$$

Intradyne reception limit

The shot-noise limited SNR for optical IQ intradyne reception and electrical IQ homodyne detection relates for both the electrical I and Q-signals of Eq. (5.143) the respective average signal powers $P_{S\,I,Q}=\frac{1}{4}i_Z^2\frac{1}{2}$ for an equally distributed random signal phase φ_s ($\overline{\cos 2\varphi_s}=0$, $\overline{\sin 2\varphi_s}=0$) to the noise power $P_{R\,I,Q}=\frac{1}{2}\overline{n_1^2}+\frac{1}{2}\overline{n_2^2}=2\times\frac{1}{2}\times 2e\left(\frac{1}{4}S\hat{E}_O^2\right)\times B$ in the baseband width B. For the SNR in both the independent I and Q-channels we therefore find

$$\gamma_{\text{intra qu}\,I,Q}=\frac{P_{S\,I,Q}}{P_{R\,I,Q}}=\frac{\frac{1}{4}S^2\hat{E}_O^2\hat{E}_s^2\frac{1}{2}}{2e\left(\frac{1}{4}S\hat{E}_O^2\right)\times B}=\frac{\eta P_s/2}{hf_O B}=\gamma_{\text{BB-het qu, 1,2}}\,. \quad (5.144)$$

Not unexpectedly, the intradyne SNR per I and Q-channel equals the incoherently detected baseband heterodyne SNR for a high IF, see Eq. (5.136) on Page 148. Because the optical signal Eq. (5.132) on Page 147 is split into its real part $\hat{E}_{s,r}=\hat{E}_s\cos\varphi_s$ and its imaginary part $\hat{E}_{s,i}=\hat{E}_s\sin\varphi_s$, the SNR of Eq. (5.144) corresponds to the SNR of Eq. (5.124) on Page 143, if the signal power is properly replaced, $P_s\to P_s/2$.

If the I and Q-channels carry the same information, we could add the I and Q-signals in Eq. (5.143) if $\varphi_{Z\,1}=\varphi_{Z\,2}$ was chosen. This doubles the signal amplitude, where again random signal phases φ_s are assumed, and it doubles the noise power. Therefore the SNR doubles as compared to Eq. (5.144),

$$\gamma_{\text{intra qu}\,I+Q}=\frac{P_{S\,I+Q}}{P_{R\,I+Q}}=\frac{4\times\frac{1}{4}S^2\hat{E}_O^2\hat{E}_s^2\frac{1}{2}}{2\times 2e\left(\frac{1}{4}S\hat{E}_O^2\right)\times B}=\frac{\eta P_s}{hf_O B}=\gamma_{\text{BB-het qu}}=2\gamma_{\text{dir qu}}\,. \quad (5.145)$$

In this case, the shot-noise limited SNR equals the case for simple optical heterodyne reception and incoherent square-law detection as in Eq. (5.124) on Page 143.

5.4.4 Signal quality metric for QAM reception

Sets of different M-ary modulation schemes such as QPSK, 8PSK, 16PSK and 1616QAM were discussed in Sect. 2.4.2 and displayed in Fig. 2.13(b) on Page 40. For these modulation formats the Q-factor and its relation to BER is not appropriate any more. A new metric has to be defined, namely the error vector magnitude (EVM), which assesses the quality of communication. The EVM expresses the difference

between the expected complex voltage of a demodulated symbol and the value of the actually received symbol. As with the Q-factor, a relation to BER can be established[46].

2198 IEEE PHOTONICS TECHNOLOGY LETTERS, VOL. 24, NO. 23, DECEMBER 1, 2012

Comments and Corrections

Corrections to "Error Vector Magnitude as a Performance Measure for Advanced Modulation Formats"

Rene Schmogrow, Bernd Nebendahl, Marcus Winter, Arne Josten, David Hillerkuss, Swen Koenig, Joachim Meyer, Michael Dreschmann, Michael Huebner, Christian Koos, Juergen Becker, Wolfgang Freude, and Juerg Leuthold

In the above paper [1], equation (4) contains an error. Equation (4) correctly reads:

$$\mathrm{BER} \approx \frac{\left(1 - L^{-1}\right)}{\log_2 L} \mathrm{erfc}\left[\sqrt{\frac{3\log_2 L}{\left(L^2 - 1\right)} \frac{1}{(k\mathrm{EVM}_m)^2 \log_2 M}}\right]. \quad (4)$$

Manuscript received September 11, 2012; accepted September 11, 2012.

R. Schmogrow, A. Josten, D. Hillerkuss, S. Koenig, J. Meyer, M. Dreschmann, M. Huebner, C. Koos, J. Becker, W. Freude, and J. Leuthold are with the Karlsruhe Institute of Technology, Karlsruhe 76131, Germany (e-mail: rene.schmogrow@kit.edu; arne.josten@student.kit.edu; david.hillerkuss@kit.edu; swen.koenig@kit.edu; joachim.meyer@kit.edu; w.freude@kit.edu; michael.huebner@kit.edu; christian.koos@kit.edu; michael.dreschmann@kit.edu; juergen.becker@kit.edu; juerg.leuthold@kit.edu).

B. Nebendahl is with Agilent Technologies, Boeblingen 71034, Germany (e-mail: bernd.nebendahl@agilent.com).

M. Winter was with the Karlsruhe Institute of Technology, Karlsruhe 76131, Germany. He is now with Polytec (e-mail: m.winter@polytec.de).

Digital Object Identifier 10.1109/LPT.2012.2219471

Additionally, the paper should include the following Appendix.

APPENDIX

The OSNR as used in our equations is the optical signal-to-noise power ratio measured in the same bandwidth B_O and in the same polarization as the signal power. This is equivalent to the definition of the electrical SNR and differs from the usual measurement practice, where the optical signal-to-noise power ratio $\mathrm{OSNR}_\mathrm{ref}$ is the total signal power (measured in both polarizations) and the total noise power from both polarizations measured in a reference bandwidth B_ref. The reference bandwidth thereby is defined by a fixed wavelength range $\Delta\lambda_\mathrm{ref} = B_\mathrm{ref} c/\lambda_\mathrm{ref}^2 = 0.1$ nm (vacuum speed of light c) centered at a reference wavelength λ_ref.

The OSNR definition in our equation and the $\mathrm{OSNR}_\mathrm{ref}$ definitions are related by

$$\mathrm{OSNR} = \frac{2B_\mathrm{ref}}{p \cdot B_\mathrm{O}} \mathrm{OSNR}_\mathrm{ref} \quad (6)$$

with $p = 1$ for a single polarization signal and $p = 2$ for a polarization multiplexed signal (see R.J. Essiambre et al; J. of Lightwave Technol., vol. 28, no. 4, pp. 662 ff., April 2011).

If there are signal or noise correlations between the two transmitted polarizations, the applicability of our formulae has to be re-checked.

REFERENCES

[1] R. Schmogrow, *et al.*, "Error vector magnitude as a performance measure for advanced modulation formats," *IEEE Photon. Technol. Lett.*, vol. 24, no. 1, pp. 61–63, Jan. 1, 2012.

[46] Shafik, R. A.; Rahman, M. S.; Islam, A. H. M. R.; Ashraf, N. S.: On the error vector magnitude as a performance metric and comparative analysis. 2nd Intern. Conf. Emerging Technologies (IEEE–ICET 2006), Peshawar, Pakistan, 13–14 November, 2006

IEEE PHOTONICS TECHNOLOGY LETTERS, VOL. 24, NO. 1, JANUARY 1, 2012 61

Error Vector Magnitude as a Performance Measure for Advanced Modulation Formats

Rene Schmogrow, Bernd Nebendahl, Marcus Winter, Arne Josten, David Hillerkuss, Swen Koenig, Joachim Meyer, Michael Dreschmann, Michael Huebner, Christian Koos, Juergen Becker, Wolfgang Freude, and Juerg Leuthold

***Abstract*—We examine the relation between optical signal-to-noise ratio (OSNR), error vector magnitude (EVM), and bit-error ratio (BER). Theoretical results and numerical simulations are compared to measured values of OSNR, EVM, and BER. We conclude that the EVM is an appropriate metric for optical channels limited by additive white Gaussian noise. Results are supported by experiments with six modulation formats at symbol rates of 20 and 25 GBd generated by a software-defined transmitter.**

***Index Terms*—Advanced modulation formats, bit-error ratio (BER), error vector magnitude (EVM), software defined transmitter.**

I. Introduction

COHERENT optical transmission systems and advanced modulation formats such as M-ary quadrature amplitude modulation (QAM) are establishing quickly [1]. To encode these formats a variety of new optical modulator concepts have been introduced [2]. Among them are modulators dedicated to a particular modulation format [3] as well as novel software-defined optical transmitters that allow encoding of many modulation formats at the push of a button [4], [5]. In light of the capabilities to encode such advanced modulation formats there is a need to reliably judge the quality of the encoded signals. In laboratory experiments so far most receivers employ offline digital signal processing (DSP) at much reduced clock rates. This offline processing makes it very time consuming to reliably compute the bit error ratio (BER), especially if the signal quality is high. As a consequence, a faster — yet reliable — performance measure is needed, in particular when investigating wavelength division multiplexing (WDM) [6] or multicarrier systems [7].

Traditionally, the Q-factor metric is well established for on-off keying (OOK) optical systems. To estimate BER from Q, marks and spaces in the detected photocurrent are assumed to be superimposed with additive white Gaussian noise (AWGN), the probability density of which is fully described by its mean and variance. A large Q leads to a small BER.

Unfortunately, the method cannot be simply transferred to QAM signals, where the optical carrier is modulated with multilevel signals both in amplitude and phase. Instead, the error vector magnitude (EVM) is employed. It describes the effective distance of the received complex symbol from its ideal position in the constellation diagram. If the received optical field is perturbed by AWGN only, the EVM can be related to BER and to the optical signal-to-noise ratio (OSNR) [8], [9]. A small EVM leads then to a small BER. The EVM metric is standard in wireless and wireline communications. However, its connection to BER and OSNR is not well established in optical communications. Especially one has to discriminate between data-aided reception, where for measurement purposes the actually sent data are known, as opposed to nondata-aided reception, where the received data are unknown. The first case is standard for BER measurements, while the second case is more common for real-world receivers (disregarding, e.g., training sequences). For strongly noisy signals, nondata-aided reception tends to underestimate the EVM, because a received symbol could be nearer to a "wrong" constellation point than to its "right" position.

In this letter we confirm experimentally and by simulations that the BER can be estimated from EVM data by an analytic relation [8]. Strictly speaking, this BER estimate is valid for data-aided reception only, but we found that the method can be also applied for nondata-aided reception if BER $< 10^{-2}$ holds. Further, the EVM can be estimated [9] if the OSNR has been measured. Both estimates are valid for systems limited by optical AWGN. To support our findings we compare measured OSNR, EVM and BER for symbol rates of 20 GBd and 25 GBd with calculated BER and EVM estimates for the modulation formats binary phase shift keying (BPSK), quadrature PSK (QPSK), 8PSK, 16QAM, 32QAM, and 64QAM.

II. Error Vector Magnitude

A. EVM Definition

Advanced modulation formats such as M-ary QAM encode a data signal in amplitude and phase of the optical electric field. The resulting complex amplitude of this field is described by points in a complex constellation plane. Fig. 1(a) depicts the

Manuscript received April 01, 2011; revised September 27, 2011; accepted October 06, 2011. Date of publication October 17, 2011; date of current version December 16, 2011. This work was supported in part by the projects EuroFOS, ACCORDANCE, and CONDOR, in part by the Xilinx University Program (XUP), in part by Micram Microelectronic GmbH, in part by the Agilent University Relations Program, and part by the Karlsruhe School of Optics & Photonics (KSOP).

R. Schmogrow is with the Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany (e-mail: rene.schmogrow@kit.edu).

B. Nebendahl is with Agilent Technologies, 71034 Boeblingen, Germany (e-mail: bernd.nebendahl@agilent.com).

M. Winter was with Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany. He is now with Polytec (e-mail: m.winter@polytec.de).

A. Josten, D. Hillerkuss, S. Koenig, J. Meyer, M. Dreschmann, M. Huebner, C. Koos, J. Becker, W. Freude, and J. Leuthold are with the Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany (e-mail: arne.josten@student.kit.edu, david.hillerkuss@kit.edu; swen.koenig@kit.edu; joachim.meyer@kit.edu; w.freude@kit.edu; michael.huebner@kit.edu; christian.koos@kit.edu; michael.dreschmann@kit.edu; juergen.becker@kit.edu; juerg.leuthold@kit.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LPT.2011.2172405

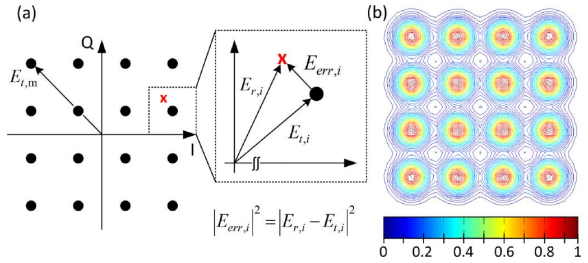


Fig. 1. Constellation diagram and error vector for a 16QAM signal. (a) Ideal constellation diagram with an actually transmitted value $\mathbf{X}$. The blowup illustrates the definition of the ith error vector $E_{\mathrm{err},i}$ in relation to the actually received signal vector $E_{r,i}$ and the vector $E_{t,i}$ of the transmitted signal. (b) Simulated constellation diagram with white Gaussian noise, EVM $= 15\%$. Color coding for reception probability of transmitted symbols.

TABLE I
MODULATION FORMAT-DEPENDENT FACTOR k^2 FROM (2)

	B/Q/8PSK	16QAM	32QAM	64QAM
k^2	1	9/5	17/10	7/3

ideal constellation points for a 16QAM signal. The actually received signal vector E_r deviates by an error vector E_{err} from the ideal transmitted vector E_t. In Fig. 1(b) a simulated noisy constellation is shown. The EVM is defined by a root mean square of E_{err} for a number of I randomly transmitted data [8] and embraces all (linear and nonlinear) impairments:

$$\mathrm{EVM}_m = \frac{\sigma_{\mathrm{err}}}{|E_{t,m}|},\ \sigma_{\mathrm{err}}^2 = \frac{1}{I}\sum_{i=1}^{I}|E_{\mathrm{err},i}|^2,\ E_{\mathrm{err},i} = E_{r,i} - E_{t,i}. \tag{1}$$

The power of the longest ideal constellation vector with magnitude $|E_{t,m}|$ serves for normalization. Other authors use the average power $|E_{t,a}|^2$ of all M symbol vectors within a constellation leading to EVM_a. The two EVM normalizations are related by a modulation format-dependent factor k,

$$\mathrm{EVM}_a = k\,\mathrm{EVM}_m,\ k^2 = \frac{|E_{t,m}|^2}{|E_{t,a}|^2},\ |E_{t,a}|^2 = \frac{1}{M}\sum_{i=1}^{M}|E_{t,i}|^2. \tag{2}$$

Table I specifies the k-values relating the two definitions for the modulation formats discussed here.

B. Relations Between OSNR, EVM, and BER

The EVM_m from (1) can be estimated from OSNR (measured for instance with an optical spectrum analyzer, OSA) [8] according to (3). The basic assumptions are that system errors are mainly due to optical AWGN (neglecting nonlinear effects and electronic noise), that reception is nondata-aided, and that quadratic M-QAM signal constellations are regarded. With (2) we find from [9]:

$$\mathrm{EVM}_m \approx \frac{1}{k}\left[\frac{1}{\mathrm{OSNR}} - \sqrt{\frac{96}{\pi(M-1)\mathrm{OSNR}}}\sum_{i=1}^{\sqrt{M}-1}\gamma_i e^{-3\beta_i^2\mathrm{OSNR}/2(M-1)} + \frac{12}{M-1}\sum_{i=1}^{\sqrt{M}-1}\gamma_i\beta_i\mathrm{erfc}\left(\sqrt{\frac{3\beta_i^2\mathrm{OSNR}}{2(M-1)}}\right)\right]^{1/2},$$
$$\gamma_i = 1 - \frac{i}{\sqrt{M}},\quad \beta_i = 2i-1. \tag{3}$$

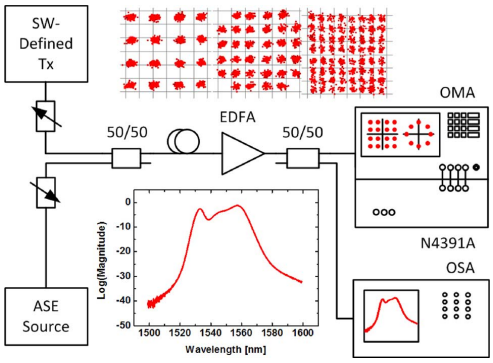


Fig. 2. Experimental setup for BER and EVM measurements. The optical signal of a software-defined transmitter [4] generates a choice of six different modulation formats (three of which are shown) for optical signal-to-noise ratios adjusted by injecting a variable amount of amplified spontaneous emission (ASE). After amplification with an erbium-doped fiber amplifier (EDFA), the OSNR is measured by an optical spectrum analyzer (OSA). The modulation is decoded by an Agilent optical modulation analyzer (OMA).

The first term in (3), i.e., $\mathrm{EVM}_m \approx 1/(k\sqrt{\mathrm{OSNR}})$, rewrites (1) and (2) for the case of data-aided reception, and if optical AWGN is the dominant source of E_{err}. The remaining terms account for nondata-aided reception and disappear for large OSNR. For large numbers of constellation points M only the first few terms in the summation need to be considered.

To estimate a BER from EVM_m we define L as the number of signal levels identical within each dimension of the (quadratic) constellation, and $\log_2 M$ as the number of bits encoded into each QAM symbol. The BER is then approximated by [8]

$$\mathrm{BER} \approx \frac{(1-L^{-1})}{\log_2 L}\mathrm{erfc}\left[\sqrt{\frac{3\log_2 L}{(L^2-1)}\,\frac{\sqrt{2}}{(k\,\mathrm{EVM}_m)^2\log_2 M}}\right]. \tag{4}$$

For (4), the same limitations as with (3) apply, but in this case data-aided reception is assumed. If the EVM_m is not derived by evaluating (3) but measured directly, the influence of electronic noise is also included.

The EVM and the Q^2-factor are related. In direct detection OOK systems assuming *electrical* AWGN with a standard deviation $\sigma_1 \propto \sigma_Q^2$ for the photocurrent $i_1 \propto |E_{t,m}|^2$ of a mark, the Q-factor in the shot-noise limited case, $\sigma_0 \approx 0$, is defined in analogy to (1) by

$$Q = \frac{i_1}{\sigma_1 + \sigma_0} \approx \frac{|E_{t,m}|^2}{\sigma_Q^2},\qquad \sigma_Q^2 = \frac{1}{I}\sum_{i=1}^{I}\left[|E_{r,i}|^2 - |E_{t,i}|^2\right]. \tag{5}$$

The Q^2-factor represents an *electrical* signal-to-noise power ratio and provides for OOK signals a good estimate of the BER $\approx (1/2)\,\mathrm{erfc}(Q/\sqrt{2})$. Conversely, the EVM is based on electrical fields and thus assesses the BER for a variety of formats accounting for both *optical* and *electrical* AWGN.

In the following, we compare the theoretical predictions (3) and (4) with numerical simulations and measurements.

III. EXPERIMENTAL SETUP

We measure OSNR, EVM and BER in a software-defined real-time multi format transmitter setup [4], Fig. 2. We sequentially generate the six complex modulation formats B/Q/8PSK

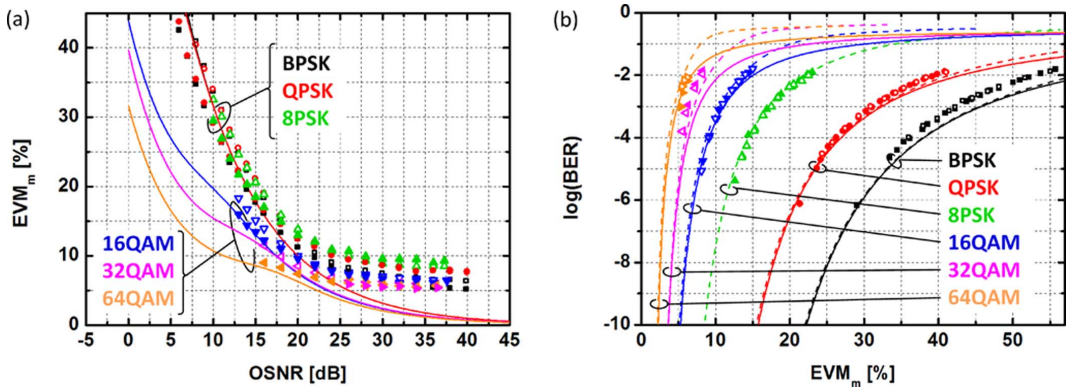


Fig. 3. Interdependencies of OSNR, EVM_m, and BER. Filled symbols represent measurements for a symbol rate of 20 GBd, open symbols for 25 GBd. (a) Measured (symbols) and calculated [8] EVM_m (solid lines) as a function of OSNR. For high OSNR levels, the measured plots have an error floor due to the electronic noise of the transmitter and receiver. The different error floors for Q/8PSK and xQAM stem from different factors k. The error floor for BPSK is lower because of transmitter specific properties. (b) Measured (symbols), simulated (dashed lines), and calculated [7] BER (solid lines) as a function of EVM_m.

and 16/32/64QAM at symbol rates of 20 GBd and 25 GBd on an external cavity laser (ECL) at 1550 nm. The modulated carrier is kept at a fixed average power and combined with a variable noise source (amplified spontaneous emission (ASE) source with attenuator) to vary the OSNR. An optical spectrum analyzer (OSA) determines the amplified signal's OSNR. An Agilent modulation analyzer (OMA) decodes the modulation and measures EVM and BER. The insets of Fig. 2 display the spectrum of the ASE source as well as constellations at 25 GBd of three selected formats and at best available OSNR.

IV. Experimental Results

In Fig. 3(a) we display the measured EVM_m for various measured OSNR values (closed symbols for 20 GBd, open symbols for 25 GBd). The solid lines represent (3). For OSNR < 20 dB the theoretical prediction coincides with the measurement. Constellations of xQAM can be recovered for OSNR > 12 dB only. For OSNR > 20 dB the electronic noise dominates so that (3) does not hold any more. If the electronic noise contribution would be less, as is the case for systems with lower symbol rate and consequently smaller bandwidth, the error floor would be seen at higher OSNR only.

Fig. 3(b) shows the measured BER as a function of the measured EVM (closed symbols: 20 GBd, open symbols: 25 GBd). The solid lines represent (4), the dashed lines result from simulations. While measurement and simulation are based on non-data-aided reception, (4) assumes data-aided detection. Still, measurement, analytical estimate and simulations coincide for a large range up to a BER of 10^{-2}.

Some more information can be extracted from Fig. 3(a) and (b). While the 32QAM constellation is not strictly quadratic, it is nearly so, and hence the estimation quality is comparable to the one for the quadratic formats. The plots also show that the EVM depends on the format, as higher-order formats are more sensitive to noise than others, as predicted by (3) and (4).

For determining the BER we use a $2^{15} - 1$ pseudo random binary sequence. The number of compared bits and the number of recorded errors were chosen according to the statistical reasoning described in [10].

V. Conclusion and Outlook

Complementing the established Q-factor evaluation for OOK systems, the EVM is a quality measure for coherent optical transmission systems with advanced modulation formats. EVM data can be used to reliably estimate the BER.

Experimental OSNR, EVM, and BER data were compared to analytical relations and to direct numerical simulations. They all showed good agreement within the specified limits.

References

[1] S. Okamoto, K. Toyoda, T. Omiya, K. Kasai, M. Yoshida, and M. Nakazawa, "512 QAM (54 Gbit/s) coherent optical transmission over 150 km with an optical bandwidth of 4.1 GHz," in *Proc. ECOC*, Torino, Italy, 2010, Paper PD2.3.

[2] S. Chandrasekhar and X. Liu, "Enabling components for future high-speed coheremt communication systems," in *Proc. OFC*, Los Angeles, CA, 2011, Paper OMU5.

[3] G. Lu *et al.*, "40-Gbaud 16-QAM transmitter using tandem IQ modulators with binary driving electronic signals," *Opt. Express*, vol. 18, no. 22, pp. 23062–23069, 2010.

[4] R. Schmogrow *et al.*, "Real-time software-defined multiformat transmitter generating 64QAM at 28 GBd," *IEEE Photon. Technol. Lett.*, vol. 22, no. 21, pp. 1601–1603, Nov. 1, 2010.

[5] R. Schmogrow *et al.*, "Real-time OFDM transmitter beyond 100 Gbit/s," *Opt. Express*, vol. 19, no. 13, pp. 12740–12749, 2011.

[6] D. Qian *et al.*, "101.7-Tb/s (370 × 294-Gb/s) PDM-128QAM-OFDM transmission over 3 × 55-km SSMF using pilot-based phase noise mitigation," in *Proc. OFC*, Los Angeles, CA, 2011, Paper PDPB5.

[7] D. Hillerkuss *et al.*, "26 Tbit s^{-1} line-rate super-channel transmission utilizing all-optical fast Fourier transform processing," *Nature Photon.*, vol. 5, no. 6, pp. 364–371, Jun. 2011.

[8] R. A. Shafik, M. S. Rahman, and A. H. M. R. Islam, "On the extended relationships among EVM, BER and SNR as performance metrics," in *Proc. 4th ICECE*, 2006, pp. 408–411.

[9] H. Arslan and H. A. Mahmoud, "Error vector magnitude to SNR conversion for nondata-aided receivers," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2694–2704, May 2009.

[10] M. Müller, R. Stephens, and R. McHugh, "Total jitter measurement at low probability levels, using optimized BERT scan method," in *White Paper*. Santa Clara, CA: Agilent Technologies, 2005, pp. 7–9.

5.5 Self-Coherent Receiver

In a self-coherent receiver the signal acts simultaneously as its own local oscillator. Because phase stability must be maintained only over one or several symbol slots, the transmitter laser requirements are much relaxed. We discuss here a specialized DPSK receiver, and a generalized form for more advanced modulation formats.

5.5.1 Differential reception

Differential phase shift keying (DPSK) as described in Fig. 2.14 on Page 41 cannot directly be received with direct detection techniques because the optical phase has to be measured. Therefore a phase-amplitude converter is employed in form of a Mach-Zehnder interferometer with an optical delay in one arm (delay interferometer, DI).

As illustrated in Fig. 3.28(b) on Page 97, any optical phase change $\Delta\vartheta$ is transformed into a change of the optical field amplitude. Figure 5.28(a) shows a setup schematic. The DI splits the phase-modulated signal into two paths, in one of which it is delayed by a symbol duration. At the output coupler, the phase modulated optical field thus interferes with its 1-symbol delayed replica.

This results in destructive (constructive) interference at the destructive (constructive) port whenever there is no phase change, but in constructive (destructive) interference when a phase change of π happened between symbols. A balanced direct detector measures the intensity and delivers an electric output signal which reflects the DPSK coding of the transmitter.

The DI as such converts the phase encoded data signal into an OOK-DB and an OOK-AMI format, see Fig. 2.15 on Page 44 and Sect. 3.3.3 on Page 100 ff., both of which can be directly detected with photodiodes.

5.5.2 Self-coherent reception

As with differential reception, a self-coherent receiver uses instead of a local oscillator laser a delayed version of the signal itself. Compared to conventional differential direct detection receivers, self-coherent receivers utilize advanced DSP algorithms in electronic circuits for signal processing and demodulation, which allows a significant reduction of complexity in the optical frontend for the receiver.

In this case, even for DPSK or PSK signals with more than two phase states, only two DI are needed, and still one can recover phase and intensity of an optical signal at low cost, at the price of a small penalty in sensitivity. The temporal resolution of these schemes is restricted by the DI time delays, which, however, can be made tunable.

With such optical DI, tunable in delay from 0 ps to 100 ps, DQPSK signals at 11.7 GBd, 28 GBd and 42.7 GBd can be received. A simplified amplitude recovery is possible when receiving DPSK or PSK signals which have a constant modulus, e. g., DBPSK, DQPSK, and D8PSK.

A separate amplitude branch method can be used to detect multi-level signals including 8QAM and 16QAM. It needs to be mentioned that a modification in the transmitter is necessary for phase preconditioning. A self-coherent system that detects signals at arbitrary states of polarization was demonstrated with a 112 Gbit/s PMSK-DQPSK signal[47].

[47] Li, J.; Billah, M. R.; Schindler, P. C.; Lauermann, M.; Schuele, S.; Hengsbach, S.; Hollenbach, U.; Mohr, J.; Koos, C.; Freude, W.; Leuthold, J.: Four-in-one interferometer for coherent and self-coherent detection. Opt. Express 21 (2013) 13293–13304

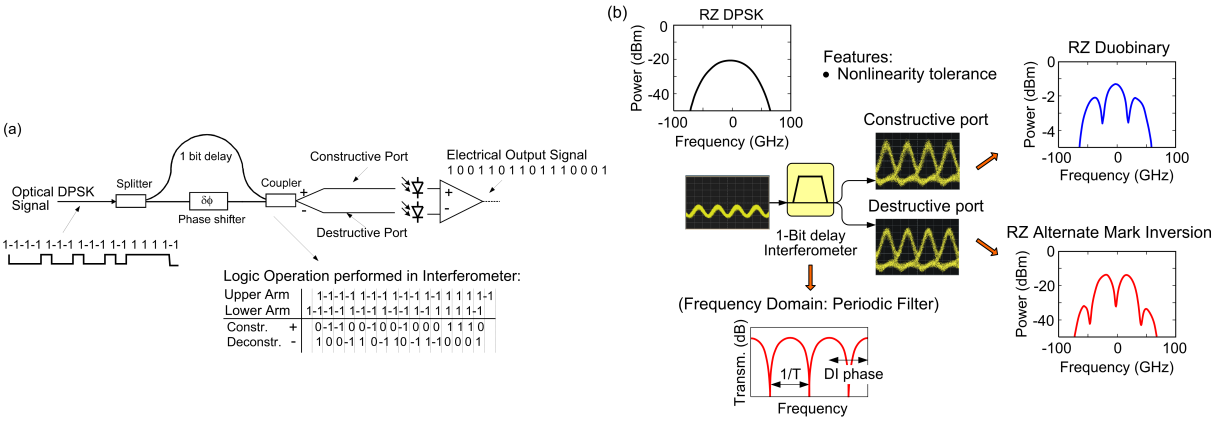


Fig. 5.28. Self-coherent receiver setup (a) Interferometer with 1-symbol delay in one arm, balanced direct photoreceiver, and differential electrical amplifier (b) Input and output spectra of a 1-symbol delay interferometer. Its transfer function is periodic. With an RZ-DPSK input spectrum, the spectra at the constructive and destructive output ports belong to an RZ-DB and an RZ-AMI modulation format, respectively. [After Fig. 2.57 of Ref. † on the Preface page]

5.6 Receiver with optical pre-amplifier

An optical pre-amplifier[48] can improve the receiver sensitivity greatly. This is true for both, direct (incoherent) and coherent receivers. However, because in the end we always detect powers (or the square root of powers), be it at optical frequencies as with a direct receiver, or at an intermediate frequency as with a heterodyne receiver, or in the baseband as with a homodyne receiver, we have the basic problem that a coherent carrier will be mixing with amplifier noise and that the noise itself becomes rectified. This will add uncertainties in interpreting the signal correctly, and thus the actual signal-to-noise ratio is impaired. As we will see, this fundamental impairment is tolerable when compared to the advantages which can be gained.

The problem of quadratically rectifying a coherent carrier embedded in noise has been discussed in detail in Appendix C on Page 197 ff., specifically in Eq. (C.24) and Fig. C.2 on Page 201.

5.6.1 Photodetection of signal and noise

An optical amplifier (OA) with single-pass power gain $\mathcal{G}_s$ increases a small received power P_e to a sufficiently large level $\mathcal{G}_s P_e$ such that any electronic amplifier noise $4kFT_0G_Q/(2eS\mathcal{G}_s P_e) \ll 1$ becomes unimportant, Eq. (5.80) on Page 132, and shot-noise limited reception would be possible. However, the amplifier itself is noisy, and this can be described with the help of Sect. 5.2.3 on Page 128 ff. We further adapt the results displayed in Fig. C.2 on Page 201 to the case of a noisy optical amplifier, Fig. 5.29.

Partial noise currents

The one-sided optical power spectrum $w_s(f)$, Fig. 5.29(a), and the photocurrent power partial spectra $w_i(f)$, Fig. 5.29(b)–(d), are displayed under the assumption of a large single-pass amplifier power gain $\mathcal{G}_s$. For the analytical relations the difference between $\mathcal{G}_s$ and $\mathcal{G}_s - 1$ is observed. From Eq. (5.67) on Page 128 the copolarized ASE noise power is characterized by an approximately constant spectral power

[48] Most often a receiver with optical pre-amplifier is referred to as a "pre-amplifed receiver". In fact, (not only) to me this sounds nonsensical: How can a receiver be amplified? Does it become larger, bigger or heavier? Will it be duplicated (as cells in genetics)? Certainly not. It is the optical signal to which the quality "amplified" or "larger" has to be attributed, and not the receiver. Therefore, you better name the device as I chose the section title to be, or call it at least a "pre-amplifier receiver". — And as an afterthought: "Literacy involves competence in reading, writing, and interpreting texts of various sorts. It involves both skill in decoding and higher levels of comprehension and interpretation." [Cited from Encyclopædia Britannica, Chicago 2008]

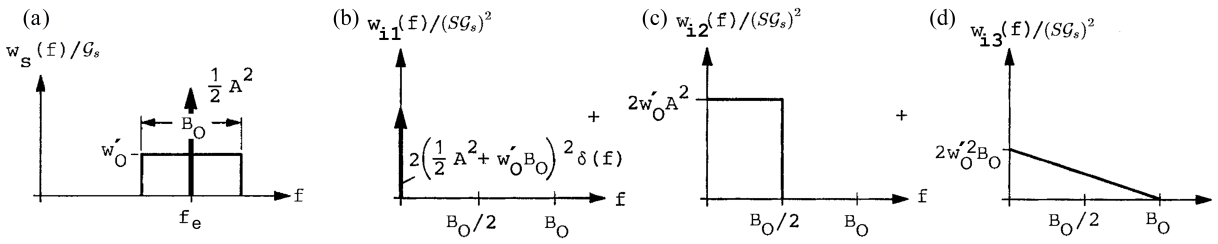


Fig. 5.29. Signal and co-polarized noise at the output of an optical amplifier with single-pass optical power gain $\mathcal{G}_s \gg 1$. (a) One-sided output power spectrum $w_s(f)$ of a sinusoidal optical signal with an input amplitude $A = \sqrt{2P_e}$ and a frequency f_e, superimposed with ASE noise of spectral density $(\mathcal{G}_s - 1)\, w'_O$ ($w'_O = n_{\mathrm{sp}} w_O$, inversion factor n_{sp}, $w_O = hf_e$). The amplifier bandwidth is narrow, $B_O \ll f_e$. Signal and narrowband noise are received with a photodetector. The optical input signal power P_e leads to a photocurrent $i = S\mathcal{G}_s P_e$ where $S = \eta e/(hf_e)$ is the sensitivity. (b) One-sided direct current (DC) power spectrum with photocurrents $i_S = S\mathcal{G}_s A^2/2$ and $i_R = S\,(\mathcal{G}_s - 1)\, w'_O B$. The total DC power is $(i_S + i_R)^2$. The integral over half a Dirac function is $\int_0^\infty \delta(f)\,\mathrm{d}f = \frac{1}{2}$. (c) Carrier-noise interference (d) Noise-noise interference. — Partial detector spectra w_{i1}, w_{i2} and w_{i3} are uncorrelated and may be added. Therefore the total power equals the sum of the partial powers. The graph corresponds to Fig. C.2 on Page 201.

density in an optical bandwidth B_O and by the inversion factor n_{sp} from Eq. (3.40) on Page 70,

$$\frac{P_{\mathrm{ASE},\,x}}{B_O} = (\mathcal{G}_s - 1)\, w'_O\,, \qquad w'_O = n_{\mathrm{sp}} w_O\,, \qquad w_O = hf_e\,. \tag{5.146}$$

The photodetection of signal and noise leads to a direct current (DC) component as displayed in Fig. 5.29(b), to a fluctuating partial photocurrent from carrier-noise interference leading to a constant spectrum Fig. 5.29(c) inside a bandwidth $B_O/2$, and to a fluctuation partial photocurrent from noise-noise interference with a triangular-shaped spectrum Fig. 5.29(d) and a maximum bandwidth B_O. In detail, we have the following one-sided partial power spectra and their associated noise currents in a signal bandwidth $B < B_O$:

1. Rectified signal and noise at $f = 0$ (this is what our eyes register). The one-sided power spectrum is (half) a Dirac function,

$$w_{i\,1}(f) = \left(S\mathcal{G}_s P_e + S\,(\mathcal{G}_s - 1)\, w'_O B_O\right)^2 2\delta(f) \overset{\mathcal{G}_s \gg 1}{=} 2\,(S\mathcal{G}_s)^2\,(P_e + w'_O B_O)^2\,\delta(f). \tag{5.147a}$$

The photocurrent fluctuations in a signal bandwidth B are determined by the coherent carrier, Eqs. (5.43) and (5.45) on Page 122, and by this part of the noise, which is copolarized with the carrier. Noise has a fundamentally different photon probability distribution[49,50] than coherent light, but can be approximated by a Poisson distribution if only a few photons occupy each mode (this is the case for a large noise bandwidth B_O, a small signal bandwidth $B \ll B_O$, and a small total noise power). Therefore we can calculate the current fluctuation from the DC term $i_{S,R} = S\mathcal{G}_s P_e + S\,(\mathcal{G}_s - 1)\, w'_O B_O$,

$$\overline{|i_{RD,1}|^2} = 2e\, i_{S,R}\, B = 2e\left[S\mathcal{G}_s P_e + S\,(\mathcal{G}_s - 1)\, w'_O B_O\right] B \overset{\mathcal{G}_s \gg 1}{=} 2eS\mathcal{G}_s\,(P_e + w'_O B_O)\, B\,. \tag{5.147b}$$

[49] See Ref. 47 on Page 89. Sect. 6.3.6

[50] The Bose-Einstein distribution for the probability $p_{N_P}(N_P)$ that in thermal equilibrium a number of N_P photons (= bosons) is measured per polarization in a total of m transverse and longitudinal modes for an average number of $\overline{N_P}$ photons (see Footnote 15 on Page 124)

$$p_{N_P}(N_P) = \frac{(N_P + m - 1)!}{N_P!(m-1)!}\,\frac{1}{(1 + \overline{N_P}/m)^m (1 + m/\overline{N_P})^{N_P}}\,, \quad \overline{\delta N_P^2} = \overline{(N_P - \overline{N_P})^2} = \underbrace{\overline{N_P}}_{\substack{\text{Particle}\\ \text{aspect:}\\ \text{Poisson}}} + \underbrace{\frac{\overline{N_P}^2}{m}}_{\substack{\text{Wave aspect:}\\ \text{Exponential for } m=1}}.$$

If only one polarization and one transverse mode is regarded (e. g., in a polarization-maintaining single-mode fibre), then m corresponds to the number of longitudinal modes $m = M_L = B_O\tau$ as calculated in Eq. (3.4) on Page 51. The observation time $\tau = 1/B$ is determined by the signal bandwidth, and usually the condition $B \ll B_O$ is fulfilled so that $m \gg 1$ holds. For moderate powers (moderate average numbers of photons $\overline{N_P}$, excluding the explosion of a fusion bomb), the condition $\overline{N_P}/m \ll 1$ is met, so that ASE noise (and also LED radiation) approximates a Poisson distribution very well.

2. Mixing of signal at f_e (optical amplitude $\sqrt{2\mathcal{G}_s P_e} = \sqrt{\mathcal{G}_s}\,A$) and noise "sidebands", which are copolarized with the signal,

$$w_{i\,2}(f) = 4S\mathcal{G}_s P_e\, S\,(\mathcal{G}_s - 1)\, w'_O \overset{\mathcal{G}_s \gg 1}{=} 4\,(S\mathcal{G}_s)^2\, P_e w'_O \qquad \text{for } 0 \le f \le B_O/2\,. \tag{5.148a}$$

The photocurrent fluctuations represent the interference of coherent light and ASE noise, $\overline{|i_{RD,2}|^2} = \int_0^B w_{i\,2}(f)\,\mathrm{d}f$,

$$\overline{|i_{RD,2}|^2} = 4S\mathcal{G}_s P_e\, S\,(\mathcal{G}_s - 1)\, w'_O B \overset{\mathcal{G}_s \gg 1}{=} 4\,(S\mathcal{G}_s)^2\, P_e w'_O B \qquad \text{for } 0 \le B \le B_O/2\,. \tag{5.148b}$$

3. As before, we regard only the noise "lines" which are copolarized with the coherent carrier. Mixing of noise "sidebands" leads to the highest spectral density $2\,(S\mathcal{G}_s)^2\, w_O^2 B_O$ at $f = 0$ because then the number of immediately adjacent noise "lines" is maximum. The spectral density decays linearly to zero, because at $f = B_O$ there is only "one fitting pair of noise lines",

$$w_{i\,3}(f) = 2S^2\,(\mathcal{G}_s - 1)^2\, w'^2_O\,(B_O - f) \overset{\mathcal{G}_s \gg 1}{=} 2\,(S\mathcal{G}_s)^2\, w'^2_O\,(B_O - f) \qquad \text{for } 0 \le f \le B_O\,. \tag{5.149a}$$

The photocurrent fluctuations represent the noise-noise interference, $\overline{|i_{RD,3}|^2} = \int_0^B w_{i\,3}(f)\,\mathrm{d}f$,

$$\overline{|i_{RD,3}|^2} = 2S^2\,(\mathcal{G}_s - 1)^2\, w'^2_O \left(B_O - \frac{B}{2}\right) B \overset{\mathcal{G}_s \gg 1}{=} 2\,(S\mathcal{G}_s)^2\, w'^2_O \left(B_O - \frac{B}{2}\right) B \qquad \text{for } 0 \le B \le B_O\,. \tag{5.149b}$$

Shot (or quantum) noise, carrier-noise mixing, and noise-noise mixing products fall into the signal baseband $0 \le f \le B$. For a sufficiently large gain $\mathcal{G}_s \gg 1$ the shot (or quantum) noise term $\overline{|i_{RD,1}|^2}$ can be neglected compared to the two interference terms, because it increases with $\mathcal{G}_s$ only, while $\overline{|i_{RD,2}|^2}, \overline{|i_{RD,3}|^2}$ increase with $\mathcal{G}_s^2$,

$$\overline{|i_{RD,1}|^2} \ll \overline{|i_{RD,2}|^2}, \overline{|i_{RD,3}|^2} \qquad \text{for} \quad \mathcal{G}_s \gg 1\,. \tag{5.150}$$

5.6.2 Direct pre-amplifier receiver

The signal bandwidth B in relation to the optical amplifier bandwidth B_O strongly influences the achievable signal-to-noise power ratio, as will be shown in the following paragraphs.

Direct reception limit with full OA bandwidth

The optical signal bandwidth is usually much smaller than the optical amplifier bandwidth, $2B \ll B_O$, and without limiting the amplifier bandwidth B_O, we find for the SNR of a pre-amplifier receiver a much smaller value than for shot-noise limited reception (Eq. (5.81 on Page 132) without an optical amplifier,

$$\gamma_{\mathrm{dir\,OA}} \approx \frac{\overbrace{\left(S\mathcal{G}_s P_e\right)^2}^{i_S^2}}{\underbrace{4\,(S\mathcal{G}_s P_e)\, S\mathcal{G}_s w'_O B}_{\overline{|i_{RD,2}|^2}} + \underbrace{2\,(S\mathcal{G}_s)^2\, w'^2_O B_O B}_{\overline{|i_{RD,3}|^2}}} = \frac{\gamma^{(1)}_{\mathrm{dir\,qu}}}{2n_{\mathrm{sp}}} \frac{1}{1 + \frac{n_{\mathrm{sp}}}{2\gamma^{(1)}_{\mathrm{dir\,qu}}} \frac{B_O}{2B}}\,, \quad 2B \ll B_O, \tag{5.151a}$$

$$\gamma_{\mathrm{dir\,OA}} \ll \frac{1}{2n_{\mathrm{sp}}} \gamma^{(1)}_{\mathrm{dir\,qu}}, \quad \gamma^{(1)}_{\mathrm{dir\,qu}} = \gamma_{\mathrm{dir\,qu}}\big|_{\eta=1} = \frac{P_e}{2hf_e B}\,, \quad w'_O = n_{\mathrm{sp}} w_O\,, \quad w_O = hf_e\,. \tag{5.151b}$$

It is remarkable that with a pre-amplifier receiver the photodetector's quantum efficiency η does not influence the SNR. Even for a large OA bandwidth, the SNR could be better than without an optical pre-amplifier, as long as electronic amplifier noise $\overline{|i'_R|^2}$ in Eq. (5.80) on Page 132 does not yet dominate, i. e., if with $\gamma_{\mathrm{dir\,OA}} \ll \gamma^{(1)}_{\mathrm{dir\,qu}}/(2n_{\mathrm{sp}})$ we still have $\overline{|i_{RD,3}|^2} \approx \overline{|i'_R|^2}$.

Direct reception limit with matched OA bandwidth

If an optical filter reduces the OA bandwidth $B_O = 2B$ to the optical signal bandwidth $2B$, the SNR improves greatly. We find the SNR from Eq. (5.151b),

$$\gamma_{\mathrm{dir\,OA\ qu}} = \frac{1}{2n_{\mathrm{sp}}}\,\gamma^{(1)}_{\mathrm{dir\,qu}}\,\frac{1}{1+\frac{n_{\mathrm{sp}}}{2\gamma^{(1)}_{\mathrm{dir\,qu}}}} \underset{\gamma^{(1)}_{\mathrm{dir\,qu}}\gg 1}{=} \frac{1}{2n_{\mathrm{sp}}}\,\gamma^{(1)}_{\mathrm{dir\,qu}} = \frac{1}{2n_{\mathrm{sp}}}\,\frac{P_e}{2hf_eB}\,, \quad 2B = B_O\,. \tag{5.151c}$$

With an ideal, fully inverted pre-amplifier ($n_{\mathrm{sp}} = 1$) a direct receiver has — even for small received optical powers P_e — an SNR, which is as high as half the theoretical quantum limit Eq. (5.81) on Page 132.

5.6.3 Coherent pre-amplifier receiver

The sensitivity of coherent receivers can be improved by an optical pre-amplifier, too, especially if the power P_O of the local oscillator is limited, and neither its relative intensity noise nor its phase noise can be ignored, not even with a balanced receiver, see Sect. 5.4.1 on Page 144. Figure 5.30 schematically displays the relevant power spectra. The optical amplifier noise spectrum (shaded blue) has a width B_O that is usually much broader than the optical signal bandwidth $2B$. For the present discussion we assume a real baseband signal so that upper and lower optical signal sidebands are correlated, see Eq. (2.36) on Page 26.

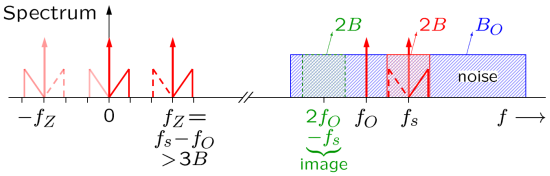


Fig. 5.30. Heterodyne spectra for a pre-amplifier receiver with an optical bandwidth B_O (also homodyne spectra for $f_O = f_s$), compare Fig. 5.25 on Page 143. A real signal with bandwidth B is modulated on an optical carrier with frequency f_s. Upper and lower optical sidebands are correlated, see Eq. (2.36) on Page 26. This optical signal together with a local oscillator (LO) at frequency f_O illuminates a photodiode and is down-converted to a current at an intermediate-frequency $f_Z = f_s - f_O$. For a direct detection of the IF signal, the condition $f_Z > B$ must be fulfilled, otherwise the IF-USB at negative frequencies overlaps with the IF-LSB for positive frequencies, and this would lead to distortions. However, it would be even better to choose an IF $f_Z > 3B$. This avoids that mixing products of the IF sidebands, which would fall into a frequency range $0 \le f \le 2B$, perturb the IF signal band. — Because $f_Z = f_s - f_O \ll B_O$ holds, the low-frequency spectra of the carrier-noise mixing and of the LO-noise mixing are virtually identical. While an IF filter limits the OA noise to the signal bandwidth $2B$, an optical signal spectrum bandpass centred at f_s would reject the optical image spectrum and thus improve the SNR.

The LO frequency f_O is either larger or smaller than the signal carrier frequency f_s. Here we choose $f_O < f_s$ for heterodyne reception. For homodyne reception, LO and carrier frequency must coincide, i. e., both sources must be optically phase-locked. As discussed earlier in Sect. 5.4.3 on Page 146, this can be also achieved by intradyne reception and subsequent signal processing.

Heterodyne reception limit

For heterodyne reception the spectra and the fluctuations of the partial photocurrents Eq. (5.147)–Eq. (5.149) on Page 157 have to modified. Signal and LO are uncorrelated and closely neighboured on an optical frequency scale, so in the mixing process their contributions add up and lead to low-frequency spectra which virtually coincide in shape and position. However, each occurrence of $\mathcal{G}_s P_e$ in Eq. (5.147)–Eq. (5.148) has to be replaced by $\mathcal{G}_s P_s + P_O$ with the substitution $P_e = P_s + P_O/\mathcal{G}_s$, because P_s is amplified while P_O is not. Further, the baseband width B must be replaced by the IF bandwidth $2B$. Other than that, the low-frequency current power spectra remain unchanged. The modified photocurrent fluctuations come – as before – from signal and LO shot noise ($\overline{|i_{RD,1\,\mathrm{OA}}|^2}$), from LO-noise and signal-noise mixing ($\overline{|i_{RD,2\,\mathrm{OA}}|^2}$), and from noise-noise mixing ($\overline{|i_{RD,3\,\mathrm{OA}}|^2}$). The third term does not represent

white noise and therefore requires special treatment. With with $\mathcal{G}_s \gg 1$ we find for the photocurrent fluctuations in the IF band $f_Z - B \le f \le f_Z + B$

$$\overline{|i_{RD,1\,\mathrm{OA}}|^2} = 2eS\mathcal{G}_s \left(P_s + \frac{P_O}{\mathcal{G}_s} + w'_O B_O \right) 2B\,, \tag{5.152a}$$

$$\overline{|i_{RD,2\,\mathrm{OA}}|^2} = 4\left(S\mathcal{G}_s\right)^2 \left(P_s + \frac{P_O}{\mathcal{G}_s} \right) w'_O 2B\,, \tag{5.152b}$$

$$\overline{|i_{RD,3\,\mathrm{OA}}|^2} = 2\left(S\mathcal{G}_s\right)^2 w'^2_O \int_{f_Z-B}^{f_Z+B} (B_O - f)\,\mathrm{d}f \tag{5.152c}$$

$$= 2\left(S\mathcal{G}_s\right)^2 w'^2_O \max\left(B_O - f_Z, 0\right) 2B = \begin{cases} 2\left(S\mathcal{G}_s\right)^2 w'^2_O \left(B_O - f_Z\right) 2B & \text{for} \quad B_O \ge f_Z\,, \\ 0 & \text{else.} \end{cases}$$

The SNR is calculated in analogy to Eq. (5.122) on Page 143,

$$\gamma = \frac{P_S}{P_R}, \quad \gamma_{\text{IF-het OA}} = \frac{i_Z^2/2}{\overline{|i_{RD,1\,\mathrm{OA}}|^2} + \overline{|i_{RD,2\,\mathrm{OA}}|^2} + \overline{|i_{RD,3\,\mathrm{OA}}|^2} + \overline{|i'_R|^2}}. \tag{5.153a}$$

We substitute the current fluctuations Eq. (5.152) and find

$$\begin{aligned} \gamma_{\text{IF-het OA}} &= \\ &= \frac{\frac{1}{2} S^2 \mathcal{G}_s \hat{E}_s^2\, 2P_O}{\left[2eS\mathcal{G}_s\left(P_s + \frac{P_O}{\mathcal{G}_s} + w'_O\, B_O\right) + 4\left(S\mathcal{G}_s\right)^2 \left(P_s + \frac{P_O}{\mathcal{G}_s}\right) w'_O + 2\left(S\mathcal{G}_s\right)^2 w'^2_O \left(B_O - f_Z\right) + 4kF'T_0 G_Q\right] 2B} \\ &= \frac{1}{2n_{\mathrm{sp}}} \frac{P_s}{hf_O\, 2B} \\ &\quad \times \frac{1}{\underbrace{\frac{1}{2}\frac{1}{\eta n_{\mathrm{sp}} \mathcal{G}_s}\left(1 + \frac{\mathcal{G}_s P_s}{P_O} + \frac{\mathcal{G}_s n_{\mathrm{sp}} h f_O B_O}{P_O}\right)}_{\overline{|i_{RD,1\,\mathrm{OA}}|^2}} + \underbrace{\left(1 + \frac{\mathcal{G}_s P_s}{P_O}\right)}_{\overline{|i_{RD,2\,\mathrm{OA}}|^2}} + \underbrace{\frac{1}{2}\frac{n_{\mathrm{sp}} \mathcal{G}_s h f_O}{P_O} \max\left(B_O - f_Z, 0\right)}_{\overline{|i_{RD,3\,\mathrm{OA}}|^2}} + \underbrace{\frac{kF'T_0}{\mathcal{G}_s P_O} \frac{h f_O}{(\eta e)^2}}_{\overline{|i'_R|^2}}}. \end{aligned} \tag{5.153b}$$

If we limit the OA bandwidth to the IF bandwidth, $B_O = 2B$, and chose an IF $f_Z = f_s - f_O > 3B$ as in Fig. 5.30, the noise-noise mixing products fall into the baseband region $0 \le f \le 2B$ outside the IF band, and the $\overline{|i_{RD,3\,\mathrm{OA}}|^2}$-term in Eq. (5.153b) can be disregarded. The signal-noise mixing products are restricted to the range $0 \le f \le B$ and need not be considered, which means that the $\overline{|i_{RD,2\,\mathrm{OA}}|^2}$-term in Eq. (5.153b) reduces to one, and only the LO-noise mixing products fall into the IF band. Finally, if the gain $\mathcal{G}_s$ is sufficiently large, both the shot noise term $\overline{|i_{RD,1\,\mathrm{OA}}|^2}$ in Eq. (5.153b) and the noise term $\overline{|i'_R|^2}$ from the electronic amplifier can be neglected. Then the SNR reduces to the shot (quantum) noise limit

$$\gamma_{\text{IF-het OA qu}} = \frac{1}{2n_{\mathrm{sp}}} \frac{P_s}{2hf_O B} = \gamma_{\text{dir OA}} = \frac{1}{2n_{\mathrm{sp}}}\, \gamma^{(1)}_{\text{dir qu}} \quad \text{for } f_Z = f_s - f_O > 3B \text{ and } \mathcal{G}_s \gg 1\,. \tag{5.153c}$$

This SNR is identical as for direct reception with an OA (Eq. (5.151c) on Page 159). As in the case of heterodyne reception without OA, Eq. (5.124) on Page 143, the SNR with OA increases by a factor of 2 if a real-valued IF signal is demodulated and transferred to the baseband (upper and lower signal sidebands are correlated, while the noise "sidebands" are not),

$$\gamma_{\text{BB-het OA qu}} = \frac{1}{2n_{\mathrm{sp}}} \frac{P_s}{hf_O B} = \frac{1}{2n_{\mathrm{sp}}}\, 2\gamma^{(1)}_{\text{dir qu}} \quad \text{for } f_Z = f_s - f_O > 3B \text{ and } \mathcal{G}_s \gg 1\,. \tag{5.154}$$

With an ideal, fully inverted optical pre-amplifier ($n_{\mathrm{sp}} = 1$), a heterodyne receiver has an SNR, which is as high as half the theoretical quantum limit Eq. (5.124) on Page 143, and double as large as with a direct pre-amplifier receiver Eq. (5.151c). This result holds true even if the LO power P_O is of the order of the amplified signal power $\mathcal{G}_s P_s$. Without an optical pre-amplifier the LO power must be sufficiently large

(and definitely much larger then the signal power) so that electronic amplifier noise becomes unimportant. However, when in a heterodyne receiver with pre-amplifier the LO power P_O becomes too small, then the shot noise term $\overline{|i_{RD,1\,\mathrm{OA}}|^2}$ in Eq. (5.153b) increases significantly. As said before, the signal-noise fluctuation $\mathcal{G}_s P_s / P_O$ in the $\overline{|i_{RD,2\,\mathrm{OA}}|^2}$-term does not fall into the IF band and can be filtered out before demodulating the IF signal.

Homodyne reception limit

For homodyne reception, the optical signal is down-converted directly to the baseband, the spectral width of which equals the signal bandwidth B. Therefore the noise currents Eq. (5.152) on Page 160 have to be modified accordingly,

$$\overline{|i_{RD,1\,\mathrm{OA}}|^2} = 2eS\mathcal{G}_s \left(P_s + \frac{P_O}{\mathcal{G}_s} + \frac{\mathcal{G}_s - 1}{\mathcal{G}_s} w'_O B_O \right) B \,, \tag{5.155a}$$

$$\overline{|i_{RD,2\,\mathrm{OA}}|^2} = 4 \left(S\mathcal{G}_s \right)^2 \left(P_s + \frac{P_O}{\mathcal{G}_s} \right) \frac{\mathcal{G}_s - 1}{\mathcal{G}_s} w'_O B \,, \tag{5.155b}$$

$$\overline{|i_{RD,3\,\mathrm{OA}}|^2} = 2 \left[S \left(\mathcal{G}_s - 1 \right) \right]^2 w'^2_O \int_0^B (B_O - f) \, \mathrm{d}f = 2 \left[S \left(\mathcal{G}_s - 1 \right) \right]^2 w'^2_O \left(B_O - \frac{B}{2} \right) B \,. \tag{5.155c}$$

The SNR is calculated in analogy to Eq. (5.153) on Page 160, but the average electrical signal power is $P_S = i_Z^2$ and not $P_S = i_Z^2/2$ as before,

$$\gamma = \frac{P_S}{P_R}, \quad \gamma_{\mathrm{hom\,OA}} = \frac{i_Z^2}{\overline{|i_{RD,1\,\mathrm{OA}}|^2} + \overline{|i_{RD,2\,\mathrm{OA}}|^2} + \overline{|i_{RD,3\,\mathrm{OA}}|^2} + \overline{|i'_R|^2}}. \tag{5.156a}$$

We substitute the current fluctuations Eq. (5.155) and find the SNR for a homodyne receiver with optical pre-amplifier,

$$\begin{aligned}
\gamma_{\mathrm{hom\,OA}} &= \\
&= \frac{S^2 \mathcal{G}_s \hat{E}_s^2 \, 2P_O}{\left[2eS\mathcal{G}_s \left(P_s + \frac{P_O}{\mathcal{G}_s} + w'_O B_O \right) + 4 \left(S\mathcal{G}_s \right)^2 \left(P_s + \frac{P_O}{\mathcal{G}_s} \right) w'_O + 2 \left(S\mathcal{G}_s \right)^2 w'^2_O \left(B_O - \frac{B}{2} \right) + 4kF'T_0 G_Q \right] B} \\
&= \frac{1}{2n_{\mathrm{sp}}} \frac{P_s}{\frac{1}{2} h f_O B} \\
&\quad \times \frac{1}{\underbrace{\frac{1}{2} \frac{1}{\eta n_{\mathrm{sp}} \mathcal{G}_s} \left(1 + \frac{\mathcal{G}_s P_s}{P_O} + \frac{n_{\mathrm{sp}} \mathcal{G}_s h f_O B_O}{P_O} \right)}_{\overline{|i_{RD,1\,\mathrm{OA}}|^2}} + \underbrace{\left(1 + \frac{\mathcal{G}_s P_s}{P_O} \right)}_{\overline{|i_{RD,2\,\mathrm{OA}}|^2}} + \underbrace{\frac{1}{2} \frac{n_{\mathrm{sp}} \mathcal{G}_s h f_O}{P_O} \left(B_O - \frac{B}{2} \right)}_{\overline{|i_{RD,3\,\mathrm{OA}}|^2}} + \underbrace{\frac{kF'T_0}{\mathcal{G}_s P_O} \frac{h f_O}{(\eta e)^2}}_{\overline{|i'_R|^2}}}.
\end{aligned} \tag{5.156b}$$

With a sufficiently large gain, $\mathcal{G}_s \gg 1$, and observing that usually $P_O \gg P_s$, i. e., that the signal power in front of the OA is much smaller than the LO power, the shot noise term $\overline{|i_{RD,1\,\mathrm{OA}}|^2}$ and the amplifier noise term $\overline{|i'_R|^2}$ become unimportant. However, because the noise-noise mixing products of the $\overline{|i_{RD,3\,\mathrm{OA}}|^2}$-term lie in the signal baseband and cannot be removed by IF filtering as with heterodyne reception, appropriate optical filtering is crucial, $B_O = 2B$. If then the LO power is much larger than the relevant ASE noise, i. e., if $P_O \gg n_{\mathrm{sp}} \mathcal{G}_s h f_O \, 2B$, the noise-noise mixing term can be also neglected. What remains is the SNR

$$\gamma_{\mathrm{hom\,OA}} = \frac{1}{2n_{\mathrm{sp}}} \frac{P_s}{\frac{1}{2} h f_O B} \frac{1}{1 + \frac{\mathcal{G}_s P_s}{P_O}} \quad \text{for } \mathcal{G}_s \gg 1,\ B_O = 2B,\ P_O \gg \max \left(P_s, n_{\mathrm{sp}} \mathcal{G}_s h f_O \, 2B \right). \tag{5.157}$$

If in addition we can guarantee that the LO power is much larger than the *amplified* signal power (this could turn out to be difficult!), i. e., if $P_O \gg \mathcal{G}_s P_s$, we end up with the maximum achievable SNR for a pre-amplifier homodyne receiver,

$$\gamma_{\mathrm{hom\,OA\ qu}} = \frac{1}{2n_{\mathrm{sp}}} \frac{P_s}{\frac{1}{2} h f_O B} \qquad \text{for } \mathcal{G}_s \gg 1,\ B_O = 2B,\ P_O \gg \max \left(\mathcal{G}_s P_s, n_{\mathrm{sp}} \mathcal{G}_s h f_O \, 2B \right). \tag{5.158}$$

Chapter 6

Optical communication systems

6.1 Transmission impairments

There are numerous influences, which impair a fibre-optic transmission. A few were mentioned already and displayed in Fig. 1.6 on Page 6. Most important is attenuation of light guided in a glass fibre. This can be compensated with (noisy) optical amplifiers. Next is dispersion as described in Sect. 2.2 on Page 18 ff., which can be equalized with dispersion compensating fibres. Modern coherent receivers in combination with digital signal processing can even mitigate impairments due to nonlinearities in the fibre. Intersymbol interference as explained in Fig. 1.5(b) on Page 5 can be avoided by proper signal shaping at the transmitter, or by equalization at the receiver as in Fig. 5.19 on Page 133.

In the following two section we discuss the noise figure of optical amplifiers and of links with concatenated amplifiers, and finally give a few results on signal shaping.

6.2 Noise figure of optical amplifiers and links

An optical amplifier (OA) — like any amplifier — degrades the signal-to-noise power ratio (SNR), because amplified spontaneous emission (ASE) adds noise[1] to its output signal. The amount of ASE noise power $P_{\mathrm{ASE},\,x}$ per mode, which is copolarized with the signal, was specified in Eq. (5.67) on Page 128. Now we derive the OA noise figure in analogy to the procedure Eq. (5.62) on Page 126 ff. for electronic amplifiers.

6.2.1 Noise figure of a single optical amplifier

The noise figure is defined as the ratio $\mathrm{SNR}_1/\,\mathrm{SNR}_2$ of SNR at the amplifier input and output. This SNR relates the optical signal power $P_{s1,2}$ and optical noise power $P_{r1,2}$ taken at the amplifier input and output, respectively. The noise figure is defined by power measurements and is thus a universal characteristic of an OA, irrespective of the type of reception (direct, heterodyne, homodyne, or intradyne). Because we disregard optical power reflection, a discrimination between an available optical $\mathrm{SNR}_{v1} = P_{sv1}/\,P_{rv1}$ and an actual optical $\mathrm{SNR}_1 = P_{s1}/\,P_{r1}$ is not required. Transducer gain and available gain according to Eq. (5.56) on Page 125 coincide with the single-pass gain, $\Gamma_{\ddot{u}} = \Gamma_v = \mathcal{G}_s = P_{s2}/P_{s1}$. The definition for the noise figure F is then

$$F = \frac{\mathrm{SNR}_1}{\mathrm{SNR}_2} = \frac{P_{s1}}{P_{r1}} \Big/ \frac{P_{s2}}{P_{r2}} = \frac{P_{r2}}{\mathcal{G}_s P_{r1}} = \frac{\text{polarized output noise power in } B_O}{\text{amplified equiv. pol. input noise power in } B_O}\,. \tag{6.1}$$

We define an SNR $\gamma^{(1)}_{\mathrm{dir\,qu}}$ and a sensitivity $S^{(1)}$ for an ideal photodetector quantum efficiency $\eta = 1$. According to Eq. (5.81) on Page 132 and following Eq. (5.151a) on Page 158, the quantum noise limited

[1] For a very brief introduction to optical amplifier noise, see Ref. 17 on Page 6. Sect. 8.1.3 p. 365

SNR_1 at the OA input is

$$\mathrm{SNR}_1 = \gamma^{(1)}_{\mathrm{dir\,qu}} = \gamma_{\mathrm{dir\,qu}}\big|_{\eta=1} = \frac{P_e}{2w_O B}, \qquad w_O = hf_e\,, \qquad S^{(1)} = S|_{\eta=1} = \frac{e}{hf_e}\,. \tag{6.2}$$

The approximate SNR at the OA output, Eq. (5.151) on Page 158 is re-written to include also the shot noise term. Taking regard of all OA noise sources $\overline{|i_{RD,1,2,3}|^2}$, we write

$$\begin{aligned}\frac{\left(S^{(1)}\mathcal{G}_s P_e\right)^2}{\mathrm{SNR}_2} &= \overline{|i_{RD,1}|^2} + \overline{|i_{RD,2}|^2} + \overline{|i_{RD,3}|^2} \\ &= 2e\big[S^{(1)}\mathcal{G}_s P_e + S^{(1)}\left(\mathcal{G}_s - 1\right) w'_O B_O\big] B + 4S^{(1)}\mathcal{G}_s P_e\, S^{(1)}\left(\mathcal{G}_s - 1\right) w'_O B \\ &\quad + 2\big[S^{(1)}\mathcal{G}_s\left(\mathcal{G}_s - 1\right) w'_O\big]^2 \left(B_O - \tfrac{B}{2}\right) B\end{aligned} \tag{6.3}$$

The noise figure F is calculated according to the definition Eq. (6.1). It consist of three terms, namely the shot noise figure F_{shot}, the noise figure F_{sn} due to mixing of signal and copolarized noise, and the noise figure F_{nn} resulting from mixing copolarized noise with noise[2],

$$F = \frac{\mathrm{SNR}_1}{\mathrm{SNR}_2} = \underbrace{\frac{1}{\mathcal{G}_s}\Big[1 + \frac{\mathcal{G}_s - 1}{\mathcal{G}_s}\,\frac{w'_O B_O}{P_e}\Big]}_{\text{shot noise: } F_{\mathrm{shot}}} + \underbrace{2\,\frac{\mathcal{G}_s - 1}{\mathcal{G}_s}\,\frac{w'_O}{w_O}}_{\text{signal-noise: } F_{\mathrm{sn}}} + \underbrace{\frac{\mathcal{G}_s - 1}{\mathcal{G}_s}\,\frac{w'_O(B_O - B/2)}{P_e}\,\frac{\mathcal{G}_s - 1}{\mathcal{G}_s}\,\frac{w'_O}{w_O}}_{\text{noise-noise: } F_{\mathrm{nn}}}\,. \tag{6.4}$$

For a transparent, i. e., nonexisting OA with $\mathcal{G}_s = 1$, there is no additional noise and $F = 1$. Spontaneous emission factor n_{sp} and gain $\mathcal{G}_s$ are linked. If the gain is larger than but close to one, then the spontaneous emission factor is very large. This expression for the noise factor depends on the input signal power P_e. However, simplifications apply for the following typical data: Signal frequency $f_e = 193.4\,\mathrm{THz}$, OA bandwidth $B_O \approx 3\,\mathrm{THz} \geq 2B$, $w'_O B_O \ll P_e$, noise power $w'_O B_O \approx 2hf_e B_O = 0.77\,\mu\mathrm{W} \mathrel{\widehat{=}} -31\,\mathrm{dBm}$ ($F \approx 4 \mathrel{\widehat{=}} 6\,\mathrm{dB}$), $P_e \gtrapprox -20\,\mathrm{dBm}$, $\mathcal{G}_s \geq 2$. Under these assumptions the red-coloured terms can be neglected, while the blue-coloured terms remain,

$$F \approx F_{\mathrm{shot}} + F_{\mathrm{sn}} \approx \frac{1}{\mathcal{G}_s} + 2\,\frac{\mathcal{G}_s - 1}{\mathcal{G}_s}\,\frac{w'_O}{w_O}\,, \quad 1 < n_{\mathrm{sp}} < \infty, \quad w'_O = n_{\mathrm{sp}} w_O\,, \quad w_O = hf_e\,. \tag{6.5}$$

If in addition the amplifier gain is large (in practice we have $\mathcal{G}_s \approx 100 \mathrel{\widehat{=}} 20\,\mathrm{dB}$), the noise figure due to shot noise can be neglected, and we end up with the simple relation for the noise figure of an optical amplifier operated at a signal frequency f_e,

$$F \approx F_{\mathrm{sn}} \approx 2n_{\mathrm{sp}}\,, \quad \mathcal{G}_s \gg 1\,, \quad 1 < n_{\mathrm{sp}} < \infty, \quad w'_O = n_{\mathrm{sp}} w_O\,, \quad w_O = hf_e\,. \tag{6.6}$$

For a fully inverted, ideal optical amplifier the minimum noise figure is $F = 2 \mathrel{\widehat{=}} 3\,\mathrm{dB}$. Real-world amplifiers have noise figures of $(4 \ldots 8)\,\mathrm{dB}$.

This noise figure as derived above cannot be calculated by referring to the ASE noise power Eq. (5.67) on Page 128 and letting $P_{r2} = P_{\mathrm{ASE},\,x}$ for $B_O = 2B$, because we cannot measure a corresponding input noise power without any optical signal: It is the signal which gives rise to quantum noise, and without signal there would be only thermal noise that is of irrelevant magnitude, $kT_0 \ll hf_e$. However, knowing the noise figure $F = 2n_{\mathrm{sp}}$, we can calculate a fictitious equivalent input power $P_{r\,\mathrm{eq},\,x}$ of copolarized noise according to the definition Eq. (6.1), $P_{r1} = P_{r2}/\left(\mathcal{G}_s F\right)$,

$$P_{r\,\mathrm{eq},\,x} = \frac{P_{\mathrm{ASE},\,x}}{\mathcal{G}_s F} = \frac{\left(\mathcal{G}_s - 1\right) n_{\mathrm{sp}} w_O B_O}{\mathcal{G}_s 2n_{\mathrm{sp}}} = \frac{\mathcal{G}_s - 1}{\mathcal{G}_s}\,\frac{w_O\,2B}{2} \overset{\mathcal{G}_s \gg 1}{=} hf_e B\,. \tag{6.7}$$

In one polarization and per signal bandwidth B we have to provide one fictitious, non-extractable noise photon with energy hf_e at the OA input to represent the unavoidable quantum fluctuations.

[2] R. Bonk, T. Vallaitis, W. Freude, J. Leuthold, R. V. Penty, A. Borghesani, I. F. Lealman: Linear semiconductor optical amplifiers. In: H. Venghaus, N. Grote (Eds.): Fibre optic communication — Key devices. Heidelberg: Springer-Verlag 2012, Chapter 12 pp. 512–571. Eq. (12.30) p. 543

We can apply the recipe Eq. (6.1), (6.2) for calculating the OA noise figure also for the case of heterodyne reception (Eq. (5.124) on Page 143, Eq. (5.154) on Page 160), and for homodyne reception (Eq. (5.130) on Page 146, Eq. (5.158) on Page 161),

$$\begin{aligned}
&\mathrm{SNR}_1: \quad \gamma^{(1)}_{\mathrm{dir\,qu}} = \frac{P_e}{2hf_eB} \qquad \gamma^{(1)}_{\mathrm{BB\text{-}het\,qu}} = \frac{P_s}{hf_OB} \qquad \gamma^{(1)}_{\mathrm{hom\,qu}} = \frac{P_s}{\frac{1}{2}hf_OB} \\
&\mathrm{SNR}_2: \quad \gamma_{\mathrm{dir\,OA\ qu}} = \frac{1}{2n_{\mathrm{sp}}}\frac{P_e}{2hf_eB} \quad \gamma_{\mathrm{BB\text{-}het\,OA\ qu}} = \frac{1}{2n_{\mathrm{sp}}}\frac{P_s}{hf_OB} \quad \gamma_{\mathrm{hom\,OA\ qu}} = \frac{1}{2n_{\mathrm{sp}}}\frac{P_s}{\frac{1}{2}hf_OB}
\end{aligned} \tag{6.8a}$$

and each time we find the same result for the noise figure of an optical amplifier,

$$F = \frac{\mathrm{SNR}_1}{\mathrm{SNR}_2} = 2n_{\mathrm{sp}}\,. \tag{6.8b}$$

6.2.2 Noise figure of an optical amplifier link

In specific network applications, it is of interest to cascade optical amplifiers in a link for compensating fibre losses along a transmission distance as in Fig. 1.6 on Page 6. We arrange a series of N links, each consisting of an amplifier with noise power spectral density $w_O^{(n)}$ and single-pass gain $\mathcal{G}_n \geq 1$, followed by a bandpass filter $B_O = 2B$ and a fibre length with a power gain $0 < g_n \leq 1$. The total gain of this link is

$$\mathcal{G}_{\mathrm{tot}} = \prod_{n=1}^{N} \mathcal{G}_n g_n\,. \tag{6.9}$$

We assume the realistic scenario $w'_O B_O \ll P_e$ and approximate the SNR of Eq. (6.3) on Page 6.3 by the first two terms. The SNR at the output of the N links is

$$\begin{aligned}
\frac{\big(S^{(1)}\mathcal{G}_{\mathrm{tot}}P_e\big)^2}{\mathrm{SNR}_N^{(\mathcal{G}_n g_n)}} &= 2eS^{(1)}\mathcal{G}_{\mathrm{tot}}P_eB + 4\big(S^{(1)}\mathcal{G}_{\mathrm{tot}}P_eB\big)S^{(1)}B \\
&\quad \times \left[(\mathcal{G}_1-1)\,w_O^{(1)}g_1\,\frac{\mathcal{G}_{\mathrm{tot}}}{\mathcal{G}_1 g_1} + (\mathcal{G}_2-1)\,w_O^{(2)}g_2\,\frac{\mathcal{G}_{\mathrm{tot}}}{\mathcal{G}_1 g_1\,\mathcal{G}_2 g_2} + \ldots + (\mathcal{G}_N-1)\,w_O^{(N)}g_N\,\frac{\mathcal{G}_{\mathrm{tot}}}{\mathcal{G}_{\mathrm{tot}}}\right].
\end{aligned} \tag{6.10}$$

Together with SNR_1 from Eq. (6.2) we find the noise figure of the concatenated links,

$$\begin{aligned}
F_N^{(\mathcal{G}_n g_n)} = \frac{\mathrm{SNR}_1}{\mathrm{SNR}_N^{(\mathcal{G}_n g_n)}} &= \frac{1}{\mathcal{G}_{\mathrm{tot}}} + \left(\frac{\mathcal{G}_1-1}{\mathcal{G}_1}\,\frac{w_O^{(1)}}{w_O} + \frac{1}{\mathcal{G}_1 g_1}\,\frac{\mathcal{G}_2-1}{\mathcal{G}_2}\,\frac{w_O^{(2)}}{w_O}\right. \\
&\quad \left. + \frac{1}{\mathcal{G}_1 g_1\,\mathcal{G}_2 g_2}\,\frac{\mathcal{G}_3-1}{\mathcal{G}_3}\,\frac{w_O^{(3)}}{w_O} + \ldots + \frac{1}{\prod_{n=1}^{N-1}\mathcal{G}_n g_n}\,\frac{\mathcal{G}_N-1}{\mathcal{G}_N}\,\frac{w_O^{(N)}}{w_O}\right).
\end{aligned} \tag{6.11}$$

With the partial noise figures $F_{sn}^{(n)}$ as used in Eq. (6.5) on Page 164, we find the total link noise figure

$$F_N^{(\mathcal{G}_n g_n)} = \frac{1}{\mathcal{G}_{\mathrm{tot}}} + \left(F_{\mathrm{sn}}^{(1)} + \frac{F_{\mathrm{sn}}^{(2)}}{\mathcal{G}_1 g_1} + \frac{F_{\mathrm{sn}}^{(3)}}{\mathcal{G}_1 g_1\,\mathcal{G}_2 g_2} + \ldots + \frac{F_{\mathrm{sn}}^{(N)}}{\prod_{n=1}^{N-1}\mathcal{G}_n g_n}\right), \quad F_{\mathrm{sn}}^{(n)} = 2\,\frac{\mathcal{G}_n-1}{\mathcal{G}_n}\,\frac{w_O^{(n)}}{w_O}\,. \tag{6.12}$$

Equation (6.12) looks very similar to Friis' formula for the noise figure of an electronic amplifier chain for given individual excess noise figures $F_{z\,n}$ and available gains $\Gamma_{v\,n}$, Eq. (5.62b), (5.65) on Page 127,

$$F_N = 1 + \left(F_{z\,1} + \frac{F_{z\,2}}{\Gamma_{v\,1}} + \frac{F_{z\,3}}{\Gamma_{v\,1}\Gamma_{v\,2}} + \ldots + \frac{F_{z\,N}}{\prod_{n=1}^{N-1}\Gamma_{v\,n}}\right) \qquad \text{(noise figure for thermal noise).} \tag{6.13}$$

The difference comes from the fact that Friis' formula Eq. (6.13) describes thermal noise, while Eq. (6.12) specifies shot (quantum) noise and ASE noise.

If the sequence of "OA, filter, lossy fibre length" was reversed to "lossy fibre length, OA, filter" in each link, the total link noise figure becomes

$$F_N^{(g_n\mathcal{G}_n)} = \frac{1}{\mathcal{G}_{\text{tot}}} + \frac{1}{g_1}\left(F_{\text{sn}}^{(1)} + \frac{F_{\text{sn}}^{(2)}}{\mathcal{G}_1 g_2} + \frac{F_{\text{sn}}^{(3)}}{\mathcal{G}_1 g_2\,\mathcal{G}_2 g_3} + \ldots + \frac{F_{\text{sn}}^{(N)}}{\prod_{n=1}^{N}\mathcal{G}_{n-1}g_n}\right), \quad F_{\text{sn}}^{(n)} = 2\,\frac{\mathcal{G}_n - 1}{\mathcal{G}_n}\,\frac{w_O^{(n)}}{w_O}\,. \tag{6.14}$$

Frequently, the individual links have virtually identical characteristics, i. e., $\mathcal{G}_n = \mathcal{G}$, $g_n = g$, $w_O^{(n)} = w_O'$, and serve to bridge a total link distance with a net gain of $\mathcal{G}_{\text{tot}} = 1$, i. e., $\mathcal{G}g = 1$. The two arrangements "OA, filter, lossy fibre length" $(\mathcal{G}g)$ and "lossy fibre length, OA, filter" $(g\mathcal{G})$ lead to total noise figures of $F_N^{(\mathcal{G}g)}$ and $F_N^{(g\mathcal{G})}$, respectively,

$$F_N^{(\mathcal{G}g)} = 1 + NF_{\text{sn}}\,, \qquad F_N^{(g\mathcal{G})} = 1 + N\mathcal{G}F_{\text{sn}}\,, \qquad \mathcal{G}_{\text{tot}} = 1\,, \qquad \mathcal{G}g = 1\,, \qquad F_{\text{sn}}^{(n)} = F_{\text{sn}}\,. \tag{6.15}$$

With an OA gain of $\mathcal{G} = 10 \mathrel{\widehat{=}} 10\,\text{dB}$ and an OA noise figure of $F_{\text{sn}} = 4$ ($F = \frac{1}{10} + 4 \approx 4 \mathrel{\widehat{=}} 6\,\text{dB}$) for $N = 10$ links, the arrangement $(\mathcal{G}g)$ has a total noise figure $F_{10}^{(\mathcal{G}g)} = 41$, while an element order with the lossy fibre length in front of the amplifiers shows a *much* larger $\mathcal{G}$-fold noise figure of $F_{10}^{(g\mathcal{G})} = 401$.

6.2.3 Noise figure of a lossy fibre

Finally, we compute the noise figure of a fibre with a power gain factor $0 < g \leq 1$. As before, the quantum-noise limited input SNR is given by Eq. (5.81) on Page 132. The fibre itself does not contribute noise (let aside a negligible amount of thermal noise, because $kT_0 \ll hf_e$), but the output power is reduced to gP_e, and this carries over to the output SNR. The noise figure F_g becomes simply

$$\text{SNR}_1 = \frac{\eta P_e}{2hf_eB}\,, \qquad \text{SNR}_2 = \frac{\eta g P_e}{2hf_eB}\,, \qquad F_g = \frac{1}{g}\,, \qquad 0 < g \leq 1\,. \tag{6.16}$$

It follows that the noise figure of a fibre, expressed as a logarithmic quantity $F_{g\,\text{dB}} = 10\lg F_g$, simply equals the power attenuation constant $a = 10\lg(1/g)$.

However, if the input SNR_1 is given by a signal embedded in classical noise, then the fibre's attenuation does not degrade the output SNR_2, which then equals the input SNR_1: Input signal and input noise are attenuated alike.

6.3 Signal shaping

The following publication describes strategies, how signal shaping could be done in the digital, electrical or optical domain. These techniqes are important to decrease the channel spacing in wavelength-division multiplexing and thereby improving the spectral efficiency, while still avoiding linear crosstalk among neigbouring channels. In addition, shaping of the transmitted pulses can also improve the nonlinear transmission performance.

2570 JOURNAL OF LIGHTWAVE TECHNOLOGY, VOL. 31, NO. 15, AUGUST 1, 2013

Pulse-Shaping With Digital, Electrical, and Optical Filters—A Comparison

Rene Schmogrow, Shalva Ben-Ezra, Philipp C. Schindler, Bernd Nebendahl, Christian Koos, Wolfgang Freude, and Juerg Leuthold

***Abstract*—We investigate the performance of sinc-shaped QPSK signal pulses generated in the digital, electrical, and optical domains. To this end an advanced transmitter with a digital pulse-shaper is compared to analog transmitters relying on pulse-shaping with electrical and optical filters, respectively. The signal quality is assessed within a single carrier setup as well as within an ultra-densely spaced WDM arrangement comprising three channels. An advanced receiver providing additional digital filtering with an adaptive equalization algorithm to approximate an ideal brick-wall Nyquist filter has been used for all schemes. It is found that at lower symbol rates, where digital processing is still feasible, digital filters with a large number of filter coefficients provide the best performance. However, transmitters equipped with only electrical or optical pulse-shapers already outperform transmitters sending plain unshaped NRZ signals, so that for higher symbol rates analog electrical and optical techniques not only save costs, but are the only adequate solution.**

***Index Terms*—Optical modulation, optical pulses, optical transmitters, pulse-shaping methods.**

I. Introduction

SHAPING the pulse envelope of M-ary quadrature amplitude modulated (QAM) signals has attracted quite some attention recently. Pulse-shaping techniques allow for instance a decrease of the channel spacing, and with this an improvement of the spectral efficiency (SE) of wavelength division multiplexed (WDM) transmission, while still avoiding linear crosstalk among neighboring channels [1]–[4]. Pulse-shaping may also improve the nonlinear transmission performance [2].

Manuscript received February 11, 2013; revised May 01, 2013 and June 24, 2013; accepted June 24, 2013. Date of publication June 27, 2013; date of current version July 10, 2013. This work was supported in part by the EU projects ACCORDANCE and Fox-C, the Helmholtz Research School of Teratronics, the Helmholtz Research School of Teratronics, and Xilinx, Micram Microelectronic, Eastern Wireless TeleComm (EWT), and the Deutsche Forschungsgemeinschaft (DFG).

R. Schmogrow was with the Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany. He is now with ETH Zurich, Institute of Electromagnetic Fields, 8092 Zurich, Switzerland (e-mail: RSchmogrow@ethz.ch).

S. Ben-Ezra is with Finisar, Nes Ziona, Israel (e-mail: shalva.ben-ezra@finisar.com).

P. C. Schindler, C. Koos, and W. Freude are with the Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany (e-mail: philipp.schindler; christian.koos; w.freude@kit.edu).

B. Nebendahl is with Agilent Technologies, 71034 Boeblingen, Germany (e-mail: bernd_nebendahl@agilent.com).

J. Leuthold is with the Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany and with ETH Zurich, 8092 Zurich, Switzerland (e-mail: JuergLeuthold@ethz.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JLT.2013.2271513

Among the many possible pulse-shapes, sinc-shaped pulses with a corresponding rectangular spectrum are of particular interest as they allow transmission in the Nyquist WDM regime [1]–[3]. For Nyquist WDM, channel spacing and symbol rate are identical. Sinc-shaped Nyquist pulses are also special in a way that their pulse form meets the Nyquist criterion, according to which impulse maxima coincide with the zeros of neighboring pulses, so that inter-symbol interference (ISI) is avoided. In practice, shaping of a pulse can be achieved by performing pulse-shaping in the digital [5], in the electrical, or optical [6], [7] domain using appropriate filters. While at lower symbol rates ($<$35 GBd) all shaping techniques are available, at higher symbol rates only analog electrical or optical [7] techniques are at hand. Each of the methods has advantages and disadvantages. Unfortunately, the various methods have never been directly compared.

In this paper we compare digital, electrical and optical pulse-shaping techniques. The comparison is performed for lower symbol rates where all techniques can be implemented. More precisely, we form a sinc-shaped pulse from a 20 GBd quadrature phase shift keyed (QPSK) signal and investigate the influence of the shaping technology on the signal quality for a single optical carrier. Sinc-shaping in the digital domain is performed by our software-defined transmitter (Tx). This allows us to create an almost perfect sinc-shaped pulse form with virtually zero roll-off. Electrical and optical filters are alternatively used to approximate sinc-shaped pulses by analog means. Last, we assess the performance of the signals in a setup with three carriers where the channel spacing of 20 GBd QPSK signals is varied from 17 GHz to 50 GHz. This way, we explore pulse-shaping for the sub-Nyquist WDM, the Nyquist WDM, and the ultra-dense WDM regimes. To make the comparison as fair as possible, we have optimized not only the Tx but also the receiver (Rx) with a sophisticated equalization technique based on Nyquist brick-wall filtering to minimize the ISI for each of the transmitters. It is found that digitally formed sinc-shaped pulses provide superior performance at low symbol rates. However, sinc-shaped pulses generated by electrical filters are not so far off, and pulses shaped by optical filters still outperform plain unshaped signals.

II. Digital, Electrical or Optical Pulse-Shapers for Single-Channel QPSK

The performance of digitally, electrically, and optically pulse-shaped signals is investigated first. As a signal source for all of the pulse-shaping techniques we use a single-polarization QPSK signal generated from a pseudo random binary sequence

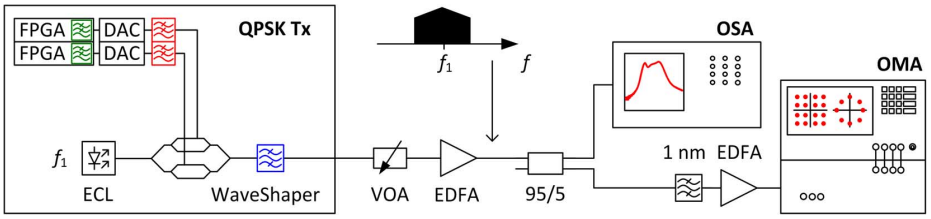


Fig. 1. Setup for single carrier pulse-shaping and measurement of its performance. A pair of field programmable gate arrays (FPGA) drives two high-speed digital-to-analog converters (DAC). The digital pulse-shaper (green) is realized within the FPGAs. Electrical image-rejection filters (red) remove either image spectra for the digital pulse-shaper or solely perform pulse-shaping of the QPSK signals. The electrical signals then modulate an external cavity laser (ECL) by means of an optical I/Q-modulator. In the case of optical pulse-shaping a Finisar WaveShaper (blue) is employed. We use a variable optical attenuator (VOA) together with an erbium doped fiber amplifier (EDFA) to adjust the optical signal-to-noise ratio (OSNR). A 95/5 splitter directs the signals to an optical spectrum analyzer (OSA) and to a coherent receiver (OMA, Agilent optical modulation analyzer).

(period: $2^{15} - 1$) and encoded onto a single-carrier. The initial pulses have a non-return-to-zero (NRZ, rectangular) pulse shape. NRZ-QPSK signals have been chosen because they are widely employed [8] and show good performance especially for long haul transmission [9]. The experimental setup is depicted in Fig. 1. To generate the NRZ-QPSK data pulses we use a versatile software-defined optical Tx comprising two Xilinx XC5VFX200T field programmable gate arrays (FPGA) and two high-speed Micram digital-to-analog converters (DAC) [10]. The DACs are operated at sampling rates up to 30 GSa/s with a physical resolution of 6 bit and an analog electrical bandwidth $f_{\mathrm{DAC}} > 18$ GHz. The respective pulse-shaping for the three schemes is implemented as follows:

- The digital filters (marked green in Fig. 1) are realized in the FPGA. The additional electrical filters (red) are then used to remove the digitally generated image spectra when sinc-pulses are generated in the digital domain.
- When the sinc-shape is approximated in the electrical domain the electrical filters alone shape the electrical drive-signals that are fed to the IQ-modulator. The IQ-modulator then encodes QPSK data onto an external cavity laser (ECL, wavelength λ_1, linewidth 100 kHz).
- When performing pulse-shaping in the optical domain the DAC output signals are directly fed to a nested $\mathrm{LiNbO_3}$ Mach-Zehnder IQ-modulator (MZM) with a modulation bandwidth of $f_{\mathrm{MZM}} > 25$ GHz. To generate the optically filtered QPSK signals, a Finisar WaveShaper [11] serves as pulse-shaper located behind the modulator (marked blue in Fig. 1). The WaveShaper comprises general imaging optics (lenses and mirrors), a diffraction grating, and liquid crystal on silicon (LCoS) cells for shaping the phase of the optical spectrum. Used as a freely programmable filter, the spectral resolution is 12.5 GHz. As an alternative for a fixed-frequency optical filter with a nearly rectangular frequency response an optical interleaver could be used [12].

A variable optical attenuator (VOA) adjusts the optical power launched into the first erbium doped fiber amplifier (EDFA), and thus varies the optical signal-to-noise ratio (OSNR in a bandwidth of 0.1 nm).

A schematic optical power spectrum centered at the ECL wavelength λ_1 is shown as an inset. The spectrum drops toward the band edges. In the case of digital filtering, this is due to the frequency response of the DAC and the image-rejection filters (an influence which could have been compensated for by digital pre-conditioning). For the electrical and optical filters this spectral drop cannot be avoided in practice, and the spectral cut-off cannot be as sharp as for the digital filter. However, this non-ideal spectral shape can be compensated in the Rx as will be explained in the next subsection.

An optical spectrum analyzer (OSA) determines the OSNR. The signal is filtered by a standard 1 nm optical filter which removes spurious EDFA noise. Finally, the signal power is leveled with the second EDFA and coherently received by the Agilent optical modulation analyzer (OMA). The OMA comprises two 90° optical hybrids (one for each polarization) and balanced photo-detectors. A free-running ECL serves as an internal local oscillator (LO). The signals are sampled by real-time oscilloscopes with 80 GSa/s each having an analog bandwidth of 32 GHz.

The following subsections describe first the digital signal processing (DSP) in the Rx irrespective of the Tx pulse-shaping technique that is employed. Subsequently, we describe the digital, analog, and optical pulse-shaping in the Tx.

A. Digital Signal Processing in the Receiver

To overcome limitations introduced by components with finite electrical bandwidth, we investigate advanced Rx processing techniques such as equalizers with finite impulse response (FIR) filters to minimize ISI, or "brick-wall" digital filtering to suppress signals outside the Nyquist frequency bandwidth f_{Nyq} that equals the symbol rate (named "Nyquist filtering" in the following). In addition to Nyquist filtering we employ an advanced clock recovery scheme [2].

First the coherently received signals are polarization de-multiplexed [13], [14] if polarization division multiplexing (PDM) [15] was applied. In this section we investigate signals on a single polarization only and thus omit the polarization de-multiplexing block. The remaining DSP blocks are shown in Fig. 2(a). The corresponding signal spectra are displayed in Fig. 2(b). The black spectrum on top (Tx) corresponds to electrically shaped QPSK signals as received and sampled by the Rx. It shows a significant roll-off within the pass-band because of the electrical image-rejection filters in the Tx, see inset Fig. 2. Due to this roll-off, the Nyquist ISI criterion is violated. The Nyquist filtering block removes the signal spectrum outside the Nyquist frequency band (Fig. 2(b), red spectrum NYQ), yet the Tx caused roll-off remains. This is done in the frequency domain using two samples for each transmitted symbol. If there is a significant frequency offset between signal carrier and LO, it should be compensated prior

2572 JOURNAL OF LIGHTWAVE TECHNOLOGY, VOL. 31, NO. 15, AUGUST 1, 2013

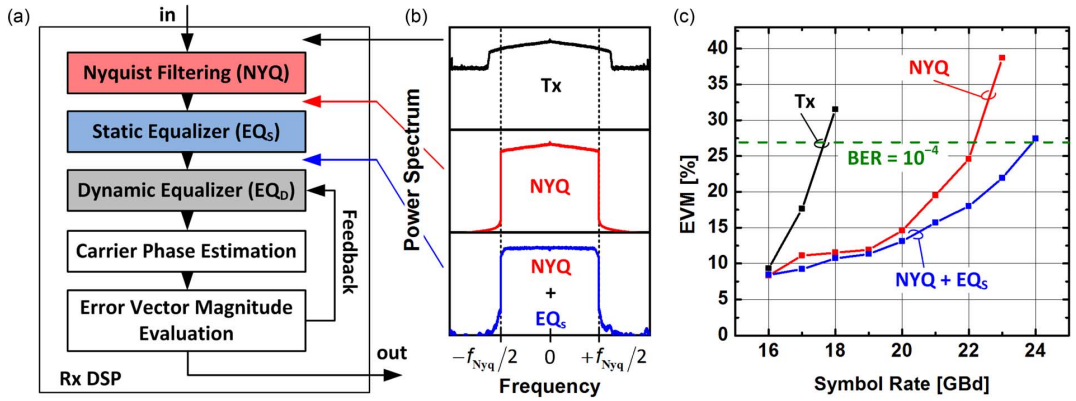


Fig. 2. Digital signal processing (DSP) blocks in the receiver together with the influence of the various equalization elements on the achievable symbol rate. (a) A Nyquist filtering block (NYQ) removes the signal spectrum outside the Nyquist frequency band and provides clock recovery. The static pre-equalizer (EQs) coarsely flattens the combined Tx and Rx transfer functions. A blind post-equalizer is dynamically adjusted by evaluating the measured error vector magnitude (EVM). Inside this control loop the carrier phase is recovered. (b) Ensemble averaged spectra for 20 GBd signals. The unfiltered spectrum (top, Tx) shows a significant roll-off. After Nyquist filtering portions outside the Nyquist frequency band are removed (middle, NYQ). The static equalizer (EQs) additionally flattens the Nyquist pass-band. (c) EVM versus symbol rate plotted for an analog Tx with cut-off frequency $f_{el} \approx 12.3$ GHz. (Black, Tx): Dynamic equalizer only. (Red, NYQ): Dynamic equalizer and Nyquist filtering. (Blue, NYQ + EQs): All DSP blocks in Subfigure (a) are active. The better a flat brick-wall spectrum is approximated, the closer the symbol rate approaches the Nyquist rate of 24.6 GBd.

to Nyquist filtering. In our case, this frequency offset (even though free-running lasers were used) was kept well below 1% of the symbol rate without any additional, more refined digital frequency offset compensation. This was achieved by adjusting the tunable laser sources. Hence, signal degradation due to carrier frequency offset is negligible. It should be noted, however, that especially for any Rx employing a matched filter (e.g. for square root raised cosine or orthogonal frequency division multiplexed signals), the carrier frequency offset should be kept at minimum prior to filtering. The described Nyquist filtering includes also the clock recovery [2] as standard clock recovery mechanisms fail for sinc-shaped Nyquist signals [16]. As an alternative, timing recovery could be performed according to [17]. Next, a static pre-equalizer with 25 coefficients coarsely flattens the combined Tx and Rx transfer functions hence mitigating ISI (Fig. 2(b), blue spectrum $\mathrm{NYQ} + \mathrm{EQ_S}$). Finally, a blind post-equalizer ($\mathrm{EQ_D}$ with 25 taps, using the least-mean square algorithm [18]) is adapted by evaluating the measured error vector magnitude (EVM) [19], [25]. It removes any residual roll-off and thus residual ISI. Inside this control loop the carrier phase is recovered. Both equalizers, static and adaptive, operate with one sample per symbol and are applied after the clock information has been recovered.

To judge the influence of the DSP blocks preceding the dynamic equalizer $\mathrm{EQ_D}$, we measured the signal quality (EVM) of the electrically shaped NRZ-QPSK as a function of the symbol rate. The LO of the coherent Rx has been tuned to approximately match the wavelength of the Tx laser (intradyne reception). For a single polarization QPSK signal we adjusted the symbol rate in 1 GBd steps from 16 GBd to 24 GBd. The Tx low-pass filters, applied to both the in-phase and quadrature of the signals, limit the analog bandwidth to a cut-off frequency $f_{\mathrm{el}} \approx 12.3$ GHz. These Tx filters also mimic a possible Rx bandwidth limitation assuming a linear transmission system. A maximum symbol rate of $2 \times 12.3\ \mathrm{GBd} = 24.6\ \mathrm{GBd}$ results [20]. The outcome is displayed in Fig. 2(c). Due to convergence issues, using solely the adaptive post-equalizer $\mathrm{EQ_D}$, a maximum symbol rate of only 17.5 GBd can be achieved for a minimum bit error ratio of $\mathrm{BER} = 10^{-4}$ (black, Tx). Activating the Nyquist filtering block (red, NYQ) enhances the possible symbol rate to 22.5 GBd. If in addition the static equalizer is turned on, a maximum symbol rate of 24 GBd is found. This comes close to the theoretical limit [20]. Said Nyquist filtering and clock recovery has already been demonstrated for M-ary QAM as high as 512QAM [21].

B. Digital Filtering in the Transmitter

Digital Nyquist pulse-shaping has proven excellent performance in ultra-densely spaced WDM networks [5]. The digitally sinc-shaped Nyquist pulses have been generated by our software defined Tx which acts as an arbitrary waveform generator (AWG), i.e., signal generation and digital pulse-shaping (Fig. 1, green) is performed offline. An FIR filter of order $R = 2048$ was used for pulse-shaping. The implications of approximating a sinc-shaped impulse response with a finite number of filter coefficients were thoroughly investigated in [2]. For practical systems, similar to the one investigated here, even a filter order as low as 32 would provide very good performance [22]. The generated signals are then stored in the FPGAs. A 6 bit DAC provides the transition from the digital to the analog domain. The electrical image-rejection low-pass filters of $f_{\mathrm{el}} \approx 12.3$ GHz (Fig. 1, red) remove spurious image spectra created by the DACs [10], and we end up with sinc-shaped Nyquist pulses with virtually zero spectral roll-off [2]. The DACs are operated at 30 GSa/s and the symbol rate is 20 GBd leading to an effective oversampling factor of $q = 30/20 = 1.5$ [23].

Ensemble averaged spectra measured with the OMA are shown in Fig. 3. Due to the high filter order R, nearly all of the signal power is confined to the Nyquist frequency band (Fig. 3, top). Thus the optical signal bandwidth is virtually 20 GHz for a 20 GBd QPSK signal. Static and dynamic equalization in the Rx flattens the spectral roll-off. The resulting spectrum shows a flat Nyquist pass-band and steep edges (Fig. 3, bottom).

We measure BER and EVM as a function of OSNR. We further estimate an equivalent BER from the measured EVM and

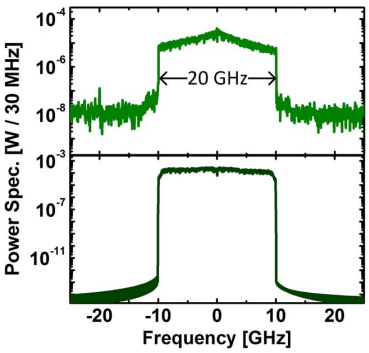


Fig. 3. Measured Tx spectra for digitally shaped single-carrier QPSK signals. Top: Nyquist-shaped Tx signal spectrum, not compensated for DAC roll-off. Bottom: Rx signal spectrum after Nyquist filtering and additional static and dynamic equalization which compensate for the roll-off introduced by Tx and Rx electronics.

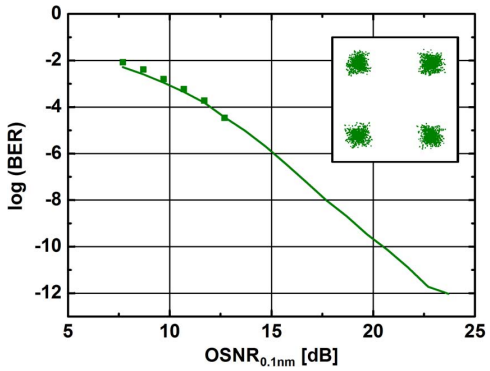


Fig. 4. Measured BER performance for digitally shaped single-carrier QPSK signals. Measured BER (squares) and estimated BER as derived from EVM measurements (solid line) as a function of OSNR for 20 GBd QPSK. The inset shows the constellation diagram at highest possible OSNR of 30 dB.

display the results in Fig. 4. Measured BER (squares) and estimated BER derived from EVM (line) coincide. An inset shows a constellation diagram for the highest achievable OSNR of 30 dB.

C. Electrical Filtering

To generate sinc-shaped Nyquist pulses in the electrical domain we use (as approximation to rectangularly shaped filters) the same low-pass image-rejection filters as before. Although we keep the DACs in the setup, they only produce two-level NRZ electrical signals. Therefore binary drivers suffice, which potentially reduces overall cost of the Tx significantly. The simulated S-parameters provided by the manufacturer and the group delay derived from the transfer function S_{21} of the electrical low-pass filters are depicted in Figs. 5 and 6. A 3 dB cut-off frequency of $f_{\mathrm{el}} = 12$ GHz can be seen from the $20\log 10|S_{21}|$ curve (blue). The reflection represented by the $20\log 10|S_{11}|$ curve (red) is suppressed by at least 20 dB throughout the pass-band, Fig. 5. The actually manufactured filters show 3 dB cut-off frequencies of $f_{\mathrm{el}} \approx 12.3$ GHz. The group delay in the pass-band stays below 0.5 ns, see Fig. 6.

A typically received and ensemble averaged power spectrum of a 20 GBd QPSK signal at an OSNR of 30 dB is shown in

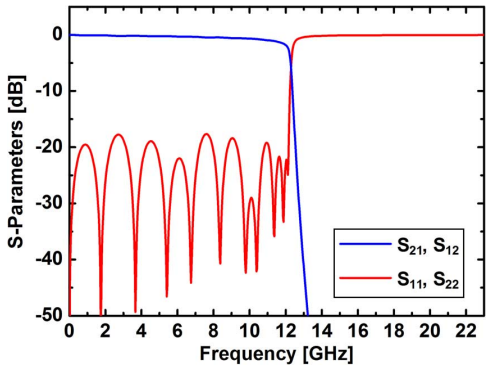


Fig. 5. Simulated S-parameters of the employed electrical low-pass filters. The measured cut-off frequency of the manufactured filters is $f_{\mathrm{el}} \approx 12.3$ GHz.

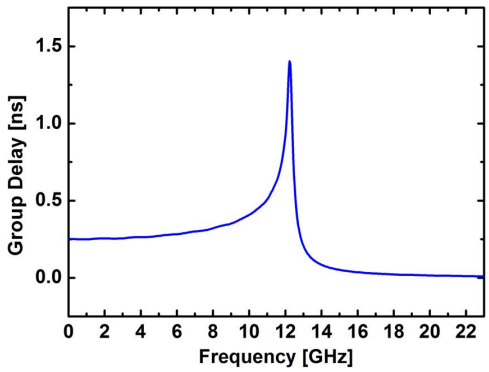


Fig. 6. Simulated group delay of the employed electrical low-pass filters. The group delay stays below 0.5 ns throughout the pass-band.

Fig. 7(a), upper row. In order to obtain the spectrum, we measure the time domain waveform with the OMA and perform a Fourier transform. The optical signal bandwidth is $f_{\mathrm{opt}} \approx 25.2$ GHz close to the $2 \times f_{\mathrm{el}} \approx 24.6$ GHz pass-band of the electrical filters. After filtering and performing the Rx DSP the signal bandwidth is digitally reduced to the Nyquist frequency band of 20 GHz, and the pass-band is flattened as to be seen in Fig. 7(a), lower row. Constellation diagrams for 20 GBd and 24 GBd QPSK are shown in Fig. 7(b).

Fig. 8 shows the measured BER (squares) and the BER estimated from EVM (solid lines) [19], [25]. Measured and estimated BER agree well. Measurements are done for single-polarization and single-carrier QPSK, for different symbol rates, and for different OSNR. As expected, the BER degrades with increasing symbol rate. For large OSNR, a BER error floor can be seen. This error floor stems from the electronic noise originating from the Tx and Rx. However, this noise is negligible compared to the optical noise of multiple EDFAs that will be picked in transmission links with multiple amplifiers. For the given optical output power of our Tx with electrical pulse-shaping, and for the given EDFA, the maximum achievable OSNR is 34 dB.

The overall BER as a function of OSNR is approximately same comparing a single optical carrier modulated with either digitally or electrically shaped sinc-pulses. However, the digital pulse-shaper produces 20 GBd signals with virtually 20 GHz bandwidth whereas electrically shaped 20 GBd signals require a bandwidth of 25.2 GHz. As a disadvantage, the maximum

2574 JOURNAL OF LIGHTWAVE TECHNOLOGY, VOL. 31, NO. 15, AUGUST 1, 2013

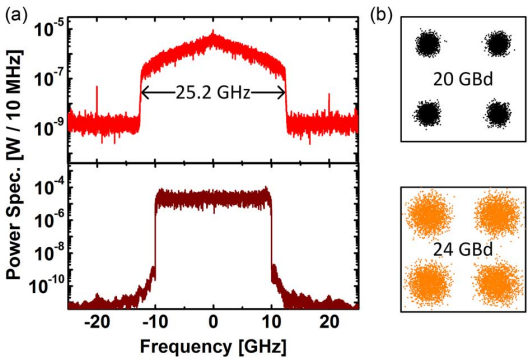


Fig. 7. (a) Measured and ensemble averaged spectrum (top) of a 20 GBd electrically generated QPSK signal. As expected, the overall signal bandwidth corresponds to two times the electrical filters' cut-off frequency. Measured spectrum (bottom) for the same signal after applying the Nyquist filtering procedure in the Rx as described in Fig. 2. The filter removes signal components outside the Nyquist frequency band and flattens the spectrum in the pass-band. (b) Received constellation diagrams for 20 GBd (top) and 24 GBd (bottom) QPSK.

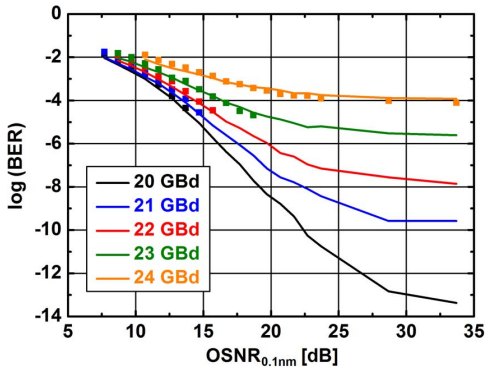


Fig. 8. Measurement results for electrically shaped QPSK signals at varying symbol rates as a function of OSNR. Measured BER (squares) and estimated BER(EVM) (solid lines) for different symbol rates and different OSNR. The error floor at high OSNR values stems from the electrical noise added by Tx and Rx. This noise has negligible influence on the measurements in Section III.

achievable OSNR for the digital pulse-shaper is 4 dB less than for the analog pulse-shaper. This is dominantly due to the increased peak-to-average power ratio of the digitally shaped signals [2] with their more pronounced side lobes as compared to the pulses shaped by analog filters (see Section II-F). For the multi-carrier experiments in Section III we limit the symbol rate to 20 GBd, since already a symbol rate of 21 GBd leads to a significant penalty.

D. Optical Filtering

Finally, to approximate sinc-shaped pulses in the optical domain we use a Finisar WaveShaper as an optical band-pass filter and apply it to conventional NRZ-QPSK signals. We adjust the WaveShaper to have a fixed optical pass-band of 12.5 GHz. Since the filter pass-band at this resolution is difficult to change we instead vary the symbol rate from 20 GBd to 28 GBd in steps of 2 GBd. We only measure signals where the signal quality remains above the quality required for state-of-the-art forward error correction (FEC).

The spectrum for a 20 GBd optically filtered QPSK signal can be seen in Fig. 9(a), upper row. As expected and due to the

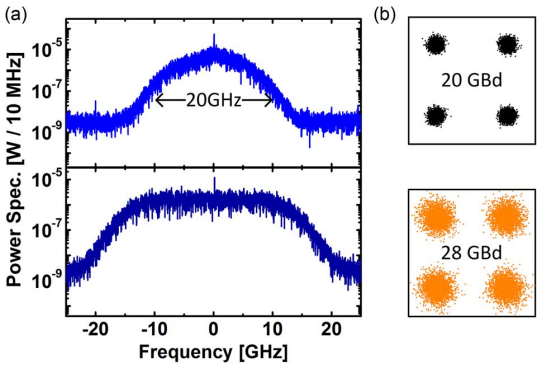


Fig. 9. (a) Measured and ensemble averaged spectrum (top) of a 20 GBd optically shaped QPSK signal. The spectral roll-off is not as steep as for electrically shaped signals. After DSP equalization at the Rx, the pass-band of the signal is flat (bottom). (b) Received constellation diagrams for 20 GBd (top) and 28 GBd (bottom) QPSK.

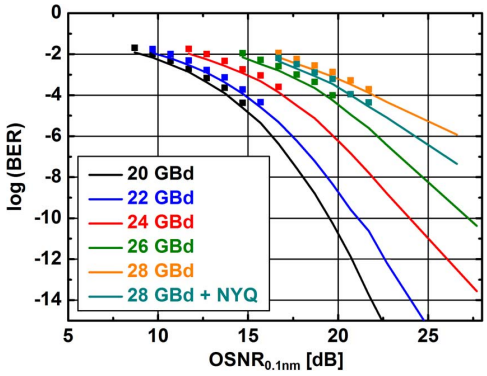


Fig. 10. Measurement results for optically shaped QPSK signals at various symbol rates as a function of OSNR. Measured BER (squares) and estimated BER(EVM) (solid lines) for different symbol rates and different OSNR. In our optical filter setup Nyquist filtering at the Rx only provides an advantage for symbol rates $\geq$28 GBd.

Lorentzian shape of the optical filter we do not see steep band edges as for the electrically shaped signal spectrum in Fig. 7(a). After DSP at the Rx we again obtain a flat pass-band of the signal spectrum (Fig. 9(a), lower row) leading to a minimum ISI. Constellation diagrams for 20 GBd and 24 GBd QPSK are shown in Fig. 9(b). Measured BER and estimated BER(EVM) for different OSNR and different symbol rates are depicted in Fig. 10. The BER increases with increasing symbol rate as the fixed filter width of 12.5 GHz significantly affects signals faster than 22 GBd. Applying the Nyquist filtering at the Rx as described in Section II-A decreases BER and EVM for signals with 28 GBd. For smaller symbol rates we find that there is no difference for a receiver with and without said electronic Nyquist filtering technique. Since 20 GBd signals showed best performance we use these signals for the WDM experiments described in Section III. The highest achievable OSNR decreases from 30 dB to 25 dB which reflects the additional insertion loss of the WaveShaper in our setup.

E. Comparing BER Performance With Theory

Since optical networks are usually operated such that received signals exhibit a BER close to the limit determined by

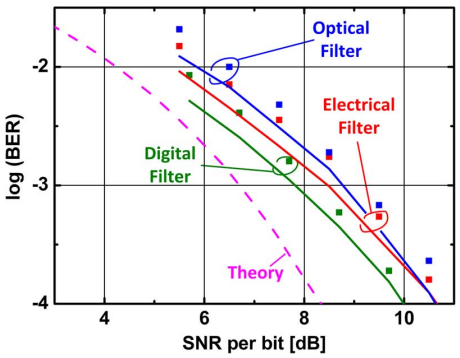


Fig. 11. BER (squares) and equivalent BER(EVM) (solid lines) as a function of SNR per bit for digitally (green), electrically (red), and optically (blue) filtered 20 GBd QPSK signals. The performance is close to what is expected from theoretical considerations (dashed curve) [22].

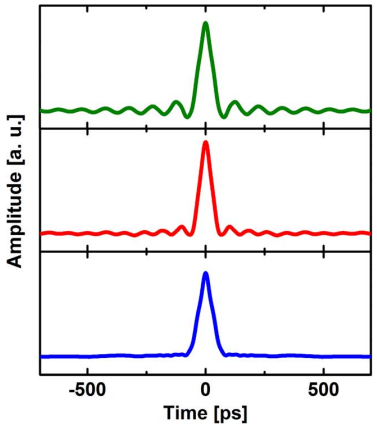


Fig. 12. Different pulse forms measured from digital (top, green), electrical (middle, red), and optical (bottom, blue) pulse-shapers. As expected, the digital pulse-shaper approximates a sinc-shaped pulse form most accurately. The electrical pulse-shaper still produces sinc-typical side lobes, whereas the optical pulse-shaper matches a sinc-function worst.

state-of-the-art FEC, we now more closely investigate the performance of signals shaped with the different filters in this very region. In addition, we compare the results to theory [24]. The outcome is depicted in Fig. 11. As before, the squares indicate the measured BER, the solid lines represent the BER(EVM) obtained from the measured EVM, and the dashed line marks the theoretically achievable performance. It can be seen that BER(EVM) deviates from direct BER measurements by a shift of only up to 0.2 on a logarithmic scale, corresponding to a factor of 1.5 on a linear scale. In order to guarantee a fair comparison, we rather plot BER as a function of SNR per bit [24] than using the $\mathrm{OSNR}_{0.1\ \mathrm{nm}}$ measure. This is because the OSA used to determine $\mathrm{OSNR}_{0.1\ \mathrm{nm}}$ does not account for the slightly different signal bandwidths of the differently shaped signals. Looking at Fig. 11 we can conclude that the BER performance obtained experimentally is close to what has been predicted theoretically [24]. It can be further seen that digitally shaped signals (green) are within 1.5 dB of the theoretical limit, and that the results for electrically (red) and optically (blue) shaped signals are closely neighbored.

F. Comparing Pulse Shapes

In order to give a better idea of how accurately a sinc-shaped impulse form is met when employing digital, electrical, and optical pulse-shapers, we use the received signal spectra from Figs. 3–9 (upper rows) without any equalization and derive the individual pulse forms. The outcome is depicted in Fig. 12. The digitally shaped pulse in Fig. 12 (top row) most accurately approximates a sinc-shaped impulse. The electrical pulse-shaper still produces sinc-typical side lobes but they decay rapidly, see Fig. 12 (middle row). The optical pulse-shaper yields the worst sinc-approximation, see Fig. 12 (bottom row). This was expected as the transfer of the optical filter is Lorentzian and not rectangular. For high symbol rates ($\geq$50 GBd), where several segments of the WaveShaper are transparent, the overall filter approximates a rectangle much better than for a 12.5 GHz single segment pass-band.

III. Performance Evaluation of the Different Pulse-Shapers in a WDM Network

The differently shaped QPSK signals are now employed in an ultra-dense WDM network scenario emulated by three channels with different, free-running carrier frequencies. A common QPSK symbol rate of 20 GBd is chosen for each channel. The channel spacing was varied between 17 GHz and 50 GHz and thus covers the WDM as well as the Nyquist WDM case. We measure single-polarization and PDM [15] signals. The quality of filtered, band-limited Tx signals is compared to the standard rectangular NRZ pulses, which are either received as is, or rectangularly filtered at the Rx.

For evaluating the performance of the different pulse-shaping techniques we investigate the transmitters discussed in Section II within a three-carrier ultra-dense WDM setup, Fig. 13. Three external cavity lasers (ECL) provide the three optical carriers with a linewidth below 100 kHz each. A fourth ECL of the same kind is used as LO within the OMA. All lasers are free-running, i.e., there is neither frequency nor phase locking. The observed frequency offsets between Tx lasers and LO laser are in the range of $\pm$100 MHz, which is only $\pm$0.5% of the symbol rate 20 GBd. Two different transmitters guarantee de-correlated data streams for the middle channel (Tx I) and the two outer channels (Tx II). The three signals are combined with equal powers. Polarization division multiplexing is emulated by splitting the signals in two arms, applying a delay of 5.3 ns in one arm, and finally combining both arms two form two orthogonal polarizations. For a worst-case linear cross-talk (where adjacent channels have the same state of polarization) we use polarization maintaining components and fibers (Fig. 13, blue). The ultra-dense WDM signal is amplified and coherently received by the OMA. By varying the carrier spacing Δf we determine the potential of the three pulse-shaping techniques for the minimum guard band and thus best spectral efficiency (SE) in a WDM network. The evaluation is based on both BER and EVM.

Measured spectra for pulses shaped with different techniques at a channel spacing of $\Delta f = 25$ GHz and for a symbol rate

2576 JOURNAL OF LIGHTWAVE TECHNOLOGY, VOL. 31, NO. 15, AUGUST 1, 2013

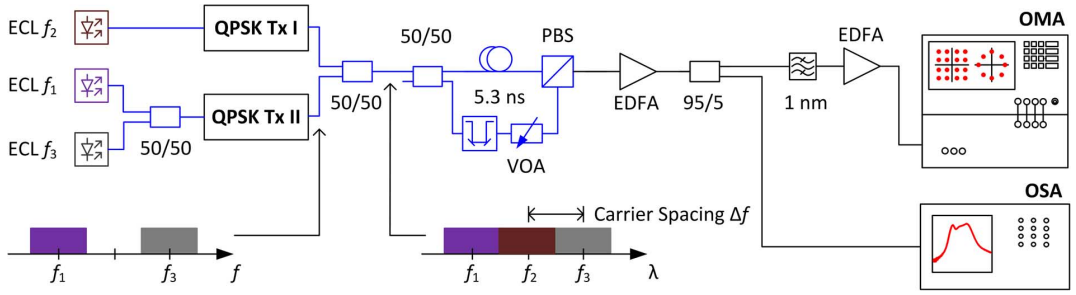


Fig. 13. Setup to test the minimum required carrier spacing Δf in an ultra-dense WDM network. Three free-running ECLs are encoded with QPSK signals and shaped in the digital, the electrical and optical domain. For testing purposes an unshaped NRZ-QPSK is tested as well. Two independent Tx guarantee uncorrelated data in adjacent channels. A worst-case scenario with polarization maintaining components (blue optical paths) maximizes inter-channel crosstalk. For PDM experiments the combined three channels are split, delayed by 5.3 ns, and combined in orthogonal polarizations. The remaining setup is identical to the ones used in Section II.

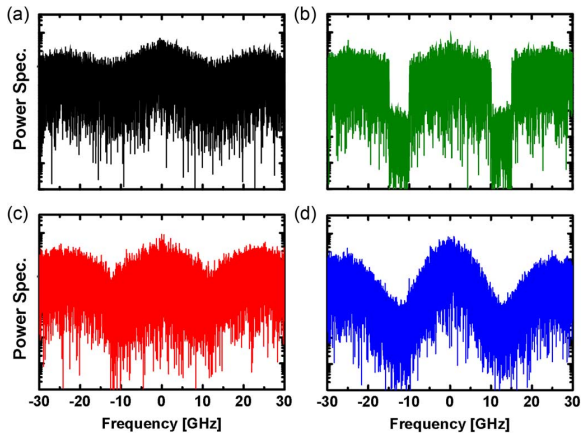


Fig. 14. Ultra-dense WDM spectrum of signals generated with different pulse-shaping techniques. The carrier spacing is $\Delta f = 25$ GHz and the carriers are QPSK encoded with single-polarization 20 GBd. (a) Unfiltered NRZ pulse-shape, only shaped by the limited electrical bandwidth of the DACs (>18 GHz) and the optical modulator (25 GHz). (b) Spectrum of digitally generated sinc-shaped QPSK signal. The digitally generated signal shows distinct spectral notches which are due to the steep-edged digital filters in the Tx. (c) Electrically pulse-shaped QPSK spectrum. (d) Optically pulse-shaped QPSK spectrum.

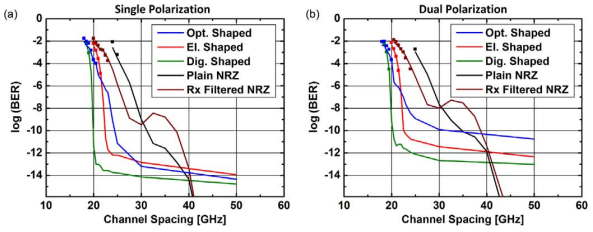


Fig. 15. Measured BER (squares) and estimated BER(EVM) (lines) for different pulse-shaping techniques as a function of varying channel spacing Δf. All channels transmit 20 GBd QPSK signals. The local maximum for Nyquist Rx-filtering near $\Delta f = 36$ GHz marks the point where the unfiltered sinc-shaped Tx spectra start overlapping. (a) Single polarization setup. (b) Polarization division multiplexing (PDM).

of 20 GBd are shown in Fig. 14. As a reference we first depict the unfiltered NRZ signal, Fig. 14(a). The NRZ signal is only shaped by the limited electrical bandwidths of DACs (>18 GHz) and the bandwidth of the optical modulator (25 GHz). For digitally pulse-shaped QPSK signals, see Fig. 14(b), the filter slopes are so steep that even notches appear in the region between the channels. For the case of the QPSK signals shaped electrically, see Fig. 14(c), and optically, see Fig. 14(d), one can see that the three channels slightly overlap.

We determine BER and BER derived from EVM measurements for all pulse-shaping schemes and for varying channel spacing Δf. All signals have a symbol rate of 20 GBd and are transmitted with highest possible OSNR. The results for the single polarization and the dual polarization experiments are depicted in Fig. 15(a) and (b), respectively. We begin with the unfiltered, plain NRZ signal (black solid lines for EVM derived BER and squares for BER measurements). As expected, the impact of inter-channel interference (ICI) for unfiltered, plain NRZ is largest. We also applied the Nyquist filtering technique to the unfiltered NRZ signal at the Rx (brown). Both signals require larger channel spacing than any of the pulse-shaped signals. At $\Delta f = 30$ GHz we see a break-even point for the unshaped NRZ signals. There is a local maximum of the Rx-filtering curve (brown) near $\Delta f = 36$ GHz where the unfiltered sinc-shaped Tx spectra start overlapping. It seems that there is a difference between detection with or without Rx Nyquist-filtering. This local maximum is much more distinct for the Rx Nyquist-filtered signals (brown) but also visible for the conventionally received NRZ (black). Optically pulse-shaped (blue) and electrically pulse-shaped (red) signals show negligible ICI for $\Delta f \geq 25$ GHz. We attribute the increased error floor for optically filtered PDM-QPSK (Fig. 15(b), blue) to the polarization de-multiplexing algorithm [13], which is sensitive to signal components of the two outer channels within the received middle channel. This is especially critical for optically pulse-shaped signals where the filter slopes are not very steep. The digitally filtered signals show negligible penalty due to ICI up to the Nyquist channel spacing of $\Delta f = 20$ GHz. This technique is clearly best suited for Nyquist WDM setups where the channel spacing is equal to the symbol rate and the symbol rate is sufficiently low. However, this digital operation comes at a price of intense signal processing so that electrical and optical pulse-shaping most likely have a CAPEX and OPEX advantage.

IV. Conclusion

We investigated the performance of digitally, electrically, and optically pulse-shaped QPSK signals for single-carrier trans-

mission and in a three-carrier ultra-dense WDM setup. For this purpose, both BER and EVM were measured. Digitally shaped sinc-pulses outperform other pulse-shaping techniques that rely on current state-of-the art electrical or optical filters. For digitally shaped signals the crosstalk is negligible even for a channel spacing of $\Delta f = 20$ GHz corresponding to the Nyquist limit for 20 GBd signals. Yet, it is important to note that both electrically and optically pulse-shaped signals always outperform unfiltered NRZ in terms of spectral efficiency. As an advantage of the analog techniques compared to digital pulse-shaping, costly DACs are not required and power consumption can be significantly reduced.

REFERENCES

[1] G. Bosco, A. Carena, V. Curri, P. Poggiolini, and F. Forghieri, "Performance limits of Nyquist-WDM and CO-OFDM in high-speed PM-QPSK systems," *IEEE Photon. Technol. Lett.*, vol. 22, no. 4, pp. 1129–1131, Apr. 2010.

[2] R. Schmogrow, M. Winter, M. Meyer, D. Hillerkuss, S. Wolf, B. Baeuerle, A. Ludwig, B. Nebendahl, S. Ben-Ezra, J. Meyer, M. Dreschmann, M. Huebner, J. Becker, C. Koos, W. Freude, and J. Leuthold, "Real-time Nyquist pulse generation beyond 100 Gbit/s and its relation to OFDM," *Opt. Express*, vol. 20, pp. 317–337, 2012.

[3] D. Hillerkuss, R. Schmogrow, M. Meyer, S. Wolf, M. Jordan, P. Kleinow, N. Lindenmann, P. Schindler, A. Melikyan, X. Yang, S. Ben-Ezra, B. Nebendahl, M. Dreschmann, J. Meyer, F. Parmigiani, P. Petropoulos, B. Resan, A. Oehler, K. Weingarten, L. Altenhain, T. Ellermeyer, M. Moeller, M. Huebner, J. Becker, C. Koos, W. Freude, and J. Leuthold, "Single-laser 32.5 Tbit/s Nyquist WDM transmission," *J. Opt. Commun. Netw.*, vol. 4, pp. 715–723, 2012.

[4] D. Qian, M. Huang, E. Ip, Y. Huang, Y. Shao, J. Hu, and T. Wang, "101.7-Tb/s (370 × 294-Gb/s) PDM-128QAM-OFDM transmission over 3 × 55-km SSMF using pilot-based phase noise mitigation," presented at the Optical Fiber Communication Conference, OSA Technical Digest (CD) (Optical Society of America, 2011), paper PDPB5.

[5] X. Zhou, L. Nelson, P. Magill, B. Zhu, and D. Peckham, "8 × 450-Gb/s, 50-GHz-spaced, PDM-32QAM transmission over 400 km and one 50 GHz-grid ROADM," presented at the Optical Fiber Communications Conference, OSA Technical Digest (CD) (Optical Society of America, 2011), paper PDPB3.

[6] Z. Dong, J. Yu, H. Chien, N. Chi, L. Chen, and G. Chang, "Ultra-dense WDM-PON delivering carrier-centralized Nyquist-WDM uplink with digital coherent detection," *Opt. Express*, vol. 19, pp. 11100–11105, 2011.

[7] T. Hirooka, P. Ruan, P. Guan, and M. Nakazawa, "Highly dispersion-tolerant 160 Gbaud optical Nyquist pulse TDM transmission over 525 km," *Opt. Express*, vol. 20, pp. 15001–15007, 2012.

[8] U. Koc, A. Leven, Y. Chen, and N. Kaneda, "Digital Coherent Quadrature Phase-Shift-Keying (QPSK)," presented at the Optical Fiber Communication Conference and Exposition and The National Fiber Optic Engineers Conference, Technical Digest (CD) (Optical Society of America, 2006), paper OThI1.

[9] J. Downie, J. Hurley, D. Pikula, and X. Zhu, "Ultra-long-haul 112 Gb/s PM-QPSK transmission systems using longer spans and Raman amplification," *Opt. Express*, vol. 20, pp. 10353–10358, 2012.

[10] R. Schmogrow, M. Winter, D. Hillerkuss, B. Nebendahl, S. Ben-Ezra, J. Meyer, M. Dreschmann, M. Huebner, J. Becker, C. Koos, W. Freude, and J. Leuthold, "Real-time OFDM transmitter beyond 100 Gbit/s," *Opt. Express*, vol. 19, pp. 12740–12749, 2011.

[11] C. Pulikkaseril, L. Stewart, M. Roelens, G. Baxter, S. Poole, and S. Frisken, "Spectral modeling of channel band shapes in wavelength selective switches," *Opt. Express*, vol. 19, pp. 8458–8470, 2011.

[12] R. Paiam, "Optical Interleaver/de-Interleaver," U.S. Patent No. 6 222 958, 2001.

[13] B. Szafraniec, B. Nebendahl, and T. Marshall, "Polarization demultiplexing in Stokes space," *Opt. Express*, vol. 18, pp. 17928–17939, 2010.

[14] R. Schmogrow, P. Schindler, C. Koos, W. Freude, and J. Leuthold, "Blind polarization demultiplexing with low computational complexity," *IEEE Photon. Technol. Lett.*, vol. 25, pp. 1230–1233, 2013.

[15] M. Sjödin, P. Johannisson, H. Wymeersch, P. Andrekson, and M. Karlsson, "Comparison of polarization-switched QPSK and polarization-multiplexed QPSK at 30 Gbit/s," *Opt. Express*, vol. 19, pp. 7839–7846, 2011.

[16] F. M. Gardner, "A BPSK/QPSK timing-error detector for sampled receivers," *IEEE Trans. Comm.*, vol. COM-34, no. 5, pp. 423–429, 1986.

[17] M. Yan, Z. Tao, L. Dou, L. Li, Y. Zhao, T. Hoshida, and J. Rasmussen, "Digital clock recovery algorithm for nyquist signal," presented at the Optical Fiber Communication Conference, OSA Technical Digest (online) (Optical Society of America, 2013), paper OTu2I.7.

[18] G. Clark, S. Mitra, and S. Parker, "Block implementation of adaptive digital filters," *IEEE Trans. Circuits Syst.*, vol. 28, pp. 584–592, 1981.

[19] R. Schmogrow, B. Nebendahl, M. Winter, A. Josten, D. Hillerkuss, J. Meyer, M. Dreschmann, M. Huebner, C. Koos, J. Becker, W. Freude, and J. Leuthold, "Error vector magnitude as a performance measure for advanced modulation formats," *IEEE Photon. Technol. Lett.*, vol. 24, pp. 61–63, 2012.

[20] H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. AIEE*, vol. 47, pp. 617–644, 1928.

[21] R. Schmogrow, D. Hillerkuss, S. Wolf, B. Bäuerle, M. Winter, P. Kleinow, B. Nebendahl, T. Dippon, P. Schindler, C. Koos, W. Freude, and J. Leuthold, "512QAM Nyquist sinc-pulse transmission at 54 Gbit/s in an optical bandwidth of 3 GHz," *Opt. Express*, vol. 20, pp. 6439–6447, 2012.

[22] R. Schmogrow, R. Bouziane, M. Meyer, P. Milder, P. Schindler, R. Killey, P. Bayvel, C. Koos, W. Freude, and J. Leuthold, "Real-time OFDM or Nyquist pulse generation—Which performs better with limited resources?," *Opt. Express*, vol. 20, pp. B543–B551, 2012.

[23] R. Schmogrow, M. Meyer, P. C. Schindler, A. Josten, S. Ben-Ezra, C. Koos, W. Freude, and J. Leuthold, "252 Gbit/s real-time nyquist pulse generation by reducing the oversampling factor to 1.33," presented at the Optical Fiber Communication Conference, OSA Technical Digest (Optical Society of America, 2013), paper OTu2I.1.

[24] R. Essiambre, G. Kramer, P. Winzer, G. Foschini, and B. Goebel, "Capacity limits of optical fiber networks," *J. Lightwave Technol.*, vol. 28, pp. 662–701, 2010.

[25] R. Schmogrow, B. Nebendahl, M. Winter, A. Josten, D. Hillerkuss, J. Meyer, M. Dreschmann, M. Huebner, C. Koos, J. Becker, W. Freude, and J. Leuthold, "Corrections to: Error vector magnitude as a performance measure for advanced modulation formats," *IEEE Photon. Technol. Lett.*, vol. 24, p. 2198, 2012, *ibid.*.

Author biographies not included at authors' request due to space constraints.

Appendix A

Linear and nonlinear fibre properties

A.1 Maxwell's equations

For describing nonlinearities in the fibre channel, we need to go back to the basic Maxwell equations for wave propagation. Electromagnetic waves, i. e., the magnetic and electric field vectors $\vec{H}$, $\vec{E}$, the electric displacement $\vec{D}$, the induced electric polarization $\vec{P}$, the magnetic induction $\vec{B}$ and the magnetic polarization $\vec{M}$ are solutions of Maxwell's equations. We assume vanishing current densities $\vec{J} = 0$, and no electric space charge densities $\varrho = 0$. The medium at the frequencies of interest be isotropic, for the time being linear and non-magnetic $\vec{M} = 0$, i. e., the medium properties are given by scalar, amplitude-independent quantities with a relative magnetic permeability $\mu_r = 1$. The dielectric constant (permittivity) and the magnetic permeability as well as the velocity of light and the wavelength λ in vacuum for a frequency f with an angular frequency $\omega = 2\pi f$ are ϵ_0, μ_0, $c = 1/\sqrt{\epsilon_0\mu_0}$, $\lambda = c/f$. The wave impedance of vacuum is $Z_0 = \sqrt{\mu_0/\epsilon_0} \approx 377\,\Omega$. With this notation, and in the International System of Units (*s*ystème *i*nternational d'unités, or SI), Maxwell's equations and the so-called constitutive or material equations are:

$$\left.\begin{aligned} \operatorname{curl}\vec{H} = \vec{J} + \frac{\partial\vec{D}}{\partial t}, \qquad & \operatorname{curl}\vec{E} = -\frac{\partial\vec{B}}{\partial t}, \\ \operatorname{div}\vec{D} = \varrho\,, \qquad & \operatorname{div}\vec{B} = 0, \end{aligned}\right\} \quad \text{Maxwell's equations}$$
$$\left.\begin{aligned} \vec{D} = \epsilon_0\vec{E} + \vec{P}, \qquad & \vec{B} = \mu_0\vec{H} + \vec{M}. \end{aligned}\right\} \quad \text{constitutive equations} \tag{A.1}$$

All vector quantities $\vec{X}$ are functions $\vec{X}(t,\vec{r})$ of time t and position vector $\vec{r} = x\vec{e}_x + y\vec{e}_y + z\vec{e}_z$ in Cartesian coordinates x, y, z (unit vectors $\vec{e}_{x,y,z}$). Assuming a positive time dependence $\exp(\mathrm{j}\,\omega t)$, the time-frequency Fourier transform relation (FT) and the inverse FT (IFT) are listed in Table 1.3 on Page 9).These functions are often discriminated only by their argument: $\Psi(t) \neq \Psi(f = t)$, $\Psi(f) := \breve{\Psi}(f)$.

A.2 Scalar optics

If the relative variations of refractive index n and spatial derivative $|\operatorname{grad} n|$ along a distance of a medium wavelength λ/n are small, $|\Delta n|_\lambda/n \ll 1$ and $|\Delta(\operatorname{grad} n)|_\lambda/|\operatorname{grad} n| \ll 1$, and if Cartesian coordinates for the vector components are used, the differential equations for the 6 scalar field components $\Psi(t,\vec{r}) := E_{x,y,z}(t,\vec{r}), H_{x,y,z}(t,\vec{r})$ are approximately decoupled as in a truly homogeneous medium,

$$\nabla^2\Psi(t,\vec{r}) = \frac{n^2(t,\vec{r})}{c^2}\frac{\partial^2\Psi(t,\vec{r})}{\partial t^2} \quad \text{for } \Psi := E_{x,y,z}, H_{x,y,z} \text{ and } \frac{|\Delta n|_\lambda}{n}, \frac{|\Delta(\operatorname{grad} n)|_\lambda}{|\operatorname{grad} n|} \ll 1\,. \tag{A.2}$$

Solving Eq. (A.2) for any of the scalar field components, say E_x, determines the total solution. This description is therefore known as the approximation of scalar optics. Naturally, the field components are

interrelated by initial and boundary conditions. Each spectral component of the vector fields can be then represented by a complex wave function $\Psi(t,\vec{r}) = \Psi(\vec{r})\exp(\mathrm{j}\,\omega t)$ with complex amplitude $\Psi(\vec{r})$, which is a solution of the so-called scalar Helmholtz equation,

$$\begin{aligned} &\Psi(t,\vec{r}) = \Psi(\vec{r})\,\mathrm{e}^{\mathrm{j}\,\omega t}, && I(\vec{r}) = \tfrac{1}{2}n(\vec{r})\;|\Psi(\vec{r})|^2\,, \\ &\left(\nabla^2 + \tfrac{\omega^2}{c^2}\epsilon_r(\vec{r})\right)\Psi(\vec{r}) = 0\,, && \tfrac{|\Delta\epsilon_r|_\lambda}{\epsilon_r},\ \tfrac{|\Delta(\mathrm{grad}\,\epsilon_r)|_\lambda}{|\,\mathrm{grad}\,\epsilon_r|} \ll 1. \end{aligned} \tag{A.3}$$

The wave amplitude $\Psi(\vec{r})$ is normalized by the optical intensity $I(\vec{r})$ (unit $\mathrm{W}\,/\,\mathrm{m}^2$).

A.3 General nonlinear medium

For a more general homogenous medium, we drop the assumptions that the medium at the frequencies of interest should be linear and isotropic, and write the (possibly nonlinear) wave equation for the electric field,

$$\begin{aligned} \mathrm{curl}\,\vec{H} = \epsilon_0\frac{\partial}{\partial t}\,\vec{E} + \frac{\partial}{\partial t}\vec{P}\,, \qquad \mathrm{curl}\,\vec{E} = -\mu_0\frac{\partial}{\partial t}\,\vec{H}\,, \qquad \mathrm{div}\,\vec{E} = 0\,, \\ \mathrm{curl}\,\mathrm{curl}\,\vec{E} = \left(-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\vec{E} + \left(-\mu_0\frac{\partial^2}{\partial t^2}\right)\vec{P}\,. \end{aligned} \tag{A.4}$$

For an isotropic medium, scalar optics as defined in Eq. (A.2) can again used for simplification.

A.3.1 Linear Polarization

A dielectric is made out of positive and negative charges, e. g., ions and electrons. When an electric field is present, it separates the charges of opposite polarity (periodically in the case of a time periodic field). This charge separation results in an additional electric field, called (induced) polarization. In real media, the polarization vector $\vec{P}(t,\vec{r})$ in Eq. (A.1) follows $\vec{E}(t,\vec{r})$ with some time-delay; it is understood that the fields are spatially local, but non-local in time. This "memory" behaviour at each position $\vec{r}$ (argument $\vec{r}$ is omitted if no ambiguity arises) may be described by a real causal impulse response $\chi_h(t)$, $\chi_h(t<0) = 0$,

$$\begin{aligned} &\vec{P}(t,\vec{r}) = \epsilon_0\int_0^\infty \chi_h(t_1,\vec{r})\,\vec{E}(t-t_1,\vec{r})\,\mathrm{d}t_1, \\ &\vec{P}(f) = \epsilon_0\underline{\chi}(f)\vec{E}(f), \qquad \underline{\chi}(f) = \int_0^\infty \chi_h(t)\,\mathrm{e}^{-\mathrm{j}\,2\pi f t}\,\mathrm{d}t, \\ &\underline{\chi}(f) = \chi(f) + \mathrm{j}\,\chi_i(f) = \epsilon_r(f) - 1 - \mathrm{j}\,\epsilon_{ri}(f), \quad \underline{\chi}(f) = \underline{\chi}^*(-f). \end{aligned} \tag{A.5}$$

The proportionality constant between the spectra $\epsilon_0\vec{E}(f)$ and $\vec{P}(f)$ is called linear electric susceptibility $\underline{\chi}$ (real part $\chi(f)$, imaginary part $\chi_i(f)$); it defines a linear complex relative dielectric constant $\bar{\epsilon}_r$ (real part $\epsilon_r(f)$, imaginary part $-\epsilon_{ri}(f)$). A linear complex refractive index $\bar{n}$ (real part $n(f)$, imaginary part $-n_i(f)$) is defined from the relation $\bar{\epsilon}_r = \bar{n}^2$.

It is possible that in certain frequency range(s) of interest $\chi(f) = \chi$ is constant (or weakly frequency dependent) having a (virtually) vanishing imaginary part, $\chi_i = 0$ (low-loss medium). For this range, the medium may be described by a real, constant (or weakly frequency dependent) relative dielectric constant ϵ_r or refractive index n. The polarization reacts *instantaneously* to the electric field. This is assumed in the usual ansatz for $\vec{D}(t)$,

$$\vec{P} = \epsilon_0\chi\vec{E}\,, \quad \vec{D} = \epsilon_0(1+\chi)\vec{E} = \epsilon_0\epsilon_r\vec{E}, \qquad n = \sqrt{\epsilon_r} \quad \text{in frequency range of interest.} \tag{A.6}$$

A.3.2 Nonlinear polarization

For large electric fields the linear relation Eq. (A.6) does not hold any more. For simplicity, we disregard here the vector nature of $\vec{E}$ and $\vec{P}$, and write for the nonlinear polarization $P(t)$ into the direction of $\vec{r} = x\vec{e}_x$ an expansion with respect to the electric field $E(t)$,

$$P(t) = \epsilon_0 \left(\chi^{(1)} E(t) + \chi^{(2)} E^2(t) + \chi^{(3)} E^3(t) + \ldots \right) \quad \text{for the frequency range of interest.} \tag{A.7}$$

The quantity $\chi^{(n)}$ is the real part of the complex susceptibility $\underline{\chi}^{(n)}$, so we assume a medium without loss or gain. Actually, in anisotropic media, the susceptibility is a tensor, which we write using Einstein's notation[1]. The susceptibility tensor[2] is also denoted by $\boldsymbol{\chi}^{(n)}$.

A.3.3 Order of nonlinearity

The coefficients $\chi^{(n)}$ in Eq. (A.7) are known as susceptibilities of order n, where $\chi^{(1)} := \chi$ in Eq. (A.5). Assuming isotropic media and a time-harmonic electric field $E(t) = \hat{E}(f_1)\cos(\omega_1 t)$, we find a polarization

$$\begin{aligned} P(t) = {} & \underbrace{\epsilon_0 \tfrac{1}{2} \chi^{(2)} \hat{E}^2}_{\text{opt. rectification}} + \underbrace{\epsilon_0 \chi^{(1)} \hat{E} \cos(\omega_1 t)}_{\text{linear optics}} + \underbrace{\epsilon_0 \tfrac{3}{4} (\chi^{(3)} \hat{E}^2) \hat{E} \cos(\omega_1 t)}_{\text{self-phase modulation}} \\ & + \underbrace{\epsilon_0 \tfrac{1}{2} \chi^{(2)} \hat{E}^2 \cos(2\omega_1 t)}_{\text{SHG}} + \underbrace{\epsilon_0 \tfrac{1}{4} \chi^{(3)} \hat{E}^3 \cos(3\omega_1 t)}_{\text{THG}} + \ldots . \end{aligned} \tag{A.10}$$

A wave of frequency f_1 generates a second wave with angular frequencies $f_2 = 0,\ f_1,\ 2f_1,\ 3f_1,\ \ldots$. Depending on the type $\chi^{(n)}$ of the nonlinearity, we identify the terms of

$\chi^{(1)}$ linear optics,

$\chi^{(2)}$ optical rectification (a DC voltage develops across the medium) and second-harmonic generation (SHG),

$\chi^{(3)}$ self-phase modulation (SPM, the effective susceptibility $\chi^{(3)}\hat{E}^2$ and therefore the refractive index are modified by the intensity $\hat{E}^2$ of the field), and finally the term of third-harmonic generation (THG).

[1] Whenever appropriate, we replace subscripts x, y, z by subscripts $1, 2, 3$. If such an index occurs two or more times in a term, it is implied, whithout any further symbols, that the terms are to be summed over all possible values of the index. Example: Let $\vec{X} = X_x\vec{e}_x + X_y\vec{e}_y + X_z\vec{e}_z$. Replacing the subscripts yields $\vec{X} = X_1\vec{e}_1 + X_2\vec{e}_2 + X_3\vec{e}_3 \mathrel{\widehat{=}} X_i$. The scalar product $\vec{X}^2 = \vec{X}\cdot\vec{X} = X_1^2 + X_2^2 + X_3^2 = \sum_{i=1}^3 X_i X_i$ may be simply written as $X_i X_i$.

[2] A tensor $\boldsymbol{\chi}^{(n)}$ of rank $(n+1)$ having 3^{n+1} elements transforms a vector $\vec{E}$ into a tensor of rank n, $\boldsymbol{\chi}^{(n)} \cdot \vec{E} = \boldsymbol{\chi}^{(n-1)}$. Common notations are (vectors $\vec{Q}, \vec{R}, \vec{S}$):

$$\begin{aligned} \boldsymbol{\chi}^{(1)} \cdot \vec{E} &= \vec{Q} \\ \boldsymbol{\chi}^{(2)} \vdots \vec{E}\vec{E} &:= (\boldsymbol{\chi}^{(2)} \cdot \vec{E}) \cdot \vec{E} = \vec{R} \\ \boldsymbol{\chi}^{(3)} \vdots \vec{E}\vec{E}\vec{E} &:= ((\boldsymbol{\chi}^{(3)} \cdot \vec{E}) \cdot \vec{E}) \cdot \vec{E} = \vec{S} \end{aligned} \tag{A.8}$$

The tensors of rank 0, 1, and 2 are more simply denoted as scalars, vectors, and tensors. Replacing the Cartesian coordinates (subscripts x, y, z) by the subscripts $1, 2, 3$, we formulate the product of a rank-2 tensor $\boldsymbol{\chi}^{(1)}$ (9 components χ_{ij}) with a vector $\vec{E}$ (3 components E_j), which results in a vector $\vec{Q}$ (3 components Q_j) by employing the Einstein summation convention (see Footnote 1 on Page 177). The same notation is used for expressions with rank-3 and rank-4 tensors:

$$\begin{aligned} \vec{Q} = \boldsymbol{\chi}^{(1)} \cdot \vec{E} \quad &\Longleftrightarrow \quad \chi^{(1)}_{ij} E_j = Q_i\,, \qquad \chi^{(1)}_{ij} E_j := \sum_{j=1}^{3} \chi^{(1)}_{ij} E_j \\ \vec{R} = \boldsymbol{\chi}^{(2)} \vdots \vec{E}\vec{E} \quad &\Longleftrightarrow \quad \chi^{(2)}_{ijk} E_j E_k = R_i\,, \qquad \chi^{(2)}_{ijk} E_j E_k := \sum_{j,k=1}^{3} \chi^{(2)}_{ijk} E_j E_k \\ \vec{S} = \boldsymbol{\chi}^{(3)} \vdots \vec{E}\vec{E}\vec{E} \quad &\Longleftrightarrow \quad \chi^{(3)}_{ijkl} E_j E_k E_l = S_i\,, \quad \chi^{(3)}_{ijkl} E_j E_k E_l := \sum_{j,k,l=1}^{3} \chi^{(3)}_{ijkl} E_j E_k E_l \end{aligned} \tag{A.9}$$

When two waves $E(t) = \hat{E}(f_1)\cos(\omega_1 t) + \hat{E}(f_2)\cos(\omega_2 t)$ are interacting through $\chi^{(2)}$ to generate a third wave of frequency f_3, the process is spoken of as three-wave mixing (TWM). Finally, when three waves $E(t) = \hat{E}(f_1)\cos(\omega_1 t) + \hat{E}(f_2)\cos(\omega_2 t) + \hat{E}(f_3)\cos(\omega_3 t)$ interact in a nonlinear $\chi^{(3)}$-medium to generate a fourth frequency f_4, this is know as four-wave mixing (FWM). All possible nth-order effects show if n incident waves with (possibly degenerate) frequencies f_1, f_2, ..., f_n interact to generate a polarization at frequency f_{n+1}. For optical quartz glass fibres, the influence of second-order nonlinear processes is small[3] compared to the importance of the $\chi^{(3)}$-nonlinearity. The general polarization Eq. (A.1) can be split into a first-order linear and a third-order nonlinear part,

$$\vec{P}(t,\vec{r}) = \vec{P}^{(1)}(t,\vec{r}) + \vec{P}^{(3)}(t,\vec{r}). \tag{A.13}$$

Using again Einstein's notation[4], we assume that the optical field maintains its initial polarization E_i, P_i along the fibre (coupling of field components is negligible), so that the scalar approach may be adopted. Then, the nonlinear wave equation (A.4) reduces to

$$\nabla^2 E_i(t,\vec{r}) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}E_i(t,\vec{r}) = \mu_0 \frac{\partial^2}{\partial t^2}\left(P_i^{(1)}(t,\vec{r}) + P_i^{(3)}(t,\vec{r})\right). \tag{A.14}$$

Because of the complexity of Eq. (A.14), several further approximations[5] become necessary.

A.4 Nonlinear Schrödinger equation

A.4.1 Separation ansatz

Assuming in Eq. (A.14) that $\vec{E} = E_x\vec{e}_x$, $\vec{P} = P_x\vec{e}_x$, the Helmholtz equation is solved[6,7] by an ansatz for $\breve{E}_x(f - f_0, \vec{r})$, where the the transverse modal field function $\tilde{F}(x,y)$, the slowly varying envelope $\breve{a}(f,z)$, the phase term $\mathrm{e}^{-\mathrm{j}\beta_{\mathrm{ref}}z}$, and a normalization constant c_p are combined in product form. The fixed reference propagation constant β_{ref} is basically arbitrary and will be determined later,

$$\breve{E}_x(f-f_0,\vec{r}) = c_p\,\tilde{F}(x,y)\,\breve{a}(f-f_0,z)\,\mathrm{e}^{-\mathrm{j}\,\beta_{\mathrm{ref}}z}, \qquad \hat{E}_x(t,\vec{r}) = \int_{-\infty}^{+\infty}\breve{E}_x(f-f_0,\vec{r})\,\mathrm{e}^{\mathrm{j}\,2\pi f t}\,\mathrm{d}f\,. \tag{A.15}$$

Because waveguide losses and nonlinearities are small, $\tilde{\epsilon}_r \approx \epsilon_r$, the reference propagation constant β_{ref} is chosen to be real. Substituting the inverse Fourier transform $\hat{E}_x(t,\vec{r})$ of Eq. (A.15) and replacing $i = x$,

[3] Frequency-doubling or SHG results from the second-order nonlinearity $\chi^{(2)}$. The electric field be oriented along the x-axis, $\vec{r} = x\,\vec{e}_x$, $\vec{E} = E\vec{e}_x$, $E(t) = \hat{E}(f_1)\cos(\omega_1 t)$, and the polarisation $\vec{P}^{(2)} = P^{(2)}\vec{e}_x$ with $P^{(2)}(t) = \hat{P}(2f_1)\cos(2\omega_1 t)$ at frequency $f_2 = 2f_1$ points into the same direction,

$$\hat{P}^{(2)}\vec{e}_x = \epsilon_0 \tfrac{1}{2}\chi^{(2)}\hat{E}^2\vec{e}_x, \qquad \hat{P}^{(2)} = \epsilon_0\tfrac{1}{2}\chi^{(2)}\hat{E}^2. \tag{A.11}$$

We maintain that the second-order susceptibility $\chi^{(2)}$ vanishes for isotropic materials and crystals with inversion symmetry (centrosymmetry; the atomic arrangement remains unchanged when mirrored at the inversion centre according to $\vec{r} = -\vec{r}$).

This may be easily proved by the following argument: If the direction of the electric field is reversed, $\vec{E} \to -\vec{E}$, the modulus of the polarisation cannot change because the physical arrangement is undistinguishable from the former one, but the direction of the polarisation should reverse, $\vec{P} \to -\vec{P}$. Looking at Eq. (A.11) we require

$$-\hat{P}^{(2)} = \epsilon_0\tfrac{1}{2}\chi^{(2)}\left(-\hat{E}\right)^2 = \epsilon_0\tfrac{1}{2}\chi^{(2)}\hat{E}^2 \qquad \left(\hat{P}^{(2)} = \epsilon_0\tfrac{1}{2}\chi^{(2)}\hat{E}^2 \text{ from Eq. (A.11)}\right), \tag{A.12}$$

which is in contradiction to Eq. (A.11). Therefore, $\chi^{(2)} = 0$ must be true for isotropic materials (glass, gases, liquids) and for crystals with inversion symmetry. Based on similar arguments, $\chi^{(2n)} = 0$ can be proved for all susceptibilities of *even* order. Fused silica (quartz *glass*) has a symmetric SiO_2 molecule, is amorphous and isotropic ($\chi^{(2n)}_{\mathrm{SiO_2\ glass}} = 0$), while a quartz *crystal* lacks this inversion symmetry ($\chi^{(2n)}_{\mathrm{SiO_2\ crystal}} \neq 0$).

In general, the susceptibility has tensor character, i. e., $\chi^{(n)}$ varies according to the vibration directions of the electric field. Depending on the symmetry class of the material, the number of different tensor coefficients reduces strongly.

[4] See Footnote 1 on Page 177

[5] See Sect. 2.3.1 Page 40 in reference Footnote 17 on Page 6

[6] See Sect. 2.3.1 Page 42 ff. in reference Footnote 17 on Page 6

[7] T. Kremp: Split-step wavelet collocation methods for linear and nonlinear optical wave propagation. PhD Thesis, Karlsruhe, February 2002. Chapter 3

we find

$$E_x(t,\vec{r}) = \frac{1}{2}\Big(\hat{E}_x(t,\vec{r})\,\mathrm{e}^{\mathrm{j}(\omega_0 t-\beta_{\mathrm{ref}}z)} + \mathrm{cc}\Big), \qquad \hat{E}_x(t,\vec{r}) = c_p\,\tilde{F}(x,y)\,a(t,z). \tag{A.16}$$

With Eq. (A.15), the wave equation for the E_x-component is

$$\Big(\nabla^2 + \tilde{\epsilon}_r(f)k_0^2\Big)\Big(c_p\,\tilde{F}(x,y)\,\breve{a}(f-f_0,z)\,\mathrm{e}^{-\mathrm{j}\,\beta_{\mathrm{ref}}z}\Big) = 0\,. \tag{A.17}$$

Introducing the transverse Laplace operator $\nabla_{\mathrm{t}}^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ in Cartesian coordinates, Eq. (A.17) leads to

$$\begin{aligned}\nabla_{\mathrm{t}}^2\tilde{F}(x,y)\,\breve{a}(f-f_0,z) + \tilde{F}(x,y)\left(\frac{\partial^2}{\partial z^2} - \mathrm{j}\,2\beta_{\mathrm{ref}}\frac{\partial}{\partial z} - \beta_{\mathrm{ref}}^2\right)\breve{a}(f-f_0,z)\\ + \tilde{\epsilon}_r(f)\,k_0^2\tilde{F}(x,y)\,\breve{a}(f-f_0,z) = 0\end{aligned} \tag{A.18}$$

For non-zero $\tilde{F}(x,y)$ and $\breve{a}(f-f_0,z)$ we divide Eq. (A.18) by $\tilde{F}(x,y)\,\breve{a}(f-f_0,z)$,

$$\underbrace{\frac{\nabla_{\mathrm{t}}^2\tilde{F}(x,y)}{\tilde{F}(x,y)} + \tilde{\epsilon}_r(f)\,k_0^2}_{g_1(f,x,y,)=\tilde{\beta}^2=\mathrm{const}_{x,y}} + \underbrace{\frac{\left(\frac{\partial^2}{\partial z^2} - \mathrm{j}\,2\beta_{\mathrm{ref}}\frac{\partial}{\partial z}\right)\breve{a}(f-f_0,z)}{\breve{a}(f-f_0,z)} - \beta_{\mathrm{ref}}^2}_{g_2(f,z)=-\tilde{\beta}^2=\mathrm{const}_z} = 0\,. \tag{A.19}$$

Both underbraced terms are represented by functions $g_1(f,x,y,) = g_2(f,z)$, which are equal, but do not depend on the same spatial variables. Therefore they must be constant with respect to their spatial spatial variables. Consequently, we define a frequency-dependent separation constant $\tilde{\beta}^2 = \mathrm{const}_{x,y,z}$,

$$\tilde{\beta}^2 = \frac{\nabla_{\mathrm{t}}^2\tilde{F}(x,y)}{\tilde{F}(x,y)} + \tilde{\epsilon}_r(f)k_0^2 = -\frac{\left(\frac{\partial^2}{\partial z^2} - \mathrm{j}\,2\beta_{\mathrm{ref}}\frac{\partial}{\partial z}\right)\breve{a}(f-f_0,z)}{\breve{a}(f-f_0,z)} + \beta_{\mathrm{ref}}^2\,. \tag{A.20}$$

This separates Eq. (A.18) in two differential equations,

$$\Big(\nabla_{\mathrm{t}}^2 + \tilde{\epsilon}_r k_0^2 - \tilde{\beta}^2\Big)\,\tilde{F}(x,y) = 0\,, \tag{A.21}$$

$$\left(\frac{\partial^2}{\partial z^2} - \mathrm{j}\,2\beta_{\mathrm{ref}}\,\frac{\partial}{\partial z} + \tilde{\beta}^2 - \beta_{\mathrm{ref}}^2\right)\breve{a}(f-f_0,z) = 0\,. \tag{A.22}$$

Without affecting the results, we could have chosen $\tilde{\beta}^2 - \mathrm{const}_{x,y,z}$ as the separation constant, e. g., $\tilde{\beta}^2 - \beta_{\mathrm{ref}}^2$. Independently of that choice, $\tilde{\beta}^2$ results by solving the eigenvalue problem Eq. (A.21) for the transverse field alone, and no solution of the differential equation (A.22) for the amplitude $\breve{a}(f-f_0,z)$ is required.

A.4.2 Slowly varying envelope approximation

We require an envelope, which varies slowly on the time scale of a period $T_0 = 1/f_0$ of an optical carrier having a frequency $f_0 = \omega_0/(2\pi)$, therefore we neglect the second z-derivative,

$$\left|\frac{\partial^2\breve{a}}{\partial z^2}\right| \ll \left|2\beta_{\mathrm{ref}}\,\frac{\partial\breve{a}}{\partial z}\right| \qquad \text{for } \Delta t \geq \frac{10}{f_0}\,. \tag{A.23}$$

This *s*lowly *v*arying *e*nvelope *a*pproximation (SVEA) neglects the backwards-propagating components of the field generated by the nonlinear polarization[8] $P_i^{(3)}(t,\vec{r})$ and is justified under the following conditions: Consider an optical carrier, the envelope of which has a temporal width Δt and a spatial width $\Delta z =$

[8] Shen, Y. R.: The principles of nonlinear optics. New York: Wiley-Intersicence 1984. Chapter xii Page 216

$\Delta t\, c/n$. Inside Δz, the maximum amplitude change for a triangular impulse is assumed to be $\Delta \breve{a} = 1$, therefore an estimate of the first and second derivatives will be $|\partial \breve{a}/\partial z| \simeq 2/\Delta z$ and $|\partial^2 \breve{a}/\partial z^2| \simeq 4/(\Delta z)^2$. The condition Eq. (A.23), $4/(\Delta z)^2 \ll (8\pi n/\lambda_0)/\Delta z$, simplifies to $\Delta t \gg 1/\omega_0$. Replacing $\gg$ by $20\pi\times$, we arrive at Eq. (A.23). For $\lambda_0 = 1.5\,\mu\text{m}$ ($f_0 = 200\,\text{THz}$), we find $\Delta t \geq 10 \times T_0 = 50\,\text{fs}$.

This assumption leads with the inverse Fourier transform $a(t,z) = \int_{-\infty}^{+\infty} \breve{a}(f - f_0, z)\, \mathrm{e}^{\mathrm{j}\, 2\pi (f-f_0)t}\, \mathrm{d}f$ and after replacing the term $(\omega - \omega_0)$ wih the differential operator $\mathrm{j}\,\partial/\partial t$ to

$$\frac{\partial a(t,z)}{\partial z} = \left(-\beta_0^{(1)} \frac{\partial}{\partial t} + \mathrm{j}\, \frac{\beta_0^{(2)}}{2!} \frac{\partial^2}{\partial t^2} + \frac{\beta_0^{(3)}}{3!} \frac{\partial^3}{\partial t^3} - \mathrm{j}\Big(\Delta\tilde{\beta}(\omega_0) + \beta_0 - \beta_{\text{ref}} \Big) \right) a(t,z). \tag{A.24}$$

With $|\breve{a}(f_0 - f_0, z)|^2 \approx |a(t,z)|^2$, and under the assumption of weak two-photon absorption α_2, relation (A.24) is known as the nonlinear Schrödinger equation (NLSE),

$$\frac{\partial a(t,z)}{\partial z} = \left(-\beta_0^{(1)} \frac{\partial}{\partial t} + \mathrm{j}\, \frac{\beta_0^{(2)}}{2} \frac{\partial^2}{\partial t^2} + \frac{\beta_0^{(3)}}{6} \frac{\partial^3}{\partial t^3} - \mathrm{j}\Big(\gamma\, |a(t,z)|^2 - \mathrm{j}\, \frac{\alpha}{2} + \beta_0 - \beta_{\text{ref}} \Big) \right) a(t,z),$$

$$\bar{\gamma} \approx \gamma = \Re\{\bar{\gamma}\} = \frac{n_2^I k_0}{A_{\text{eff}}}\,, \qquad \frac{|\alpha_2^I|}{A_{\text{eff}}}\, |a|^2 \ll \frac{|\alpha|}{2}\,, \qquad A_{\text{eff}} = \frac{\Big(\iint_{-\infty}^{+\infty} |F(x,y)|^2\, \mathrm{d}x\, \mathrm{d}y \Big)^2}{\iint_{-\infty}^{+\infty} |F(x,y)|^4\, \mathrm{d}x\, \mathrm{d}y}\,. \tag{A.25}$$

If for any $z = z_0$ the initial value $a(t, z_0)$ is given, the envelope $a(t,z)$ for all z can be determined using Eq. (A.25).

A.4.3 Transformation of variables

To simplify Eq. (A.25), we introduce new variables. The time in a reference frame moving with the group velocity $v_g = 1/\beta_0^{(1)}$ is denoted by T, and the distance is temporarily called Z,

$$T = T(t,z) := t - \beta_0^{(1)} z = t - z/v_g\,, \qquad Z = Z(t,z) := z\,. \tag{A.26}$$

In the coordinate system (T, z), we define an envelope $A(T, Z)$,

$$A\big(T(t,z), Z(t,z)\big) := a(t,z). \tag{A.27}$$

Differentiating Eq. (A.27) yields

$$\frac{\partial a(t,z)}{\partial z} = \frac{\partial A(T,Z)}{\partial T} \frac{\partial T}{\partial z} + \frac{\partial A(T,Z)}{\partial Z} \frac{\partial Z}{\partial z} = -\beta_0^{(1)} \frac{\partial A(T,Z)}{\partial T} + \frac{\partial A(T,Z)}{\partial Z}\,, \tag{A.28}$$

$$\frac{\partial a(t,z)}{\partial t} = \frac{\partial A(T,Z)}{\partial T} \frac{\partial T}{\partial t} + \frac{\partial A(T,Z)}{\partial Z} \frac{\partial Z}{\partial t} = \frac{\partial A(T,Z)}{\partial T}\,. \tag{A.29}$$

For the higher derivatives $n \geq 2$ with respect to t, we see with Eq. (A.29)

$$\begin{aligned} \frac{\partial^n a(t,z)}{\partial t^n} &= \frac{\partial a^{(n-1)}(t,z)}{\partial t} = \frac{\partial A^{(n-1)}(T,Z)}{\partial T} \frac{\partial T}{\partial t} + \frac{\partial A^{(n-1)}(T,Z)}{\partial Z} \frac{\partial Z}{\partial t} \\ &= \frac{\partial A^{(n-1)}(T,Z)}{\partial T} = \frac{\partial^n A(T,Z)}{\partial T^n}\,. \end{aligned} \tag{A.30}$$

Substituting Eq. (A.27)–(A.30) in Eq. (A.25), we find

$$\frac{\partial A(T,Z)}{\partial Z} = \left(\mathrm{j}\, \frac{\beta_0^{(2)}}{2} \frac{\partial^2}{\partial T^2} + \frac{\beta_0^{(3)}}{6} \frac{\partial^3}{\partial T^3} - \mathrm{j}\big(\gamma\, |A(T,Z)|^2 - \mathrm{j}\, \frac{\alpha}{2} + \beta_0 - \beta_{\text{ref}} \big) \right) A(T,Z)\,. \tag{A.31}$$

Because of Eq. (A.26), we re-substitute Z by z in the argument of the envelope A. If we further choose

$$\beta_{\text{ref}} = \beta_0\,, \tag{A.32}$$

equation (A.31) simply becomes

$$\frac{\partial A(T,z)}{\partial z} = \left(\mathrm{j}\,\frac{\beta_0^{(2)}}{2}\frac{\partial^2}{\partial T^2} + \frac{\beta_0^{(3)}}{6}\frac{\partial^3}{\partial T^3} - \mathrm{j}\,\gamma\,|A(T,z)|^2 - \frac{\alpha}{2} \right) A(T,z). \tag{A.33}$$

Due to the quasi-monochromacy of optical signals, the higher temporal derivatives $\partial^n/\partial T^n$ for $n \geq 3$ in Eq. (A.33) can be neglected if $\beta_0^{(2)} \neq 0$, i. e., if f_0 does not lie in the vicinity of a zero-dispersion wavelength of the waveguide,

$$\frac{\partial A(T,z)}{\partial z} = \mathrm{j}\,\frac{\beta_0^{(2)}}{2}\frac{\partial^2 A(T,z)}{\partial T^2} - \mathrm{j}\,\gamma\,|A(T,z)|^2\,A(T,z) - \frac{\alpha}{2}A(T,z). \tag{A.34}$$

In the case of zero linear attenuation $\alpha = 0$, Eq. (A.34) resembles the well-known Schrödinger equation of quantum mechanics with a nonlinear (quadratic) potential term. Thus, it is called the *nonlinear Schrödinger equation*[9,10] (NLSE). If during the propagation of a light signal its loss is continuously compensated by gain, then the power loss constant can be set actually to zero, $\alpha = 0$. For including random perturbations by, e. g., ASE noise of optical amplifiers, a random field[11] $-\mathrm{j}\,N_{\mathrm{ASE}}(T,z)$ can be added on the right-hand side of Eq. (A.34).

[9] See Sect. 2.3.1 Eq. (2.3.27) Page 43 in reference Footnote 17 on Page 6

[10] Boyd, R. W: Nonlinear optics. 3. Ed. San Diego: Academic Press 2008. Section 7.5.2, Eq. (7.5.32)

[11] R.-J. Essiambre, G. Kramer, P. J. Winzer, G. J. Foschini, B. Goebel: Capacity limits of optical fiber networks. J. Lightw. Technol. 28 (2010) 662–701

Appendix B

Sampling, quantizing and discrete Fourier transform

For digital signal processing, real-world continuous signals must be discretized, i. e., the signals have to be sampled with respect to time and frequency, and their amplitudes have to be quantized. As a consequence, the continuous Fourier transform, which is involved in signal analysis very often, has to be replaced by the discrete Fourier transform, or, equivalently, by the numerically more efficient fast Fourier transform.

Physically it is not possible to sample or to quantize a signal in infinitely small, δ-sized intervalls. Instead, we integrate the signal over a periodically repeated window called a "bin". Sometimes the bins are small enough to approximate an ideal δ-sampling, but especially in quantizing amplitudes the bins have a non-negligible width. The following sections treat a few aspects in this context. Finally, we discuss some properties of the discrete Fourier transform.

B.1 Sampling with a finite temporal bin size

In Sect. 2.1.1 on Page 13 ff. we discussed the case of ideal sampling with $\delta\,(t)$-functions in time. According to Eq. (2.1) we found that a complex signal $\Psi(t)$ with a spectrum $\breve{\Psi}(f)$ that is limited to a bandwidth B can be reconstructed from samples $\Psi(iT_s)$ $(i = 0, \pm1, \pm2, \ldots)$, if $T_s = 1/F_s \leq 1/B$ holds, i. e., if the sampling frequency F_s is as large as or larger than the signal's bandwidth B.

In a more practical case the signal $\Psi(t)$ is first integrated in a finite temporal window (a "bin") with a width of T_s, thus forming the moving average

$$\Psi^{(\mathrm{b})}(t) = \frac{1}{T_s} \int_{t-T_s/2}^{t+T_s/2} \Psi(t_1)\,\mathrm{d}t_1\,. \tag{B.1}$$

This averaged signal is then sampled. The procedure is equivalent to sampling with a window of finite width T_s. Here we choose a bin width identical to the sampling interval T_s, but a smaller size would be also possible. For this so-called "bin-sampled" function $\Psi_s^{(\mathrm{b})}(t)$ and its spectrum $\breve{\Psi}_s^{(\mathrm{b})}(f)$ we find

$$\begin{aligned}\Psi_s^{(\mathrm{b})}(t) &= \frac{1}{T_s} \int_{t-T_s/2}^{t+T_s/2} \Psi(t_1)\,\mathrm{d}t_1\; T_s \sum_{i=-\infty}^{+\infty} \delta\,(t - iT_s) = \frac{1}{T_s} \int_{-\infty}^{+\infty} \Psi(t_1)\,\mathrm{rect}\Big(\frac{t-t_1}{T_s}\Big)\mathrm{d}t_1\; T_s \sum_{i=-\infty}^{+\infty} \delta\,(t - iT_s) \\ &= \Big(\Psi\,(t) * \frac{1}{T_s}\,\mathrm{rect}\Big(\frac{t}{T_s}\Big)\Big)(t)\; T_s \sum_{i=-\infty}^{+\infty} \delta\,(t - iT_s)\,, \end{aligned} \tag{B.2a}$$

$$\breve{\Psi}_s^{(\mathrm{b})}(f) = \Big(\breve{\Psi}(f)\;\mathrm{sinc}\Big(\frac{f}{F_s}\Big)\Big) * \sum_{i=-\infty}^{+\infty} \delta\Big(f - iF_s\Big) = \sum_{i=-\infty}^{+\infty} \breve{\Psi}\Big(f - iF_s\Big)\;\mathrm{sinc}\Big(\frac{f - iF_s}{F_s}\Big)\,. \tag{B.2b}$$

The spectrum $\breve{\Psi}_s^{(\mathrm{b})}(f)$ of the bin-sampled function $\Psi_s^{(\mathrm{b})}(t)$ differs from the spectrum $\breve{\Psi}_s(f)$ in Eq. (2.1) of a δ-sampled signal $\Psi_s(t)$ by the weighing function $\operatorname{sinc}(f/F_s - i)$.

If for a real[1] signal $\Psi(t)$ the sampling theorem is fulfilled, i. e., if $T_s \leq 1/(2B)$ or $F_s \geq 2B$ holds, then the central partial spectrum $\breve{\Psi}(f)\,\operatorname{sinc}(f/F_s)$ with $i = 0$, which vanishes for $|f| > B$, is *not* superimposed by neigbouring image spectra with $i \neq 0$ (aliasing does not occur), and a baseband filter with bandwidth B reconstructs the original moving-average signal,

$$\Psi^{(\mathrm{b})}(t) = \int_{-\infty}^{+\infty} \breve{\Psi}(f)\,\operatorname{sinc}\Big(\frac{f}{F_s}\Big)\,\mathrm{e}^{\mathrm{j}\,2\pi f t}\,\mathrm{d}f = \bigg(\Psi(t) * \frac{1}{T_s}\operatorname{rect}\Big(\frac{t}{T_s}\Big)\bigg)(t) = \frac{1}{T_s}\int_{t-T_s/2}^{t+T_s/2} \Psi(t_1)\,\mathrm{d}t_1\,. \quad \text{(B.3)}$$

B.2 Quantizing with an analogue-to-digital converter

Sampling with a finite temporal bin size, i. e., quantizing in time, and quantizing of other physical quantities like voltages are described with the same formalism. An essential characteristic of an analogue-to-digital converter (ADC) is the error introduced by the quantizing process. Closely connected to this so-called quantization noise is the effective number of bits (ENOB), which we can attribute to an ADC. Before we enter the discussion of quantization noise and ENOB, we need to summarize a few elements of probability theory. If you are not interested in these subtleties, or if you are familiar with the topic, you can just jump to Sect. B.2.2 on Page 187, where we discuss how a functional dependence $y = f(x)$ of an output quantity y on an input x changes the statistics of x. The characteristics of an ADC are then detailed on Page 189 ff.

B.2.1 Elements of probability theory

Random variables

A random variable (RV, *German* Zufallsvariable) is a number $\mathbf{x}(\xi)$, which is associated with a certain result (outcome) ξ of an experiment, i. e., it is associated with a certain elementary event[2,3,4]. This number $\mathbf{x}(\xi)$ could be the price in a game of chance, or the momentary electrical voltage of a noisy resistor. For instance, if a randomly thrown die shows the six (outcome $\xi = 6$), then let the winning be $\mathbf{x}(6) = 1.50\,€$. However, it would be also possible to associate with the outcome $\xi = 6$ meaning "the face with six dots lies on top", for instance the number 17. If a certain noise voltage $\xi = 10\,\mu\mathrm{V}$ is measured (outcome), the associated number could be $\mathbf{x}(\xi = 10\,\mu\mathrm{V}) = 1$; naturally, we could also agree on $\mathbf{x}(\xi) = \xi = 10\,\mu\mathrm{V}$.

In general, the following statement holds: A random variable is a function $\mathbf{x}(\xi)$, the independent variable ξ of which belongs to a set (*German* Menge), which is named the *domain* $\mathcal{X}$ of the function, and which is defined by the experiment. Conversely, the set of values $\mathbf{x}(\xi)$ belonging to ξ is termed the image (or range, *German* Bereich) of the function. According to the usual interpretation of the theory of functions, the symbol $\mathbf{x}(\xi)$ denotes the number (value) associated with the outcome ξ. Ignoring the existence of complex numbers for the moment, the domain of $\cos\xi$ is the set of all real numbers $\xi \in \mathbb{R}$, while its image is the set of real numbers $|\cos\xi| \leq 1$.

Random variables are preferably written with boldface[5] symbols $\mathbf{x}$, while the *values* of random variables in sums and integrals are denoted by non-bold symbols x. Sometimes these conventions are not practicable, and therefore ignored in these cases.

---

[1] "Real" understood as opposed to "complex"

[2] Fisz, M.: Wahrscheinlichkeitsrechnung und mathematische Statistik. Berlin: VEB Deutscher Verlag der Wissenschaften 1989. Chapter 2 Sect. 2.1 Page 47

[3] Papoulis, A.: Probability, random variables, and stochastic processes. 3rd Ed. New York: McGraw-Hill 1991. Chapter 4 Sect. 4.1 Page 63

[4] Jondral, F.; Wiesler, A.: Wahrscheinlichkeitsrechnung und stochastische Prozesse, 2nd Ed. Stuttgart: B. G. Teubner-Verlag 2002. Sect. 4.1 Page 38

[5] See Chapter 4 Sect. 4.1 Page 64 in Ref. 3 on Page 184

Stochastic process A stochastic process $\mathbf{x}(t,\xi)$ represents a family of random variables, where the parameter t is interpreted as a continuous quantity that mostly stands for time. For each fixed t_i the quantity $\mathbf{x} := \mathbf{x}(t_i,\xi)$ is a random variable. For a a fixed ξ_j the quantity $\mathbf{x}(t) := \mathbf{x}(t,\xi_j)$ is named a random function (*German* Zufallsfunktion, Musterfunktion, Realisierung, Pfad des Prozesses).

An example: The electrical noise voltages $\mathbf{x}(t_i,\xi_1)$, $\mathbf{x}(t_i,\xi_2)$, ..., $\mathbf{x}(t_i,\xi_j)$, ... of an ensemble of equal resistors R_1, R_2, R_j, ... at the same temperature at a fixed time t_i define the range of the random variable $\mathbf{x} := \mathbf{x}(t_i,\xi)$, while the time-dependent noise voltage $\mathbf{x} = \mathbf{x}(t,\xi_j)$ of a certain resistor R_j represents a random function $\mathbf{x}_j(t) := \mathbf{x}(t,\xi_j)$.

Discrete random variables, probability, moments

For a random variable $\mathbf{x}$ we take N measurements. Each of the N_n observations result in a value x_n ($\sum_n N_n = N$). The probability for the result x_n is the (empirical) limit of the relative frequency (*German* Häufigkeit),

$$p_x(x_n) = \lim_{N\to\infty} \frac{N_n}{N}, \qquad \sum_n p_{\mathbf{x}}(x_n) = 1\,. \tag{B.4}$$

In general, we denote as the mth moment of the discrete random variable $\mathbf{x}$ the expression

$$\overline{\mathbf{x}^m} = \mathcal{E}(\mathbf{x}^m) = \sum_n x_n^m p_x(x_n)\,. \tag{B.5}$$

The expectation $\overline{\mathbf{x}} = \mathcal{E}(\mathbf{x})$ is called a first moment, while the variance σ_x^2 is the second central moment (standard deviation σ_x; also named effective fluctuation),

$$\begin{aligned} \overline{\mathbf{x}} &= \mathcal{E}(\mathbf{x}) = \textstyle\sum_n x_n p_x(x_n), \\ \sigma_x^2 &= \overline{\left(\mathbf{x}-\overline{\mathbf{x}}\right)^2} = \mathcal{E}[(\mathbf{x}-\mathcal{E}(\mathbf{x}))^2] = \textstyle\sum_n \left(x_n-\overline{\mathbf{x}}\right)^2 p_x(x_n) = \overline{\delta\mathbf{x}^2}. \end{aligned} \tag{B.6}$$

The random quantity $\delta\mathbf{x} = \mathbf{x} - \overline{\mathbf{x}}$ is called fluctuation, because its expectation is zero, $\overline{\delta\mathbf{x}} = 0$. The variance σ_x^2 is also named mean squared fluctuation. Central mth momente are calculated following the definition $\overline{\left(\mathbf{x}-\overline{\mathbf{x}}\right)^m}$.

From N observations of the RV $\mathbf{x},\mathbf{y}$ we measure (N_{kl})-times the pair x_k, y_l, where $\sum_{k,l} N_{kl} = N$ holds. From there we find the joint probability

$$p_{xy}(x_k,y_l) = \lim_{N\to\infty} \frac{N_{kl}}{N}, \qquad \sum_k \sum_l p_{xy}(x_k,y_l) = 1\,. \tag{B.7}$$

The probabilities $p_x(x_k)$ and $p_y(y_l)$ are calculated by summing over all l and k. Besides the expectations $\overline{\mathbf{x}}, \overline{\mathbf{y}}$ and the variances σ_x^2, σ_y^2, the covariance is especially important,

$$\overline{\left(\mathbf{x}-\overline{\mathbf{x}}\right)\left(\mathbf{y}-\overline{\mathbf{y}}\right)} = \sum_k \sum_l \left(x_k-\overline{x}\right)\left(y_l-\overline{y}\right) p_{xy}(x_k,y_l) = \overline{\mathbf{xy}} - \overline{\mathbf{x}}\,\overline{\mathbf{y}}\,. \tag{B.8}$$

The covariance disappears when the random variables $\mathbf{x}$, $\mathbf{y}$ are uncorrelated. If $p_{xy}(x_k,y_l) = p_x(x_k)\,p_y(y_l)$ holds, then the RV are statistically independent and therefore by necessity uncorrelated. However, it is not possible to conclude from a zero covariance that the underlying random variables are statistically independent. The covariance coefficient

$$\rho_{xy} = \frac{\overline{\left(\mathbf{x}-\overline{\mathbf{x}}\right)\left(\mathbf{y}-\overline{\mathbf{y}}\right)}}{\sigma_x \sigma_y} \tag{B.9}$$

is zero, if $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated. If $\mathbf{x}$ and $\mathbf{y}$ are statistically dependent, $\mathbf{y} = k\mathbf{x}$, we find $\rho_{xy} \pm 1$ according to the sign of k.

Continuous random variables, probability density function, moments

The relations for discrete random variables can be transferred to continues variables, where the sums have to be replaced by integrals. The functions $w_x(x)$, $w_y(y)$ are denoted as probability density functions (PDF), and $p(x \le \mathbf{x} \le x + \mathrm{d}\,x) = w_x(x)\,\mathrm{d}x$ is the probability that we observe for the RV $\mathbf{x}$ values in the differential interval $x \le \mathbf{x} \le x + \mathrm{d}\,x$. The probability that $\mathbf{x}$ takes any value x is therefore 1,

$$p(-\infty < \mathbf{x} < +\infty) = \int_{-\infty}^{+\infty} w_x(x)\,\mathrm{d}x = 1\,. \tag{B.10}$$

In general, the mth moment of a continuous random variable $\mathbf{x}$ is expressed by

$$\overline{\mathbf{x}^m} = \mathcal{E}(\mathbf{x}^m) = \int_{-\infty}^{+\infty} x^m w_x(x)\,\mathrm{d}x\,. \tag{B.11}$$

As before, we calculate the mth central moments according to the definition $\overline{(\mathbf{x} - \overline{x})^m}$.

Further, we define the joint probability density $w_{xy}(x, y)$ for the simultaneous observation of $\mathbf{x} = x$ *and* $\mathbf{y} = y$. The conditional probability density function $w_x(x|y)$ is the probability density for an observation of $\mathbf{x} = x$ given that $\mathbf{y} = y$ was already measured,

$$w_{xy}(x, y) = w_x(x|y)w_y(y) = w_y(y|x)w_x(x), \qquad \int_{-\infty}^{+\infty} w_x(x|y)\,\mathrm{d}x = \int_{-\infty}^{+\infty} w_y(y|x)\,\mathrm{d}y = 1\,. \tag{B.12}$$

For statistically independent RV $\mathbf{x}$ and $\mathbf{y}$, the conditional probability densities are

$$\begin{array}{l} w_y(y|x) = w_y(y), \\ w_x(x|y) = w_x(x), \end{array} \quad \text{i. e.,} \quad w_{xy}(x, y) = w_x(x)w_y(y)\,, \quad \overline{xy} = \int_{-\infty}^{+\infty} w_{xy}(x, y)\,\mathrm{d}x = \overline{x}\,\overline{y}\,. \tag{B.13}$$

From Eq. (B.13), Eq. (B.8) it follows again that statistical independence implies zero correlation. We denote the inverse of the function $f(x)$ by

$$\mathbf{y} = f(\mathbf{x}), \quad \text{inverse function} \quad \mathbf{x} = f^{-1}(\mathbf{y})\,. \tag{B.14}$$

For statistically dependent RV $\mathbf{x}, \mathbf{y}$ we then find the conditional probability densities

$$w_y(y|x) = \delta\big[y - f(x)\big], \qquad w_x(x|y) = \delta\big[x - f^{-1}(y)\big]. \tag{B.15}$$

Characteristic function and moments

The characteristic function (CF) of a random variable $\mathbf{x}$ with value x is defined as the expectation

$$C_x\,(\xi) = \overline{\mathrm{e}^{-\mathrm{j}\,2\pi\xi x}} = \int_{-\infty}^{+\infty} w_x(x)\,\mathrm{e}^{-\mathrm{j}\,2\pi\xi x}\,\mathrm{d}x\,, \qquad |C_x\,(\xi)| \le C_x\,(0) = 1\,. \tag{B.16}$$

The function is maximum at $\xi = 0$ because for a PDF the relation $w_x(x) \ge 0$ holds. Both, CF and PDF, are a Fourier pair. In this context the quantity ξ does not represent an event as in Sect. B.2.1 on Page 184, but rather the Fourier variable ξ corresponding to the value x of the RV $\mathbf{x}$.

The CF is useful if the PDF $w_z(z)$ of a sum $\mathbf{z} = \mathbf{x} + \mathbf{y}$ of statistically independent RV $\mathbf{x}$ and $\mathbf{y}$ has to be calculated. Instead of computing the convolution

$$w_z(z) = \int_{-\infty}^{+\infty} w_x(x)\,w_y(z - x)\,\mathrm{d}x = \big(w_x(x) * w_y(x)\big)\,(z)\,, \tag{B.17}$$

the CF $C_z\,(\zeta)$ can be calculated simply by multiplying the CF $C_x\,(\zeta)$ and $C_y\,(\zeta)$, and by performing a Fourier back-transform,

$$C_z(\zeta) = C_x(\zeta)\,C_y(\zeta), \qquad w_z(z) = \int_{-\infty}^{+\infty} C_z(\zeta)\,\mathrm{e}^{+\mathrm{j}\,2\pi\zeta x}\,\mathrm{d}z\,. \tag{B.18}$$

Taking the mth derivative of the CF $C_x\,(\xi)$ at $\xi = 0$ results in the mth moment of the RV $\mathbf{x}$,

$$\overline{\mathbf{x}^m} = \frac{1}{(-\mathrm{j}\,2\pi)^m}\,\frac{\mathrm{d}^m C_x\,(\xi)}{\mathrm{d}\xi^m}\bigg|_{\xi=0} = \frac{1}{(-\mathrm{j}\,2\pi a)^m}\,\frac{\mathrm{d}^m C_x\,(a\xi)}{\mathrm{d}\xi^m}\bigg|_{\xi=0} = \int_{-\infty}^{+\infty} x^m w_x(x)\,\mathrm{d}x \tag{B.19}$$

B.2.2 Transformation of random variables

Consider the probability density function $w_x(x)$ of the random variable $\mathbf{x}$. We define another random variable $\mathbf{y}$ by the memoryless strictly monotonic function $\mathbf{y} = f(\mathbf{x})$, and we look for the transformed probability density function $w_y(y)$. The transforming function f is assumed to be continuously differentiable having the derivative $f'(x) = \mathrm{d}f/\,\mathrm{d}x$, see Fig. B.1(a). Obviously, the probability p that the RV $\mathbf{y}$ has an outcome y in an interval $y \ldots y + \mathrm{d}\,y$ $(\mathrm{d}\,y > 0)$ is

$$\begin{aligned} w_y(y)\ \mathrm{d}\,y = p(y \le \mathbf{y} \le y + \mathrm{d}\,y) &= p(x_1 \le \mathbf{x} \le x_1 + \mathrm{d}\,x_1) + \\ &+ p(x_2 + \mathrm{d}\,x_2 \le \mathbf{x} \le x_2) + p(x_{31} \le \mathbf{x} \le x_3 + \mathrm{d}\,x_3) \end{aligned} \tag{B.20}$$

Inside the strictly monotonic partial intervals of a non-monotonic function $f(x)$ as displayed in Fig. B.1(a) we find the probabilities

$$\begin{aligned} p(x_1 \le \mathbf{x} \le x_1 + \mathrm{d}\,x_1) &= w_x(x_1)\ \mathrm{d}\,x_1\,, & \mathrm{d}\,x_1 &= \mathrm{d}\,y/f'(x_1)\,, \\ p(x_2 + \mathrm{d}\,x_2 \le \mathbf{x} \le x_2) &= w_x(x_2)\,|\,\mathrm{d}\,x_2|\,, & \mathrm{d}\,x_2 &= \mathrm{d}\,y/f'(x_2)\,, \\ p(x_3 \le \mathbf{x} \le x_3 + \mathrm{d}\,x_3) &= w_x(x_3)\ \mathrm{d}\,x_3\,, & \mathrm{d}\,x_3 &= \mathrm{d}\,y/f'(x_3)\,, \end{aligned} \tag{B.21}$$

leading to the result

$$w_y(y)\ \mathrm{d}\,y = \left(\frac{w_x(x_1)}{f'(x_1)} + \frac{w_x(x_2)}{|f'(x_2)|} + \frac{w_x(x_3)}{f'(x_3)} \right) \mathrm{d}\,y\,. \tag{B.22}$$

In general we find: If $x_n = x_n(y)$ $(n = 1, 2, \ldots)$ are the real roots of the equation $y = f(x)$, then the probability density function of the transformed random variable $\mathbf{y}$ is

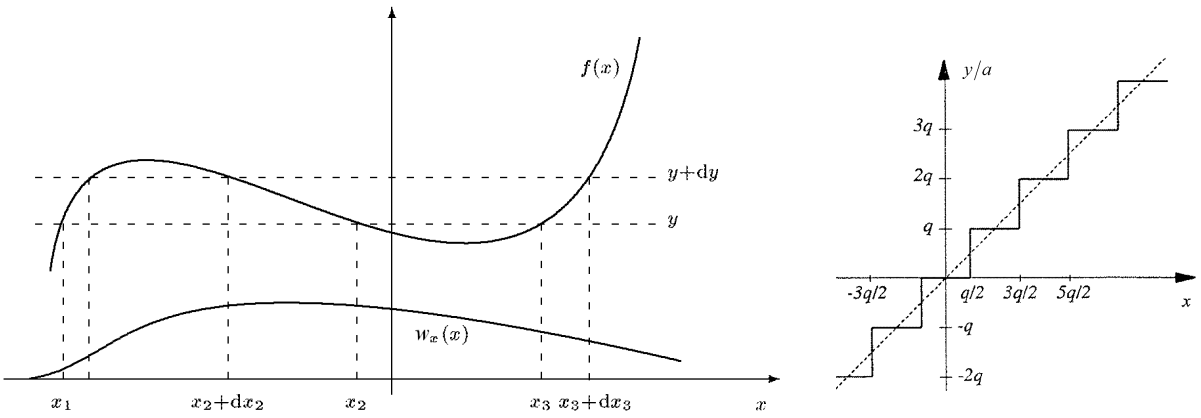
$$w_y(y) = \sum_n \frac{w_x(x_n)}{|f'(x_n)|} \qquad \text{for} \quad f(x_n) = y \quad \text{and} \quad f'(x) = \frac{\mathrm{d}f}{\mathrm{d}x}\,. \tag{B.23}$$

As an example, we discuss the transformation of random variables for two nonlinear functions, which model two types of rectifiers (detectors).

Linear envelope detector

The memory-less linear envelope detector is defined by the straight-line characteristic ($H(x)$ is the Heaviside function, see Table 1.3 on Page 9)

$$y = f(x) = x\,H(x) = \begin{cases} x & \text{for} \quad x \ge 0\,, \\ 0 & \text{for} \quad x < 0\,, \end{cases} \qquad f'(x) = x\delta(x) + H(x) = H(x)\,. \tag{B.24}$$



(a) Strictly monotonic transformation function $f(x)$ (b) Quantizer characteristic

Fig. B.1. Transformation of probability density function $w_x(x)$ for a random variable $\mathbf{x}$ by a memory-less function $y = f(x)$

All $x \le 0$ lead to $y = 0$. All $x \ge 0$ are mapped to values $y = x$. For $y < 0$, both the PDF $w_y(y < 0) = 0$ and the probability distribution $w_y(y < 0) = \int_{-\infty}^{y} w_y(y_1)\,\mathrm{d}\,y_1 = 0$ are zero. For $y > 0$ the function $y = x\,H(x)$ has only one solution $x_3 = y$, therefore:

$$\begin{aligned} w_y(y) &= 0 \qquad & p_y(y) &= 0 \qquad & \text{for } y < 0 \\ w_y(y) &= w_x(y) \qquad & p_y(y) &= p_x(y) \qquad & \text{for } y > 0 \end{aligned} \tag{B.25}$$

Obvioulsy, the probability distribution $p_y(y)$ is discontinuous at $y = 0$. The total probability must be normalized to 1, $p_y(y < \infty) = \int_{-\infty}^{+\infty} w_y(y_1)\,\mathrm{d}\,y_1 = \int_{0}^{\infty} w_y(y_1)\,\mathrm{d}\,y_1 = 1$. Therefore, and because of the continuity of $w_x(y)$ at $y = 0$ and the discontinuity of $w_y(y)$ at $y = 0$ we find:

$$1 = \int_0^\infty w_y(y)\,\mathrm{d}\,y = \overbrace{\int_{-\infty}^{0^-} w_y(y)\;\mathrm{d}\,y}^{y<0} + \overbrace{\int_{0^+}^{+\infty} w_y(y)\;\mathrm{d}\,y}^{y>0} + \overbrace{\int_{0^-}^{0^+} w_y(y)\;\mathrm{d}\,y}^{y=0}$$

$$1 = \int_0^\infty w_y(y)\,\mathrm{d}\,y = \underbrace{\int_{-\infty}^{0^-} 0\cdot\mathrm{d}\,y}_{0} + \underbrace{\int_{0^+}^{+\infty} w_x(y)\,\mathrm{d}\,y}_{1-p_x(0)} + \underbrace{\int_{0^-}^{0^+} w_y(y)\,\mathrm{d}\,y}_{p_y(0^+)-p_y(0^-)\,\stackrel{!}{=}\,w_x(0)}$$

$$= \int_0^\infty w_x(y)\;\mathrm{d}\,y + p_x(0) = \int_0^\infty w_x(y)\;\mathrm{d}\,y + 2\int_0^{+\infty} \delta(y)\,p_x(0)\;\mathrm{d}\,y$$

From equal integrals we can conclude that the integrands are also identical if the transformation is unique. Therefore the PDF is a symbolic function:

$$w_y(y) = w_x(y) + 2\delta(y)\,p_x(0) = w_x(y) + 2\delta(y)\int_{-\infty}^{0} w_x(x)\,\mathrm{d}\,x\,, \quad y \ge 0 \tag{B.26}$$

If specifically a Gaussian PDF is chosen for the RV $\mathbf{x}$,

$$w_x(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(-\frac{x^2}{2\sigma_x^2}\right), \tag{B.27}$$

the transformed probability density function and the expectation for $\mathbf{x}$ are

$$\begin{aligned} w_y(y) &= \delta(y) + \frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(-\frac{y^2}{2\sigma_x^2}\right), \\ \overline{\mathbf{y}} &= \int_0^\infty y\cdot\left[\delta(y) + \frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(-\frac{y^2}{2\sigma_x^2}\right)\right]\mathrm{d}\,y = \frac{\sigma_x}{\sqrt{2\pi}}, \quad y \ge 0. \end{aligned} \tag{B.28}$$

Quadratic rectifier

The quadratic rectifier is defined by its parabolic characteristic $y = x^2$. With Eq. (B.23) we calculate

$$\begin{aligned} w_x(-x) + w_x(x) = w_y(y)\cdot 2x, \qquad f'(x) = 2x, \qquad x_{1,2} = \pm\sqrt{y}\,, \\ w_y(y) = \frac{w_x\left(+\sqrt{y}\right) + w_x\left(-\sqrt{y}\right)}{2\sqrt{y}}\,. \end{aligned} \tag{B.29}$$

If again a Gaussian PDF is with expectation $\overline{\mathbf{x}} = 0$ is chosen, Eq. (B.27), we end up with

$$w_y(y) = \frac{1}{\sqrt{2\pi y\sigma_x^2}}\exp\left[-\frac{y}{2\sigma_x^2}\right], \qquad y \ge 0\,. \tag{B.30}$$

Analogue-to-digital converter

An analogue-to-digital converter (ADC) quantizes an analogue input signal x and transforms it into a quantized output[6] according to the memory-less function $y = af(x)$, see Fig. B.1(b) on Page 187. Because of the non-monotonic nature of the quantizer characteristic, Eq. (B.23) on Page 187 cannot be applied directly, and the calulation follows the strategy developed for the linear envelope detector in Sect. B.2.2 on Page 187 ff.

Probability density function and moments In the following, we refer frequently to the relations in Table 1.3 on Page 9 without further mentioning. First we note that only discrete output values y are admitted. They are related to the input values x by a gain factor a and by the quantization interval $q > 0$, which describes the range of x-values leading to a certain y,

$$y = a \begin{cases} kq & \text{for} \quad kq - q/2 < x < kq + q/2 \\ 0 & \text{else} \end{cases}, \qquad k = 0, \pm 1, \pm 2, \ldots \tag{B.31}$$

The probability of finding the RV $\mathbf{y}$ in an infinitesimally small intervall $\pm\varepsilon$ around $y_k = akq$ is p_{y_k} and equals the probability p_x that $\mathbf{x}$ takes on values in the finite intervall $kq - q/2 < \mathbf{x} < kq + q/2$,

$$\begin{aligned} p_{y_k}\left(akq - \varepsilon < \mathbf{y} < akq + \varepsilon\right) &= \int_{akq-\varepsilon}^{akq+\varepsilon} w_{y_k}\left(y_1\right) \mathrm{d}y_1 \\ &= p_x\left(kq - q/2 < \mathbf{x} < kq + q/2\right) = \int_{kq-q/2}^{kq+q/2} w_x\left(x_1\right) \mathrm{d}x_1 . \end{aligned} \tag{B.32}$$

From the definition of the Dirac function in Table 1.3 on Page 9, $\int_{-\infty}^{+\infty} \delta\left(t\right) \Psi\left(t\right) \mathrm{d}t = \int_{-\varepsilon}^{+\varepsilon} \delta\left(t\right) \Psi\left(t\right) \mathrm{d}t = \Psi\left(0\right)$, it can be seen that the probability density function $w_{y_k}(y)$ for output level y_k can be written as

$$w_{y_k}(y) = \delta\left(y - akq\right)\, p_x\left(y/a - q/2 < \mathbf{x} < y/a + q/2\right) = \delta\left(y - akq\right) \int_{y/a-q/2}^{y/a+q/2} w_x\left(x_1\right) \mathrm{d}x_1 . \tag{B.33}$$

The probability of finding any output signal value y is $p_y\left(-\infty < \mathbf{y} < +\infty\right) = \sum_{k=-\infty}^{+\infty} p_{y_k}$, and therefore

$$\begin{aligned} w_y(y) &= p_x\left(y/a - q/2 < \mathbf{x} < y/a + q/2\right) \sum_{k=-\infty}^{+\infty} \delta\left(y - akq\right) \\ &= \int_{y/a-q/2}^{y/a+q/2} w_x\left(x_1\right) \mathrm{d}x_1 \sum_{k=-\infty}^{+\infty} \delta\left(y - akq\right) . \end{aligned} \tag{B.34}$$

It can be easily verified that $p_y\left(-\infty < \mathbf{y} < +\infty\right) = 1$ holds true,

$$\begin{aligned} p_y\left(-\infty < \mathbf{y} < +\infty\right) &= \int_{-\infty}^{+\infty} w_y(y_1) \mathrm{d}y_1 = \int_{-\infty}^{+\infty} \mathrm{d}y_1 \int_{y_1/a-q/2}^{y_1/a+q/2} w_x\left(x_1\right) \mathrm{d}x_1 \sum_{k=-\infty}^{+\infty} \delta\left(y_1 - akq\right) \\ &= \sum_{k=-\infty}^{+\infty} \int_{akq/a-q/2}^{akq/a+q/2} w_x\left(x_1\right) \mathrm{d}x_1 = \int_{-\infty}^{+\infty} w_x\left(x_1\right) \mathrm{d}x_1 = 1 . \end{aligned} \tag{B.35}$$

We now calulate the moments $\overline{y}$ and $\overline{y^2}$ at the ADC output in terms of scaled moments $a\,\overline{x}$ and $a^2\,\overline{x^2}$ at the ADC input. For the expectation we find

$$\begin{aligned} \overline{\mathbf{y}} &= \int_{-\infty}^{+\infty} y_1\, w_y(y_1) \mathrm{d}y_1 = \sum_{k=-\infty}^{+\infty} \int_{-\infty}^{+\infty} y_1\, \delta(y_1 - akq)\, p_x\left(y_1/a - q/2 < \mathbf{x} < y_1/a + q/2\right) \mathrm{d}y_1 \\ &= a \sum_{k=-\infty}^{+\infty} kq\, p_x\left(kq - q/2 < \mathbf{x} < kq + q/2\right) = a\,\overline{kq} \underset{q\to 0}{=} a\,\overline{x}. \end{aligned} \tag{B.36a}$$

[6] Widrow, B.; Kollár, I.; Liu, M.-C.: Statistical theory of quantization. IEEE Trans. Instrum. Meas. 45 (1996) 353–361

The second moment reads

$$\overline{\mathbf{y}^2} = \int_{-\infty}^{+\infty} y_1^2\, w_y(y_1)\,\mathrm{d}y_1 = \sum_{k=-\infty}^{+\infty} \int_{-\infty}^{+\infty} y_1^2\, \delta(y_1 - akq)\, p_x\left(y_1/a - q/2 < \mathbf{x} < y_1/a + q/2\right) \mathrm{d}y_1$$

$$= a^2 \sum_{k=-\infty}^{+\infty} (kq)^2\, p_x\left(kq - q/2 < \mathbf{x} < kq + q/2\right) = a^2\, \overline{(k\mathbf{q})^2} \underset{q\to 0}{=} a^2\, \overline{x^2}. \tag{B.36b}$$

Only for an infinitely small quantization $q \to 0$ such that $kq \to x$, the moments $\overline{\mathbf{y}}$ and $\overline{\mathbf{y}^2}$ at the ADC output reproduce the scaled moments $a\,\overline{\mathbf{x}}$ and $a^2\,\overline{\mathbf{x}^2}$ at the ADC input. Otherwise, an error will occur.

As an example, consider the continuous RV $\mathbf{x}$ that is equally distributed in the intervall $-\left(K+\frac{1}{2}\right)q \le \mathbf{x} < +\left(K+\frac{1}{2}\right)q$ for $K = 0, 1, 2, \ldots$ The first moments are $\overline{\mathbf{x}}, \overline{\mathbf{y}} = 0$, and the second moment $\overline{\mathbf{x}^2}$ is

$$w_x(x) = \frac{1}{2\left(K+\frac{1}{2}\right)q}\,\mathrm{rect}\Big(\frac{x}{2\left(K+\frac{1}{2}\right)q}\Big), \quad \overline{\mathbf{x}^2} = \int_{-(K+\frac{1}{2})q}^{+(K+\frac{1}{2})q} \frac{x^2\,\mathrm{d}x}{2\left(K+\frac{1}{2}\right)q} = \frac{K^2q^2}{3} + \frac{Kq^2}{3} + \frac{q^2}{12}\,. \tag{B.37a}$$

Assume that the quantizer has $2K+1$ levels so that it spans the full range of the input PDF $w_x(x)$. The resulting second moment[7] of the quantized $\mathbf{y}$ is *smaller* than the scaled second moment of the input $\mathbf{x}$,

$$\overline{\mathbf{y}^2} = 2a^2q^2 \sum_{k=1}^{K} k^2 \frac{q}{2\left(K+\frac{1}{2}\right)q} + a^2q^2 \sum_{k=0}^{0} k^2 \frac{q}{2\left(K+\frac{1}{2}\right)q} = a^2\Big(\frac{K^2q^2}{3} + \frac{Kq^2}{3}\Big) = a^2\Big(\overline{\mathbf{x}^2} - \frac{q^2}{12}\Big). \tag{B.37b}$$

From the quantizer characteristic Fig. B.1(b) on Page 187 we see that for $K = 0$ the output RV $\mathbf{y}$ (and all its moments) always assumes the value $y = 0$. Therefore, and for this specific PDF $w_x(x)$, the down-scaled output power $\overline{\mathbf{y}^2}/a^2$ is always smaller than the input power $\overline{\mathbf{x}^2}$. Further it is obvious that for $Kq = \text{const}$ and $q \to 0$, i. e., for very small and very many quantizing bins $K \to \infty$, the down-scaled second output moment equals the second input moment, $\overline{\mathbf{y}^2}/a^2 = \overline{\mathbf{x}^2}$.

While for a given input PDF the correct second moments $\overline{\mathbf{y}^2}$ and $\overline{\mathbf{x}^2}$ can be calculated from Eq. (B.36), it is useful to derive a more general statement by applying a constraint, which is at least approximately valid in real-world problems. This will be done in the following.

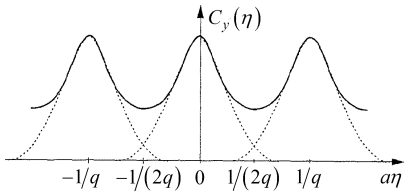
Band-limited characteristic functions and moments Equation (B.34) on Page 189 resembles[8,9] the temporal sampling as described in Eq. (2.1) on Page 13, and it is especially close to Eq. (B.2) on Page 183: The probability density function $w_x(x)$ of the input signal is sampled at equidistant values $x = y/a = qk$. Each sampled δ-shaped partial function $w_{y_k}(y)$ corresponds to the probability $p_x\left(qk - q/2 < \mathbf{x} < qk + q/2\right)$ of finding the RV $\mathbf{x}$ in an interval $qk - q/2 < \mathbf{x} < qk + q/2$.

To see this more clearly we calculate the characteristic function of $w_y(y)$ according to the definition Eq. (B.16) on Page 186. As before, we refer frequently to the relations in Table 1.3 on Page 9 without mentioning. Further, the δ-comb transformations of Eq. (2.1) on Page 13 are used. First we write Eq. (B.34) with the help of the symmetric rect-function $\mathrm{rect}\,(x/q) = \mathrm{rect}\,(-x/q)$ in form of a convolution, and then

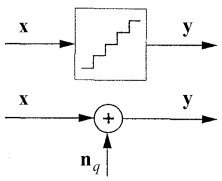
[7] Gradstein, I.; Ryshik, I.: Summen-, Produkt- und Integral-Tafeln, 5th Russian Ed., 1st German-English Ed. Volume 1 and 2. Thun: Harri Deutsch 1981. Formula 0.121.2.: $\sum_{k=1}^{n} k^2 = n\,(n+1)\,(2n+1)/6$

[8] See Ref. 6 on Page 189

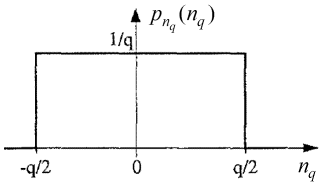
[9] Kiencke, U.; Eger, R.: Messtechnik. Systemtheorie für Elektrotechniker. 6th Ed. Berlin: Springer 2005. Sect. 7.2.3 Page 263



(a) Characteristic ADC output function $C_y(\eta)$ (——) with weighed CF $C_x(\eta)$ (. . . . .) of the continuous input signal $\mathbf{x}$



(b) ADC representation and linear model with random noise $\mathbf{n}_q$



(c) Probability density function $w_{nq}(n_q)$ of input quantization noise $\mathbf{n}_q$

Fig. B.2. Analogue-to-digital converter (ADC). Characteristic function $C_y(\eta)$ of the output $\mathbf{y}$ representing the quantized continuous input signal $\mathbf{x}$, linear ADC model with random noise $\mathbf{n}_q$, and probability density function of $\mathbf{n}_q$, equally distributed in $-q/2 \leq \mathbf{n}_q < +q/2$. — If the quantizing theorem QT I of Eq. (B.39) is fulfilled, i. e., if the continuous input signal $\mathbf{x}$ is "band-limited" such that $C_x\big(|\eta| > 1/(2aq)\big) = 0$ holds, adjacent "partial spectra" in (a) do not overlap, and the PDF $w_x(x)$ of the continuous input signal can be reconstructed error-free by interpolation (or by "filtering" in the "spectral" domain). — The real factor a describes the average slope of the analogue-to-digital converter, see Fig. B.1(b) on Page 187. [Modified after Ref. 9 on Page 190, Fig. 7.16, 7.12, and 7.17]

perform the Fourier transform $C_y(\eta) = \mathcal{F}\{w_y(y)\}$,

$$w_y(y) = \int_{y/a-q/2}^{y/a+q/2} w_x(x_1)\,\mathrm{d}x_1 \sum_{k=-\infty}^{+\infty} \delta(y - akq) = \int_{-\infty}^{+\infty} w_x(x_1)\,\mathrm{rect}\Big(\frac{y/a - x_1}{q}\Big)\mathrm{d}x_1 \sum_{k=-\infty}^{+\infty} \delta(y - akq)$$

$$= \Big(w_x(x) * \mathrm{rect}\Big(\frac{x}{q}\Big)\Big)\Big(\frac{y}{a}\Big) \sum_{k=-\infty}^{+\infty} \delta(y - akq)\,, \tag{B.38a}$$

$$C_y(\eta) = a\Big(C_x(a\eta)\; q\,\mathrm{sinc}(aq\eta)\Big) * \frac{1}{aq} \sum_{k=-\infty}^{+\infty} \delta\Big(\eta - \frac{k}{aq}\Big) = \sum_{k=-\infty}^{+\infty} C_x\Big(a\eta - \frac{k}{q}\Big)\;\mathrm{sinc}\left(q\Big(a\eta - \frac{k}{q}\Big)\right). \tag{B.38b}$$

The CF $C_y(\eta)$ of the quantized ADC output signal $\mathbf{y}$ in Fig. B.2(a) (solid line, ——) contains the CF $C_x(a\eta)$ of the continuous ADC input signal $\mathbf{x}$ (dotted line, ...), weighed with the scaled Fourier transform $\mathrm{sinc}(aq\eta)$ of the input bin size q. Because of the periodic "sampling" at positions $x = kq$, the weighed input CF $C_x(a\eta)$ is periodically repeated at integer multiples $a\eta = 0, \pm\frac{1}{q}, \pm\frac{2}{q}, \ldots$ of the scaled "sampling frequency" $a\eta_s = \frac{1}{q}$.

At these "frequencies", one central sinc-function is maximum with a value of 1, while all adjacent sinc-functions have the value 0, leading to $C_y(0) = C_x(0) = 1$. However, the *slope* of $C_y(\eta)$ at $\eta = 0$, which is responsible for the moments of $\mathbf{y}$, *does* depend on neigbouring repetitions of the weighed input CF. In this sense the schematic of the quantizer's output signal CF $C_y(\eta)$ in Fig. B.2(a) is equivalent to Fig. 2.2(b) on Page 14.

The CF $C_x(\eta)$ and the PDF $p_x(x)$ of the continuous input signal can be uniquely recovered, if $C_x(\eta)$ is "band-limited" such that the quantizing theorem QT I holds[10],

$$C_x\big(|\eta| > 1/(2aq)\big) = 0\,. \tag{B.39}$$

Under this condition, no overlap ("aliasing") of adjacent periodically repeated input CF occurs. If Eq. (B.39) is not fulfilled, because the ADC has too large a bin size q compared to the span of the input signal $\mathbf{x}$, an added auxiliary random signal with limited "bandwidth" acting as a "dither" spreads the input over more quantization levels and helps in fulfilling Eq. (B.39) for the composite signal[11]. However, it has to be noted that real-world characteristic functions cannot be band-limited, because the span of the input signal $\mathbf{x}$ as well as its PDF are limited (for an extreme example see Eq. (B.37a) on Page 190), and the Fourier transform of a truncated PDF is always unlimited in η.

[10] See Ref. 6 on Page 189, Eq. (4)
[11] See Ref. 6 on Page 189, end of Sect. II

If only moments as in Eq. (B.19) on Page 186 are of interest, it suffices to know the derivatives of the CF at $\eta = 0$ only, and therefore the weaker quantizing theorem QT II has to be fulfilled[12] for avoiding "moment aliasing", as can be immediately verified by inspecting Fig. B.2(a),

$$C_x\big(|\eta| > 1/(aq)\big) = 0\,. \tag{B.40}$$

None of the random variables which occur in practice have a perfectly band-limited CF. However, most of them are approximately band-limited, and a fine enough quantization q (large enough "sampling frequency" $\eta_s = \frac{1}{aq}$) assures acceptable fulfillment of QT I or II as formulated in Eq. (B.39) and (B.40), by allowing the input CF to be wide[13].

B.2.3 Quantization noise

The quantized ADC output $\mathbf{y}$ cannot reproduce the input signal $\mathbf{x}$ perfectly, and there remains a deterministic quantization error which can, under the conditon of QT II, be interpreted as uniformly distributed quantization noise $\mathbf{n}_q$, statistically independent from $\mathbf{x}$. An example of this deterministic error signal $\mathbf{n}_q$ is displayed as a blue curve in Fig. 2.3 (upper row, central subfigure) on Page 16.

We assume that QT II holds and start by computing the expectation $\overline{\mathbf{y}}$. We evaluate Eq. (B.38b) on Page 191 for $k = 0$ with the help of Eq. (B.19) on Page 186,

$$\begin{aligned}\overline{\mathbf{y}} &= \frac{1}{-\mathrm{j}\,2\pi}\,\frac{\mathrm{d}\,C_y\,(\eta, k=0)}{\mathrm{d}\eta}\bigg|_{\eta=0} = \left[\frac{\mathrm{d}\,C_x\,(a\eta)}{\mathrm{d}\eta}\,\operatorname{sinc}\,(aq\eta) + C_x\,(\eta)\,\frac{\pi aq\eta\cos\,(\pi aq\eta) - \sin\,(\pi aq\eta)}{\pi aq\eta^2}\right]_{\eta=0} \\ &= \frac{1}{-\mathrm{j}\,2\pi}\,\frac{\mathrm{d}\,C_x\,(a\eta)}{\mathrm{d}\eta}\bigg|_{\eta=0} = a\,\overline{\mathbf{x}}\,.\end{aligned} \tag{B.41a}$$

The quantized output $\overline{\mathbf{y}} = a\,\overline{\mathbf{x}}$ is scaled but statistically unbiased[14]. Proceeding in an analogue fashion, we calculate the second moment $\overline{\mathbf{y}^2}$,

$$\begin{aligned}\overline{\mathbf{y}^2} &= \frac{1}{(-\mathrm{j}\,2\pi)^2}\,\frac{\mathrm{d}^2 C_y\,(\eta, k=0)}{\mathrm{d}\eta^2}\bigg|_{\eta=0} \\ &= \frac{1}{(-\mathrm{j}\,2\pi)^2}\left[\frac{\mathrm{d}^2 C_x\,(a\eta)}{\mathrm{d}\eta^2}\,\operatorname{sinc}\,(aq\eta) + 2\frac{\mathrm{d}\,C_x\,(a\eta)}{\mathrm{d}\eta}\,\frac{\pi aq\eta\cos\,(\pi aq\eta) - \sin\,(\pi aq\eta)}{\pi aq\eta^2}\right. \\ &\qquad\left. + C_x\,(a\eta)\,\frac{[2 - (\pi aq\eta)^2]\sin\,(\pi aq\eta) - 2\pi aq\eta\cos\,(\pi aq\eta)}{\pi aq\eta^3}\right]_{\eta=0} \\ &= \frac{1}{(-\mathrm{j}\,2\pi)^2}\,\frac{\mathrm{d}^2 C_x\,(a\eta)}{\mathrm{d}\eta^2}\bigg|_{\eta=0} + \frac{(aq)^2}{12} = a^2\,\overline{\mathbf{x}^2} + a^2\,\frac{q^2}{12} = a^2\big(\overline{\mathbf{x}^2} + \overline{\mathbf{n}_q^2}\big), \quad \overline{\mathbf{n}_q^2} = \frac{q^2}{12}\,.\end{aligned} \tag{B.41b}$$

The result Eq. (B.41b) contradicts the second moment as calculated for a rectangular input PDF in Eq. (B.37) on Page 190, because in this case the condition of QT II was severely violated.

Linear quantizer model

Provided that the quantizing theorem QT II Eq. (B.40) on Page B.40 holds, the moment $\overline{\mathbf{y}^2} = a^2\big(\overline{\mathbf{x}^2} + \overline{\mathbf{n}_q^2}\big)$ at the output of the physical ADC as depicted in the upper half of Fig. B.2(b) on Page 191 can be interpreted as resulting from a linear superposition of two statistically independent RV $\mathbf{x}$ (input signal) and $\mathbf{n}_q$ (equivalent quantization noise at input with $\overline{\mathbf{n}_q} = 0$), lower half of Fig. B.2(b) and Fig. 2.3 on Page 16.

[12] See Ref. 6 on Page 189, Eq. (8)
[13] This sentence quoted after Ref. 6 on Page 189, Sect. VI, after Eq. (9)
[14] statistically unbiased: *German* erwartungstreu

Assuming further a rectangular PDF for $w_{n_q}(n_q)$ as in Fig. B.2(c) and observing $2\,\overline{\mathbf{x}\,\mathbf{n}_q} = 2\,\overline{\mathbf{x}}\,\overline{\mathbf{n}_q} = 0$ for statistically independent RV (Eq. (B.13) on Page 186), we find

$$\mathbf{y} = a\,(\mathbf{x}+\mathbf{n}_q)\;, \quad \overline{\mathbf{y}^2} = a^2\big(\overline{\mathbf{x}^2}+\overline{\mathbf{n}_q^2}\big), \quad w_{n_q}(n_q) = \frac{1}{q}\,\mathrm{rect}\Big(\frac{n_q}{q}\Big), \quad \overline{\mathbf{n}_q^2} = \frac{1}{q}\int_{-q/2}^{+q/2} n_q^2\,\mathrm{d}n_q = \frac{q^2}{12}\,. \tag{B.42a}$$

Signal and noise variances at input and output of the ADC are

$$\sigma_x^2 = \overline{\big(\mathbf{x}-\overline{\mathbf{x}}\big)^2} = \overline{\mathbf{x}^2} - \overline{\mathbf{x}}^2, \quad \sigma_y^2 = a^2\overline{\big(\mathbf{x}+\mathbf{n}_q-\overline{\mathbf{x}}\big)^2} = a^2\sigma_x^2 + a^2\sigma_{n_q}^2\,, \quad \sigma_{n_q}^2 = \frac{q^2}{12}\,. \tag{B.42b}$$

Equations (B.41) establish for the ADC a linear signal and noise model as shown in the lower half of Fig. B.2(b). We are now in a position to calculate the signal-to-noise power ratio (SNR_q) and the effective number of bits (ENOB).

Signal-to-noise power ratio

The signal-to-noise power ratio SNR_q of an ADC is defined as the ratio of signal power P_S and noise power P_N. For definiteness, we assume an input sinusoidal $\mathbf{x} = \hat{x}\cos(\omega_0 t + \varphi)$, the full range $2\hat{x}$ of which is quantized with an ADC having $M = 2^r \gg 1$ levels corresponding to an ADC with r bit, therefore $2\hat{x} = Mq$. The signal power is $P_S = \sigma_x^2 = \frac{1}{2}\hat{x}^2$, and the noise power amounts to $P_R = \sigma_{n_q}^2 = q^2/12$ from Eq. (B.41). For SNR_q and its logarithmic equivalent $\mathrm{SNR}_{q,\mathrm{dB}} = 10\lg\mathrm{SNR}_b$ we then find

$$\mathrm{SNR}_q^{(\sin)} = \frac{P_S}{P_R} = \frac{\sigma_x^2}{\sigma_{n_q}^2} = \frac{\frac{1}{2}\,\frac{1}{4}\,2^{2r}q^2}{\frac{1}{12}\,q^2} = \frac{3}{2}\,2^{2r} \quad \text{for } M = 2^r \gg 1, \qquad \mathrm{SNR}_{q,\mathrm{dB}}^{(\sin)} = 6.02\,r + 1.76\,. \tag{B.43}$$

Effective number of bits

Given the signal power P_S of a sinusoidal and the noise power P_R corresponding to the quantizing error, the appropriate number of bits r can be extracted from $\mathrm{SNR}_{q,\mathrm{dB}}^{(\sin)}$ in Eq. (B.43),

$$r = \frac{\mathrm{SNR}_{q,\mathrm{dB}}^{(\sin)}}{6.02} - 0.293\,, \qquad r_e = \mathrm{ENOB} = \frac{\mathrm{SNDR}_{q,\mathrm{dB}}}{6.02} - 0.293\,. \tag{B.44}$$

For arbitrary sources of noise, nonlinear distortion and timing jitter, characterized by a general logarithmic signal-to-noise and distortion power ratio $\mathrm{SNDR}_{q,\mathrm{dB}} = 10\lg(P_S/P_R) = 20\lg(\sigma_x/\sigma_n)$, this relation can be generalized[15] to define the effective number of bits $r_e = \mathrm{ENOB}$. The factor of $6.02 = 20\lg 2$ reflects the notion that to improve ENOB by 1 bit, either the full-scale swing of the input signal $\mathbf{x}$ must be increased by a factor of 2, or the effective distortion and noise σ_n must be reduced by a factor of 2.

Usually, the ENOB decreases for wide-bandwidth signals, e. g., for an ADC with 40 GSa/s and 6 bit we find a reduction[16] from ENOB = 5.5 for a full-scale sinusoidal at $f_0 = 0.5$ GHz to ENOB = 4 for a sinusoidal at $f_0 = 18$ GHz.

B.3 The discrete Fourier transform

The discrete Fourier transform can be axiomatically defined, but it also approximates the continuous Fourier transform

$$\Psi(t) = \int_{-\infty}^{+\infty} \breve{\Psi}(f)\,\mathrm{e}^{+\mathrm{j}\,2\pi f t}\,\mathrm{d}f\,, \qquad \breve{\Psi}(f) = \int_{-\infty}^{+\infty} \Psi(t)\,\mathrm{e}^{-\mathrm{j}\,2\pi f t}\,\mathrm{d}t\,. \tag{B.45}$$

[15] See Ref. 6 on Page 15, Eq. (3)

[16] Laperle, C.; O'Sullivan, M.: High-speed DACs and ADCs for next generation flexible transceivers. Proc. Advanced Photonics for Communications (APC'14), San Diego (CA), USA, July 13–17, 2014. Paper SM3E.1

To see this we must discretize time t and frequency f, and we have to impose a limitation on the maximum time and frequency values which are admitted. According to Eq. (2.1) on Page 13 we assume a sampling interval T_s and a maximum band-limiting frequency F_s, beyond which the spectrum $\breve{\Psi}(f)$ is zero. It is convenient to choose the number of samples N to be a power of two,

$$t = nT_s\,, \qquad f = \nu\frac{F_s}{N}\,, \qquad T_s = \frac{1}{F_s}\,, \qquad n,\nu = 0,1,2,\ldots,N-1\,, \quad N = 2^r, \quad r = 2,3,4,\ldots \tag{B.46a}$$

Combining Eq. (B.46a) and (B.45), we find the approximations

$$\Psi(nT_s) \approx \frac{F_s}{N}\sum_{\nu=-N/2}^{N/2-1}\breve{\Psi}(\nu F_s/N)\,\mathrm{e}^{+\mathrm{j}\,2\pi\frac{\nu n}{N}}, \qquad \breve{\Psi}(\nu F_s/N) \approx T_s\sum_{n=-N/2}^{N/2-1}\Psi(nT_s)\,\mathrm{e}^{-\mathrm{j}\,2\pi\frac{\nu n}{N}}, \tag{B.46b}$$

which become the better, the larger N and the smaller T_s are. The axiomatic definition of the discrete N-point Fourier transform $\breve{c}_\nu$ (DFT) and its inverse c_n (IDFT),

$$c_n = \sum_{\nu=-N/2}^{N/2-1}\breve{c}_\nu\,\mathrm{e}^{+\mathrm{j}\,2\pi\frac{n\nu}{N}} \approx \Psi(nT_s)\,, \qquad\qquad \breve{c}_\nu = \frac{1}{N}\sum_{n=-N/2}^{N/2-1}c_n\,\mathrm{e}^{-\mathrm{j}\,2\pi\frac{n\nu}{N}} \approx \frac{F_s}{N}\,\breve{\Psi}(\nu F_s/N)\,, \tag{B.46c}$$

relates the approximations Eq. (B.46b) to the IDFT and DFT Eq. (B.46c). In numerical routines, a different DFT definition is frequently used (coefficients C_n, $\breve{C}_\nu$), which can be translated to Eq. (B.46c) with the shift theorem of Fourier theory,

$$C_n = \sum_{\nu=0}^{N-1}\breve{C}_\nu\,\mathrm{e}^{+\mathrm{j}\,2\pi\frac{n\nu}{N}} = c_{n-N/2}\,\mathrm{e}^{\mathrm{j}\,\pi n}, \qquad \breve{C}_\nu = \frac{1}{N}\sum_{n=0}^{N-1}C_n\,\mathrm{e}^{-\mathrm{j}\,2\pi\frac{n\nu}{N}} = \breve{c}_{\nu-N/2}\,\mathrm{e}^{-\mathrm{j}\,\pi\nu}\,. \tag{B.47}$$

Because the DFT operates on sampled data, the spectrum $\breve{c}_\nu$ repeats periodically, see Eq. (2.1) and Fig. 2.2(b) on Page 13. In addition, the data length c_n is finite (implicitly meaning a periodic repetition in contrast to assuming zero data outside the interval $-N/2 \le n \le +N/2-1$), therefore the spectrum is discrete. As a consequence of the periodicities in time and frequency domain, the following coefficients are identical and are therefore excluded from the sums: $c_{+N/2} \equiv c_{-N/2}$, $\breve{c}_{+N/2} \equiv \breve{c}_{-N/2}$, $C_N \equiv C_0$, $\breve{C}_N \equiv \breve{C}_0$.

B.3.1 Pareseval's theorem

For a better understanding of the physical meaning of the DFT coefficients, a look at Parseval's[17] identity is useful. With the help of the orthogonality relation for infinitely extended harmonic functions it states that the total energy in time and frequency domain must be equal,

$$\int_{-\infty}^{+\infty}|\Psi(t)\,|^2\,\mathrm{d}t = \int_{-\infty}^{+\infty}|\breve{\Psi}(f)\,|^2\,\mathrm{d}f\,, \qquad\qquad \lim_{k\to\infty}\int_{-k}^{+k}\mathrm{e}^{\pm\mathrm{j}\,2\pi f(t-t')}\,\mathrm{d}f\; = \delta\,(t-t')\,. \tag{B.48a}$$

Applying Parseval' theorem to Eq. (B.46b), and with the help of an orthogonality relation for time-limited harmonic functions we find (Kronecker symbol $\delta_{nn'}$, Eq. (12) in Table 1.3 on Page 9)

$$T_s\sum_{n=-N/2}^{N/2-1}|\Psi(nT_s)\,|^2 = \frac{F_s}{N}\sum_{\nu=-N/2}^{N/2-1}|\Psi(\nu F_s/N)\,|^2, \qquad \frac{1}{N}\sum_{\nu=-N/2}^{N/2-1}\mathrm{e}^{\pm\mathrm{j}\,2\pi\frac{\nu(n-n')}{N}} = \delta_{nn'}\,. \tag{B.48b}$$

Parseval's theorem for the DFT definitions Eq. (B.46c) or (B.47), however, reads

$$\sum_{n=-N/2}^{N/2-1}|c_n|^2 = N\sum_{\nu=-N/2}^{N/2-1}|\breve{c}_\nu|^2 \qquad \text{or} \qquad \sum_{n=0}^{N-1}|C_n|^2 = N\sum_{\nu=0}^{N-1}|\breve{C}_\nu|^2. \tag{B.48c}$$

[17] Marc-Antoine Parseval des Chênes, ⋆1755, †1836. A French mathematician best known for his theorem in Fourier analysis.

The weighing factor N in Eq. (B.48c) comes from the different normalizations for unlimited harmonics, Eq. (B.48a), and for time-limited harmonics, Eq. (B.48b).

From Eq. (B.48a) we conclude that the power $|\Psi(nT_s)|^2$ in a time slot T_s and the spectral power $|\Psi(\nu F_s/N)|^2(F_s/N)^2$ in a resolution bandwidth F_s/n can be approximated by

$$|\Psi(nT_s)|^2 \approx |c_n|^2, \qquad |\Psi(\nu F_s/N)|^2(F_s/N)^2 = |\breve{c}_\nu|^2. \tag{B.49}$$

B.3.2 Zero padding and interpolation

As we have seen in Sect. 2.1.1 and Eq. (2.2) on Page 14, the reconstruction of Nyquist-sampled data limited to a maximum spectral frequency F_s requires an ideal rectangular (and therefore non-causal) filter with infinitely steep slopes, a so-called "brick wall" filter. Obviously, such a filter is not realizable in practice.

For relaxing the filter requirements, the sampling frequency has to be increased. This so-called up-sampling is most easily done by "in-between zero padding" (zero padding, *German* Nullpolsterung): Between every two original time samples, a number of $w-1$ samples with zero values are inserted. This increases the sampling rate to $F_s^{(w)} = wF_s$, thereby creating a spectral gap of $(w-1)\,F_s$, which then accomodates also finite, physically realizable filter slopes.

Naturally, the sampled signal itself must not be changed by this in-between zero padding. Indeed, the spectral information up to F_s, i. e., the Fourier coefficients $\breve{c}_\nu$ for $-N/2 \le \nu \le +N/2-1$, remain unchanged, but additional spurious coefficients $\breve{c}_\nu$ are created in the "spectral gap" intervals $-wN/2 \le \nu < -N/2$ and $N/2 \le \nu < wN/2-1$. However, implicit in applying the DFT is the assumption of a band limited complex signal (no spectral components beyond $f = F_s$), and therefore the spurious coefficients $\breve{c}_\nu$ due to the in-between zero padding can be ignored, i. e., set to zero. This again is called zero padding and could have been done also in the first place.

An IDFT operating on the adjusted, zero-padded $\breve{c}_\nu$ then creates a set of coefficients c'_n with interpolated values in-between the original samples. These interpolated coefficients replace the primarily inserted zeros. Had we not inserted zeros during the in-between zero padding process, but had we chosen any other arbitrary values, then the Fourier coefficients $\breve{c}_\nu$ for $-N/2 \le \nu \le +N/2-1$ *would have changed*!

Quite often this leads to a some confusion. We therefore discriminate between "in-between zero padding" and "end zero padding", which both can be done either in the time domain or in the frequency domain. In-between zero padding in one domain corresponds to up-sampling the available data, and in the other domain it adds irrelevant data at both ends of the transformed available data. End zero padding in one domain interpolates the data in the other domain, thus performing an up-sampling, too.

The following examples illustrate both processes. Because operations involving zeros need not be executed, these techniques are computationally very efficient in digital signal processing (DSP).

In-between zero padding in the time domain (up-sampling)

Consider Eq. (B.46c) in a notation, where we introduce complex so-called "twiddle factors" $W_N^{\nu n}$, which in fact represent a square matrix $(W_N^{\nu n})$,

$$c_n = \sum_{\nu=-N/2}^{N/2-1} W_N^{\nu n}\,\breve{c}_\nu\,, \qquad \breve{c}_\nu = \sum_{n=-N/2}^{N/2-1} W_N^{\nu n\,*}\,c_n\,, \qquad W_N^{\nu n} = \mathrm{e}^{\mathrm{j}\,2\pi\frac{\nu n}{N}}, \quad W_N^{\nu n\,*} = \mathrm{e}^{-\mathrm{j}\,2\pi\frac{\nu n}{N}}\,. \tag{B.50}$$

For the case $N = 4$, these twiddle factors are ± 1 and $\mathrm{e}^{\pm\mathrm{j}\,\pi/2}$, and we find the relations

$$\breve{c}_\nu = \sum_{n=-2}^{1} W_4^{\nu n\,*}\,c_n\,, \qquad \begin{array}{rcl} \breve{c}_{-2} &=& c_{-2} - c_{-1} + c_0 - c_1 \\ \breve{c}_{-1} &=& -c_{-2} + \mathrm{e}^{-\mathrm{j}\,\pi\frac{1}{2}}\,c_{-1} + c_0 + \mathrm{e}^{+\mathrm{j}\,\pi\frac{1}{2}}\,c_1 \\ \breve{c}_0 &=& c_{-2} + c_{-1} + c_0 + c_1 \\ \breve{c}_1 &=& -c_{-2} + \mathrm{e}^{+\mathrm{j}\,\pi\frac{1}{2}}\,c_{-1} + c_0 + \mathrm{e}^{-\mathrm{j}\,\pi\frac{1}{2}}\,c_1 \end{array}. \tag{B.51}$$

Now let us insert one zero ($w = 2$) after every other time-domain sample c_n. This doubles the number of new coefficients c'_n to $wN = 8$, halves the sampling interval to T_s/w, and doubles the sampling rate to

wF_s, but leaves the frequency step size $wF_S/(wN) = F_s/N$ unchanged. For avoiding too abstract sum formulae, we write the following results in full, where $(\cdots)^T$ denotes the transpose of a row matrix, i. e., it represents a colum matrix,

$$\breve{c}_\nu = \sum_{n=-4}^{3} W_8^{\nu n\,*}\, c'_n\,, \quad \left(\breve{c}_{-4}\ \breve{c}_{-3}\ \breve{c}_{-2}\ \breve{c}_{-1}\ \breve{c}_0\ \breve{c}_1\ \breve{c}_2\ \breve{c}_3\right)^T = (W_8^{\nu n\,*})\left(c_{-2}\ 0\ c_{-1}\ 0\ c_0\ 0\ c_1\ 0\right)^T. \quad \text{(B.52a)}$$

After performing the matrix multiplication, we end up with a set of equations for $w = 2$, $wN = 8$:

$$w=2,\ wN=8 \begin{cases} \breve{c}_{-4} = \ c_{-2} + & c_{-1} + c_0 + & c_1 \\ \breve{c}_{-3} = -c_{-2} + \mathrm{e}^{-\mathrm{j}\pi\frac{3}{2}} c_{-1} + & c_0 + \mathrm{e}^{+\mathrm{j}\pi\frac{3}{2}} c_1 & \\ \breve{c}_{-2} = \ c_{-2} - & c_{-1} + c_0 - & c_1 \\ \breve{c}_{-1} = -c_{-2} + \mathrm{e}^{-\mathrm{j}\pi\frac{1}{2}} c_{-1} + & c_0 + \mathrm{e}^{+\mathrm{j}\pi\frac{1}{2}} c_1 & \\ \breve{c}_0 \ = \ c_{-2} + & c_{-1} + c_0 + & c_1 \\ \breve{c}_1 \ = -c_{-2} + \mathrm{e}^{+\mathrm{j}\pi\frac{1}{2}} c_{-1} + & c_0 + \mathrm{e}^{-\mathrm{j}\pi\frac{1}{2}} c_1 & \\ \breve{c}_2 \ = \ c_{-2} + & c_{-1} + c_0 + & c_1 \\ \breve{c}_3 \ = -c_{-2} + \mathrm{e}^{+\mathrm{j}\pi\frac{3}{2}} c_{-1} + & c_0 + \mathrm{e}^{-\mathrm{j}\pi\frac{3}{2}} c_1 & \end{cases} \quad N=4 \begin{cases} \breve{c}_{-2} = \ c_{-2} - & c_{-1} + c_0 - & c_1 \\ \breve{c}_{-1} = -c_{-2} + \mathrm{e}^{-\mathrm{j}\pi\frac{1}{2}} c_{-1} + & c_0 + \mathrm{e}^{+\mathrm{j}\pi\frac{1}{2}} c_1 & \\ \breve{c}_0 \ = \ c_{-2} + & c_{-1} + c_0 + & c_1 \\ \breve{c}_1 \ = -c_{-2} + \mathrm{e}^{+\mathrm{j}\pi\frac{1}{2}} c_{-1} + & c_0 + \mathrm{e}^{-\mathrm{j}\pi\frac{1}{2}} c_1 & \end{cases} \quad \text{(B.52b)}$$

For an easy comparison, we duplicate here Eq. (B.51) for $N = 4$. Because every wth coefficient c'_n is set to zero, the original spectral coefficients $\left(\breve{c}_{-2}\ \breve{c}_{-1}\ \breve{c}_0\ \breve{c}_1\right)^T$ remain unchanged, but spurious coefficients $\breve{c}_{-4}$, $\breve{c}_{-3}$ and $\breve{c}_2$, $\breve{c}_3$ are generated. If the stop band of a subsequent digital filter suppresses these spurious coefficients sufficiently, the effect is an end zero padding in the frequency domain.

End zero padding in the frequency domain (interpolation)

Due to the primary assumption of a bandlimited signal, we know that any Fourier coefficient subscripted with $-wN/2 \le \nu < -N/2$ and $N/2 \le \nu < wN/2 - 1$ must be zero, so we modify these coefficients $\breve{c}_\nu$ Eq. (B.52b) with end zero padding and write the IDFT

$$c'_n = \sum_{\nu=-4}^{3} W_8^{\nu n}\, \breve{c}_\nu\,, \quad \left(c'_{-4}\ c'_{-3}\ c'_{-2}\ c'_{-1}\ c'_0\ c'_1\ c'_2\ c'_3\right) = (W_8^{\nu n})\left(0\ 0\ \breve{c}_{-2}\ \breve{c}_{-1}\ \breve{c}_0\ \breve{c}_1\ 0\ 0\right)^T. \quad \text{(B.53a)}$$

In performing the matrix multiplication, the first and the last column pair of the twiddle factor matrix $(W_8^{\nu n})$ can be disregarded, so only the Fourier coefficients $\left(\breve{c}_{-2}\ \breve{c}_{-1}\ \breve{c}_0\ \breve{c}_1\right)^T$ of the original signal $\left(c_{-2}\ c_{-1}\ c_0\ c_1\right)^T$ come into play, and the result for $w = 2$, $wN = 8$ is:

$$w=2,\ wN=8 \begin{cases} c'_{-4} = & \breve{c}_{-2} - & \breve{c}_{-1} + \breve{c}_0 - & \breve{c}_1 \\ c'_{-3} = & \mathrm{e}^{+\mathrm{j}\pi\frac{6}{4}} \breve{c}_{-2} + \mathrm{e}^{-\mathrm{j}\pi\frac{3}{4}} \breve{c}_{-1} + & \breve{c}_0 + \mathrm{e}^{-\mathrm{j}\pi\frac{3}{4}} \breve{c}_1 & \\ c'_{-2} = & -\breve{c}_{-2} + & \breve{c}_{-1} + \breve{c}_0 + \mathrm{e}^{-\mathrm{j}\pi\frac{2}{4}} \breve{c}_1 & \\ c'_{-1} = & -\mathrm{e}^{+\mathrm{j}\pi\frac{2}{4}} \breve{c}_{-2} + & \breve{c}_{-1} + \breve{c}_0 + \mathrm{e}^{-\mathrm{j}\pi\frac{1}{4}} \breve{c}_1 & \\ c'_0 \ = & \breve{c}_{-2} + & \breve{c}_{-1} + \breve{c}_0 + & \breve{c}_1 \\ c'_1 \ = & \mathrm{e}^{-\mathrm{j}\pi\frac{2}{4}} \breve{c}_{-2} + \mathrm{e}^{-\mathrm{j}\pi\frac{1}{4}} \breve{c}_{-1} + & \breve{c}_0 + \mathrm{e}^{+\mathrm{j}\pi\frac{1}{4}} \breve{c}_1 & \\ c'_2 \ = & -\breve{c}_{-2} + \mathrm{e}^{-\mathrm{j}\pi\frac{2}{4}} \breve{c}_{-1} + & \breve{c}_0 + \mathrm{e}^{+\mathrm{j}\pi\frac{2}{4}} \breve{c}_1 & \\ c'_3 \ = & \mathrm{e}^{-\mathrm{j}\pi\frac{6}{4}} \breve{c}_{-2} + \mathrm{e}^{-\mathrm{j}\pi\frac{3}{4}} \breve{c}_{-1} + & \breve{c}_0 + \mathrm{e}^{+\mathrm{j}\pi\frac{3}{4}} \breve{c}_1 & \end{cases} \quad N=4 \begin{cases} c_{-2} = & \breve{c}_{-2} - & \breve{c}_{-1} + \breve{c}_0 - & \breve{c}_1 \\ c_{-1} = & -\breve{c}_{-2} + & \breve{c}_{-1} + \breve{c}_0 + \mathrm{e}^{-\mathrm{j}\pi\frac{1}{2}} \breve{c}_1 & \\ c_0 \ = & \breve{c}_{-2} + & \breve{c}_{-1} + \breve{c}_0 + & \breve{c}_1 \\ c_1 \ = & -\breve{c}_{-2} + \mathrm{e}^{-\mathrm{j}\pi\frac{1}{2}} \breve{c}_{-1} + & \breve{c}_0 + \mathrm{e}^{+\mathrm{j}\pi\frac{1}{2}} \breve{c}_1 & \end{cases} \quad \text{(B.53b)}$$

For a better comparison, we also specify in Eq. (B.53b) the IDFT as calculated from the unmodified Fourier coefficients $\left(\breve{c}_{-2}\ \breve{c}_{-1}\ \breve{c}_0\ \breve{c}_1\right)^T$ for $N = 4$. It is obvious that the original coefficients c_n are recovered, and that in c'_n interpolated coefficients are to be found. They do not represent any new information, because these intermediate values could have been also inferred from the sinc-interpolated continuous data according to Eq. (2.3) on Page 15. However, end zero padding in the frequency domain with a subsequent IDFT is numerically much simpler than any sinc-interpolation.

Appendix C

Coherent signal and noise

C.1 Signal representation

We seek expressions for the detector current of a square-law detected (rectified) coherent signal that is embedded in narrowband noise. To this end, the following signal representation is especially useful.

C.1.1 Narrowband noise

Noise in a narrow bandwidth $B \ll f_0$ centred at frequency f_0 is described as a sum of sinusoidals[1,2],

$$s(t) = \sum_{n=-N}^{+N} a_n \cos\left[(\omega_0 + n\Delta\omega)\, t + \varphi_n\right], \qquad \tfrac{1}{2}\overline{a_n^2} = w_s(f_0 + n\Delta f)\Delta f. \tag{C.1}$$

The phases φ_n are independent random variables, which are equally distributed in an interval $0 \le \varphi_n < 2\pi$. With $\cos(x \pm y) = \cos x \cos y \mp \sin x \sin y$ we find

$$s(t) = x(t)\cos(\omega_0 t) - y(t)\sin(\omega_0 t) = r(t)\cos\left[\omega_0 t + \varphi(t)\right] \tag{C.2}$$

with the abbreviations

$$\begin{aligned} x(t) &= \sum_n a_n \cos(n\Delta\omega t + \varphi_n), & r(t) &= [x^2(t) + y^2(t)]^{1/2}, \\ y(t) &= \sum_n a_n \sin(n\Delta\omega t + \varphi_n), & \varphi(t) &= \arctan\frac{y(t)}{x(t)}. \end{aligned} \tag{C.3}$$

Applying the central limit theorem[3] for sufficiently large N, the quantities $x(t)$, $y(t)$ are independent Gaussian random variables with variances

$$\frac{1}{2}\sum_n \overline{a_n^2} = \sigma_x^2 = \sigma_y^2 = \sigma^2, \qquad \overline{x} = \overline{y} = 0, \qquad \overline{xy} = 0. \tag{C.4}$$

For narrowband noise, the quantities r and φ can be interpreted as slowly varying amplitude and slowly varying phase, respectively.

If $\Theta_s(f)$ denotes the two-sided power spectrum of $s(t)$, we find the low-frequency part of the two-sided power spectra of $x(t)$ and $y(t)$ through shifting $\Theta_s(f)$ by $\pm f_0$, where we regard only the low-frequency contributions in the vicinity of $f = 0$, see Fig. C.1,

$$\Theta_x(f) = \Theta_y(f) = \Theta_s(f - f_0)|_{\text{at } f\,=\,0} + \Theta_s(f + f_0)|_{\text{at } f\,=\,0}. \tag{C.5}$$

The two-sided power spectrum $\Theta_x(f) = \Theta_x(-f)$ is real and symmetric, because the covariance of real ran-

[1] See Ref. 10 on Page 122. Eq. (1.7-1)

[2] Rice, S. O.: Mathematical analysis of random noise. Bell Syst. Techn. J. 24 (1945) 46–156. Eq. (2.8-1), (3.7-2)

[3] See Ref. 10 on Page 122. Sect. 2.10

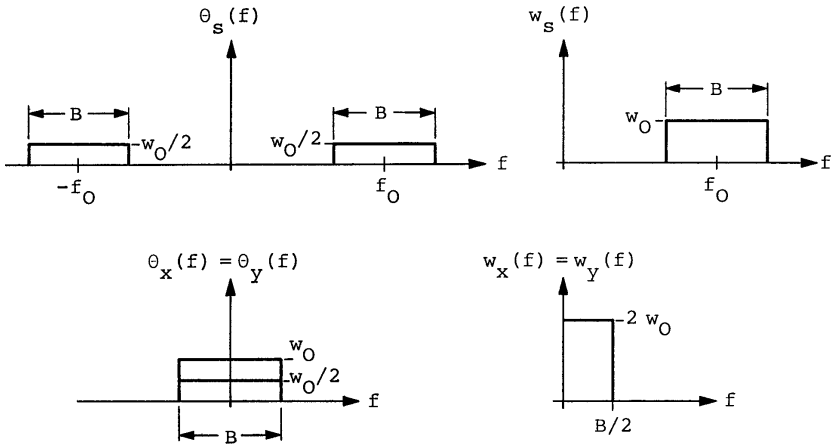


Fig. C.1. Two-sided power spectra $\Theta_s\,(f)$, $\Theta_x\,(f)$, $\Theta_y\,(f)$ and one-sided power spectra $w_s\,(f)$, $w_x\,(f)$, $w_y\,(f)$ of narrowband noise

dom variables $\mathbf{x}$ is real and symmetric, $K_x\,(\tau) = K_x\,(-\tau)$ (equals the auto-correlation function $\vartheta_x\,(\tau)$ for ergodic processes). Covariance or auto-correlation function and power spectrum form a Fourier pair. This famous theorem from 1930 is named after Wiener[4] and Khintchine[5],

$$K_x(\tau) = \overline{\mathbf{x}(t)\,\mathbf{x}(t-\tau)} = \int\limits_{-\infty}^{+\infty} \Theta_x(f)\,\mathrm{e}^{\mathrm{j}\,2\pi f\tau}\,\mathrm{d}f\,, \tag{C.6}$$

$$\vartheta_x(\tau) = \langle \mathbf{x}(t)\,\mathbf{x}(t-\tau)\rangle = K_x\,(\tau) \ \text{(for ergodic } \mathbf{x}), \tag{C.7}$$

$$\vartheta_x(0) = K_x(0) = P = \overline{\mathbf{x}^2} = \int\limits_{-\infty}^{+\infty} \Theta_x(f)\,\mathrm{d}f = \int\limits_{0}^{+\infty} 2\Theta_x(f)\,\mathrm{d}\,f = \int\limits_{0}^{\infty} w_x(f)\,\mathrm{d}f. \tag{C.8}$$

The one-sided (real) power spectrum $w_x(f)$ according to Eq. (C.8) is defined by

$$w_x(f) = \begin{cases} 0 & \text{for } f < 0\,, \\ 2\Theta_x(f) & \text{for } f > 0\,. \end{cases} \tag{C.9}$$

Any one-sided power P_0 at $f = 0$ is described by the term $2P_0\delta(f)$, because an integral over half of the symmetric Dirac function results in $\frac{1}{2}$, and $2\int_0^\infty \delta\,(f)\,\mathrm{d}f = 1$.

We now specialize to the case of "white" noise with a constant spectral power density w_0 inside the

[4]Norbert Wiener, American mathematician, ⋆Columbia (Missouri) †26.11.1894, Stockholm 18.3.1964. Since 1932 professor in Cambridge (Massachusetts), numerous research visits in Europa, China, India, Mexico. Wiener formulated some of the most important contributions to mathematics in the 20th century. During the 1920s Wiener did highly innovative and fundamental work on what are now called stochastic processes and, in particular, on the theory of Brownian motion and on generalized harmonic analysis, as well as significant work on other problems of mathematical analysis. During World War II Wiener worked on the problem of aiming gunfire at a moving target. The ideas that evolved led to "Extrapolation, Interpolation, and Smoothing of Stationary Time Series" (1949), which first appeared as a classified report and established Wiener as a codiscoverer, with the Russian mathematician Andrey Kolmogorov, of the theory on the prediction of stationary time series. It introduced certain statistical methods into control and communications engineering and exerted great influence in these areas. This work also led him to formulate the concept of cybernetics.

[5]Chintschin, A. J.: Korrelationstheorie der stationären stochastischen Prozesse. Math. Annalen 109 (1934) 604 — Aleksandr Yakovlevich Khinchin (pronounced ['xɪntʃɪn], in Western Europe also transcribed by "Chintschin" or "Chinčin"), Russian mathematician and statistician, ⋆Kondrowo (region Kaluga) 19.7.1894, †Moskow 18.11.1959. Worked on probability theory and its applications. Published several important works on statistical physics, where he used the methods of probability theory, and on information theory, queuing theory and mathematical analysis.

bandwidth B. The power density outside disappears. The average noise power from Eq. (C.1)–(C.5) is

$$\begin{aligned}\overline{P} &= \overline{s^2(t)} = \tfrac{1}{2}\sum_n \overline{a_n^2} = \int_{f_0-B/2}^{f_0+B/2} w_s(f)\,\mathrm{d}f = \int_{f_0-B/2}^{f_0+B/2} w_0\,\mathrm{d}f \\ &= \tfrac{1}{2}\,\overline{r^2(t)} = \tfrac{1}{2}\,\overline{x^2(t)} + \tfrac{1}{2}\,\overline{y^2(t)} = \overline{x^2(t)} = \overline{y^2(t)} \\ &= \sigma_x^2 = \sigma_y^2 = \sigma^2 = \int_0^{B/2} w_x(f)\,\mathrm{d}f = \int_0^{B/2} w_y(f)\,\mathrm{d}f = w_0 B\,.\end{aligned} \tag{C.10}$$

C.1.2 Signal and narrowband noise

We consider a coherent signal with frequency f_0 and constant amplitude A, which is embedded in narrowband noise. The resulting signal $s\,(t)$ is represented by

$$\begin{aligned}s(t) &= [A + x(t)]\cos(\omega_0 t) - y(t)\sin(\omega_0 t) = z(t)\cos(\omega_0 t) - y(t)\sin(\omega_0 t) \\ &= r(t)\cos[\omega_0 t + \varphi(t)], \\ z(t) &= A + x(t), \quad \overline{z} = A, \\ r(t) &= \sqrt{z^2(t) + y^2(t)}, \qquad \sigma_z^2 = \sigma_x^2 = \sigma_y^2 = \sigma^2 = Bw_0,\ \ \overline{xy} = 0\,.\end{aligned} \tag{C.11}$$

The one-sided power spectrum is shown in Fig. C.2(a) on Page 201.

C.2 Quadratic detection of signal and narrowband noise

We follow the derivation by Rice[6] and consider a signal comprising a coherent carrier in a certain polarization with narrowband noise in the same polarization as defined in Eq. (C.11). The one-sided power spectrum is displayed in Fig. C.2 on Page 201. A quadratic rectifier (detector) generates a current in proportion to $s^2(t)$. The low-frequency part of the detector current is denoted by $i(t)$. It is calculated by averaging over a carrier period. The direct current (DC) part is $\overline{i(t)}$. With $\cos x \sin x = \frac{1}{2}\sin 2x$, $\cos^2 x = \frac{1}{2}\,(1 + \cos 2x)$, $\sin^2 x = \frac{1}{2}\,(1 - \cos 2x)$ we find

$$i(t) = \frac{1}{2}\left[A^2 + 2Ax(t) + x^2(t) + y^2(t)\right], \qquad \overline{i(t)} = \frac{1}{2}\left[A^2 + \overline{x^2} + \overline{y^2}\right] = \frac{1}{2}A^2 + \sigma^2. \tag{C.12}$$

C.2.1 Auto-correlation function of detector current

For determining the low-frequency power spectrum $\Theta_i(f)$ of the current $i\,(t)$, we first calculate its covariance $K_i(\tau) = \overline{i(t)\,i(t-\tau)}$ (= auto-correlation function $\vartheta_i\,(\tau)$, ACF). In the following, we abbreviate $x := x(t)$ and $x_\tau := x(t-\tau)$, and proceed similarly for other relevant time functions. The noise process is assumed to be stationary and ergodic. Expectations of odd powers of x and y vanish because their probability density functions are symmetric Gaussians with $\overline{x} = 0$, $\overline{y} = 0$. We find

$$\begin{aligned}\overline{i(t)\,i(t-\tau)} = \overline{i\,i_\tau} &= \tfrac{1}{4}\,\overline{(A^2 + 2Ax + x^2 + y^2)(A^2 + 2Ax_\tau + x_\tau^2 + y_\tau^2)} \\ &= \tfrac{1}{4}\,\overline{(A^4 + 2A^3x_\tau + A^2x_\tau^2 + A^2y_\tau^2)} \\ &\quad + \tfrac{1}{4}\,\overline{(2A^3x + 4A^2xx_\tau + 2Axx_\tau^2 + 2Axy_\tau^2)} \\ &\quad + \tfrac{1}{4}\,\overline{(A^2x^2 + 2Ax^2x_\tau + x^2x_\tau^2 + x^2y_\tau^2)} \\ &\quad + \tfrac{1}{4}\,\overline{(A^2y^2 + 2Ax_\tau y^2 + x_\tau^2y^2 + y^2y_\tau^2)} \\ &= \tfrac{1}{4}\,(A^4 + A^2\overline{x^2} + A^2\overline{x^2}) + \tfrac{1}{4}\,(4A^2\overline{xx_\tau}) \\ &\quad + \tfrac{1}{4}\,(A^2\overline{x^2} + \overline{x^2x_\tau^2} + \overline{x^2}\,\overline{x_\tau^2}) + \tfrac{1}{4}\,(A^2\overline{x^2} + \overline{x_\tau^2}\,\overline{x^2} + \overline{x^2x_\tau^2}) \\ &= \tfrac{1}{4}\left[A^4 + 4A^2\overline{x^2} + 4A^2\underbrace{\overline{xx_\tau}} + 2\big(\overline{x^2}\big)^2 + 2\underbrace{\overline{x^2x_\tau^2}}\right].\end{aligned} \tag{C.13}$$

[6] See Ref. 2 on Page 197. Sect. 4.5 Noise through square law device

The terms $\overline{xx_\tau}$ and $\overline{x^2x_\tau^2}$ need to be evaluated in the following.

Term $\overline{xx_\tau}$ The two-sided power spectrum $\Theta_x(f=\pm f_0)$ in Fig. C.1 on Page 198 has a height of $w_0/2$. For the low-frequency spectra near $f=0$ we therefore require a height of $w_0/2+w_0/2=w_0$, because the power must be conserved. The auto-correlation $\overline{xx_\tau}$ reads

$$\overline{xx_\tau} = \vartheta_x(\tau) = \int_{-B/2}^{+B/2} \underbrace{\Theta_x(f)}_{w_0(f)} \mathrm{e}^{\mathrm{j}\,2\pi f\tau}\,\mathrm{d}f = \int_{-B/2}^{+B/2} w_0\,\mathrm{e}^{\mathrm{j}\,2\pi f\tau}\,\mathrm{d}f = \frac{w_0}{\pi\tau}\,\sin(\pi B\tau) = w_0 B\,\frac{\sin(\pi B\tau)}{\pi B\tau}\,, \tag{C.14}$$

$$\overline{x^2} = \sigma^2 = w_0 B.$$

This results in an expression for the normalized auto-correlation function,

$$\rho = \frac{\overline{xx_\tau}}{\overline{x^2}} = \frac{\sin(\pi B\tau)}{(\pi B\tau)}\,. \tag{C.15}$$

Term $\overline{x^2x_\tau^2}$ For calculating $\overline{x^2x_\tau^2}$ we separate x_τ in one part that is correlated with x, and in another contribution z that is statistically independent from x,

$$x_\tau = \rho x + z, \qquad \overline{x^n z^m} = \overline{x^n}\;\overline{z^m}. \tag{C.16}$$

From Eq. (C.16) and (C.15) we find

$$\overline{z^2} = \overline{(x_\tau-\rho x)^2} = \overline{x_\tau^2} - 2\rho\underbrace{\overline{xx_\tau}}_{=\rho\overline{x^2}} + \rho^2\overline{x^2} = \overline{x^2} - 2\rho^2\overline{x^2} + \rho^2\overline{x^2} = \overline{x^2}(1-\rho^2), \tag{C.17}$$

which leads together with an expression for the moments[7] of a Gaussian, especially for $\overline{x^4}$, to the required result for the term $\overline{x^2x_\tau^2}$,

$$\begin{aligned}\overline{x^2x_\tau^2} = \overline{x^2(\rho x+z)^2} &= \rho^2\overline{x^4} + 2\rho\overline{x^3}\times\overbrace{\overline{z}}^{0} + \overbrace{\overline{x^2z^2}}^{\text{stat. indep.}} = \rho^2\overline{x^4} + \overline{x^2}\;\overline{z^2}\\ &= \rho^2\overline{x^4} + (\overline{x^2})^2(1-\rho^2) = \rho^2\times\Big[1\times3\times(\overline{x^2})^2\Big] + (\overline{x^2})^2(1-\rho^2)\\ &= 2\rho^2(\overline{x^2})^2 + (\overline{x^2})^2 = 2\rho^2\sigma^4+\sigma^4\end{aligned} \tag{C.20}$$

Substitution of terms The expressions Eq. (C.15) and (C.20),

$$\overline{xx_\tau} = \rho\overline{x^2} = \rho\sigma^2 \quad\text{and}\quad \overline{x^2x_\tau^2} = 2\rho^2(\overline{x^2})^2 + (\overline{x^2})^2 = 2\rho^2\sigma^4+\sigma^4,$$

are substituted in Eq. (C.13),

$$\overline{i(t)\,i(t-\tau)} = \frac{1}{4}\Big[A^4 + 4A^2\overline{x^2} + 4A^2\underbrace{\overline{xx_\tau}} + 2\overbrace{\left(\overline{x^2}\right)^2}^{\sigma^4} + 2\underbrace{\overline{x^2x_\tau^2}}\Big],$$

and with Eq. (C.14) and (C.15) the covariance reads

$$\begin{aligned}K_i(\tau) = \overline{i(t)\,i(t-\tau)} &= \left(\tfrac{1}{2}A^2+\sigma^2\right)^2 + A^2\rho\sigma^2 + \rho^2\sigma^4\\ &= \left(\frac{1}{2}A^2 + w_0B\right)^2 + A^2w_0B\,\frac{\sin(\pi B\tau)}{(\pi B\tau)} + w_0^2B^2\,\frac{\sin^2(\pi B\tau)}{(\pi B\tau)^2}\,.\end{aligned} \tag{C.21}$$

[7] Gaussian probability density function and its moments:

$$w_x(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\Big[-\frac{(x-a)^2}{2\sigma_x^2}\Big], \quad \overline{x}=a, \quad \overline{(x-\overline{x})^2} = \overline{\delta x^2} = \sigma_x^2\,, \tag{C.18}$$

$$\text{For } \overline{x}=0\text{:}\quad \overline{x^{2n+1}} = 0, \quad \overline{x^{2n}} = (2n-1)!!\,\sigma_x^{2n} = 1\cdot3\cdot5\cdot\ldots\cdot(2n-1)\sigma_x^{2n}. \tag{C.19}$$

C.2.2 Power spectrum of detector current

From Eq. (C.21) and with the Wiener-Khintchine theorem Eq. (C.6) on Page 198 we compute the required low-frequency detector current power spectrum,

$$\Theta_i(f) = \int_{-\infty}^{+\infty} K_i(\tau)\,\mathrm{e}^{-\mathrm{j}\,2\pi f\tau}\,\mathrm{d}\tau = \Big(\frac{1}{2}A^2 + w_0 B\Big)^2\delta(f) + 2A^2 w_0 B \int_0^{\infty} \frac{\sin(\pi B\tau)}{(\pi B\tau)}\cos(2\pi f\tau)\,\mathrm{d}\tau + 2w_0^2 B^2 \int_0^{\infty} \frac{\sin^2(\pi B\tau)}{(\pi B\tau)^2}\cos(2\pi f\tau)\,\mathrm{d}\tau\,. \quad \text{(C.22)}$$

Having solved the integrals[8] in Eq. (C.22), we find the one-sided power spectra $w_i(f)$ of a square-law detector when demodulating a coherent carrier and noise. We follow the recipe Eq. (C.9) on Page 198,

$$\begin{aligned} w_i(f) = {} & 2\left(\tfrac{1}{2}A^2 + w_0 B\right)^2\delta(f) \\ & + 2w_0 A^2[H(f) - H(f - B/2)] + 2w_0^2(B - f)[H(f) - H(f - B)]. \end{aligned} \quad \text{(C.24)}$$

The first term represents the rectified carrier and the rectified noise, Fig. C.2(b). The second term stems from the low-frequency mixing products of polarized carrier and identically polarized noise "sidebands"

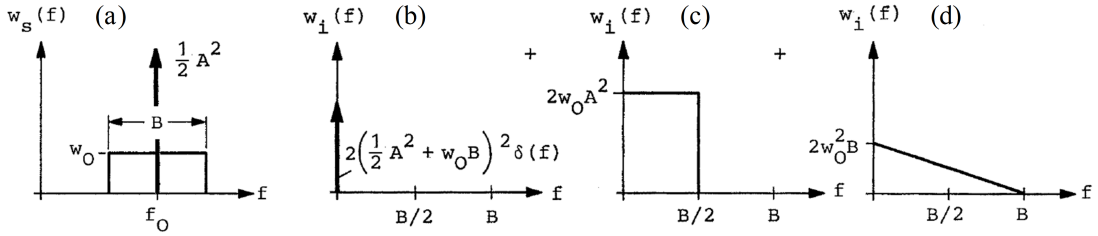


Fig. C.2. Quadratic rectification of a coherent carrier embedded in narrowband noise. (a) One-sided power spectrum $w_s(f)$ of signal $s(t) = [A + x(t)]\cos(\omega_0 t) - y(t)\sin(\omega_0 t)$ (b) One-sided direct current power spectrum with $i_S \sim A^2/2$ and $i_R \sim w_0 B$, total detector current power $(i_S + i_R)^2$. The integral over half a Dirac function is $\int_0^{\infty}\delta(f)\,\mathrm{d}f = \frac{1}{2}$. (c) Carrier-noise interference (d) Noise-noise interference. — Partial detector spectra are uncorrelated and may be added. Therefore the total power equals the sum of the partial powers.

(carrier-noise interference), Fig. C.2(c). The third term finally describes the low-frequency part of the mixing of identically polarized noise "sidebands" among themselves (noise-noise interference), Fig. C.2(d). These contributions to $w_i(f)$ are displayed in Fig. C.2(b)–(d).

If the superposition of a modulated carrier and narrowband noise in a bandwidth B comprises a message with bandwidth $\mathcal{B} \ll B$, this message will be veiled by the low-frequency noise-noise interference $2w_0^2 B$, Fig. C.2(d). This has to be avoided by a proper filtering, because an optical amplifier has a bandwidth of about $B = 12\,\mathrm{THz}$, see Eq. (3.61) on Page 76. For large coherent carriers the carrier-noise interference $2w_0 A^2$ dominates in the low-frequency part of the detector current spectrum $w_i(f)$, Fig. C.2(c).

[8] Two definite integrals:

$$\int_0^{\infty}\frac{\sin(ax)}{x}\cos(bx)\,\mathrm{d}x = \begin{cases} \pi/2, & a > |b|, \\ \pi/4, & a = |b| > 0, \\ 0, & b > |a|, \end{cases} \qquad \int_0^{\infty}\frac{\sin^2(ax)}{x^2}\cos(2bx)\,\mathrm{d}x = \begin{cases} \frac{\pi}{2}(|a| - |b|), & |b| < |a|, \\ 0, & |b| > |a|. \end{cases} \quad \text{(C.23)}$$