Modern Methods of Statistical Data Analysis

From parameter estimation to deep learning — A guided tour of probability

Lecture 8
-
Limit Setting & Unfolding

P.-D. Dr. Roger Wolf
roger.wolf@kit.edu

Dr. Pablo Goldenzweig
pablo.goldenzweig@kit.edu

Dr. Jan Kieseler
jan.kieseler@kit.edu

Dr. Slavomira Stefkova
slavomira.stefkova@kit.edu

Today

- Limits near a physical boundary
  - Shifted and Bayesian approaches
  - Example: Upper limit on the mean of a Poisson variable with background
- Unfolding
  - Formulation of the problem
  - Matrix inversion
  - Method of correction factors
  - Regularized unfolding

Higgs Challenge

Please mail **Sally Stefkova** if you plan to do it! We just want to gauge how many groups are working on this.

Evaluations: Lecture & Computerpraktikum.

Please take a few minutes to fill them out. Your feedback is greatly appreciated. We will take your comments into consideration in trying to improve the course.

Evaluation period: through 22 June (lecture) & 15 July (Computerpraktikum)

We’ve come a long way so far

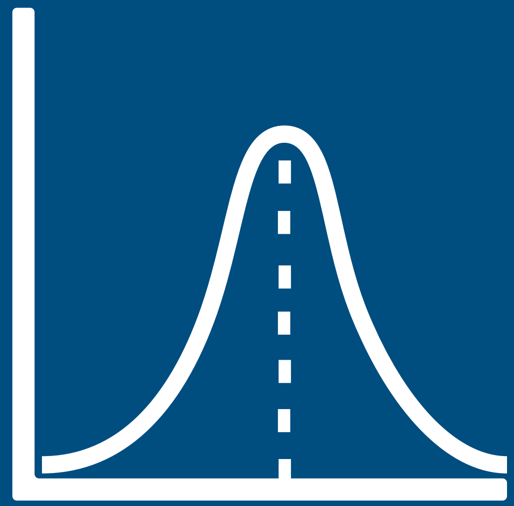
#	Lecture date	Lecture Topic
1	21.4	Fundamental concepts I
2	28.4	Fundamental concepts II
3	5.5	Monte Carlo method & production of random distributions
4	12.5	Parameter estimation & maximum likelihood
5	19.5	Chi-square method
6	26.5	Hypothesis tests & Neyman Pearson
7	9.6	Confidence intervals
8	16.6	Limit setting & unfolding
9	23.6	Event classification - Introduction and perceptron
10	30.6	Classification with the multilayer perceptron
11	7.7	Neural network training
12	14.7	Training algorithms & regularization methods
13	21.7	Training validation
14	28.7	Advanced neural networks

P.-D. Dr. Roger Wolf

Dr. Jan Kieseler

Review

Confidence Intervals



Statistical errors, confidence intervals and limits

Up to now: when discussing 'error analysis' we focused on estimating the (co)variances of estimators. This is not always adequate and other ways of communicating the statistical uncertainty of measurements have to be found.

$$\hat{\theta}_{\rm obs} \pm \hat{\sigma}_{\hat{\theta}}$$

Classical confidence intervals (CI)

Review

Alternative (& often =) method of reporting the statistical uncertainty of a measurement

- Suppose you have n observations of a random variable X, which can be used to evaluate an estimator for an unknown true parameter θ:

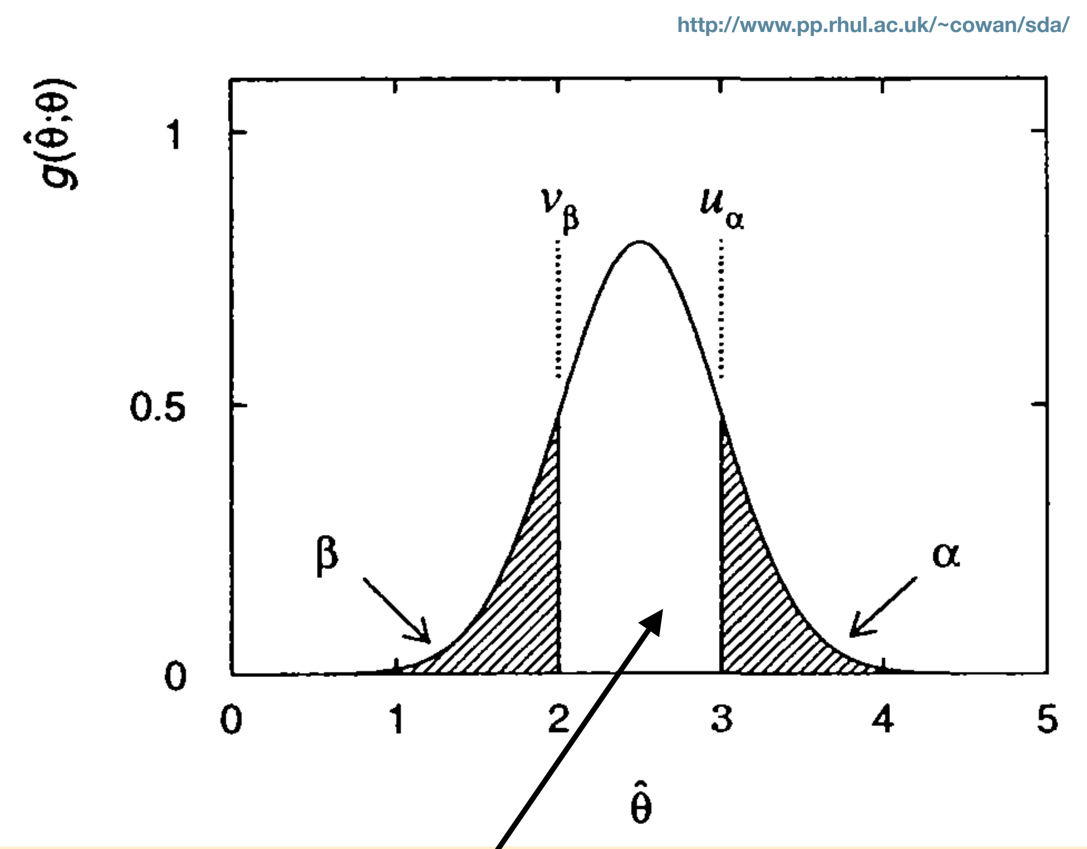
$$\hat{\theta}(x_1, \ldots, x_n) = \hat{\theta}_{\text{obs}}$$

value obtained

- Furthermore, suppose we know the PDF of $\hat{\theta}$ denoted by $g(\hat{\theta}; \theta)$

  Real value of θ unknown, BUT for a given θ one knows what the PDF of $\hat{\theta}$ would be

- From $g(\hat{\theta}; \theta)$, can determine ν_β and u_α such that there are fixed probabilities β and α to observe $\hat{\theta} < \nu_\beta$ or $\hat{\theta} > u_\alpha$



$$\boxed{P(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)) = 1 - \alpha - \beta.}$$

Shows the probability density for an estimator $\hat{\theta}$ for a particular value of the true parameter θ

u_α and u_β depend on the true value θ and are thus determined by

$$\boxed{\beta = P(\hat{\theta} \leq v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta}; \theta) d\hat{\theta} = G(v_\beta(\theta); \theta),}$$

$$\boxed{\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta}; \theta) d\hat{\theta} = 1 - G(u_\alpha(\theta); \theta),}$$

CDF

… so α and β are the probabilities!

Next: lets build the CI step by step…

Classical confidence intervals (CI)

Review

Alternative (& often =) method of reporting the statistical uncertainty of a measurement

- Suppose you have n observations of a random variable X, which can be used to evaluate an estimator for an unknown true parameter θ:

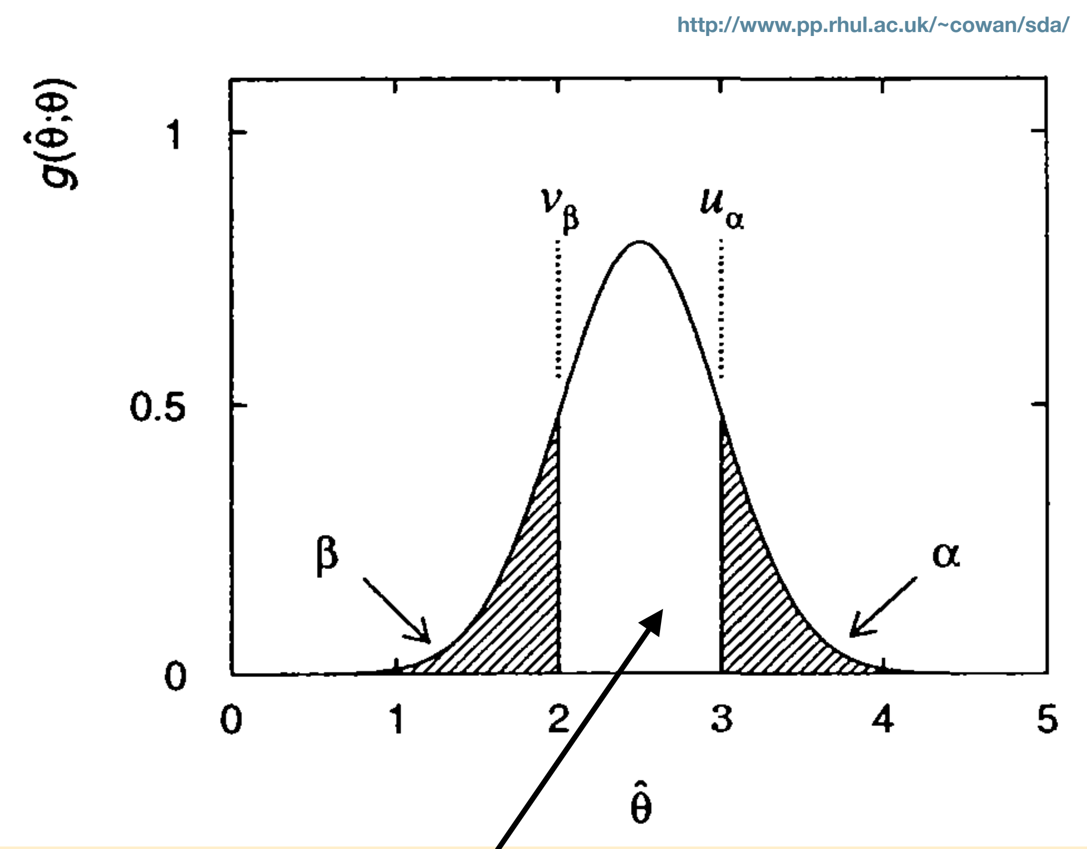
$$\hat{\theta}(x_1, \ldots, x_n) = \hat{\theta}_{\text{obs}}$$

value obtained

- Furthermore, suppose we know the PDF of $\hat{\theta}$ denoted by $g(\hat{\theta}; \theta)$

  Real value of θ unknown, BUT for a given θ one knows what the PDF of $\hat{\theta}$ would be

- From $g(\hat{\theta}; \theta)$, can determine ν_β and u_α such that there are fixed probabilities β and α to observe $\hat{\theta} < \nu_\beta$ or $\hat{\theta} > u_\alpha$



$$\boxed{P(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)) = 1 - \alpha - \beta.}$$

Shows the probability density for an estimator $\hat{\theta}$ for a particular value of the true parameter θ

u_α and u_β depend on the true value θ and are thus determined by

$$\boxed{\beta = P(\hat{\theta} \leq v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta}; \theta) d\hat{\theta} = G(v_\beta(\theta); \theta),}$$

$$\boxed{\alpha = P(\hat{\theta} \geq u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta}; \theta) d\hat{\theta} = 1 - G(u_\alpha(\theta); \theta),}$$

CDF

… so α and β are the probabilities!

Next: lets build the CI step by step…

Classical confidence intervals (CI)

Review

Alternative (& often =) method of reporting the statistical uncertainty of a measurement

- Suppose you have n observations of a random variable X, which can be used to evaluate an estimator for an unknown true parameter θ:

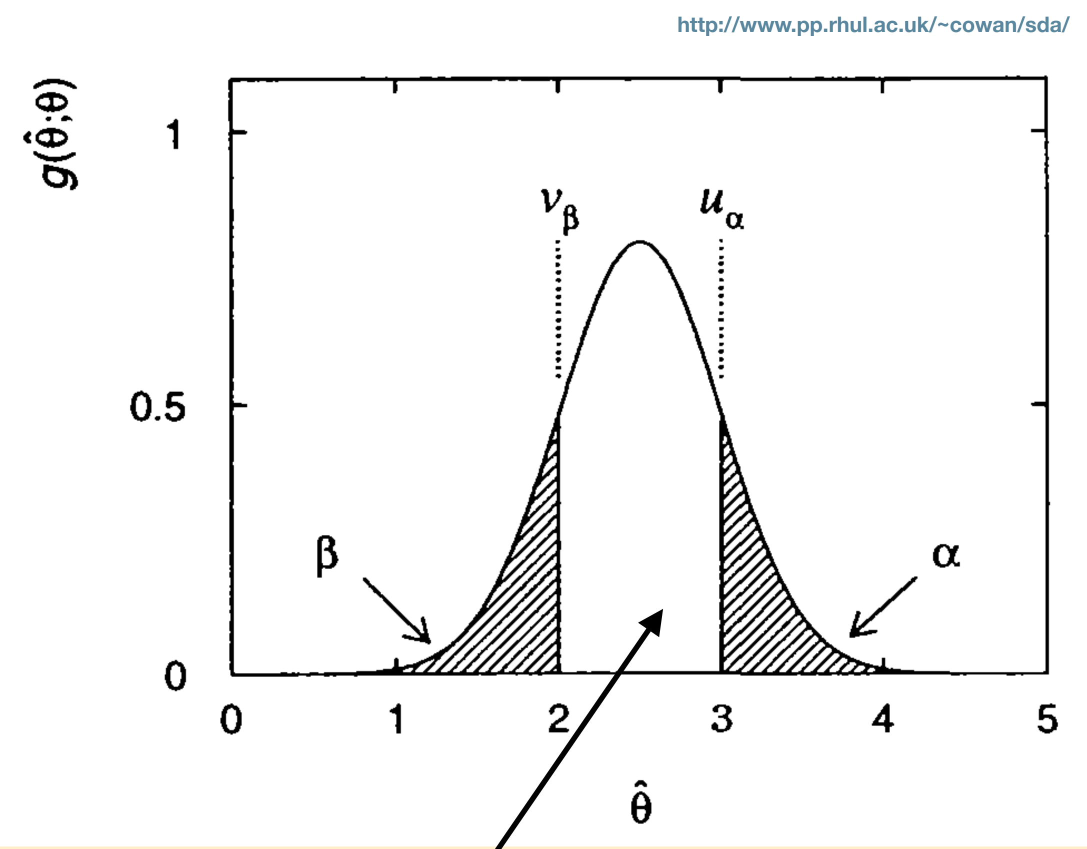
$$\hat{\theta}(x_1, \ldots, x_n) = \hat{\theta}_{\text{obs}}$$

value obtained

- Furthermore, suppose we know the PDF of $\hat{\theta}$ denoted by $g(\hat{\theta}; \theta)$

  Real value of θ unknown, BUT for a given θ one knows what the PDF of $\hat{\theta}$ would be

- From $g(\hat{\theta}; \theta)$, can determine ν_β and u_α such that there are fixed probabilities β and α to observe $\hat{\theta} < \nu_\beta$ or $\hat{\theta} > u_\alpha$



$$P(v_\beta(\theta) \le \hat{\theta} \le u_\alpha(\theta)) = 1 - \alpha - \beta.$$

Shows the probability density for an estimator $\hat{\theta}$ for a particular value of the true parameter θ

u_α and u_β depend on the true value θ and are thus determined by

$$\boxed{\beta = P(\hat{\theta} \le v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta}; \theta) d\hat{\theta} = G(v_\beta(\theta); \theta),}$$

$$\boxed{\alpha = P(\hat{\theta} \ge u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta}; \theta) d\hat{\theta} = 1 - G(u_\alpha(\theta); \theta),}$$

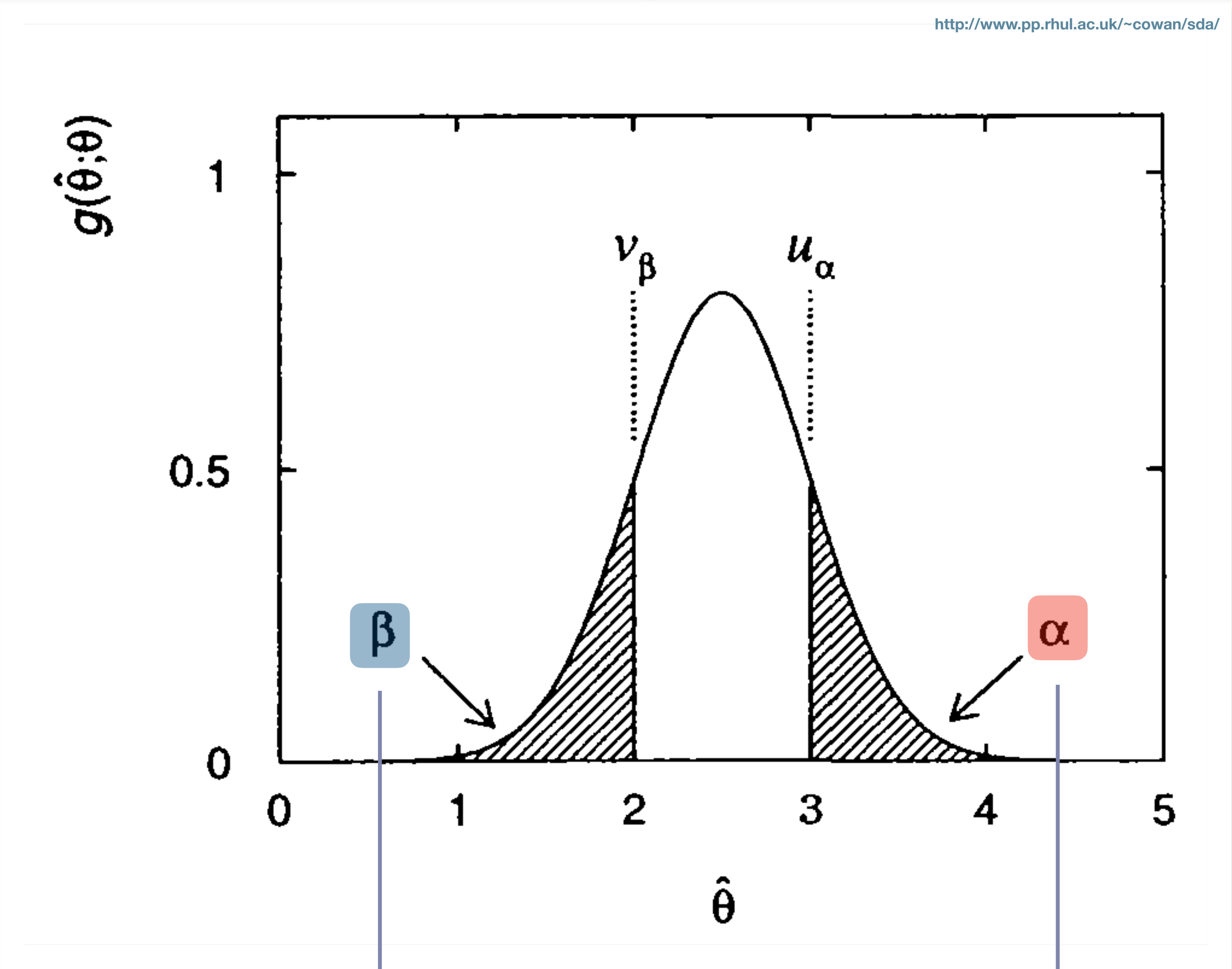
CDF

... so α and β are the probabilities!

Next: lets build the CI step by step…

Confidence Belt (i)

Review

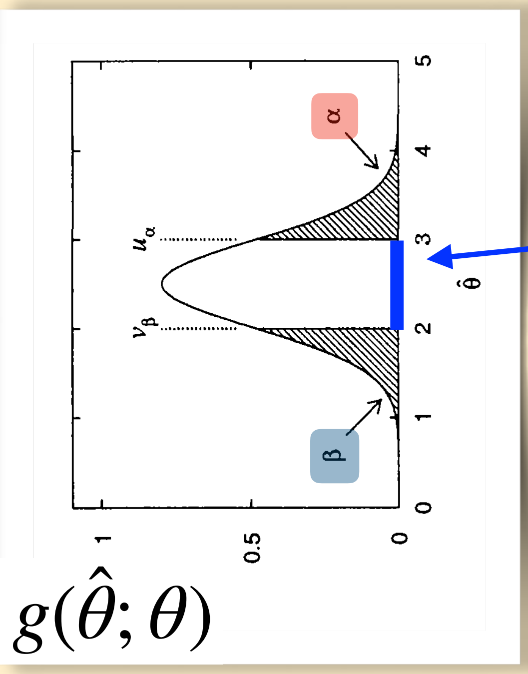


$$\beta = P(\hat{\theta} \le v_\beta(\theta)) = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\,d\hat{\theta} = G(v_\beta(\theta);\theta),$$

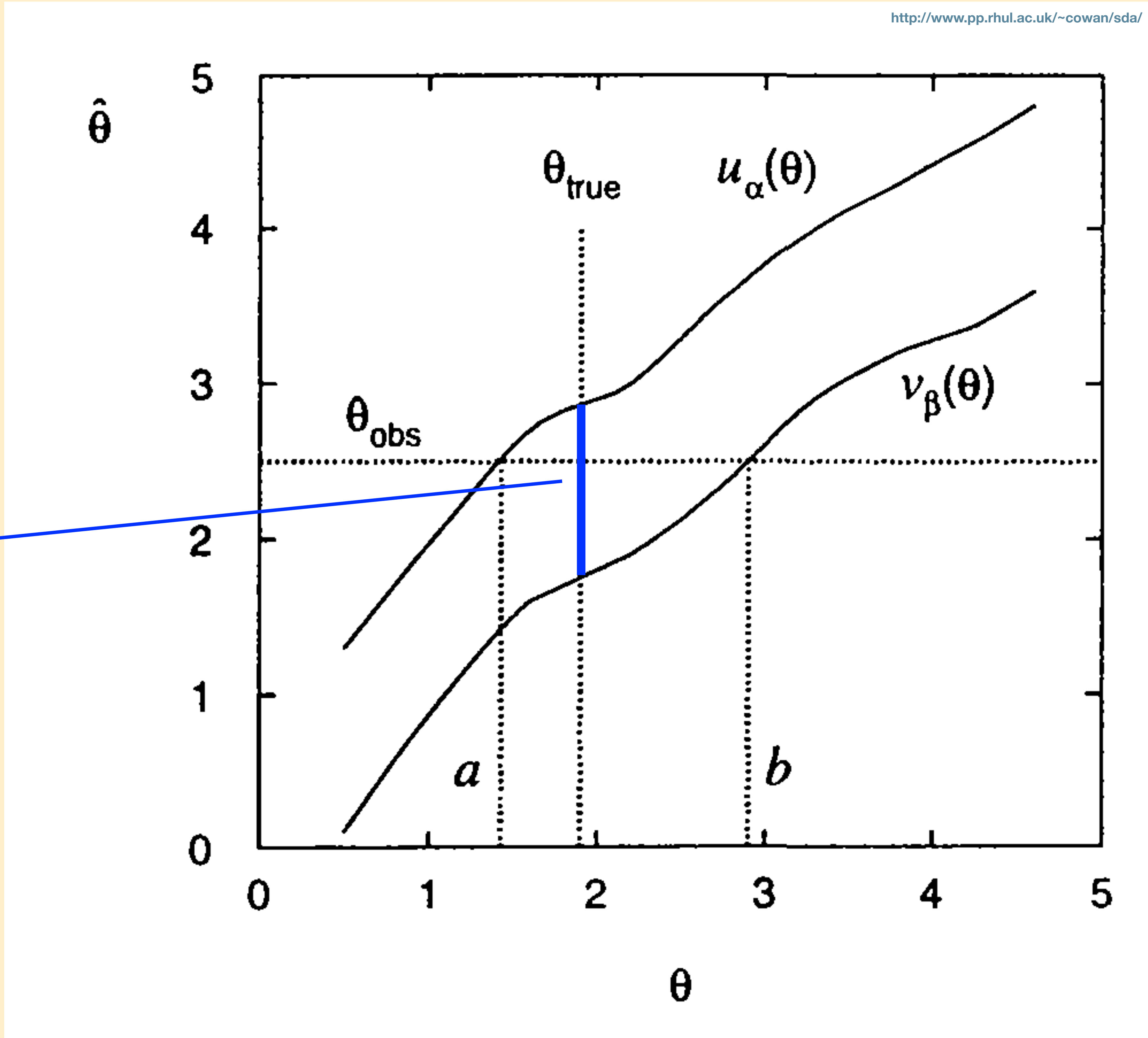
$$\alpha = P(\hat{\theta} \ge u_\alpha(\theta)) = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\,d\hat{\theta} = 1 - G(u_\alpha(\theta);\theta),$$

Confidence Belt (ii)

Review

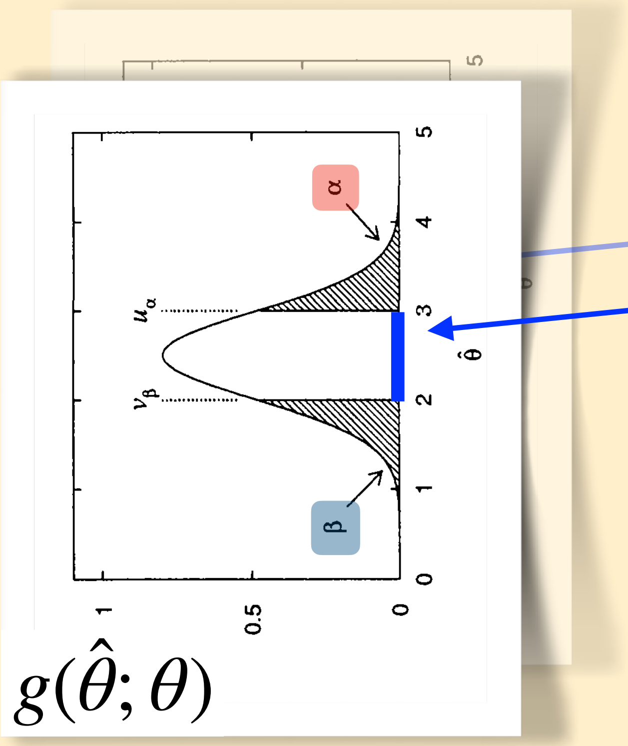


for a given value of (true) θ

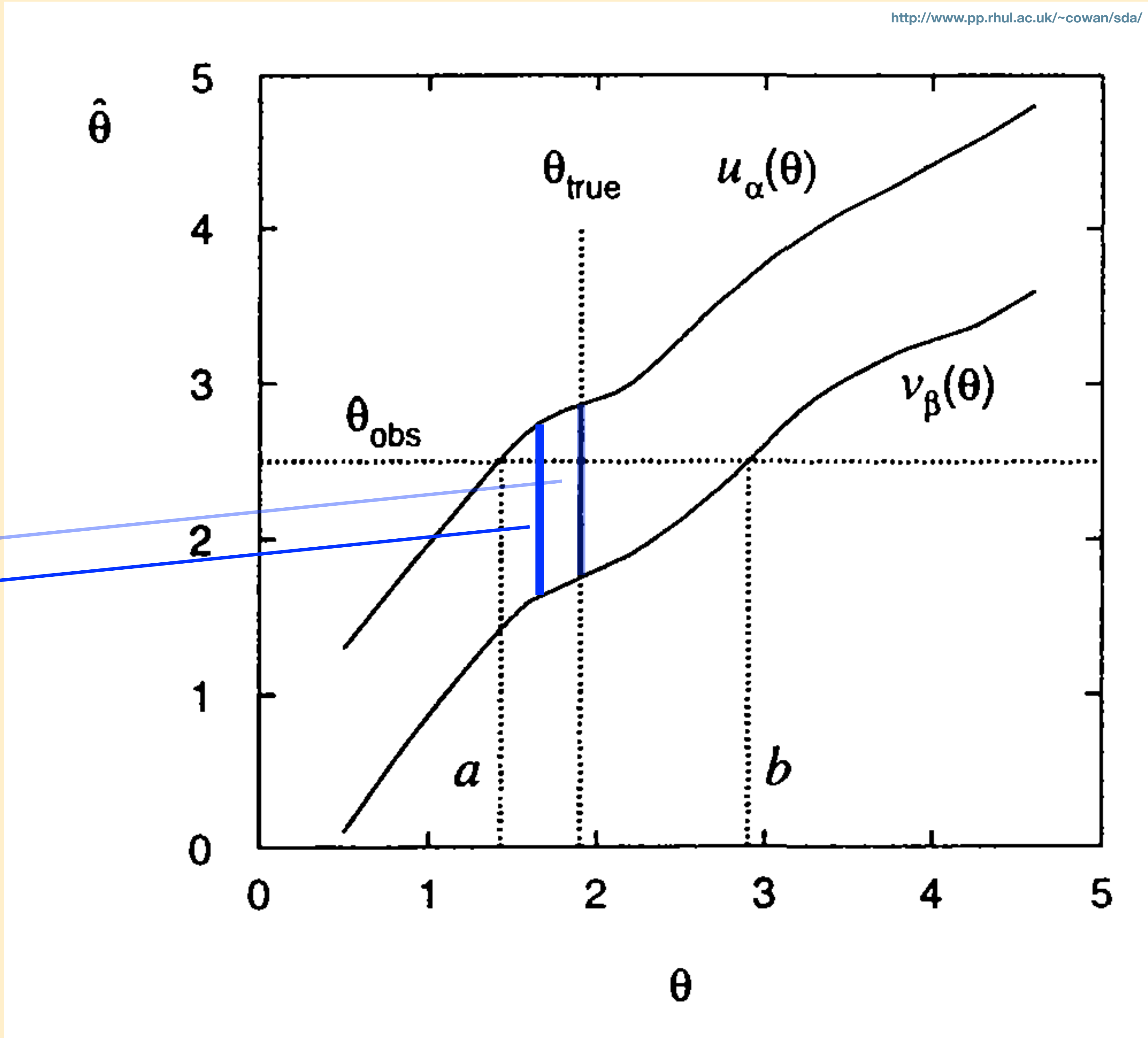


Confidence Belt (iii)

Review



for *another* given value of (true) θ



Confidence Belt (iv)

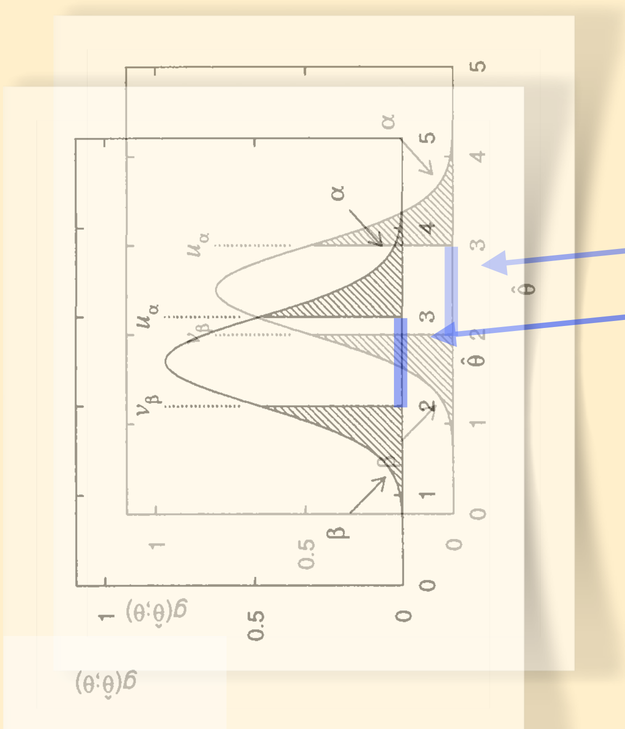
Region between the curves:

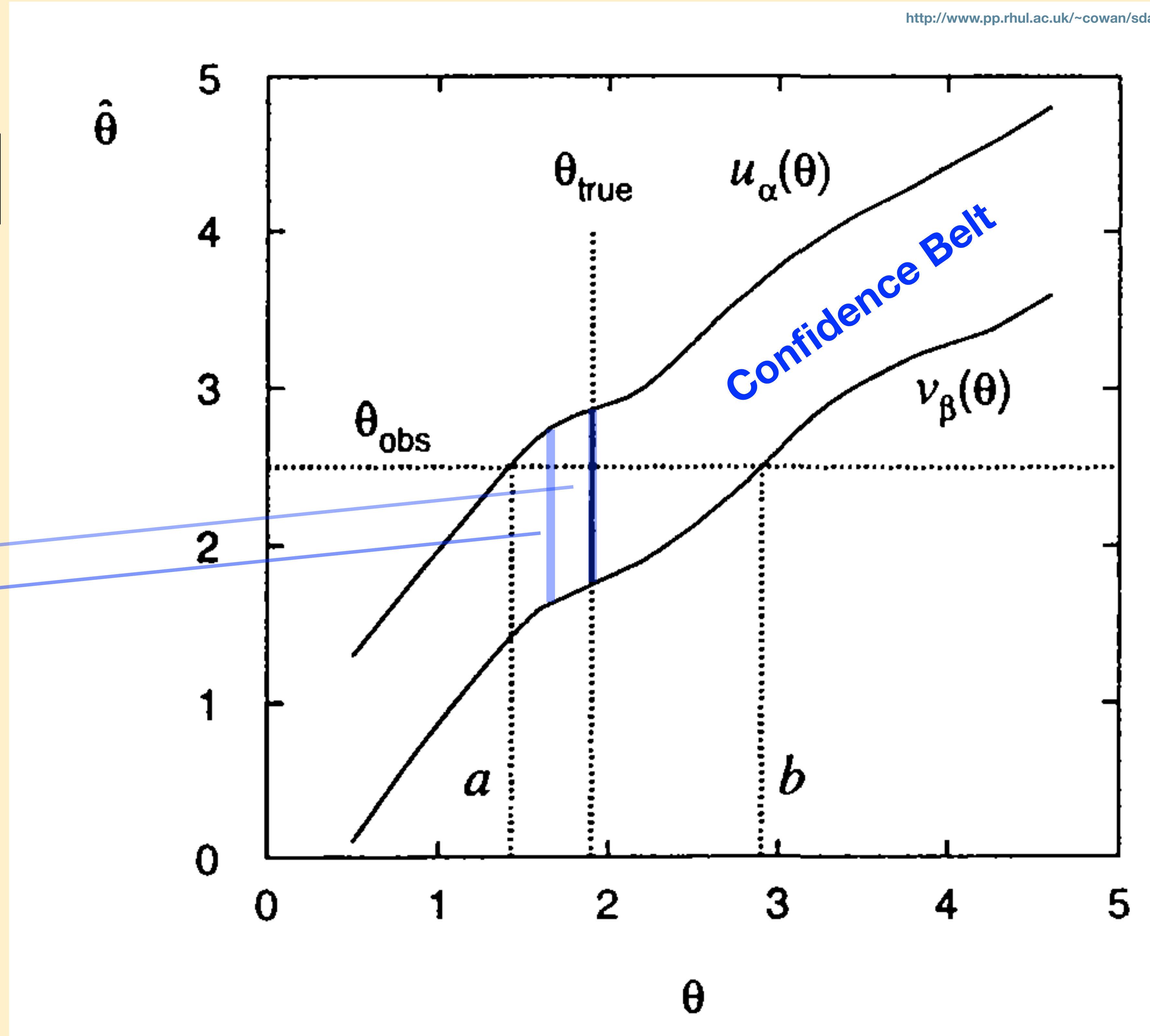
Confidence Belt

(Neyman Belt)

$$P(v_\beta(\theta) \leq \hat{\theta} \leq u_\alpha(\theta)) = 1 - \alpha - \beta.$$

The probability for the estimator $\hat{\theta}$ to be inside the belt, regardless of the value of θ





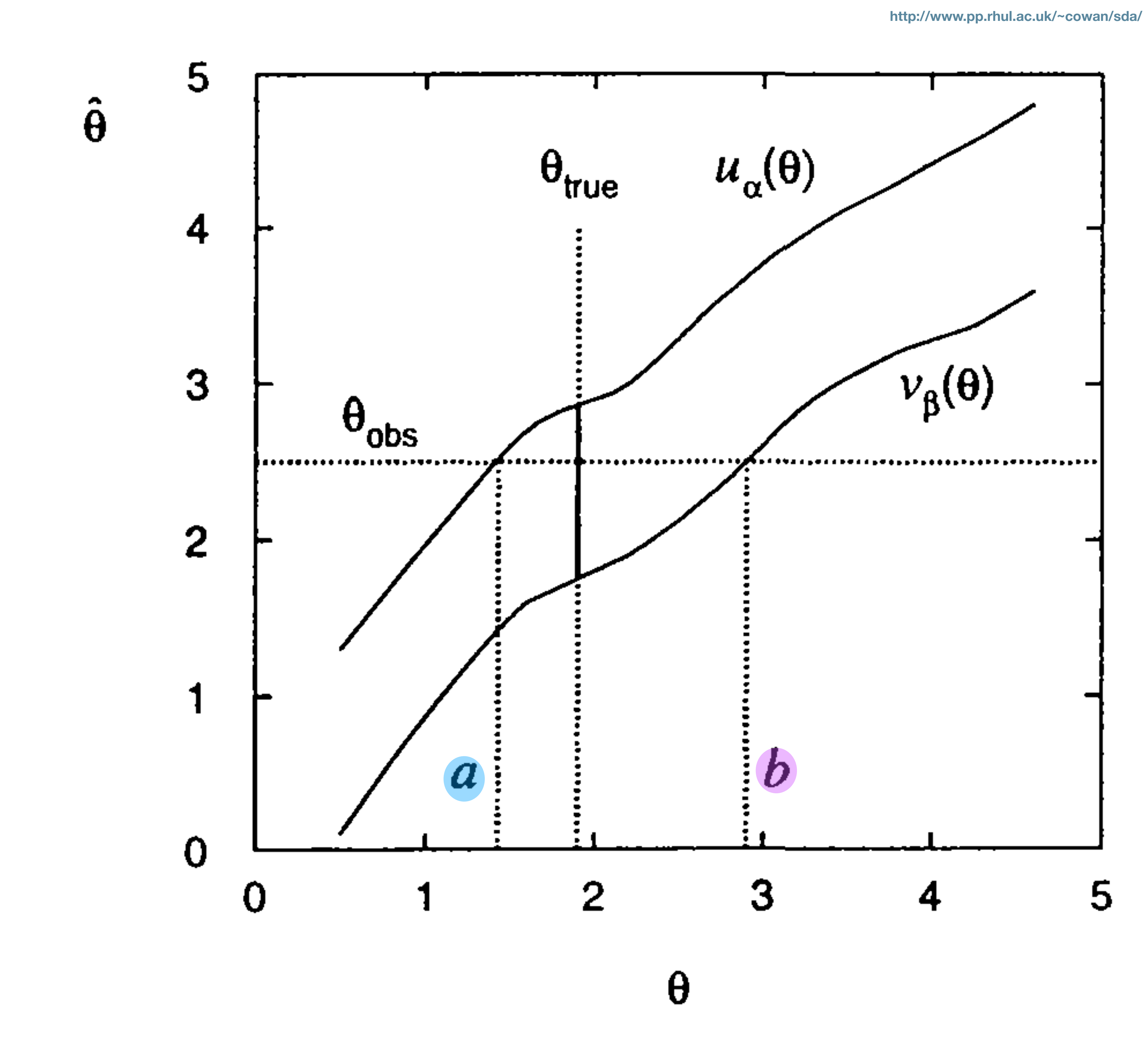
Confidence Interval (i)

Review

If $u_\alpha(\theta)$ and $\nu_\beta(\theta)$ are monotonically increasing functions of θ, then one can determine the inverse functions

$$a(\hat{\theta}) \equiv u_\alpha^{-1}(\hat{\theta}),$$
$$b(\hat{\theta}) \equiv v_\beta^{-1}(\hat{\theta}).$$

(Should be the case if $\hat{\theta}$ is a good estimator for θ)



Confidence Interval (ii)

Review

$$a(\hat\theta) \equiv u_\alpha^{-1}(\hat\theta),$$
$$b(\hat\theta) \equiv v_\beta^{-1}(\hat\theta).$$

This then implies:

$$\boxed{\begin{aligned}\hat\theta &\geq u_\alpha(\theta),\\ \hat\theta &\leq v_\beta(\theta),\end{aligned}} \xrightarrow[\text{invert}]{} \boxed{\begin{aligned}a(\hat\theta) &\geq \theta,\\ b(\hat\theta) &\leq \theta.\end{aligned}} \boxed{\begin{aligned}P(a(\hat\theta) \geq \theta) &= \alpha,\\ P(b(\hat\theta) \leq \theta) &= \beta,\end{aligned}}$$

or

$$\boxed{P(v_\beta(\theta) \leq \hat\theta \leq u_\alpha(\theta)) = 1 - \alpha - \beta.}$$

$$\boxed{P(a(\hat\theta) \leq \theta \leq b(\hat\theta)) = 1 - \alpha - \beta.}$$

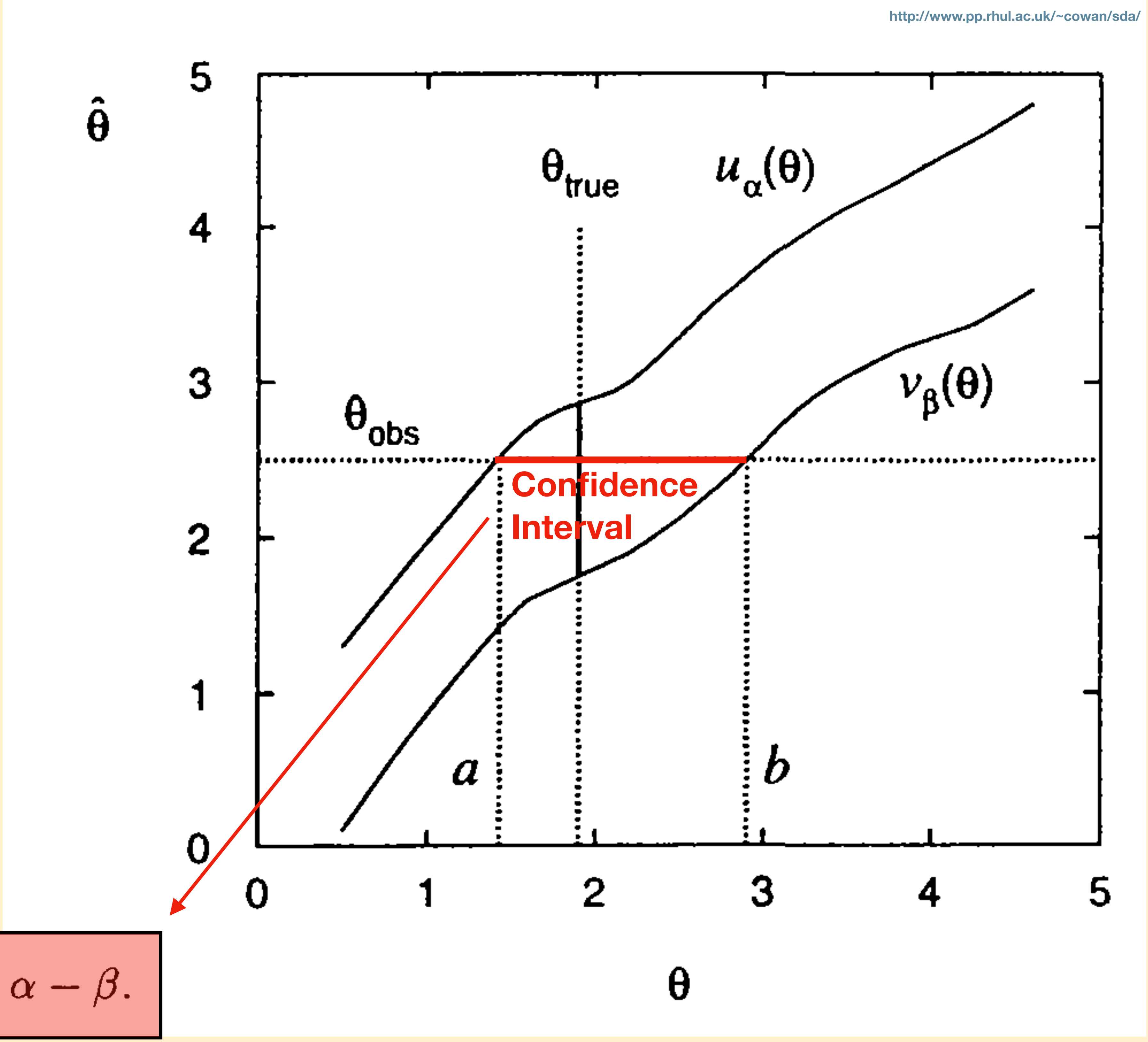
If the functions $a(\hat\theta)$ and $b(\hat\theta)$ are evaluated with the value of the estimator obtained in the experiment ($\hat\theta_{\mathrm{obs}}$), then this determines 2 values $[a, b]$

Confidence Interval (iii)

Review

Often chooses $\alpha = \beta = \dfrac{\gamma}{2}$

giving a so-called **central CI** with probability $= 1 - \gamma$



$[a, b]$: **Confidence Interval**, at a **confidence level** (or **coverage probability**) of $\mathbf{1 - \alpha - \beta}$

$$P(a(\hat{\theta}) \leq \theta \leq b(\hat{\theta})) = 1 - \alpha - \beta.$$

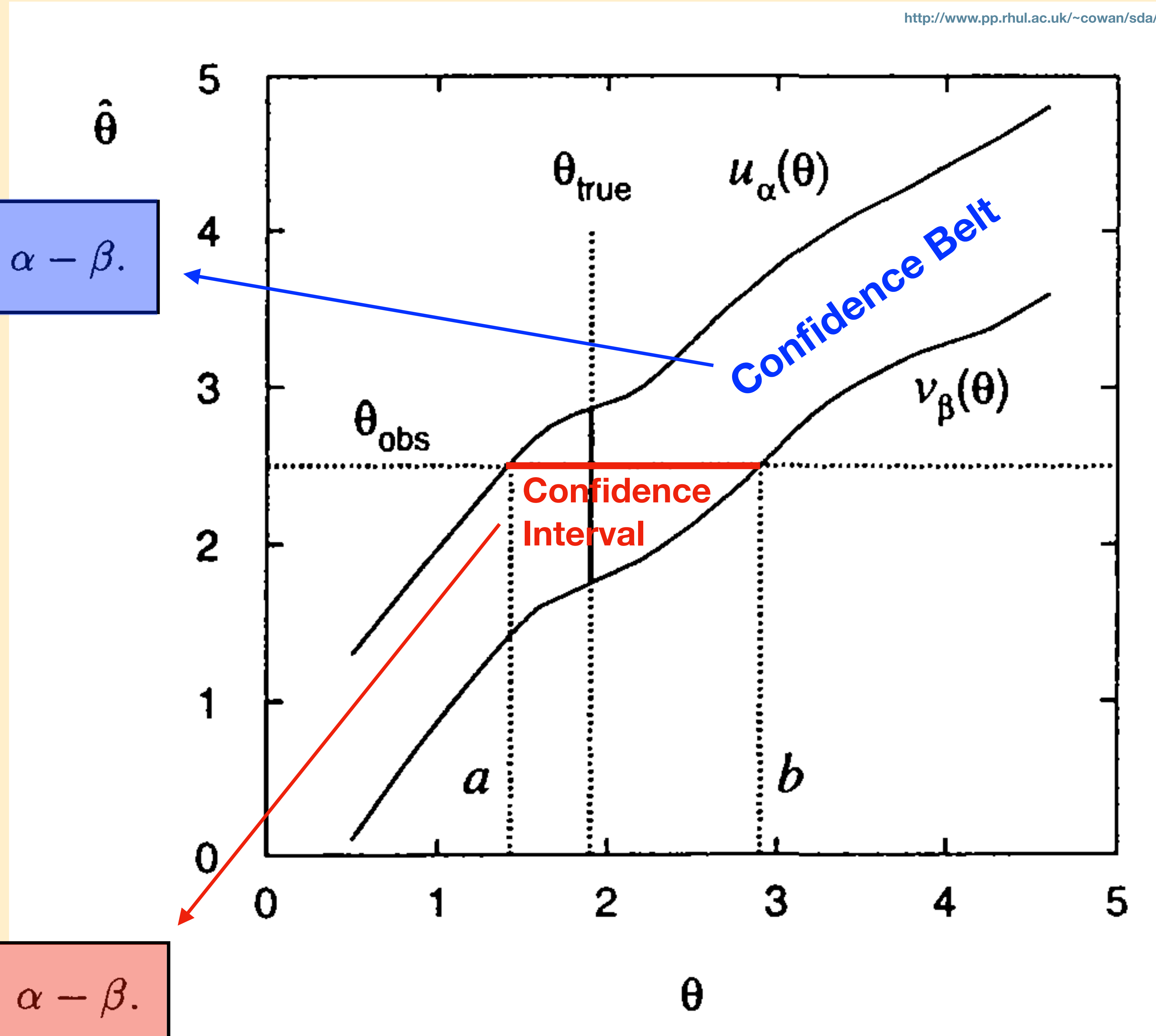
All together now

Review

$$P(v_\beta(\theta) \le \hat\theta \le u_\alpha(\theta)) = 1 - \alpha - \beta.$$

Note where the ^s are in the 2 equations!

$[a, b]$: **Confidence Interval**, at a **confidence level** (or **coverage probability**) of $1 - \alpha - \beta$

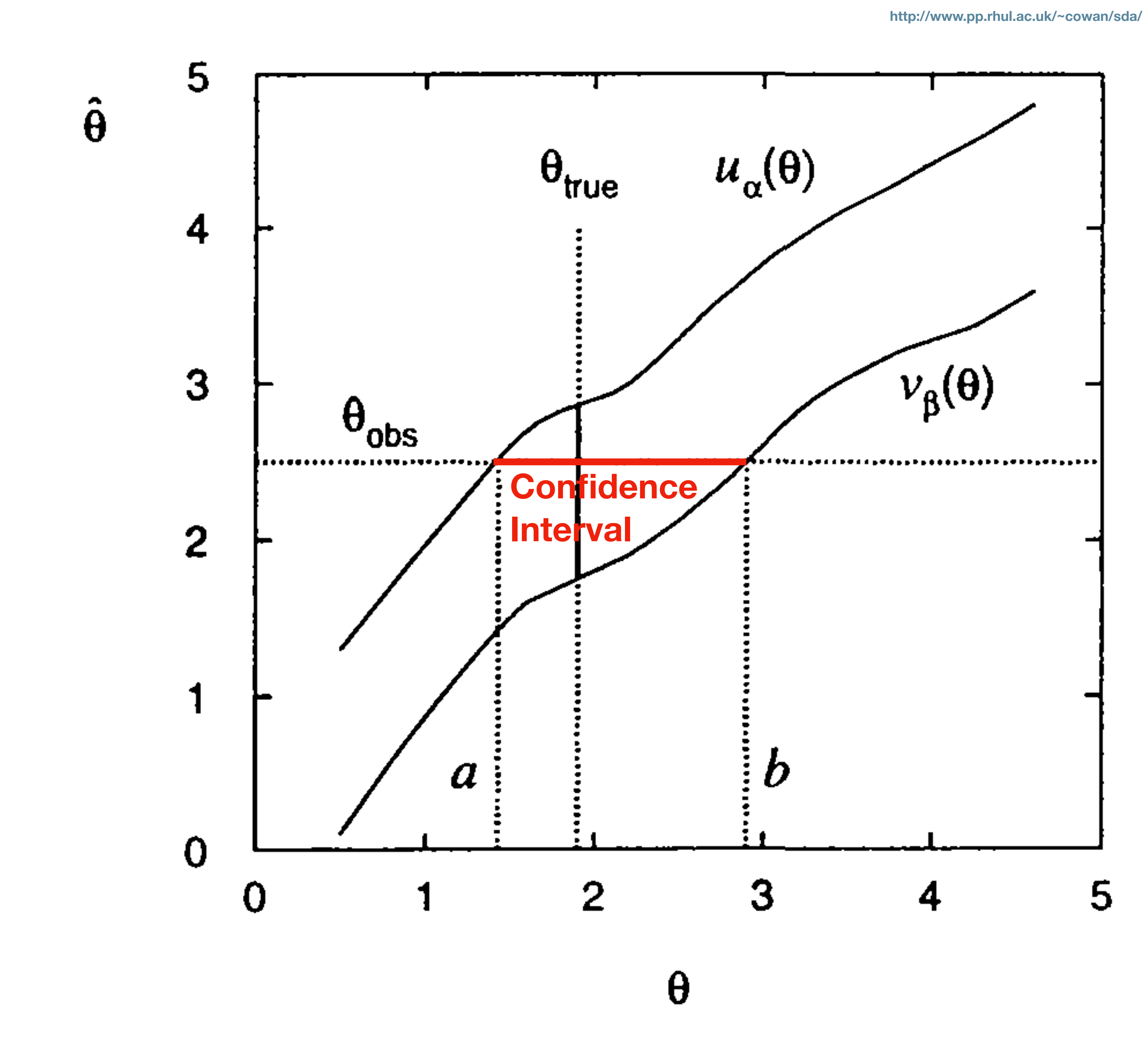


$$P(a(\hat\theta) \le \theta \le b(\hat\theta)) = 1 - \alpha - \beta.$$

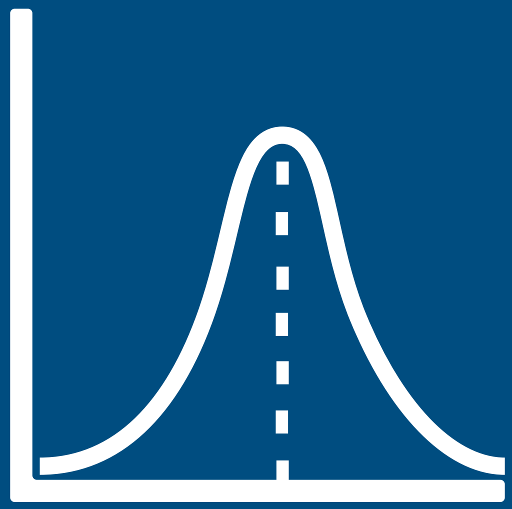
Take home message

Review

If the experiment were repeated many times, the interval $[a, b]$ would include the true value of the parameter θ in a fraction $1 - \alpha - \beta$ of the experiments



Gaussian Confidence Intervals



Let’s apply what we’ve built up to the Gaussian limit

CI for Gaussian distributed estimators (i)

Review

- **Simple and very important application:**
  - $\hat{\theta}$ is Gaussian with mean θ and standard deviation $\sigma_{\hat{\theta}}$
  - Cumulative distribution of $\hat{\theta}$ is then

Commonly occurring situation, since according to the CLT, any estimator that is a linear function of a sum of RVs becomes Gaussian in the large sample limit

$$G(\hat{\theta};\theta,\sigma_{\hat{\theta}}) = \int_{-\infty}^{\hat{\theta}} \frac{1}{\sqrt{2\pi\sigma_{\hat{\theta}}^2}} \exp\left(\frac{-(\hat{\theta}'-\theta)^2}{2\sigma_{\hat{\theta}}^2}\right) d\hat{\theta}' .$$

  - Suppose that the standard deviation is known and that the experiment resulted in an estimate $\hat{\theta}_{\mathrm{obs}}$. Then we can determine the confidence interval $[a,b]$ by solving

$$\alpha = 1 - G(\hat{\theta}_{\mathrm{obs}}; a, \sigma_{\hat{\theta}}) = 1 - \Phi\left(\frac{\hat{\theta}_{\mathrm{obs}} - a}{\sigma_{\hat{\theta}}}\right),$$

$$\beta = G(\hat{\theta}_{\mathrm{obs}}; b, \sigma_{\hat{\theta}}) = \Phi\left(\frac{\hat{\theta}_{\mathrm{obs}} - b}{\sigma_{\hat{\theta}}}\right),$$

standard normal CDF
$\Phi = G(\hat{\mu}\ ;\ \mu = 0,\ \sigma = 1)$

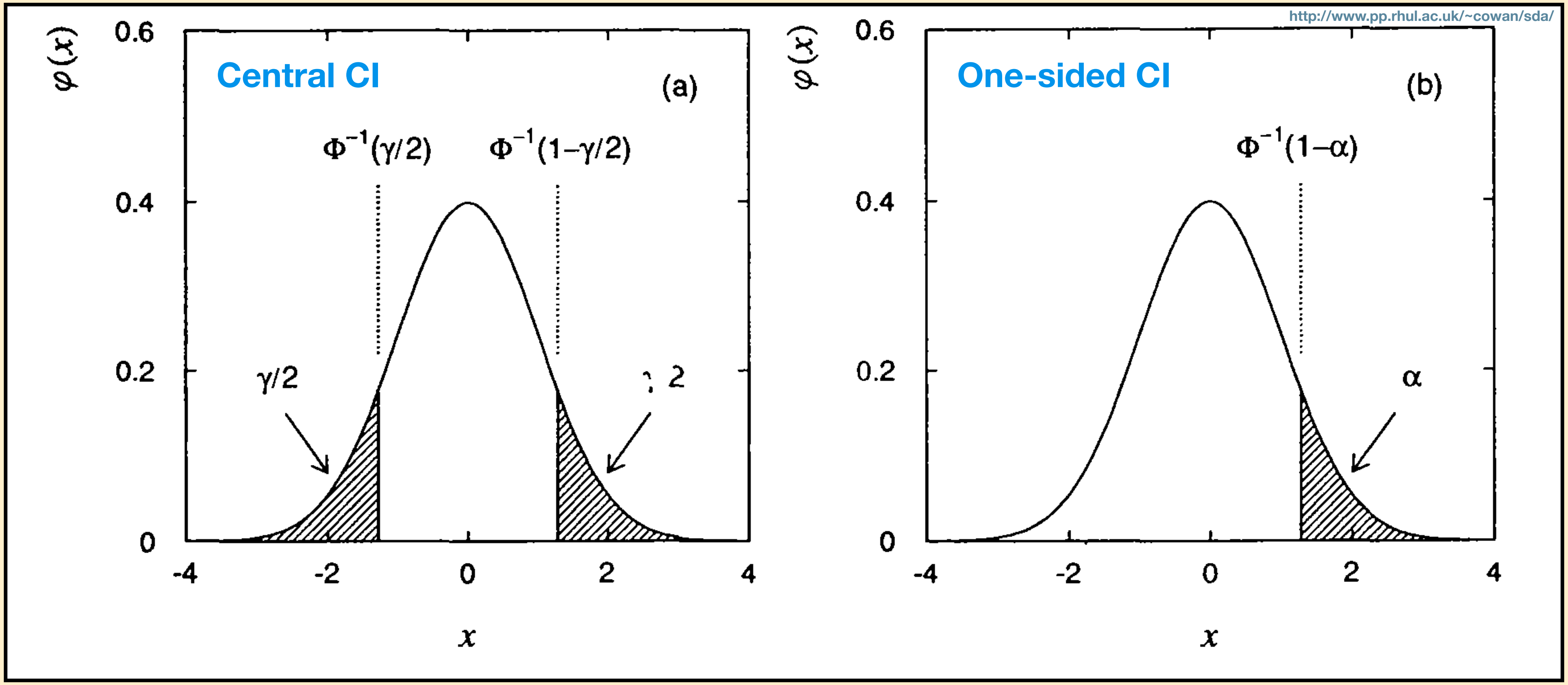
CI for Gaussian distributed estimators *Review* (ii)

- This results in

$$a = \hat{\theta}_{\text{obs}} - \sigma_{\hat{\theta}}\, \Phi^{-1}(1-\alpha),$$
$$b = \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}}\, \Phi^{-1}(1-\beta).$$

inverse of standard normal CDF

i.e., the inverse function of Φ equals the quantile of the std. Gaussian



The relationship between the quantiles of the std. Gaussian distribution and the CI

CI for Gaussian distributed estimators (iii)

Review

- Consider a central confidence interval with $\alpha = \beta = \gamma/2$
  - The confidence level $(1-\gamma)$ is often chosen, such that $\Phi^{-1}(1-\gamma/2)$ is a small integer (e.g., 1,2,3)
  - Similarly, one-sided intervals are often small integer values
    - Sometimes one also prefers to use a round value for $1-\alpha$ or $1-\gamma$

$\Phi^{-1}(1-\gamma/2)$	$1-\gamma$	$\Phi^{-1}(1-\alpha)$	$1-\alpha$
1	0.6827	1	0.8413
2	0.9544	2	0.9772
3	0.9973	3	0.9987

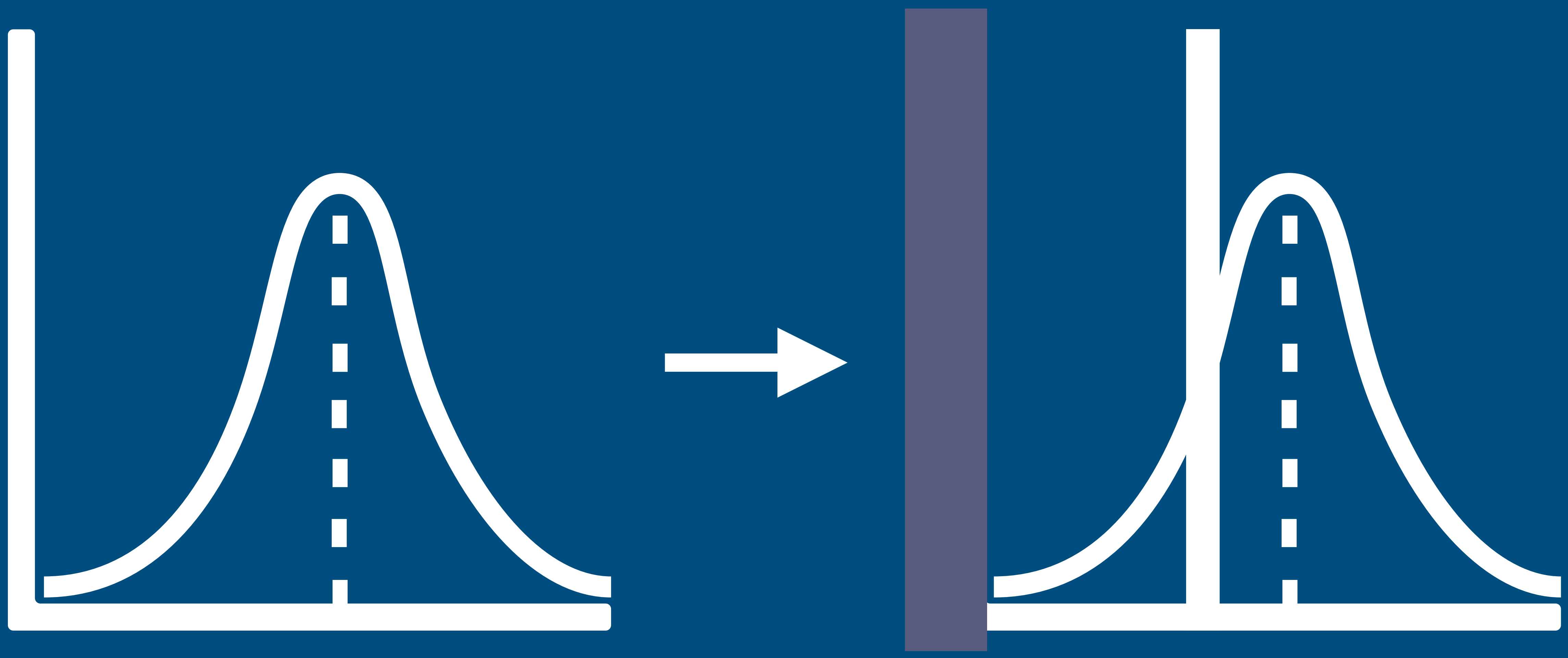
$1-\gamma$	$\Phi^{-1}(1-\gamma/2)$	$1-\alpha$	$\Phi^{-1}(1-\alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326

- For conventional 68.3% CI one has

$$[a, b] = [\hat{\theta}_{\mathrm{obs}} - \sigma_{\hat{\theta}}, \hat{\theta}_{\mathrm{obs}} + \sigma_{\hat{\theta}}].$$

- All of this is valid, if $\sigma_{\hat{\theta}}$ is known
  - Often not the case, *but **in large n limit** can use* $\sigma_{\hat{\theta}} \to \hat{\sigma}_{\hat{\theta}}$

Limits near a physical boundary



Limits near a physical boundary

- Often the **purpose** of an **experiment** is to **search** for a **new effect**
  - E.g. **measure** the **mass of the neutrino**, which in the Standard Model is massless
    - If the data yield a **value** of the **parameter** significantly **different from zero**, then the new effect has been discovered (Hooray!)
      - We know how to quote such a result
    - If, on the other hand, the **data** result in a fitted value of the **parameter** that is **consistent with zero**, then the result of the experiment is **often reported** by giving an **upper** or **lower limit**
  - Difficulties arise though when an estimator can take a value in an unphysical region
    - This can occur if the estimator $\hat{\theta}$ for a parameter θ is of the form

$$\hat{\theta} = x - y \longrightarrow \hat{m}^2 = E^2 - p^2$$

Both RVs

e.g. measure energy and momentum independently

can result in negative masses

Limits near a physical boundary

- How to place a limit on m^2 when the estimate is near an excluded or unphysical region?
- Let's make this more concrete with an example:
  - $\hat{\theta} = x - y$ with x, y Gaussian RVs with mean and variances $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2$
  - The difference is also a Gaussian variable with $\theta = \mu_x - \mu_y$ and $\sigma_{\hat{\theta}}^2 = \sigma_x^2 + \sigma_y^2$
    (see proof in characteristic functions chapter 10 Cowan)
  - Assume that θ is known a priori to be non-negative (e.g. like the mass squared) and suppose the experiment resulted in a value $\hat{\theta}_{\mathrm{obs}}$ for the estimator $\hat{\theta}$
  - According to what we discussed (S17 ***Review***), the upper limit θ_{up} at CL $1-\beta$ is

$$\theta_{\mathrm{up}} = \hat{\theta}_{\mathrm{obs}} + \sigma_{\hat{\theta}}\,\Phi^{-1}(1-\beta)$$

inverse of standard normal CDF

Limits near a physical boundary

- For the commonly used 95% CL one obtains the quantile

$$\Phi^{-1}(0.95) = 1.645$$

Table 9.2 (L07, S33)
http://www.pp.rhul.ac.uk/~cowan/sda/

$1-\gamma$	$\Phi^{-1}(1-\gamma/2)$	$1-\alpha$	$\Phi^{-1}(1-\alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326

- The interval $(-\infty, \theta_{\mathrm{up}}]$ is constructed to include the true value θ with a probability of 95%, *independent of the true value.*
- Let's now suppose the standard deviation $\sigma_{\hat{\theta}} = 1$ and the observed value from the experiment is $\hat{\theta}_{\mathrm{obs}} = -2.0$
  - Using $\theta_{\mathrm{up}} = \hat{\theta}_{\mathrm{obs}} + \sigma_{\hat{\theta}}\,\Phi^{-1}(1-\beta)$ we obtain $\boxed{\theta_{\mathrm{up}} = -0.355 \text{ at 95\% CL}}$
- Not only is the observed value in the unphysical region (half of the estimates actually should be if θ is zero), *but the upper limit is below zero as well*
  - Not particularly unusual; we expect 5% of all experiments to report this if θ is zero.

Nothing went wrong!

- As far as the definition of CL is concerned, **nothing fundamental has gone wrong**.
    - The **interval** was **designed** to **cover** the **true value of θ** in a **certain fraction of repeated experiments**, and we have obviously encountered one of those experiments where θ is not in the interval
    - But many people **don't find this very satisfying**, since we already know from physical reasons that θ is greater than zero (and certainly greater than $\theta_{\text{up}} = -0.355$) without having to perform an experiment.
- **Regardless of** the **upper limit**, it is **important to report** the **actual value** of the estimate obtained and its standard deviation, i.e.

$$\hat{\theta}_{\text{obs}} \pm \sigma_{\hat{\theta}}$$ **or if Errors are non-Gaussian: the likelihood function $\mathscr{L}(\boldsymbol{\theta})$**

    - In this way, the **average of many experiments will converge to the correct value** as long as the estimator is unbiased.

Upper Limit Bonanza

Table 9.2 (L07, S33)

http://www.pp.rhul.ac.uk/~cowan/sda/

$1-\gamma$	$\Phi^{-1}(1-\gamma/2)$	$1-\alpha$	$\Phi^{-1}(1-\alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326

- **Nevertheless**, most experimenters want to report some sort of upper limit, that takes into account the knowledge of the unphysical region.
  - Many different solutions have been proposed, but there is no established convention on how this should be done. **So it's imperative to state what procedure you used. Otherwise people will not be able to combine or use your result.**
- To come back to our example: $\hat{\theta}_{\text{obs}} = -2.0, \sigma_{\hat{\theta}} = 1$
  - One might **feel tempted** to just **quote** a **limit at a higher CL**, e.g. 99% would result in $\theta_{\text{up}} = 0.326$ ($\Phi^{-1}(0.99) = 2.326$)
  - This would lead to an **upper limit better than** the **intrinsic resolution** of our **experiment** ($\sigma_{\hat{\theta}} = 1$) at a v**ery high confidence level** of 99%
    - *This is a bit misleading…*
  - But even worse would be to adjust the CL to give an arbitrary small limit, $\theta_{\text{up}} = 10^{-5}$ at 97.725 CL%

Alternative approaches: Max method

- In order to avoid such difficulties, a commonly used technique is to simply shift a negative estimate to zero before determining the value, i.e.

$$\theta_{\text{up}} = \max(\hat{\theta}_{\text{obs}}, 0) + \sigma_{\hat{\theta}} \, \Phi^{-1}(1 - \beta)$$

  - This way the upper limit is always at least the same order of magnitude as the resolution of the experiment
    - If $\hat{\theta}_{\text{obs}}$ is positive, nothing changes and the upper limit coincides with the classical procedure. (See Fig. on slide 30.)
  - This technique has a certain intuitive appeal and is often used, but the **interpretation** as an **interval that will cover the true parameter** with a probability $1 - \beta$ **no longer applies**.
    - The coverage probability is clearly larger than $1 - \beta$ (one speaks of **over-coverage**)

Alternative approaches: Bayesian limit

- Another alternative is to report an interval based on the **Bayesian posterior PDF** $p(\theta|\boldsymbol{x})$, obtained via

Likelihood function

Prior PDF of θ

(Reflects the state of knowledge of θ before consideration of the data)

$$p(\theta|\boldsymbol{x}) = \frac{\mathscr{L}(\boldsymbol{x}\,|\,\theta)\,\pi(\theta)}{\int \mathscr{L}(\boldsymbol{x}\,|\,\theta')\,\pi(\theta')\,\mathrm{d}\theta'}$$

Observed data

- We now can use $p(\theta|\boldsymbol{x})$ to determine an interval $[a, b]$ such that for given probabilities α and β one has

$$\boxed{\alpha = \int_{-\infty}^{a} p(\theta|\mathbf{x})\,d\theta} \qquad \boxed{\beta = \int_{b}^{\infty} p(\theta|\mathbf{x})\,d\theta.}$$

Alternative approaches: Bayesian limit

- Choosing $\alpha = \beta$ gives a central interval with e.g. $1 - \alpha - \beta = 68.3\,\%$
- Another possibility is to choose α and β s.t. all **values of** $p(\theta\,|\,\boldsymbol{x})$ inside the interval $[a, b]$ are **higher than any values outside**, which implies $p(a\,|\,\boldsymbol{x}) = p(b\,|\,\boldsymbol{x})$. One can show that this gives the **shortest possible interval**.
- One advantage of the Bayesian interval, is that the prior knowledge, e.g. $\theta \geq 0$ can easily be incorporated by setting the prior PDF to zero in the excluded region.
  - Bayes' Theorem then gives a posterior probability $p(\theta\,|\,\boldsymbol{x})$ with $p(\theta\,|\,\boldsymbol{x}) = 0$ for $\theta < 0$. The upper limit thus is given by

$$1 - \beta = \int_{-\infty}^{\theta_{\mathrm{up}}} p(\theta|\mathbf{x})d\theta = \frac{\int_{-\infty}^{\theta_{\mathrm{up}}} L(\mathbf{x}|\theta)\,\pi(\theta)\,d\theta}{\int_{-\infty}^{\infty} L(\mathbf{x}|\theta)\,\pi(\theta)\,d\theta}.$$

Bayesian limit: constant prior

- The difficulties with this approach is that there is **no unique way** to **specify** the **prior density** $\pi(\theta)$. A **common choice** is:

$$\boxed{\pi(\theta) = \begin{cases} 0 & \theta < 0 \\ 1 & \theta \geq 0. \end{cases}}$$

- I.e.: Normalize the likelihood function to unit area in the physical region, and then integrate it out to θ_{up} s.t. the fraction of the area covered is $1 - \beta$.
  - **Although the method is simple, it has some conceptual drawbacks:**
    - For the case where one knows $\theta \geq 0$ (e.g. Neutrino mass), one does not really believe that $0 < \theta < 1$ has the same prior probability as $10^{40} < \theta < 10^{40} + 1$
    - Furthermore the upper limit derived from $\pi(\theta)$ **= const.** is **not invariant with respect to a nonlinear transformation of the parameter**.

Bayesian limit: Jeffreys prior

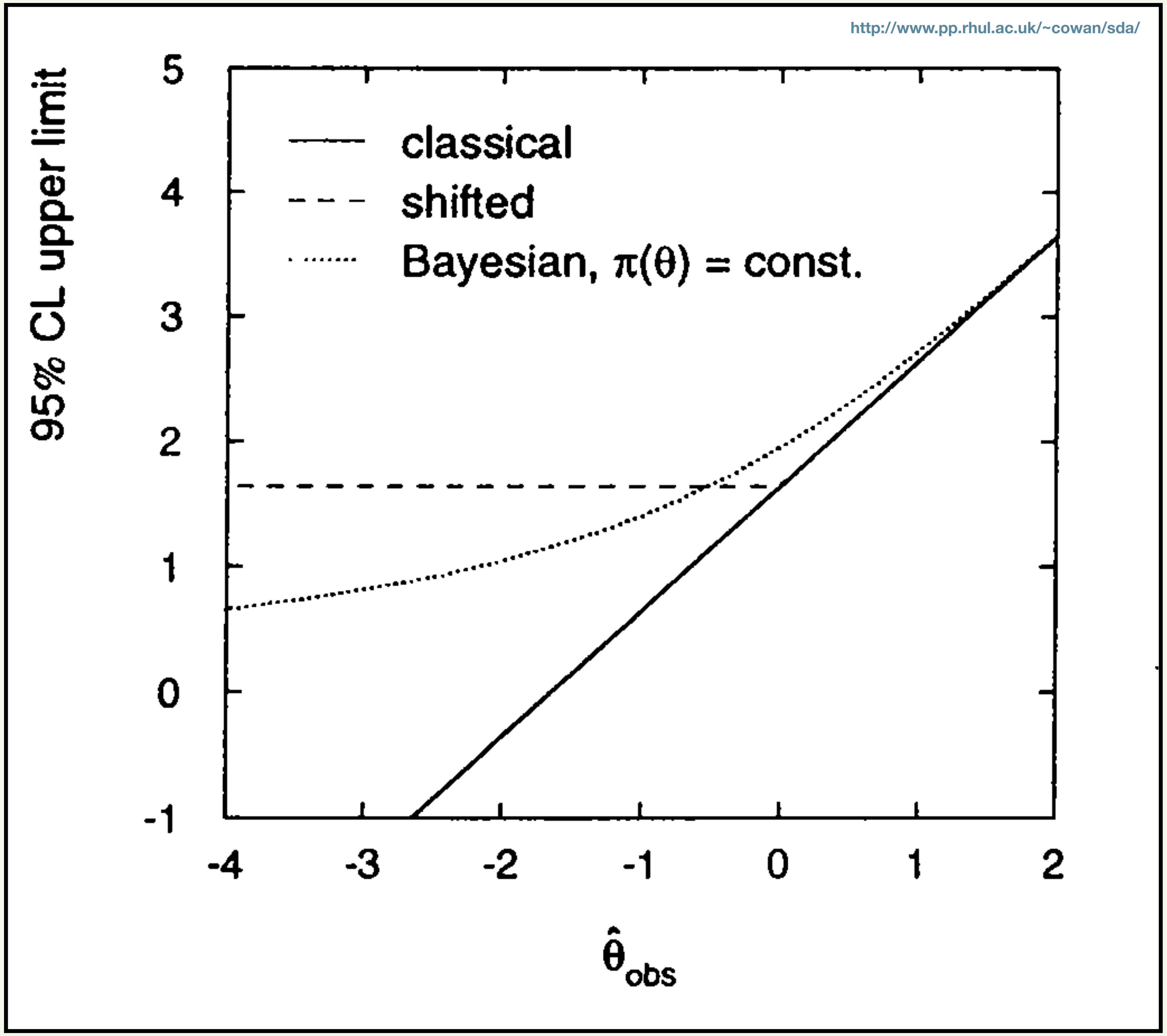
- It has been argued that in cases where $\theta \geq 0$ *but no further information*, one should use

$$\pi(\theta) = \begin{cases} 0 & \theta \leq 0 \\ \frac{1}{\theta} & \theta > 0. \end{cases}$$

  - This has the advantage that upper limits are invariant with respect to a transformation of the parameter by raising to an arbitrary power. This is equivalent to a uniform (improper) prior of previous form for $\log \theta$.
    - For this to be usable, however, the **likelihood function** must go to **zero for $\theta \to 0$ and $\theta \to \infty$,** or else the **integrals diverge**. Thus this description is **often not applicable**.
    - Therefore the uniform prior density (previous slide) is the most commonly used choice for setting limits on parameters.

Different approaches compared

- Comparison of the three methods:
  - Classical and shifted are equal for $\hat{\theta}_{\text{obs}} \geq 0$;
  - The Bayesian limit (here a constant prior is used) is always positive, and is always $>$ the classical limit;
  - As the observed value grows, all limits approach each other.



From one of our papers

PHYSICAL REVIEW D **101,** 032007 (2020)

Search for $B^+ \to \mu^+\nu_\mu$ and $B^+ \to \mu^+ N$ with inclusive tagging

M. T. Prim,[32] F. U. Bernlochner,[3] P. Goldenzweig,[32] M. Heck,[32] I. Adachi,[18,15] K. Adamczyk,[62] H. Aihara,[85] S. Al Said,[79,34] D. M. Asner,[4] H. Atmacan,[76] V. Aulchenko,[5,65] T. Aushev,[54] R. Ayad,[79] V. Babu,[9] A. M. Bakich,[78] V. Bansal,[67] P. Behera,[25] C. Beleño,[14] V. Bhardwaj,[22] B. Bhuyan,[23] T. Bilka,[6] J. Biswal,[31] A. Bobrov,[5,65] A. Bozek,[62] M. Bračko,[48,31] N. Braun,[32] T. E. Browder,[17] M. Campajola,[29,57] L. Cao,[32] D. Červenkov,[6] P. Chang,[61] V. Chekelian,[49] A. Chen,[59] B. G. Cheon,[16] K. Chilikin,[42] H. E. Cho,[16] K. Cho,[36] Y. Choi,[77] S. Choudhury,[24] D. Cinabro,[89] S. Cunliffe,[9] Z. Doležal,[6] S. Eidelman,[5,65,42] D. Epifanov,[5,65] J. E. Fast,[67] T. Ferber,[9] B. G. Fulsom,[67] R. Garg,[68] V. Gaur,[88] A. Garmash,[5,65] A. Giri,[24] O. Grzymkowska,[62] Y. Guan,[8] J. Haba,[18,15] T. Hara,[18,15] K. Hayasaka,[64] H. Hayashii,[58] W.-S. Hou,[61] T. Iijima,[56,55] K. Inami,[55] G. Inguglia,[9] A. Ishikawa,[18] M. Iwasaki,[66] Y. Iwasaki,[18] S. Jia,[2] Y. Jin,[85] D. Joffe,[33] K. K. Joo,[7] A. B. Kaliyar,[25] G. Karyan,[9] T. Kawasaki,[35] H. Kichimi,[18] C. Kiesling,[49] C. H. Kim,[16] D. Y. Kim,[75] K. T. Kim,[37] S. H. Kim,[16] K. Kinoshita,[8] P. Kodyš,[6] S. Korpar,[48,31] D. Kotchetkov,[17] P. Križan,[44,31] R. Kroeger,[51] P. Krokovny,[5,65] T. Kuhr,[45] R. Kulasiri,[33] R. Kumar,[70] T. Kumita,[87] A. Kuzmin,[5,65] Y.-J. Kwon,[91] K. Lalwani,[47] J. S. Lange,[12] I. S. Lee,[16] J. K. Lee,[73] J. Y. Lee,[73] S. C. Lee,[39] P. Lewis,[17] C. H. Li,[43] L. Li Gioi,[49] J. Libby,[25] K. Lieret,[45] D. Liventsev,[88,18] P.-C. Lu,[61] T. Luo,[11] J. MacNaughton,[52] M. Masuda,[84] T. Matsuda,[52] D. Matvienko,[5,65,42] M. Merola,[29,57] F. Metzner,[32] K. Miyabayashi,[58] R. Mizuk,[42,54] G. B. Mohanty,[80] R. Mussa,[30] M. Nakao,[18,15] G. De Nardo,[29,57] K. J. Nath,[23] Z. Natkaniec,[62] M. Nayak,[89,18] M. Niiyama,[38] N. K. Nisar,[69] S. Nishida,[18,15] K. Nishimura,[17] S. Ogawa,[82] H. Ono,[63,64] Y. Onuki,[85] P. Pakhlov,[42,53] G. Pakhlova,[42,54] B. Pal,[4] S. Pardi,[29] H. Park,[39] S.-H. Park,[91] S. Patra,[22] S. Paul,[81] T. K. Pedlar,[46] R. Pestotnik,[31] L. E. Piilonen,[88] V. Popov,[42,54] E. Prencipe,[20] M. Ritter,[45] A. Rostomyan,[9] M. Rozanska,[62] G. Russo,[57] D. Sahoo,[80] Y. Sakai,[18,15] L. Santelj,[18] V. Savinov,[69] O. Schneider,[41] G. Schnell,[1,21] J. Schueler,[17] C. Schwanda,[27] Y. Seino,[64] K. Senyo,[90] M. E. Sevior,[50] V. Shebalin,[17] J.-G. Shiu,[61] B. Shwartz,[5,65] F. Simon,[49] A. Sokolov,[28] E. Solovieva,[42] M. Starič,[31] J. F. Strube,[67] M. Sumihama,[13] T. Sumiyoshi,[87] W. Sutcliffe,[32] M. Takizawa,[74,19,71] U. Tamponi,[30] Y. Tao,[10] F. Tenchini,[9] K. Trabelsi,[40] M. Uchida,[86] T. Uglov,[42,54] Y. Unno,[16] S. Uno,[18,15] Y. Ushiroda,[18,15] Y. Usov,[5,65] S. E. Vahsen,[17] R. Van Tonder,[32] G. Varner,[17] K. E. Varvell,[78] A. Vinokurova,[5,65] B. Wang,[49] C. H. Wang,[60] M.-Z. Wang,[61] P. Wang,[26] S. Watanuki,[83] E. Won,[37] S. B. Yang,[37] H. Ye,[9] J. H. Yin,[26] Y. Yusa,[64] Z. P. Zhang,[72] V. Zhilich,[5,65] V. Zhukova,[42] and V. Zhulanov[5,65]

(Belle Collaboration)

[1]*University of the Basque Country UPV/EHU, 48080 Bilbao*
[2]*Beihang University, Beijing 100191*
[3]*University of Bonn, 53115 Bonn*
[4]*Brookhaven National Laboratory, Upton, New York 11973*
[5]*Budker Institute of Nuclear Physics SB RAS, Novosibirsk 630090*
[6]*Faculty of Mathematics and Physics, Charles University, 121 16 Prague*
[7]*Chonnam National University, Kwangju 660-701*
[8]*University of Cincinnati, Cincinnati, Ohio 45221*
[9]*Deutsches Elektronen–Synchrotron, 22607 Hamburg*
[10]*University of Florida, Gainesville, Florida 32611*
[11]*Key Laboratory of Nuclear Physics and Ion-beam Application (MOE) and Institute of Modern Physics, Fudan University, Shanghai 200443*
[12]*Justus-Liebig-Universität Gießen, 35392 Gießen*
[13]*Gifu University, Gifu 501-1193*
[14]*II. Physikalisches Institut, Georg-August-Universität Göttingen, 37073 Göttingen*
[15]*SOKENDAI (The Graduate University for Advanced Studies), Hayama 240-0193*
[16]*Hanyang University, Seoul 133-791*
[17]*University of Hawaii, Honolulu, Hawaii 96822*
[18]*High Energy Accelerator Research Organization (KEK), Tsukuba 305-0801*
[19]*J-PARC Branch, KEK Theory Center, High Energy Accelerator Research Organization (KEK), Tsukuba 305-0801*
[20]*Forschungszentrum Jülich, 52425 Jülich*
[21]*IKERBASQUE, Basque Foundation for Science, 48013 Bilbao*
[22]*Indian Institute of Science Education and Research Mohali, SAS Nagar, 140306*
[23]*Indian Institute of Technology Guwahati, Assam 781039*
[24]*Indian Institute of Technology Hyderabad, Telangana 502285*
[25]*Indian Institute of Technology Madras, Chennai 600036*

2470-0010/2020/101(3)/032007(20) 032007-1 Published by the American Physical Society

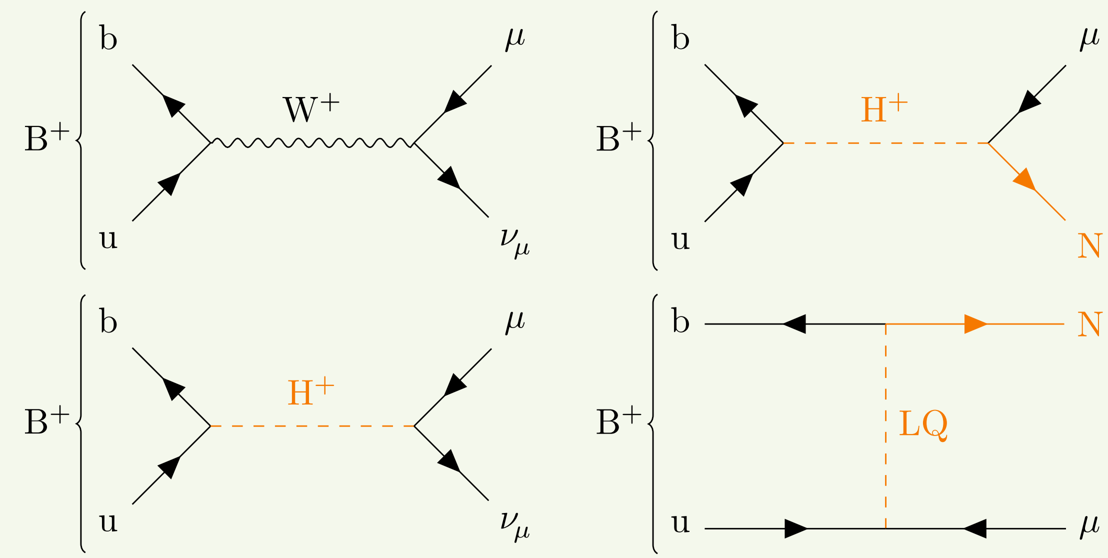
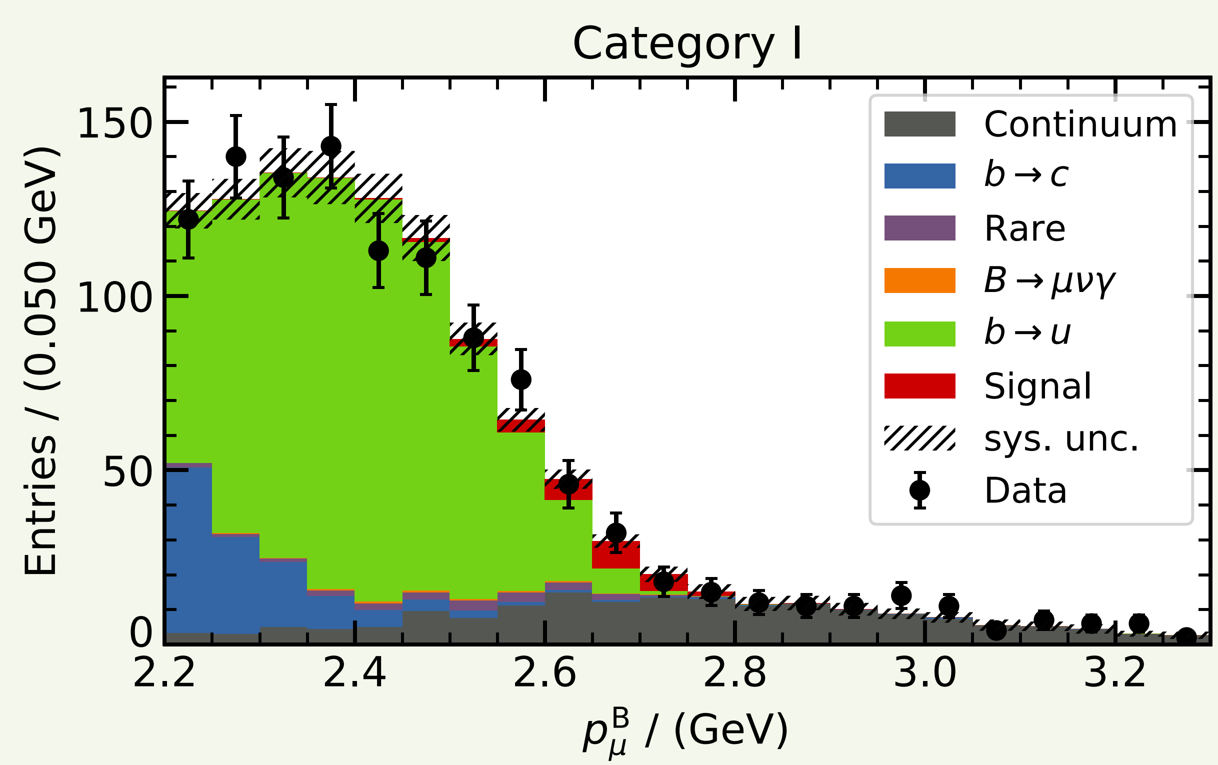
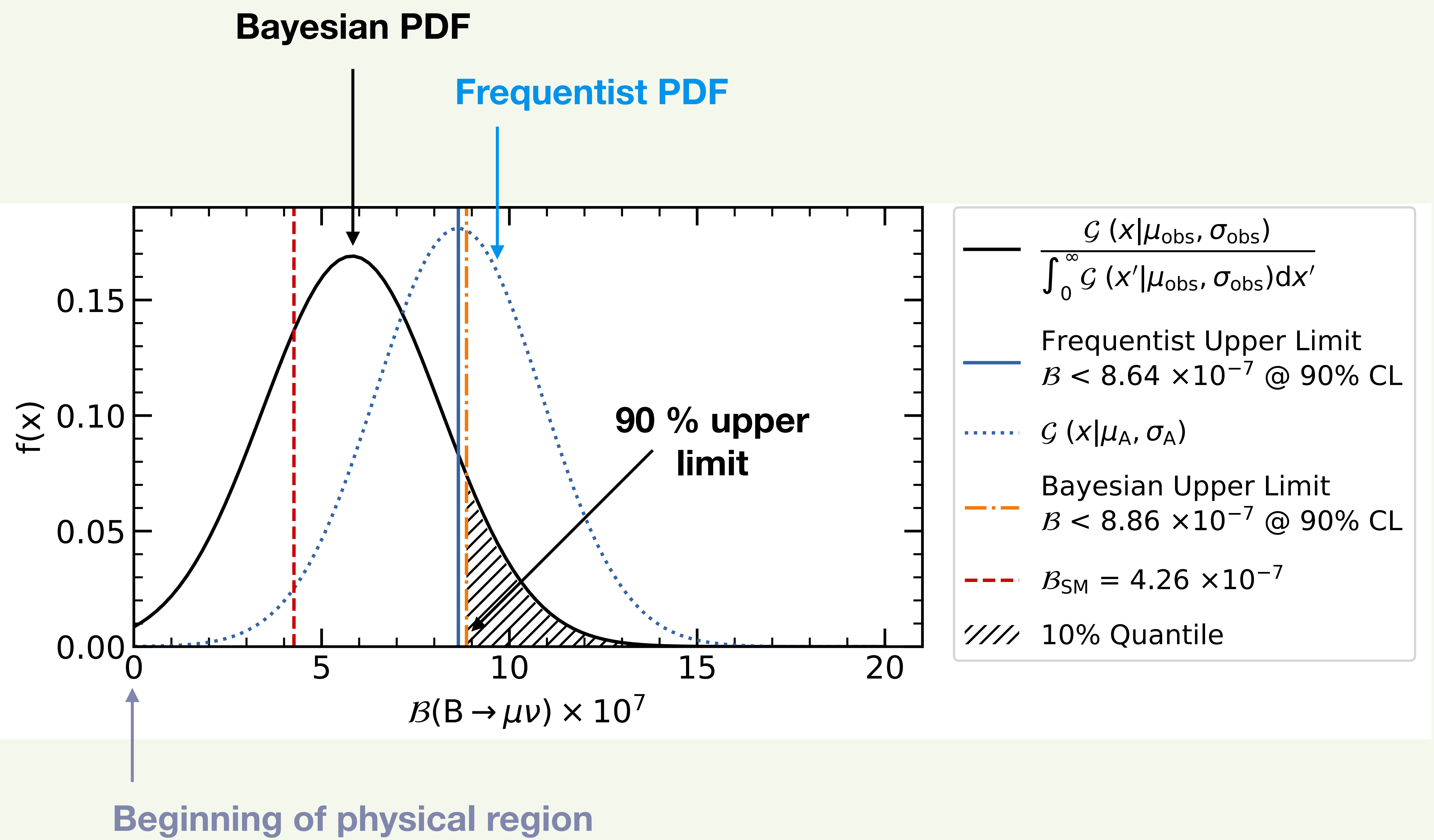


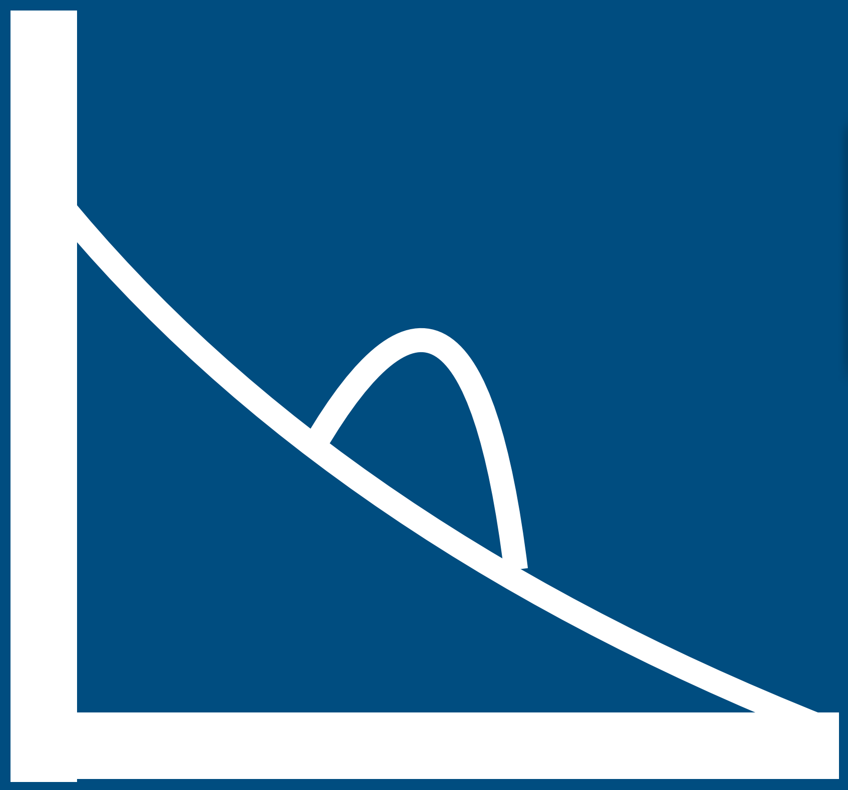
FIG. 1. The SM leptonic $B^+ \to \mu^+\nu_\mu$ decay process and possible BSM processes with and without a sterile neutrino N in the final state are shown.



From one of our papers



Upper limit on the mean of Poisson variable with background



Read at home

UL of mean of Poisson variable with bkg

- Recall from last lecture the UL we placed on the mean ν of a Poisson variable n. *(Last week we considered signal only though.)*
  - Often one faces a somewhat more complicated situation, where the observed value of n is the sum of the desired signal n_s, as well as the background events n_b,
    - $n = n_s + n_b$ where both n_s and n_b can be regarded as Poisson variables with means ν_s and ν_b, respectively.
    - Suppose for the moment, that the **mean** of the **background** ν_b is **known without** any **uncertainty**.
    - For ν_s one only **knows** a priori that $\nu_s \geq 0$.
  - Our goal is to construct an UL for the signal parameter ν_s given a measured value of n.

Upper Limit

- Since n is the sum of two Poisson variables, one can show that it itself is a Poisson variable with the probability function

$$\boxed{f(n;\nu_s,\nu_b) = \frac{(\nu_s+\nu_b)^n}{n!}\, e^{-(\nu_s+\nu_b)}.}$$

- The ML estimator for ν_s is $\boxed{\hat{\nu}_s = n - \nu_b,}$
  - It has zero bias since $E[n] = \nu_s + \nu_b$,
- The equations determining the confidence interval become

$$\boxed{\begin{aligned}
\alpha &= P(\hat{\nu}_s \geq \hat{\nu}_s^{obs};\nu_s^{lo}) = \sum_{n\geq n_{obs}} \frac{(\nu_s^{lo}+\nu_b)^n\, e^{-(\nu_s^{lo}+\nu_b)}}{n!},\\
\beta &= P(\hat{\nu}_s \leq \hat{\nu}_s^{obs};\nu_s^{up}) = \sum_{n\leq n_{obs}} \frac{(\nu_s^{up}+\nu_b)^n\, e^{-(\nu_s^{up}+\nu_b)}}{n!}.
\end{aligned}}$$

Can solve numerically for ν_s^{lo} and ν_s^{up}

Comparison with no background result

- **Comparing to** our **previous expressions**, we see that these **limits** are **related** to **the ones without background** by

$$\boxed{\begin{aligned} \nu_s^{lo} &= \nu_s^{lo}(\text{no background}) - \nu_b, \\ \nu_s^{up} &= \nu_s^{up}(\text{no background}) - \nu_b. \end{aligned}}$$

- The **difficulties** that can arise here are **similar** to **the example without background**, i.e. when the total number of events observed is not large compared to the expected number of background events.
    - **Because of these difficulties**, the **classical limit often causes problems**
        - *As previously mentioned, one should always report $\hat{\nu}_s$ and an estimate for its variance to allow for meaningful combinations later*

Bayesian Limit

- The **Bayesian method** can be **used here** as well, with for example a uniform prior. The **likelihood function** and **posterior probability** are given by

$$L(n_{\rm obs}|\nu_{\rm s}) = \frac{(\nu_{\rm s}+\nu_{\rm b})^{n_{\rm obs}}}{n_{\rm obs}!} e^{-(\nu_{\rm s}+\nu_{\rm b})}.$$

$$p(\nu_{\rm s}|n_{\rm obs}) = \frac{L(n_{\rm obs}|\nu_{\rm s})\,\pi(\nu_{\rm s})}{\int_{-\infty}^{\infty} L(n_{\rm obs}|\nu'_{\rm s})\,\pi(\nu'_{\rm s})\,d\nu'_{\rm s}}.$$

- Taking $\pi(\nu_s)$ **= const. for** $\nu_s > 0$ and **zero otherwise**, the **upper limit** $\nu_s^{\rm up}$ **at CL** $1-\beta$ is

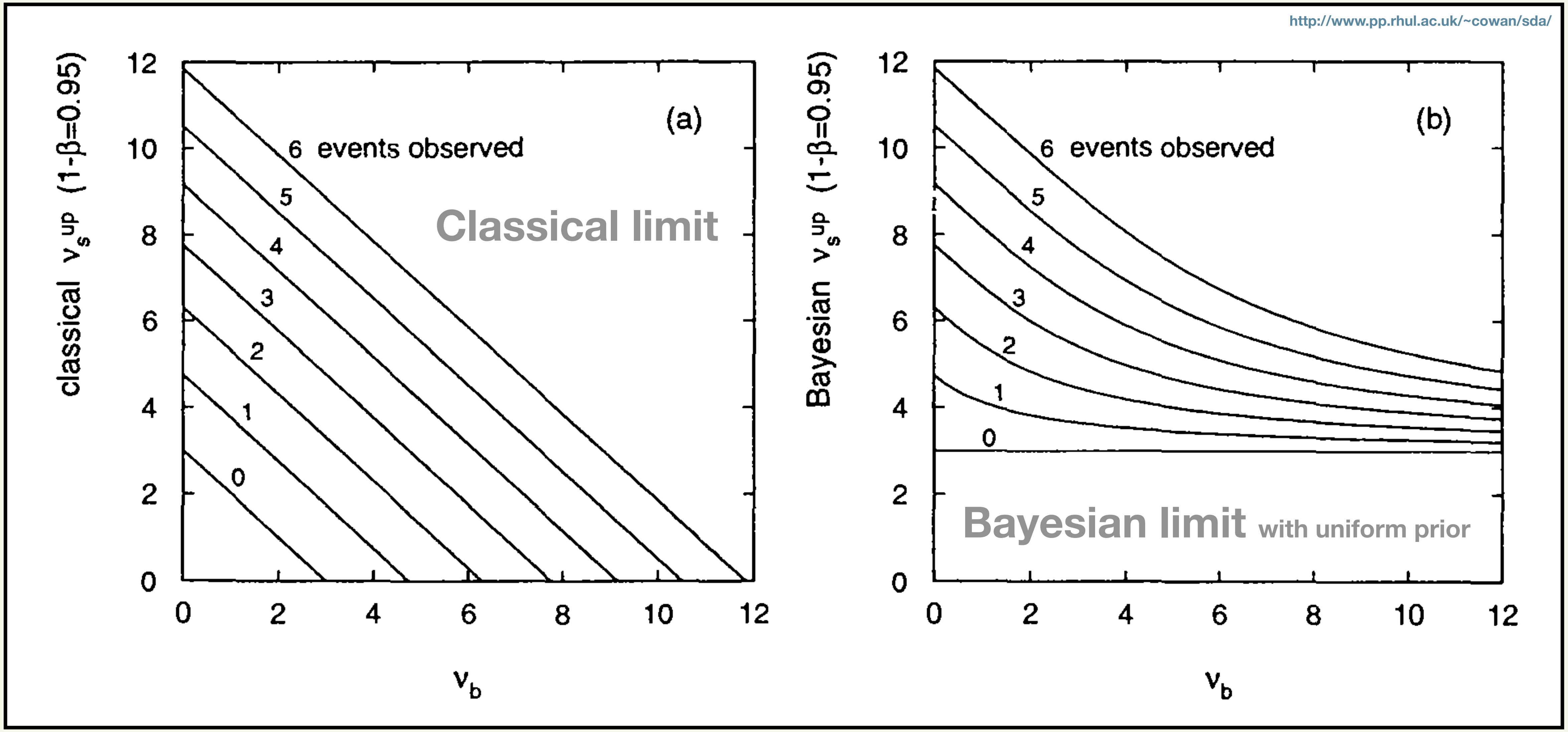
$$\begin{aligned} 1-\beta &= \frac{\int_0^{\nu_{\rm s}^{\rm up}} L(n_{\rm obs}|\nu_{\rm s})\,d\nu_{\rm s}}{\int_0^{\infty} L(n_{\rm obs}|\nu_{\rm s})\,d\nu_{\rm s}} \\ &= \frac{\int_0^{\nu_{\rm s}^{\rm up}} (\nu_{\rm s}+\nu_{\rm b})^{n_{\rm obs}}\,e^{-(\nu_{\rm s}+\nu_{\rm b})}\,d\nu_{\rm s}}{\int_0^{\infty} (\nu_{\rm s}+\nu_{\rm b})^{n_{\rm obs}}\,e^{-(\nu_{\rm s}+\nu_{\rm b})}\,d\nu_{\rm s}}. \end{aligned}$$

Integrals can be related to incomplete gamma functions and one obtains:

$$\beta = \frac{e^{-(\nu_{\rm s}^{\rm up}+\nu_{\rm b})}\sum_{n=0}^{n_{\rm obs}}\frac{(\nu_{\rm s}^{\rm up}+\nu_{\rm b})^n}{n!}}{e^{-\nu_{\rm b}}\sum_{n=0}^{n_{\rm obs}}\frac{\nu_{\rm b}^n}{n!}}.$$

Bayesian Limit

- Upper limits at CL $1 - \beta = 0.95$ for different number of observed events and as a function of the expected number of background events.



More realistic scenario

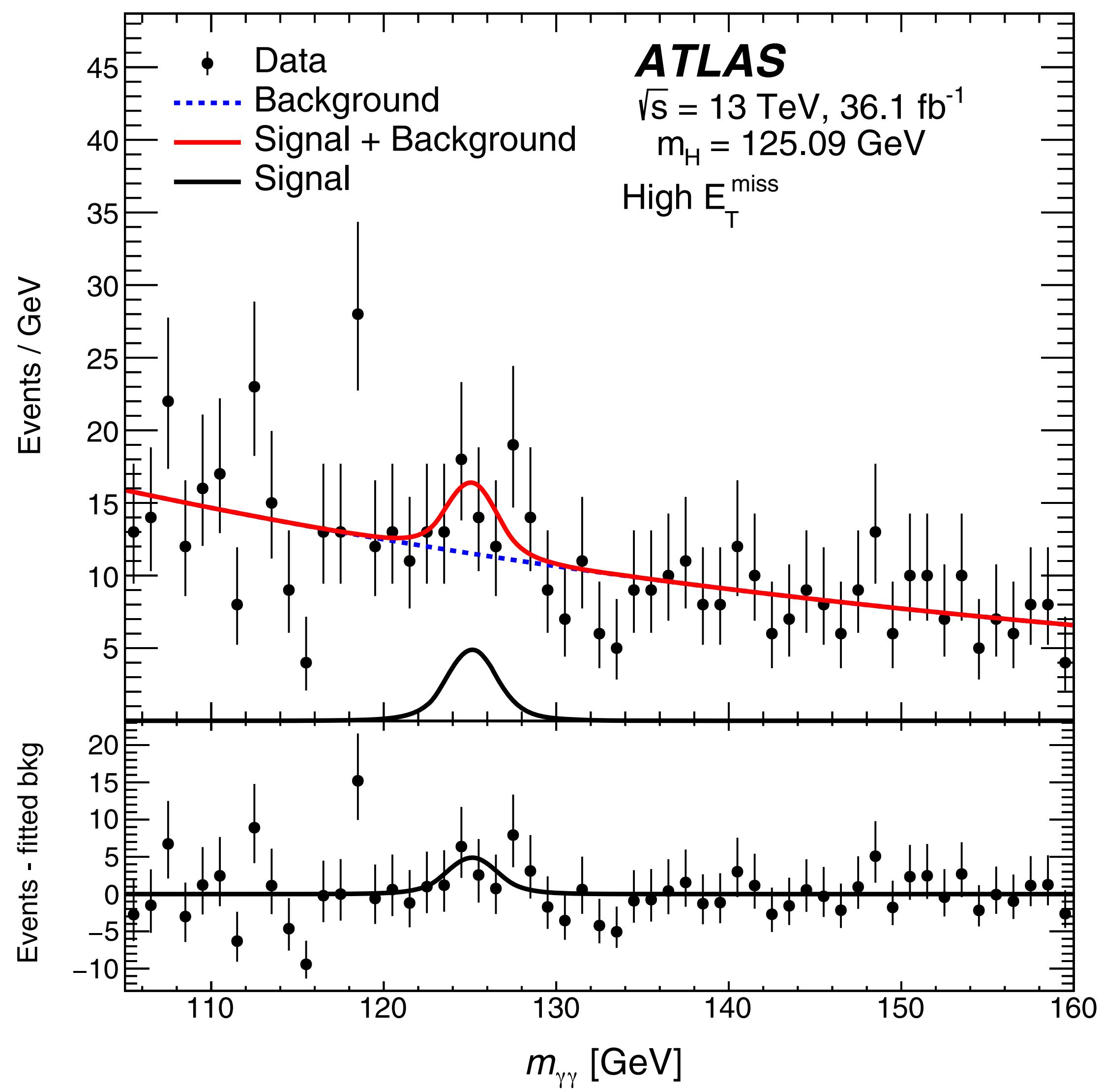
- Often the result of an experiment is not simply the number of n observed events, but includes in addition measured values $x_1, x_2, \ldots, x_n$ of some property of the events (e.g. mass).
  - Suppose the probability density for x is

$$\boxed{f(x;\nu_\mathrm{s},\nu_\mathrm{b}) = \frac{\nu_\mathrm{s} f_\mathrm{s}(x) + \nu_\mathrm{b} f_\mathrm{b}(x)}{\nu_\mathrm{s} + \nu_\mathrm{b}},}$$

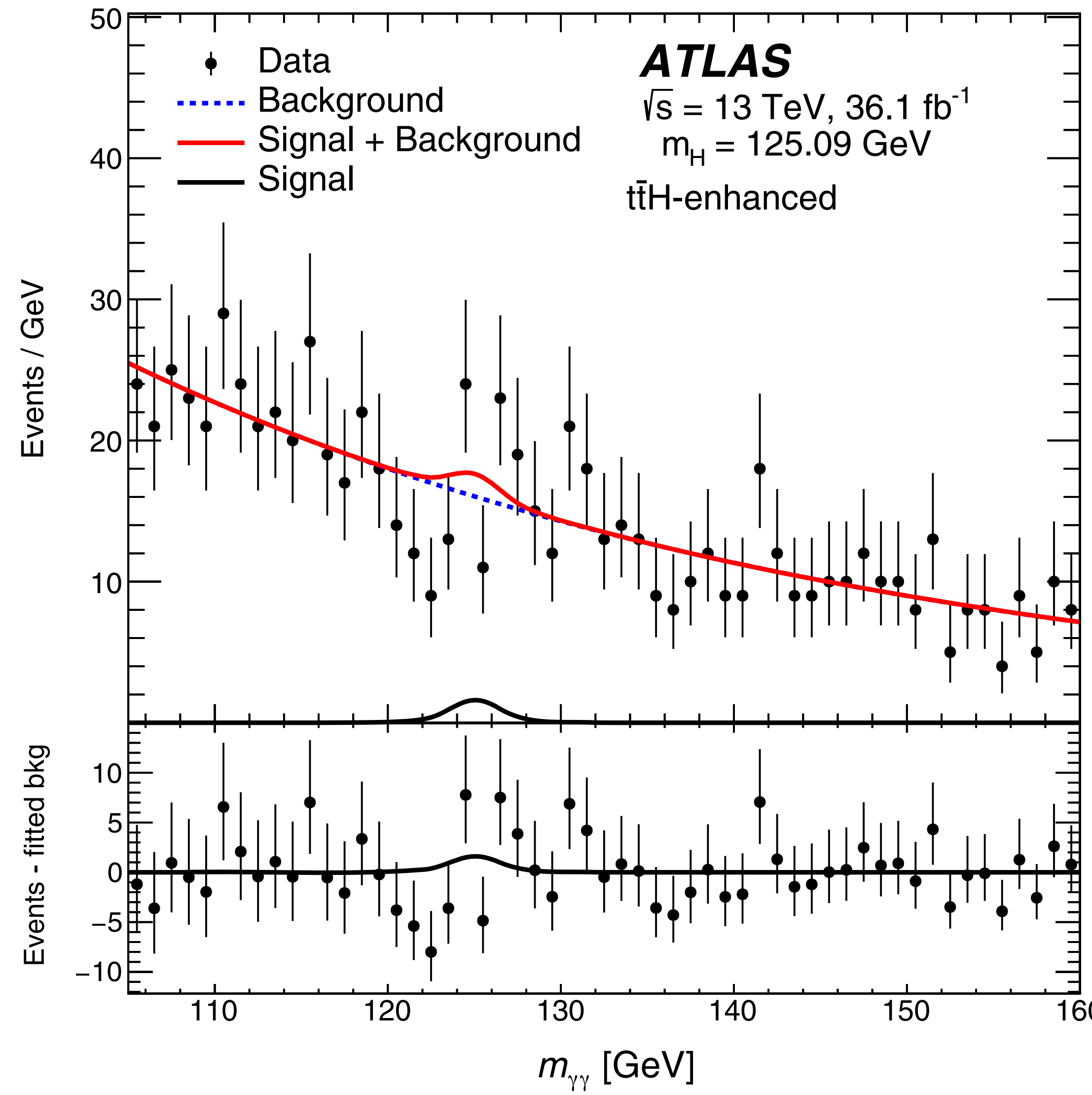
  - This information can be incorporated into the limit ν_s by using the extended likelihood function

$$\begin{aligned} L(\nu_\mathrm{s}) &= \frac{(\nu_\mathrm{s}+\nu_\mathrm{b})^n}{n!} e^{-(\nu_\mathrm{s}+\nu_\mathrm{b})} \prod_{i=1}^{n} \frac{\nu_\mathrm{s} f_\mathrm{s}(x_i) + \nu_\mathrm{b} f_\mathrm{b}(x_i)}{\nu_\mathrm{s} + \nu_\mathrm{b}} \\ &= \frac{e^{-(\nu_\mathrm{s}+\nu_\mathrm{b})}}{n!} \prod_{i=1}^{n} [\nu_\mathrm{s} f_\mathrm{s}(x_i) + \nu_\mathrm{b} f_\mathrm{b}(x_i)], \end{aligned}$$

$\longrightarrow$ *limits in general must be determined numerically or via MC methods*



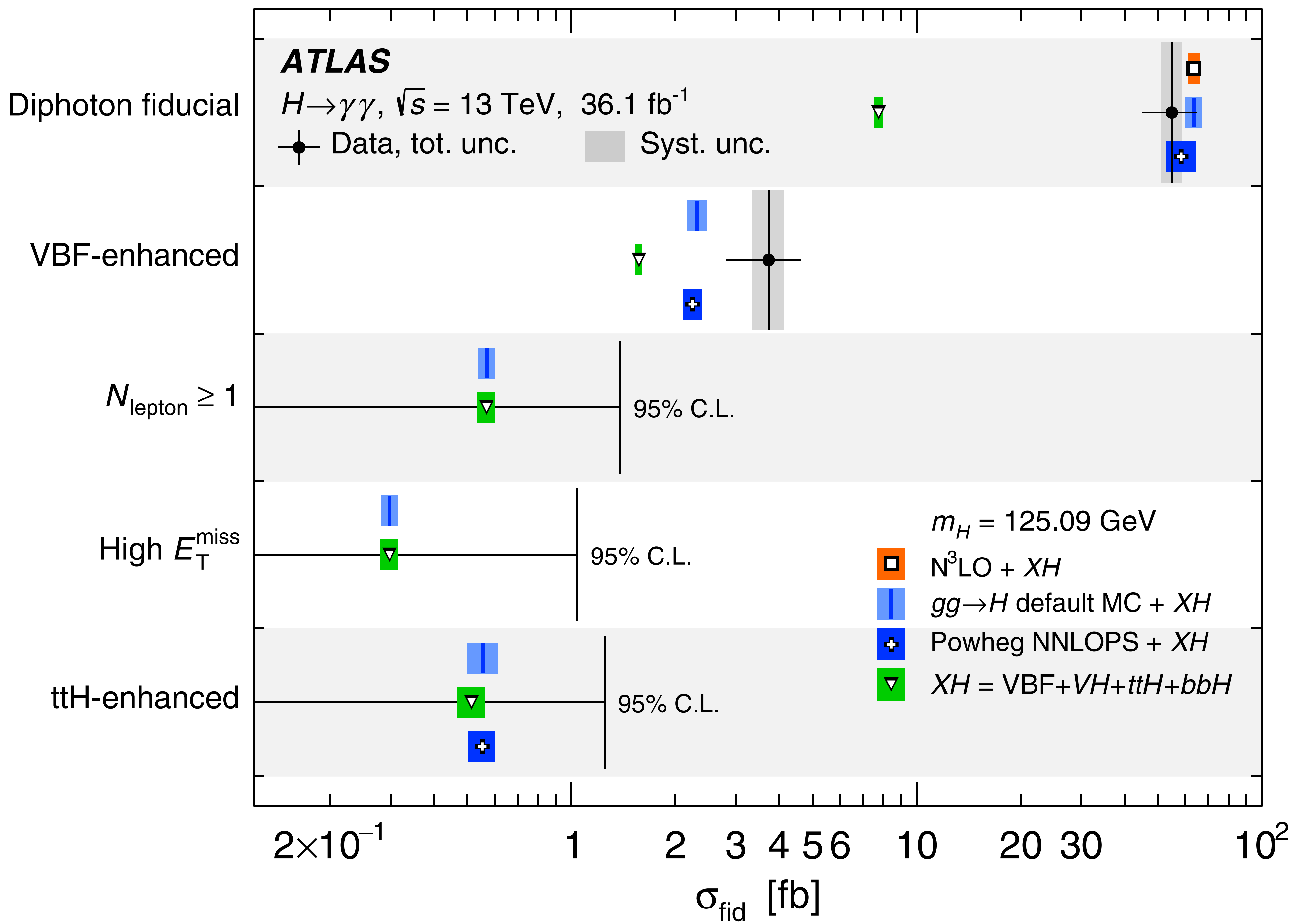
(c) High $E_{\mathrm{T}}^{\mathrm{miss}}$



(d) $t\bar{t}H$-enhanced

(4) High $E_{\mathrm{T}}^{\mathrm{miss}}$: This region retains events with missing transverse momentum $E_{\mathrm{T}}^{\mathrm{miss}} > 80$ GeV and $p_{\mathrm{T}}^{\gamma\gamma} > 80$ GeV is defined to study VH production and possible contributions of Higgs boson production with dark matter particles. The simultaneous requirement that the Higgs boson system balances the missing transverse momentum reduces the fraction of selected events at detector level without particle-level $E_{\mathrm{T}}^{\mathrm{miss}} > 80$ GeV.

(5) $t\bar{t}H$-enhanced: This region retains events with either at least one lepton and three jets or no leptons and four jets to study Higgs boson production in association with top quarks. In addition, one of the jets needs to be identified as originating from a bottom quark.

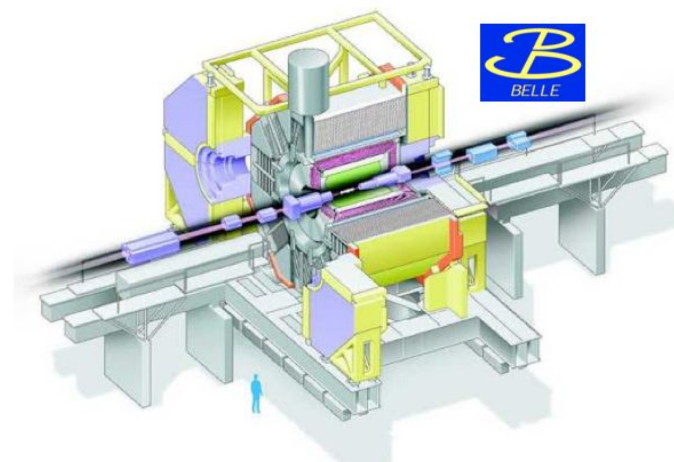


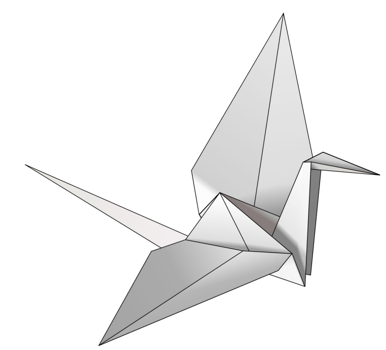
$$\theta_{\text{up}} = \hat{\theta}_{\text{obs}} + \sigma_{\hat{\theta}}\,\Phi^{-1}(1-\beta) \qquad \Phi^{-1}(0.95) = 1.645$$

Take 5

Unfolding

In a nutshell

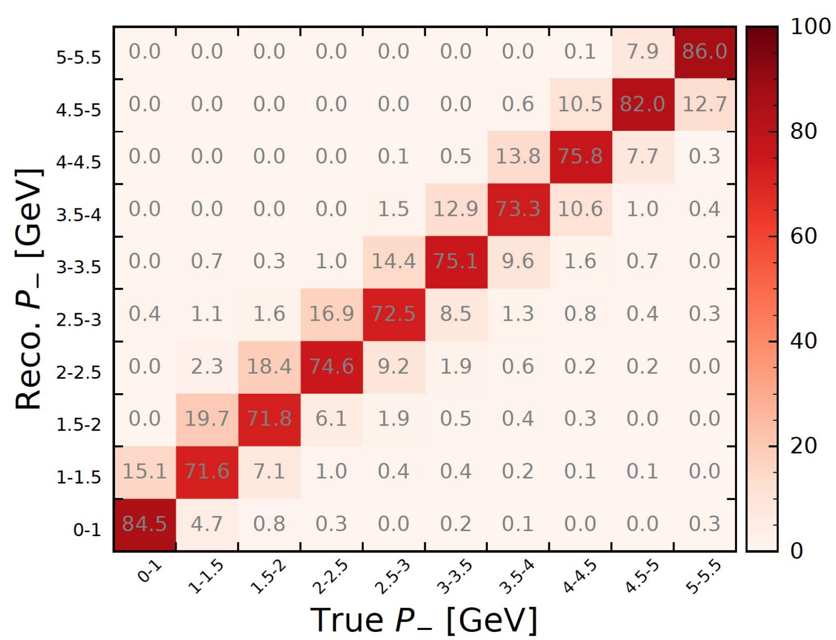




Y: **true distribution**

R: **detector response**

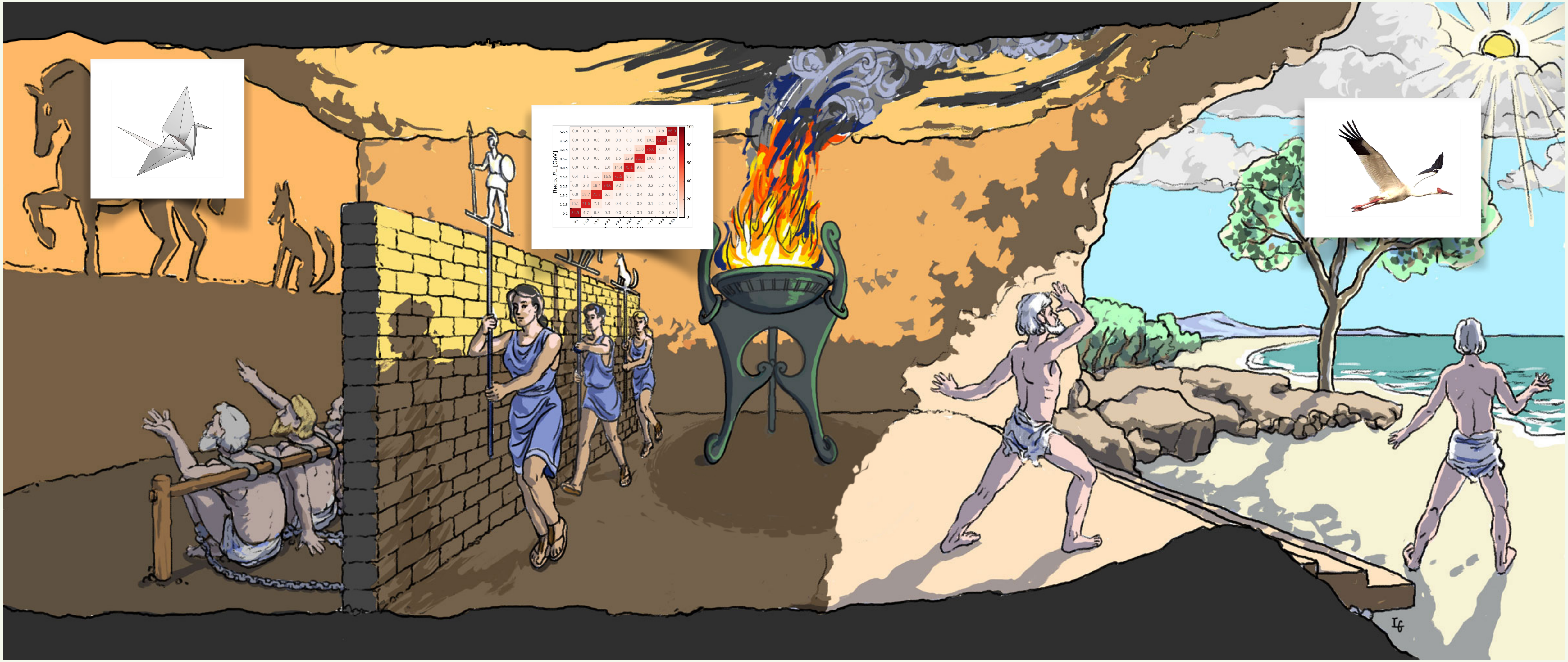
X: **measured distribution**



The detector response is represented by a migration (response) matrix R.

$R(i,j)$ indicates the probability to observe an event in bin i if it had generator-level value in bin j.

Allegory of The Cave (Plato's Republic)



By 4edges - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=73850232

The unfolding problem

- Up till now:
  - Have considered that RVs such as particle energies, decay times, etc., can be measured with **absolute precision**.
- In reality:
  - Every experimental apparatus has **finite resolution**.
    - This distorts measurements.
      - Correct for this = **Unfolding**

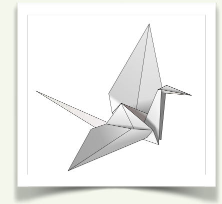
Derivation (i)

- $f_{\text{true}}(y)$ = PDF of true value 'y' 
  - To construct a usable estimator for $f_{\text{true}}(y)$, must represent it by means of some finite set of parameters.
  - If no functional form for $f_{\text{true}}(y)$ is known a priori, then it can still be represented as a normalized histogram with M bins.

$$\boxed{p_j = \int_{\text{bin } j} f_{\text{true}}(y)dy \quad \text{is the probability to find } y \text{ in bin } j}$$

- μ_{tot} = expectation value of total # of events.
  - $\mu_j = \mu_{\text{tot}} \; p_j$ is the expected # of events in bin j
  - The vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_M)$ is the 'true histogram'
    - Careful: not the actual number of events in each bin, but the *expectation values*

Derivation (ii)

- Begin with a sample of measured values x 
  - Entered into a histogram of N bins: $\boldsymbol{n} = (n_1, n_2, \ldots, n_N)$
    - The # of bins N can be $>$, $<$, $=$ to the # of bins in the true histogram M
- Regard n_i as independent Poisson variables with expectation value ν_i

$$P(n_i;\nu_i) = \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!} \qquad \nu_i = E[n_i]$$

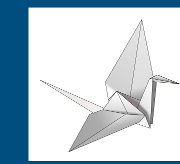
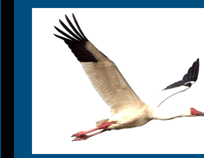
- From the law of total probability:

$$\begin{aligned}
\nu_i &= \mu_{\text{tot}}\, P(\text{event observed in bin } i) \\
&= \mu_{\text{tot}} \int dy\, P\left(\begin{array}{l} \text{observed} \\ \text{in bin } i \end{array} \middle| \begin{array}{l} \text{true value } y \text{ and} \\ \text{event detected} \end{array} \right) \varepsilon(y)\, f_{\text{true}}(y) \\
&= \mu_{\text{tot}} \int_{\text{bin } i} dx \int dy\, s(x|y)\, \varepsilon(y)\, f_{\text{true}}(y).
\end{aligned}$$

Resolution function
(point spread function in imaging applications)

Detection efficiency =
the probability that an event leads to some measured value

Derivation (iii)

$s($  $|$  $)$

- The **resolution function** $s(x \mid y)$ is a conditional PDF:
  - For the measured value x, given the true value y
    - Probability that an even leads to some measured value

- Sometimes also incorporates the **detection efficiency** $\epsilon(y)$
  - $r(x \mid y) = s(x \mid y)\ \epsilon(y)$

Response function: includes the effect of limited efficiency

- One says that the true distribution is **folded** with the response function
  - i.e., expressing ν_i as a function of $s(x \mid y)$ = folding

- **Unfolding** = the task of estimating f_{true}

Derivation (iv)

True: $j,\ y,\ \mu,\ M$
Measured: $i,\ x,\ \nu,\ N$

- Take our integral for ν_i (from slide 48)

$$= \mu_{\text{tot}} \int_{\text{bin } i} dx \int dy\, s(x|y)\, \varepsilon(y)\, f_{\text{true}}(y)$$

Break up the integral over y into a sum over bins j

Multiply numerator and denominator by μ_j

$$= \sum_{j=1}^{M} \frac{\int_{\text{bin } i} dx \int_{\text{bin } j} dy\, s(x|y)\, \varepsilon(y)\, f_{\text{true}}(y)}{(\mu_j/\mu_{\text{tot}})}\, \mu_j$$

$$= \sum_{j=1}^{M} R_{ij}\, \mu_j ,$$

P(observed in bin i and true value in bin j)

P(true value in bin j)

Response matrix =
The conditional probability that an event will be found with measured value x in bin i, given that the true value y was in bin j

Derivation (iv)

True: $j,\ y,\ \mu,\ M$
Measured: $i,\ x,\ \nu,\ N$

- Take our integral for ν_i (from slide 48)

$$= \mu_{\text{tot}} \int_{\text{bin } i} dx \int dy\, s(x|y)\, \varepsilon(y)\, f_{\text{true}}(y)$$

Break up the integral over y into a sum over bins j

Multiply numerator and denominator by μ_j

$$= \sum_{j=1}^{M} \frac{\int_{\text{bin } i} dx \int_{\text{bin } j} dy\, s(x|y)\, \varepsilon(y)\, f_{\text{true}}(y)}{(\mu_j/\mu_{\text{tot}})} \mu_j$$

$$= \sum_{j=1}^{M} R_{ij}\, \mu_j ,$$

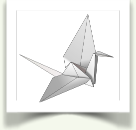
$$\frac{P(\text{observed in bin } i \text{ and true value in bin } j)}{P(\text{true value in bin } j)}$$

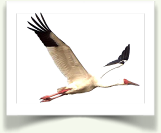
$$=$$

$$P(\text{observed in bin } i \mid \text{true value in bin } j)$$

Response matrix =

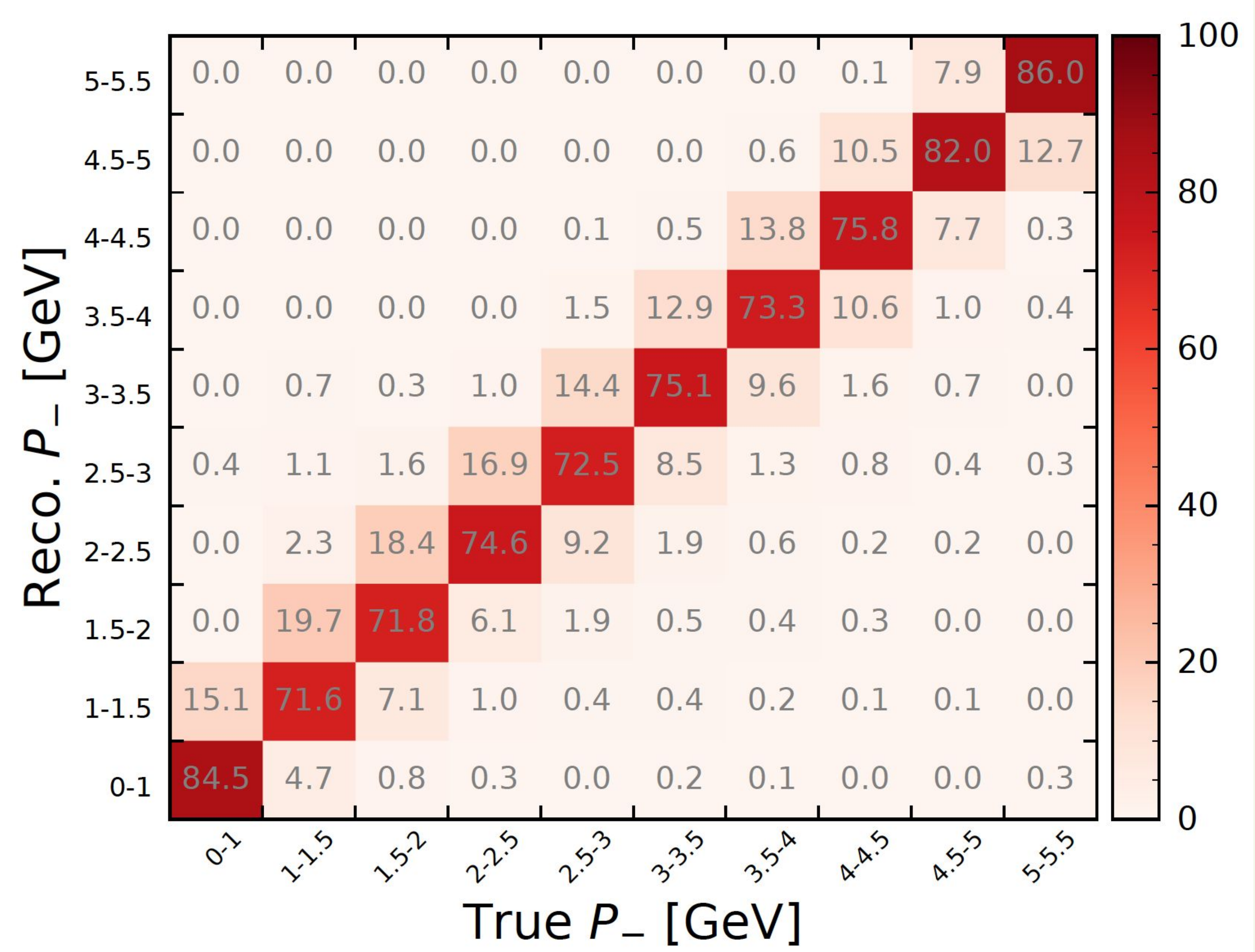
The conditional probability that an event will be found with measured value x in bin i, given that the true value y was in bin j





Response matrix

- The effect of the off-diagonal elements in R is to smear out any fine structure
  - A peak in the true histogram concentrated mainly in 1 bin will be observed over several bins
  - 2 peaks separated by less than several bins will be merged into a single broad peak



R doesn't need to be symmetric

Efficiencies

True: $j,\ y,\ \mu,\ M$

Measured: $i,\ x,\ \nu,\ N$

- Sum over the 'measured' index i and use $\int s(x\,|\,y)dx = 1$

$$\sum_{i=1}^{N} R_{ij} = \sum_{i=1}^{N} \frac{\int_{\text{bin } i} dx \int_{\text{bin } j} dy\, s(x|y)\, \varepsilon(y)\, f_{\text{true}}(y)}{(\mu_j/\mu_{\text{tot}})}$$

$$= \frac{\int_{\text{bin } j} dy\, \varepsilon(y)\, f_{\text{true}}(y)}{\int_{\text{bin } j} f_{\text{true}}(y)\, dy}$$

$$\equiv \varepsilon_j\,,$$

The average value of the efficiency over bin j

Include background

- In addition to limited resolution and efficiency, must allow for the possibility of **background** processes
    - Measuring device produces a value when no true event of the type under study occurred
        - E.g., for β-decay, background = spurious signals in the detector, the presence of other radioactive nuclei in the sample, interactions due to cosmic rays, etc.

$$\nu_i = \sum_{j=1}^{m} R_{ij}\mu_j + \beta_i$$

The # of entries in bin i which originate from **background** processes

The uncertainty from the background is a source of **systematic error** in the unfolded result

To summarize:

True: $j,\ y,\ \mu,\ M$
Measured: $i,\ x,\ \nu,\ N$

- The vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$ is the 'true histogram' *→ Expectation values of true # of entries in each bin*
- The normalized true histogram $\boldsymbol{p} = (p_1, \ldots, p_M) = \boldsymbol{\mu}/\mu_{\text{tot}}$ *→ Probabilities*
- The expectation values of the observed # of entries $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$
- The actual # of entries observed $\boldsymbol{n} = (n_1, \ldots, n_N)$ *→ The data*
- Efficiencies $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_M)$
- Expected background values $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$
- Response matrix R_{ij},
  - $i = 1,\ldots,N$ represents the bin of the observed histogram
  - $j = 1,\ldots,M$ gives the bin of the true histogram

Related by

$$\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$$

Assume known

Goal

Construct estimators $\hat{\boldsymbol{\mu}}$ for the true histogram, or estimators $\hat{\boldsymbol{p}}$ for the probabilities

Assume we either:

Know the form of the probability distribution for the data $\boldsymbol{n}$ $\Rightarrow$ *Allow us to construct the $\mathscr{L}$ function*

Have the covariance matrix $V_{ij} = \text{cov}[n_i, n_j]$ $\Rightarrow$ *Used to construct a χ^2 function*

Method 1: Invert the response matrix

True: $j,\ y,\ \mu,\ M$
Measured: $i,\ x,\ \nu,\ N$

- Start with the matrix form (with $M = N$)

$$\boldsymbol{\nu} = R\boldsymbol{\mu} + \boldsymbol{\beta}$$

  - Invert it to obtain

$$\boldsymbol{\mu} = R^{-1}(\boldsymbol{\nu} - \boldsymbol{\beta})$$

    - Set the estimators for $\boldsymbol{\nu}$ to be the data values $\boldsymbol{n}$

$$\hat{\boldsymbol{\nu}} = \boldsymbol{n}$$

      - The estimators for the $\boldsymbol{\mu}$ are then

$$\hat{\boldsymbol{\mu}} = R^{-1}(\boldsymbol{n} - \boldsymbol{\beta})$$

Properties

True: $j,\ y,\ \mu,\ M$
Measured: $i,\ x,\ \nu,\ N$

- Expectation value of inversion:

$$\boxed{\begin{aligned} E[\hat{\mu}_j] &= \sum_{i=1}^{N} (R^{-1})_{ji}\, E[n_i - \beta_i] = \sum_{i=1}^{N} (R^{-1})_{ji}\, (\nu_i - \beta_i) \\ &= \mu_j, \end{aligned}}$$

Estimators $\hat{\mu}_j$ are unbiased
(Since by assumption $\hat{\nu}_i = n_i$ is unbiased)

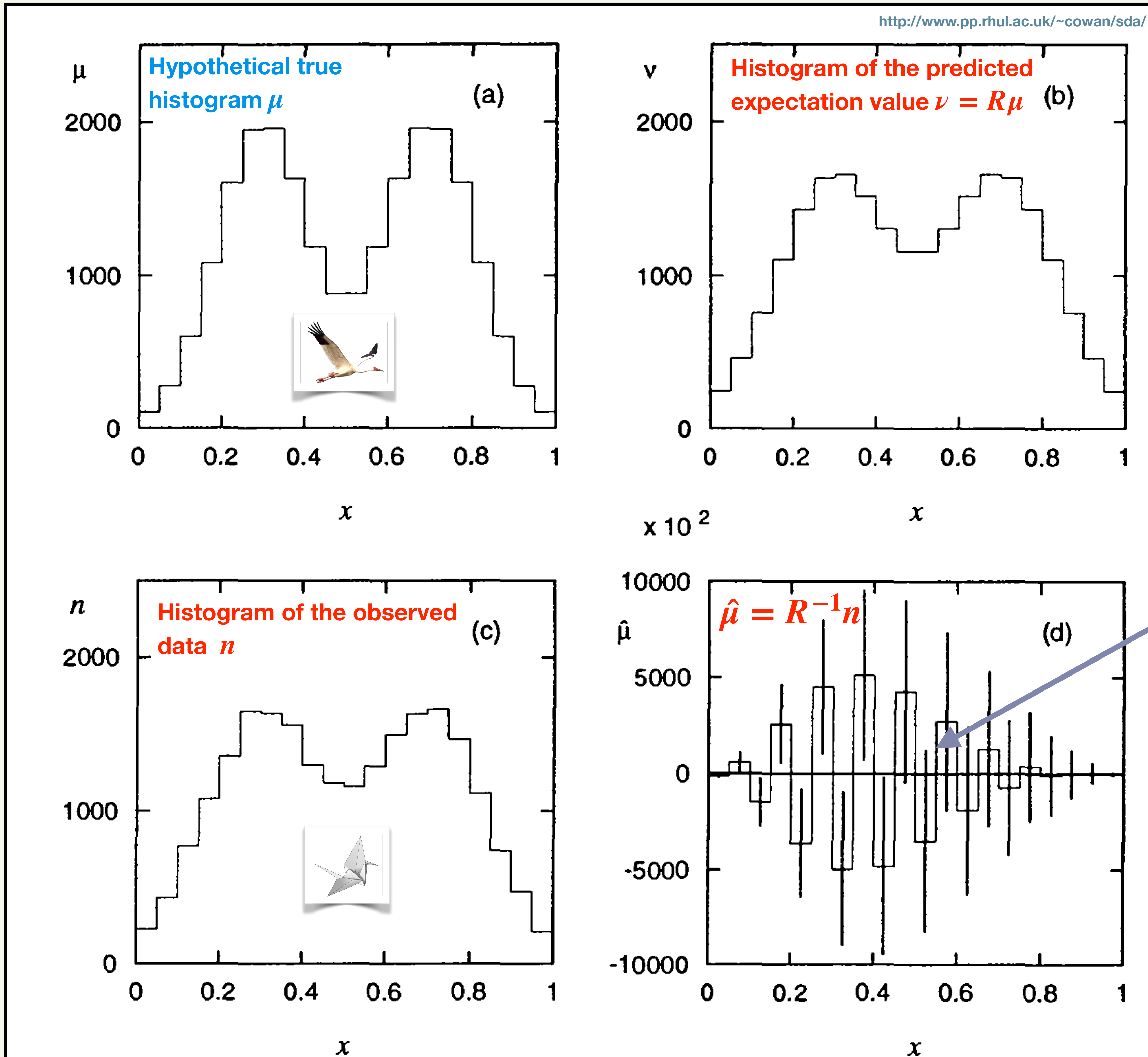
- Covariance of uncorrelated Poisson variables:

$$\boxed{\begin{aligned} \mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j] &= \sum_{k,l=1}^{N} (R^{-1})_{ik}\, (R^{-1})_{jl}\, \mathrm{cov}[n_k, n_l] \\ &= \sum_{k=1}^{N} (R^{-1})_{ik}\, (R^{-1})_{jk}\, \nu_k, \end{aligned}}$$

- Covariance of correlated Gaussian variables:

$$\boxed{U = R^{-1}\, V\, (R^{-1})^T.}$$

Ex. where matrix inversion goes "wrong"



Assume 0 background, $\beta = 0$

Response matrix (R) based on a Gaussian resolution function with $\sigma = 1.5 \times$ bin width

Assume $\epsilon_i = 1$

Results in

P(event to remain in the bin created) = 26%
P(event to migrate 1 bin) = 21%
P(event to migrate 2 bins) = 16%

Very large anti-correlations!

- Applying the response matrix R smears out fine structure
- Applying R^{-1} creates (often unwanted) structure

- We do not have the expectation values ν
- Only have the data n, which are RVs and subject to statistical fluctuations

So what went "wrong"?

- **Nothing really**
  - The resulting unfolded yields are **unbiased**, but **heavily correlated**
  - They can be used to test hypothesis, given one takes into account the **full set of correlations**

$$\boxed{\chi^2 = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)^T \, U^{-1} \, (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0),}$$

Use to test the compatibility of the estimators $\hat{\mu}$ with the hypothesis μ_0

    - Can reduce such oscillations considerably by **making bin widths larger than** the **width** of the **resolution function**

- **Alternatives:**
  - Either incorporate **prior knowledge** or do not rely on neighboring bins to determine resolution correction
    - Both come at a price: **trading variance for bias**

Method 2: Correction factors (i)

- Assume the bins of the true distribution (μ) are the same as the data (n)
  - Determine the **correction factor** for each bin (e.g., from MC simulation)

$$\hat{\mu}_i = C_i(n_i - \beta_i) \qquad C_i = \frac{\mu_i^{MC}}{\nu_i^{MC}}$$

Run MC program w/out detector simulation

Run MC program with detector simulation

- Works well if bin-to-bin sharing (smearing) is negligible $R_{ij} = \delta_{ij}\epsilon_j$

$$\nu_i^{\text{sig}} = \nu_i - \beta_i = \epsilon_i \mu_i$$

- Expectation value for corrected data

$$E[\hat{\mu}_i] = C_i E[n_i - \beta_i] = C_i(\nu_i - \beta_i) = \frac{\mu_i^{MC}}{\nu_i^{MC}} \nu_i^{\text{sig}}$$

Method 2: Correction factors (ii)

- Rearrange to make the bias explicit (identical to previous expression for $E[\hat{\mu}_i]$)

$$E[\hat{\mu}_i] = \frac{\mu_i^{MC}}{\nu_i^{MC}} \nu_i^{\mathrm{sig}} = \underbrace{\left(\frac{\mu_i^{MC}}{\nu_i^{MC}} - \frac{\mu_i}{\nu_i^{\mathrm{sig}}} \right) \nu_i^{\mathrm{sig}}}_{\text{Bias}} + \mu_i$$

Bias = 0 if MC = nature

- Covariance matrix for the estimators

$$\mathrm{cov}[\hat{\mu}_i, \hat{\mu}_j] = C_i^2 \mathrm{cov}[n_i, n_j] = C_i^2 \delta_{ij} \nu_i$$

Smearing fluctuations independent between bins

- Iterative bin-by-bin method:
    - Begin with (plausible) guess of the true spectrum
    - Apply correction to measurement
    - Generate new C_i from corrected spectrum of previous iteration
    - Repeat (for a few iterations)

<u>Drawback</u>: Highly model dependent

Method 3: Regularized unfolding

- **Regularization** = impose a measure of **smoothness** on the estimators of the true histogram $\boldsymbol{\mu}$
  - Matrix inversion **IS** the maximum likelihood solution (see page 162)

Independent Poisson fluctuations

$$\log \mathscr{L}(\boldsymbol{\mu}) = \sum_{i=1}^{N} (n_i \log \nu_i - \nu_i)$$

ML estimator (same as s56)

$$\hat{\boldsymbol{\nu}} = \boldsymbol{n}$$

$$\hat{\boldsymbol{\mu}} = R^{-1}(\boldsymbol{n} - \boldsymbol{\beta})$$

  - Accept solutions that are **close** to the ML estimate

$$\log \mathscr{L}(\boldsymbol{\mu}) \geq \log \mathscr{L}(\boldsymbol{\mu}_{\mathrm{max}}) - \Delta \log \mathscr{L}(\boldsymbol{\mu})$$

determines trade-off between bias and variance in unfolded histogram

  - Define a **regularization (aka smoothness) function** S that increases when the unfolded solution becomes smoother
    - Task: choose the solution with the highest degree of smoothness out of the acceptable solutions determined by above inequality
    - Must maximize

$$\Phi(\boldsymbol{\mu}) = \alpha \log \mathscr{L}(\boldsymbol{\mu}) + S(\boldsymbol{\mu})$$

Regularization parameter which depends on $\Delta \log \mathscr{L}(\boldsymbol{\mu})$

$\alpha \to \infty$ gives ML solution

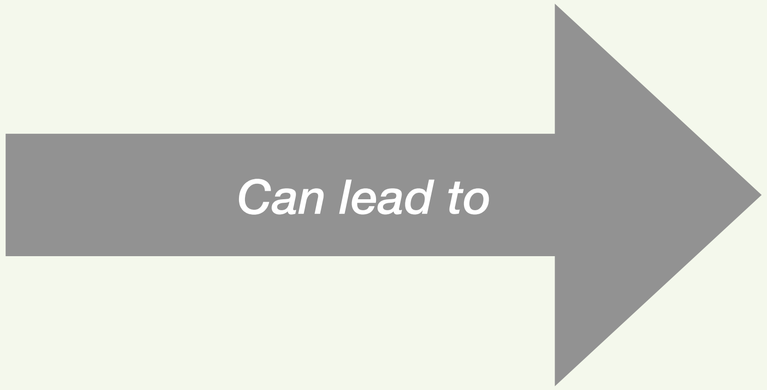
Regularization functions

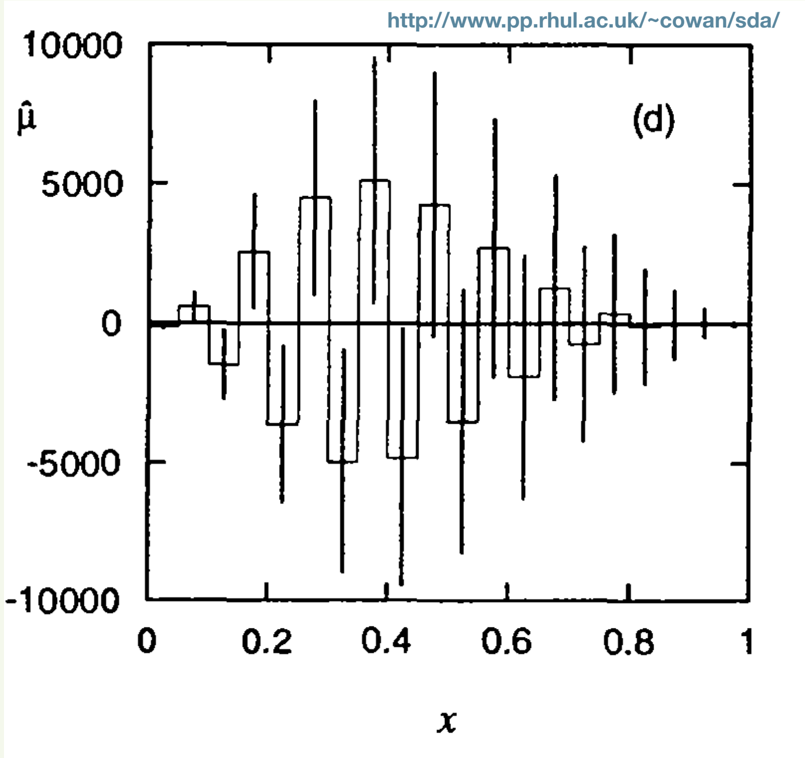
- There are several options (sections 11.5.1-11.5.4)
  - Tikhonov regularization:
    - Measure of smoothness is the mean value of the square of some derivative of the true distribution. Tikhonov regularization using the second derivative (so that $S(\boldsymbol{\mu})$ is related to the avg. curvature) is widely used in particle physics.
  - Regularization functions based on entropy:
    - Interpret the entropy as a measure of the smoothness of a histogram. Estimators are constructed according to the principle of maximum entropy. Often developed in the framework of Bayesian statistics.
  - Regularization function based on cross-entropy:
    - Useful if we have prior knowledge that the true events approximately follow some distribution.

Choice of α

$$\Phi(\boldsymbol{\mu}) = \alpha \log \mathscr{L}(\boldsymbol{\mu}) + S(\boldsymbol{\mu})$$

- The choice of α determines the trade off between the **bias** and **variance** of the estimators $\hat{\boldsymbol{\mu}}$
  - If α is very large, solution is dominated by the likelihood function and one has $\log \mathscr{L}(\boldsymbol{\mu}) = \log \mathscr{L}_{\max}$ and very **large variances**





  - If α is small, leads to a perfectly smooth distribution (since all of the weight is put on the regularization function S)

Choice of α

- Recall the mean square error from L04, S11:

$$\mathrm{MSE} = E\left[\left(\hat{\theta} - \theta\right)^2\right] = E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right] + \left(E[\hat{\theta} - \theta]\right)^2$$

$$= V[\hat{\theta}] + b^2$$

i.e., sum of variance and bias^2

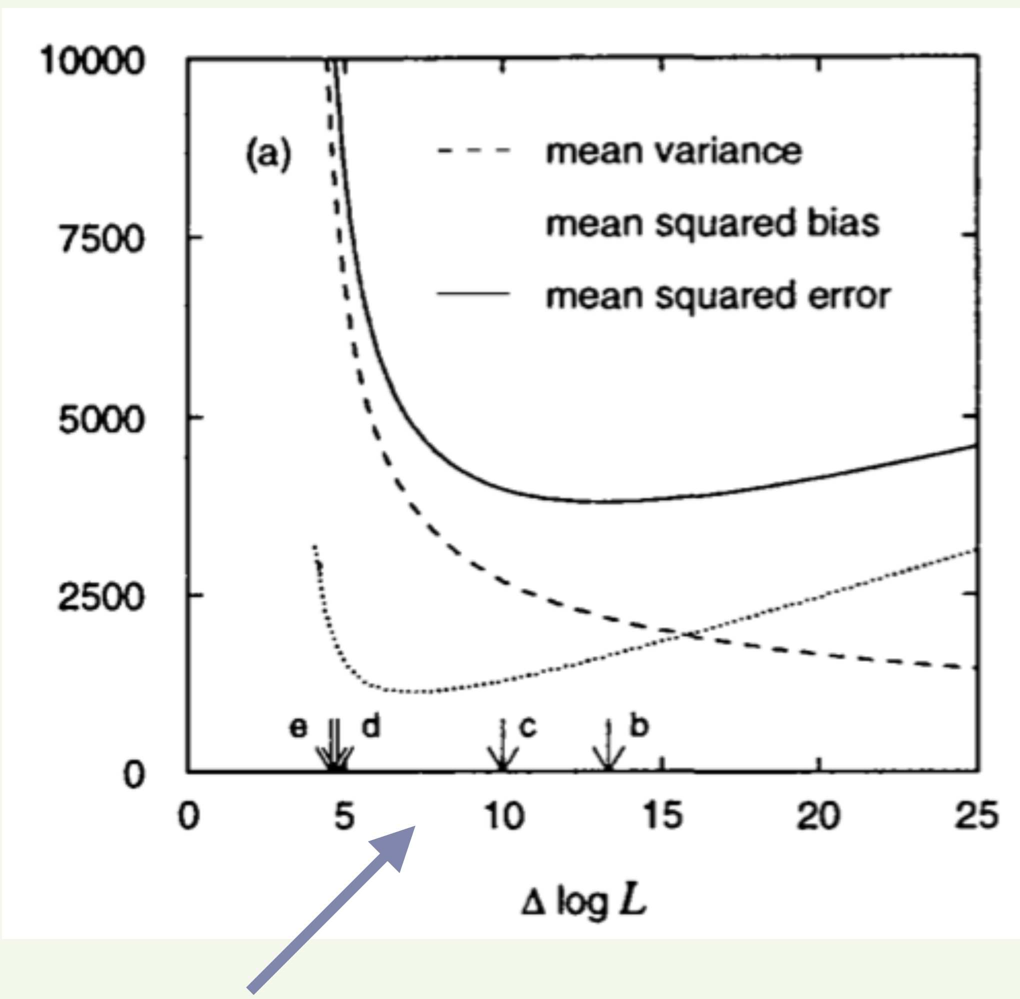
***Interpret*: sum of squares of statistical and systematic uncertainties**

  - Take the **MSE averaged over all bins** as the measure of the goodness of the final result. *One can determine α so as to obtain a particular value of the MSE.*

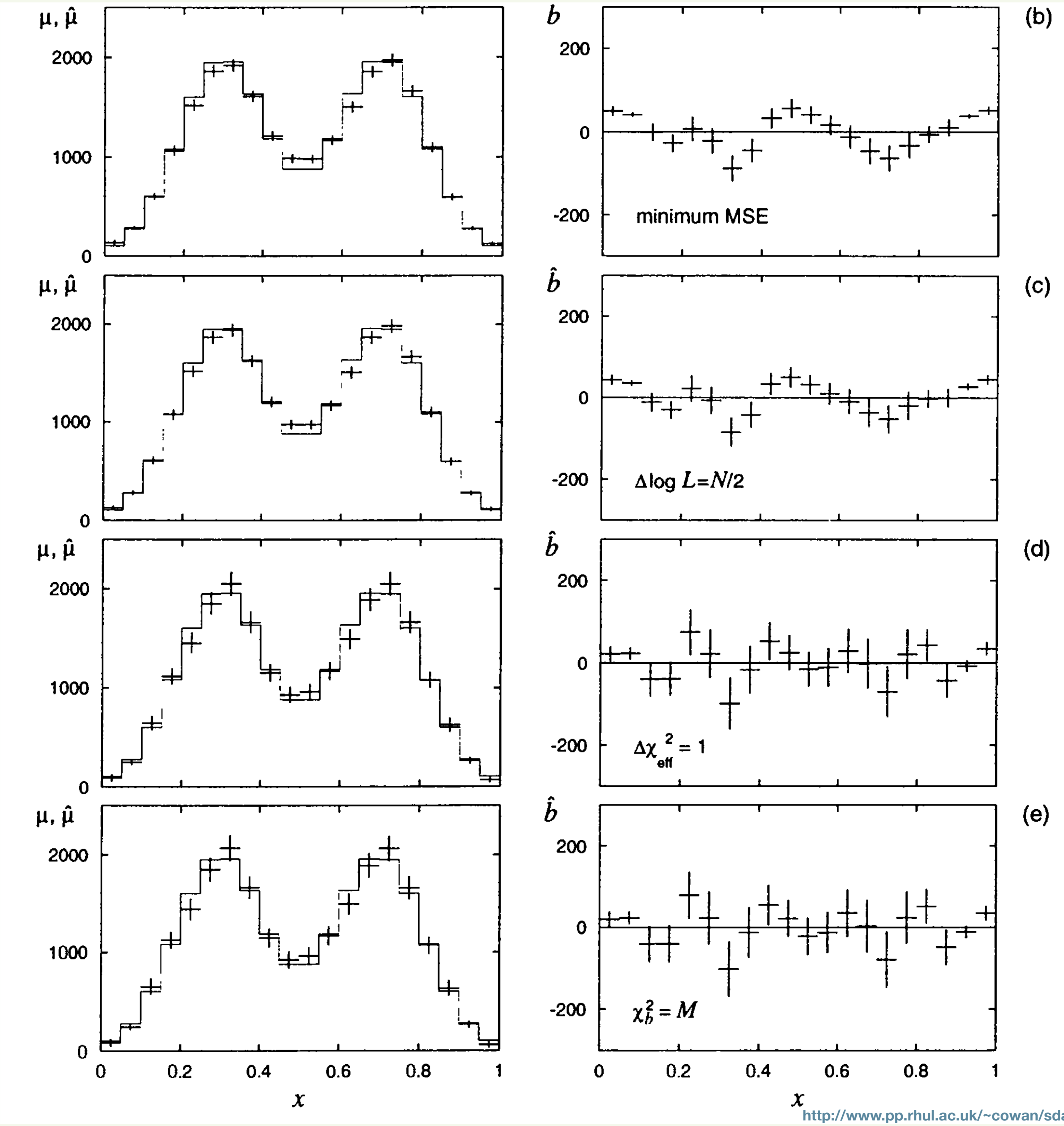
- Can also use:
  - $\Delta \log \mathscr{L} = \log \mathscr{L}_{\max} - \log \mathscr{L} = N/2$
  - $\Delta \chi^2_{\mathrm{eff}} = 1$
  - $\chi^2_b = M$

Example with Maximum Entropy

- Return to our original example (which was unfolded using matrix inversion in s58)
- Now try with Maximum Entropy regularization

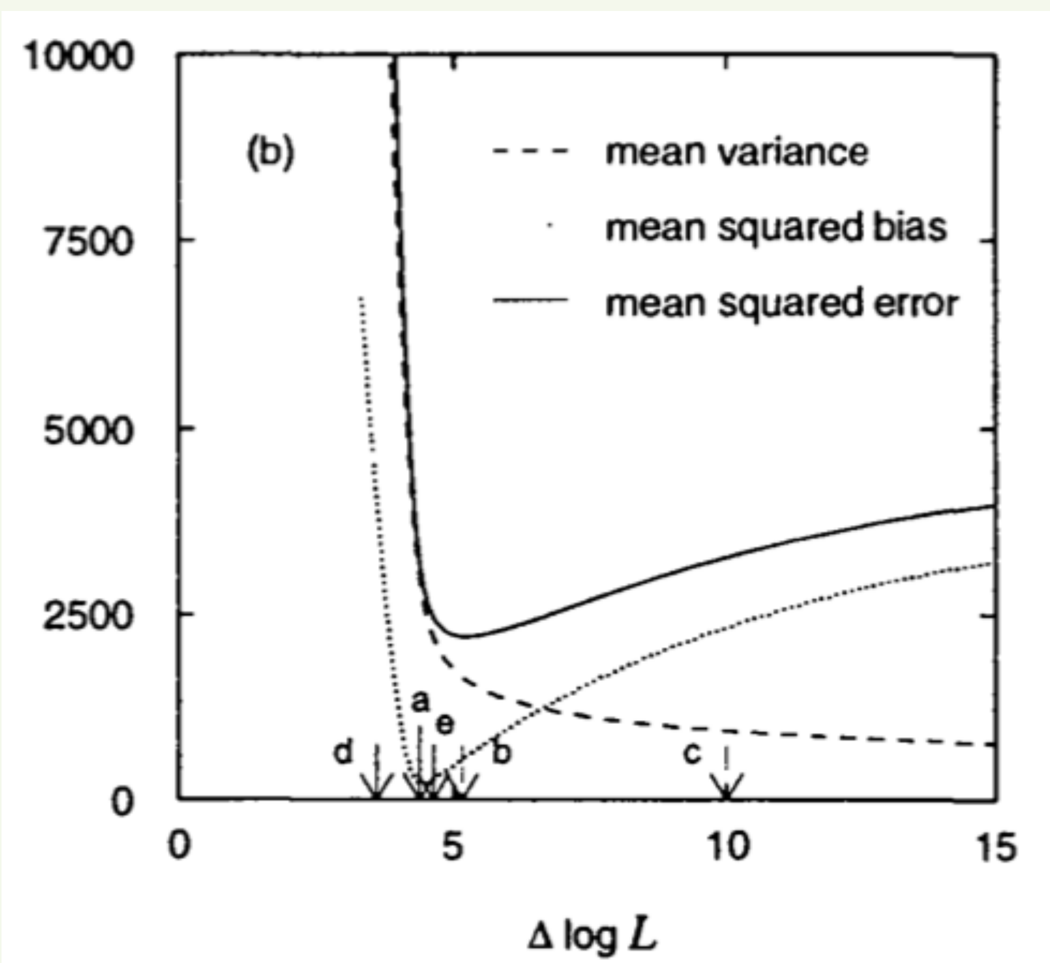


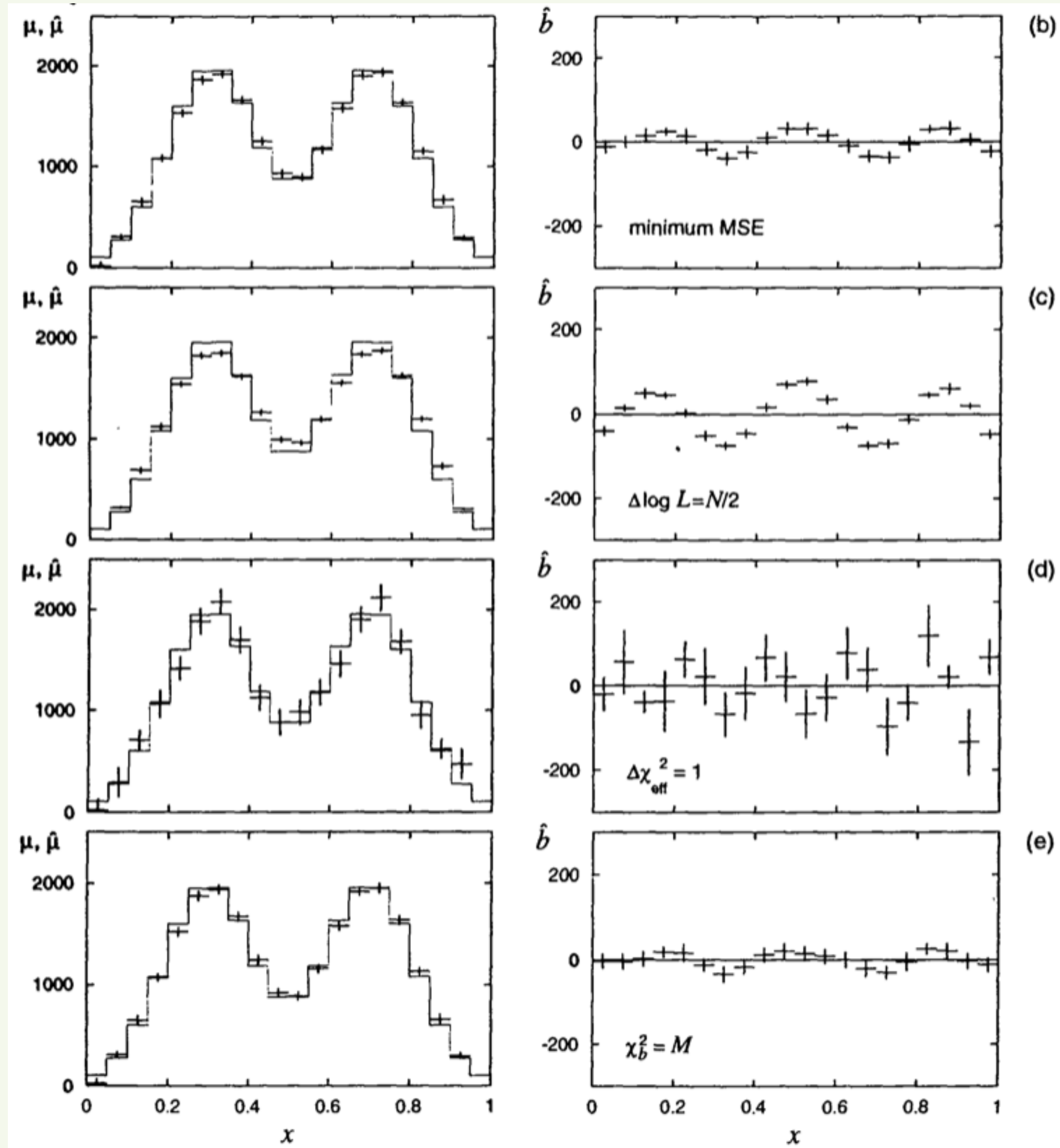
Arrows indicate solutions based on the criteria in the last slide



Example with Tikhonov

- Return to our original example (which was unfolded using matrix inversion in s58)
- Now try with Tikhonov regularization





For next time

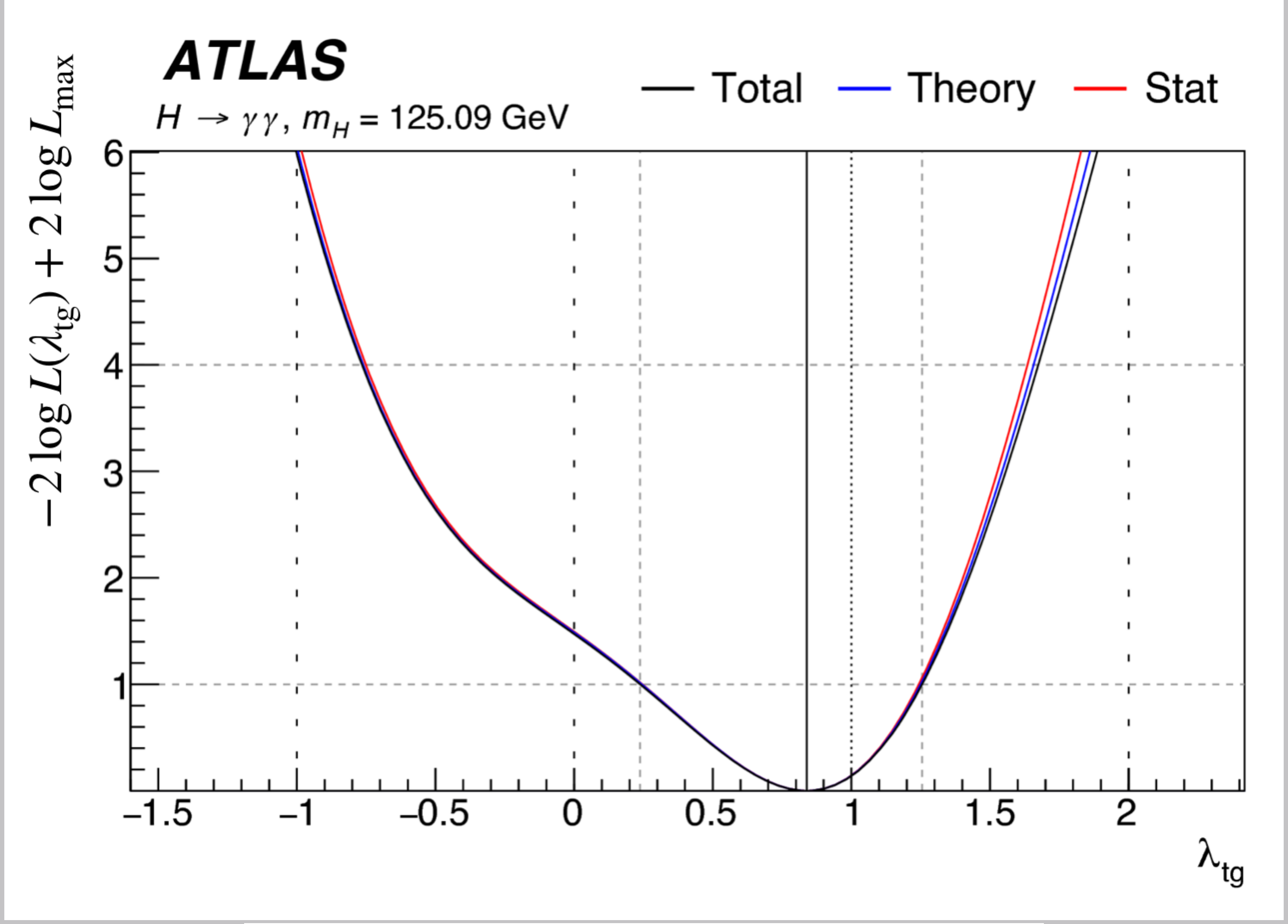
- Required reading
  - Cowan textbook: chapters 9 (9.8-9.9), 10, and 11

- Extra reading for fun: /Reading material / L08 /
  - Search for $B^+ \to \mu^+\nu_\mu$ and $B^+ \to \mu^+N$ at Belle

Quiz Time: 8th Round

Lower and upper limit

1. Determine the lower and upper limit at 90% CL of the parameter λ_{tg} using the scan of the likelihood function and the provided values for Q_γ. Note that the likelihood is already normalised with respect to its maximal value.



$1-\gamma$	Q_γ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

The problem with priors

2. Write down the posterior probability density function of a parameter θ as a function of the Likelihood of some data x and prior probability density function $\pi(x)$. What is the problem of using Bayesian priors when you quote a limit and which functional form for the prior somehow remedies them?

Comprehension about unfolding

1. What is the unfolding problem? Write down the relevant equations for measurement $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)$ with background $\boldsymbol{\beta} = (\beta_1 \ldots \beta_N)$, which should be unfolded in yields $\boldsymbol{\mu} = (\mu_1 \ldots \mu_N)$ using a response matrix R.

2. Describe what the response matrix element R_{ij} means in terms of a conditional probability. Does the response matrix need to be a square or maybe even a symmetric matrix?

3. Describe two methods to solve the unfolding problem that do not involve regularization. Sketch out in detail what steps need to be taken.

Correction factors and regularized unfolding

4. What is the method of correction factors? What are the drawbacks of using this method?

5. What is the idea behind regularized unfolding?

KCETA Colloquium

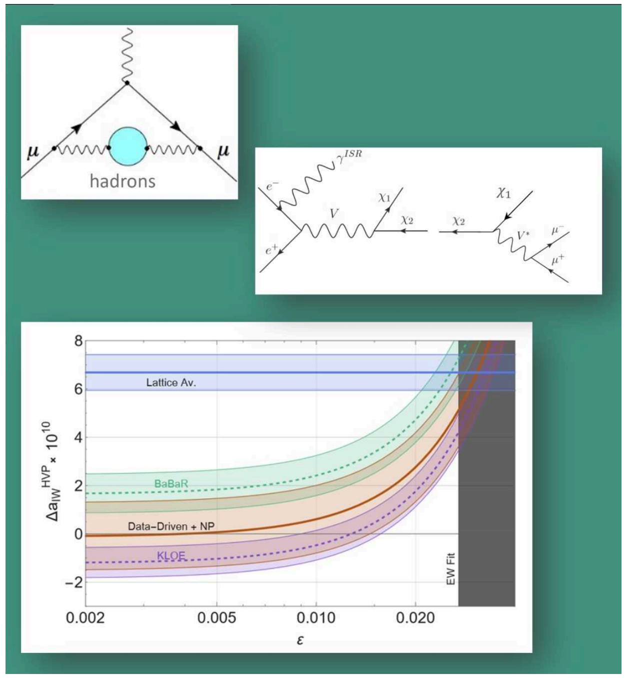
The muon g-2 window discrepancy and GeV-scale new physics

Thursday, June 22, 2023
Kleiner Hörsaal A (CS) 15:45 - 17:00

Dr. Luc Darmé
(Institut de Physique des 2 Infinis de Lyon (IP2I), CNRS/IN2P3)

The decade-old discrepancy between the Standard Model prediction of the muon anomalous magnetic moment and the experimental results has seen striking developments in the past two years. In particular, recent lattice determinations of the hadronic vacuum polarization contribution deviate from the established data-driven ones at almost 5σ. This new anomaly can be also seen as a tension between ab-initio lattice calculations and experimental measurements of e+e-→ hadrons processes at and below the GeV scale.

We will review this puzzling situation and show how new processes beyond the standard model can affect indirectly the hadronic data around this scale, reconciling the lattice and data-driven results while complying with current phenomenological constraints. We will finally present a simple dark matter-motivated model as an explicit example.



Please note:
The colloquium will also be live-streamed to B402 SR 224 (CN).

KIT Center Elementary Particle and Astroparticle Physics (KCETA)
www.kceta.kit.edu

Karlsruher Institut für Technologie

Bibliography

- Part of the material presented in this lecture is taken from the following sources. See the active links (when available) for a complete reference
  - **Statistical Data Analysis** textbook by G. Cowan (U. London): all figures & equations with white background