Statistical Methods for Data Analysis in Particle Physics

Luca Lista

Università Federico II, Napoli

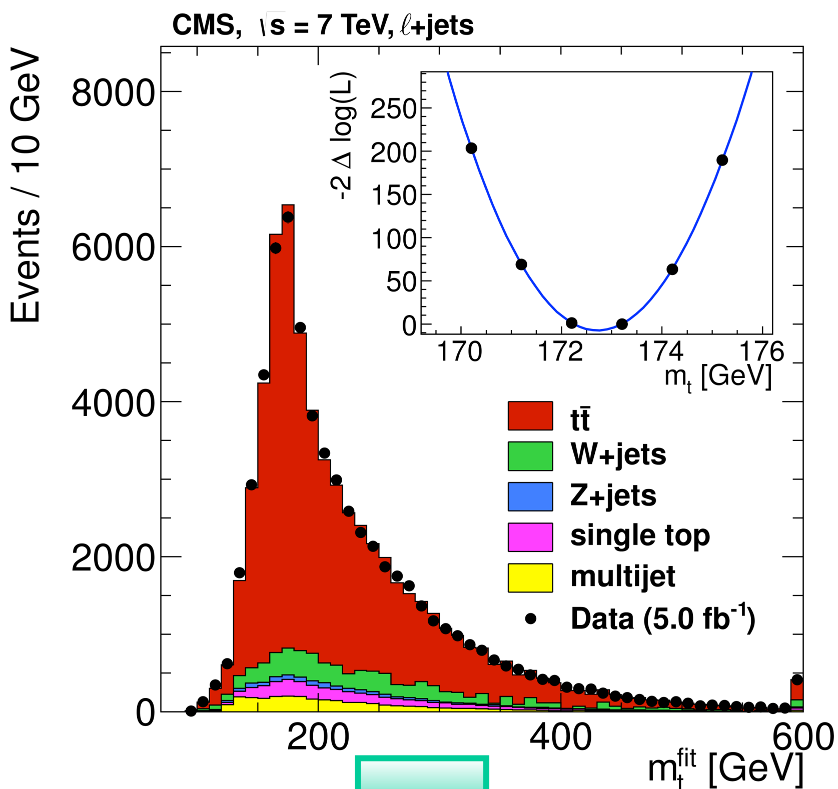
INFN Sezione di Napoli

Lecture 1

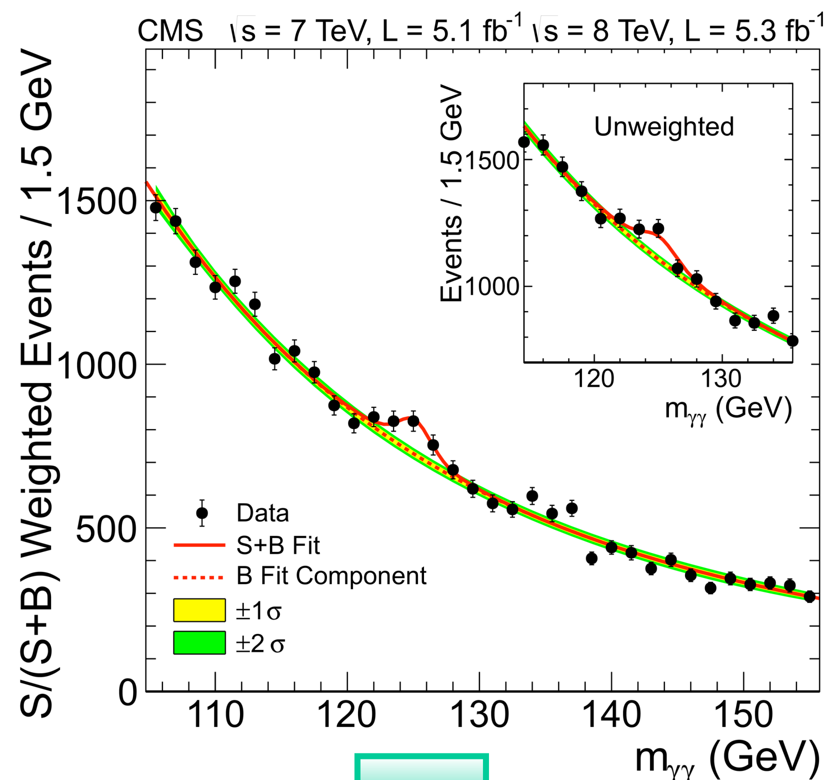
- Introduction to probability
- Subjective/Bayesian probability vs Frequentist probability
- Kolmogorov axiomatic approach to probability
- Probability distributions: discrete, continuous, in more dimensions
- Conditional probability, independent events and variables
- The Bayes theorem
- Examples of application of Bayes theorem
- Bayes rule and likelihood function
- Bayesian approach to probability as learning process
- Inference with the Bayesian approaches
- Choice of credible and confidence intervals: upper and lower limits
- Error propagation with Bayesian estimates
- Issues with prior PDF with the Bayesian approach

The goals of a physicist

- Measurements
- Discoveries



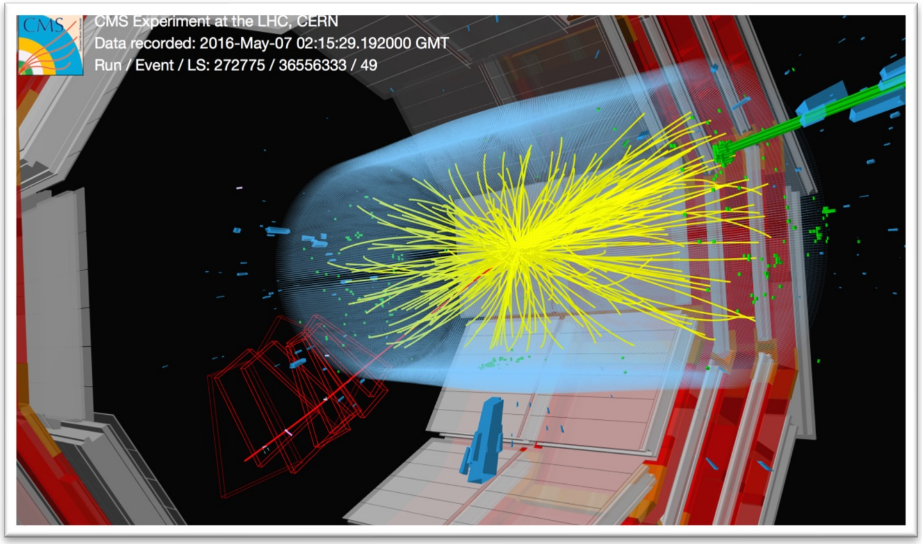
$m_{\mathrm{t}} = 173.49 \pm 1.07$ GeV



Higgs boson!

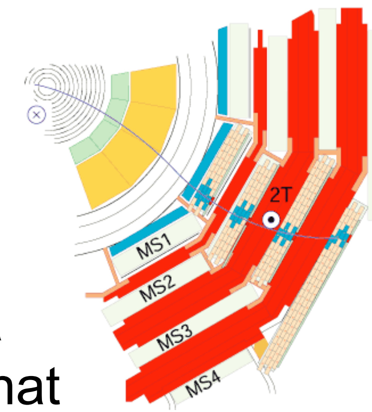
Probability & measurements

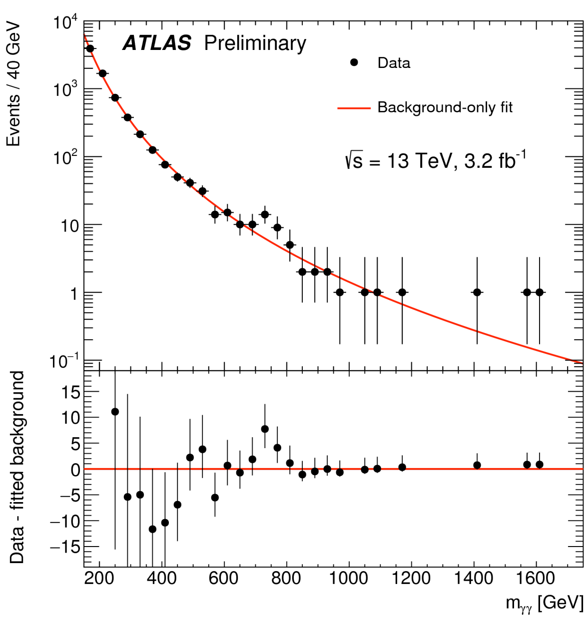
- Measurements are closely related to random processes and probability
- Repeated measurements always give the same result only for trivial cases
- Many physical processes have intrinsic randomness
  - Quantum Mechanics: $\mathcal{P} \propto |\mathcal{A}|^2$
- Detector response is somewhat random
  - Fluctuations, resolution, efficiency, …



Repeatable experiments

- What's the probability to extract one ace in a deck of cards?
- What is the probability to win a lottery (or bingo, …)?
- What is the probability that a pion is incorrectly identified as a muon in CMS?





- What is the probability that a fluctuation in the background can produce a peak in the $\gamma\gamma$ spectrum with a magnitude at least equal to what has been observed by ATLAS?
- **Note**: different question w.r.t.: what is the probability that the peak is due to a background fluctuation? (non repeatable!)

Non repeatable claims

DAY	HIGH/LOW	DESCRIPTION	PRECIP	WIND	HUMIDITY
TODAY May 20	69°/47°	Partly Cloudy	0%	E 5 mph	54%
SAT May 21	78°/51°	Sunny	0%	SSE 8 mph	54%
SUN May 22	73°/45°	Partly Cloudy	20%	SW 9 mph	58%
MON May 23	57°/43°	Rain	90%	SW 9 mph	67%
TUE May 24	66°/46°	Mostly Sunny	20%	NE 7 mph	57%
WED May 25	71°/51°	Partly Cloudy	20%	ESE 6 mph	60%

- Could be about future events:
  - what is the probability that tomorrow it will rain in Geneva?
  - what's the probability that your favorite team will win next championship?
- But also past events:
  - What's the probability that dinosaurs went extinct because of an asteroid?

- More in general, it's about unknown events:
  - What is the probability that dark matter is made of particles heavier than 1 TeV?
  - What is the probability that climate changes are mainly due to human intervention?

Introduction to probability

- Probability can be defined in different ways
- The applicability of each definition depends on the kind of claim we are considering to applying the concept of probability
- One subjective approach expresses the degree of belief/credibility of the claim, which may vary from subject to subject
- For repeatable experiments, probability may be a measure of how frequently the claim is true

Classical probability

- Probability determined by symmetry properties of a random device
- "Equally undecided" about event outcome, according to Laplace definition

Pierre Simon Laplace (1749-1827)

$$\text{Probability: } P = \frac{\text{Number of favorable cases}}{\text{Number of total cases}}$$

P = 1/2

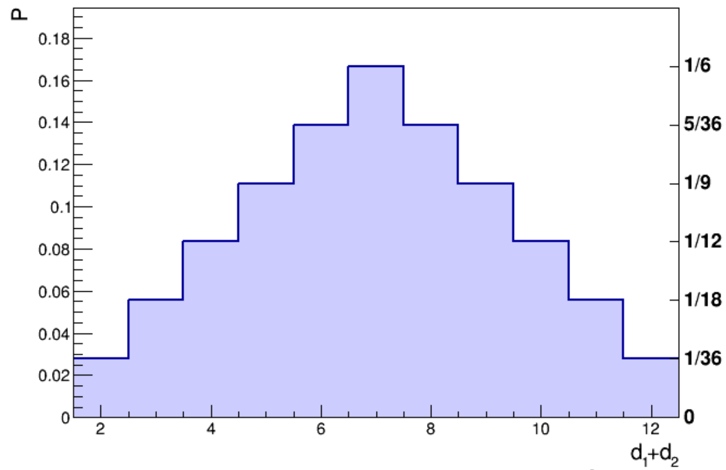
P = 1/6 (each dice)

P = 1/10

P = 1/4

Composite cases

- Composite cases are managed via combinatorial analysis
- Reduce the (composite) event of interest into elementary equiprobable events (sample space)



E.g:

$2 = \{(1,1)\}$
$3 = \{(1,2), (2,1)\}$
$4 = \{(1,3), (2,2), (3,1)\}$
$5 = \{(1,4), (2,3), (3,2), (4,1)\}$
etc. …

- Statements about an event can be defined via set algebra
  - **and/or/not** ⇒ intersection/union/complement
  - E.g.:

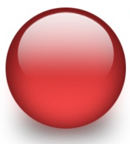
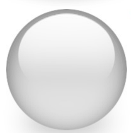
    "*sum of two dices is even and greater than four*"

    ⇕

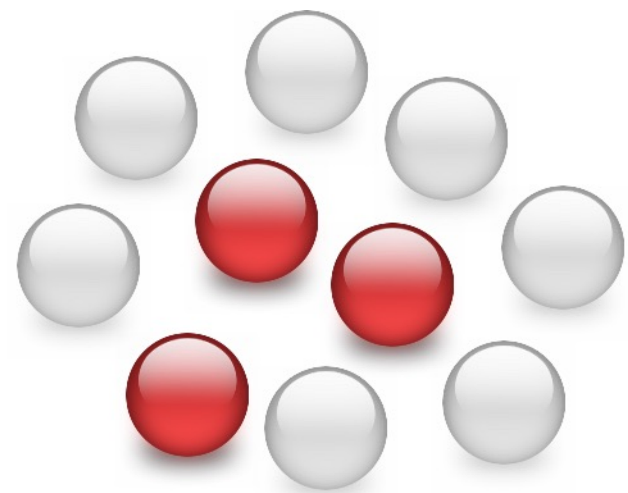
    $\{(d_1, d_2)\text{: mod}(d_1 + d_2, 2) = 0\} \cap \{(d_1, d_2)\text{: } d_1 + d_2 > 4\}$

Binomial distribution

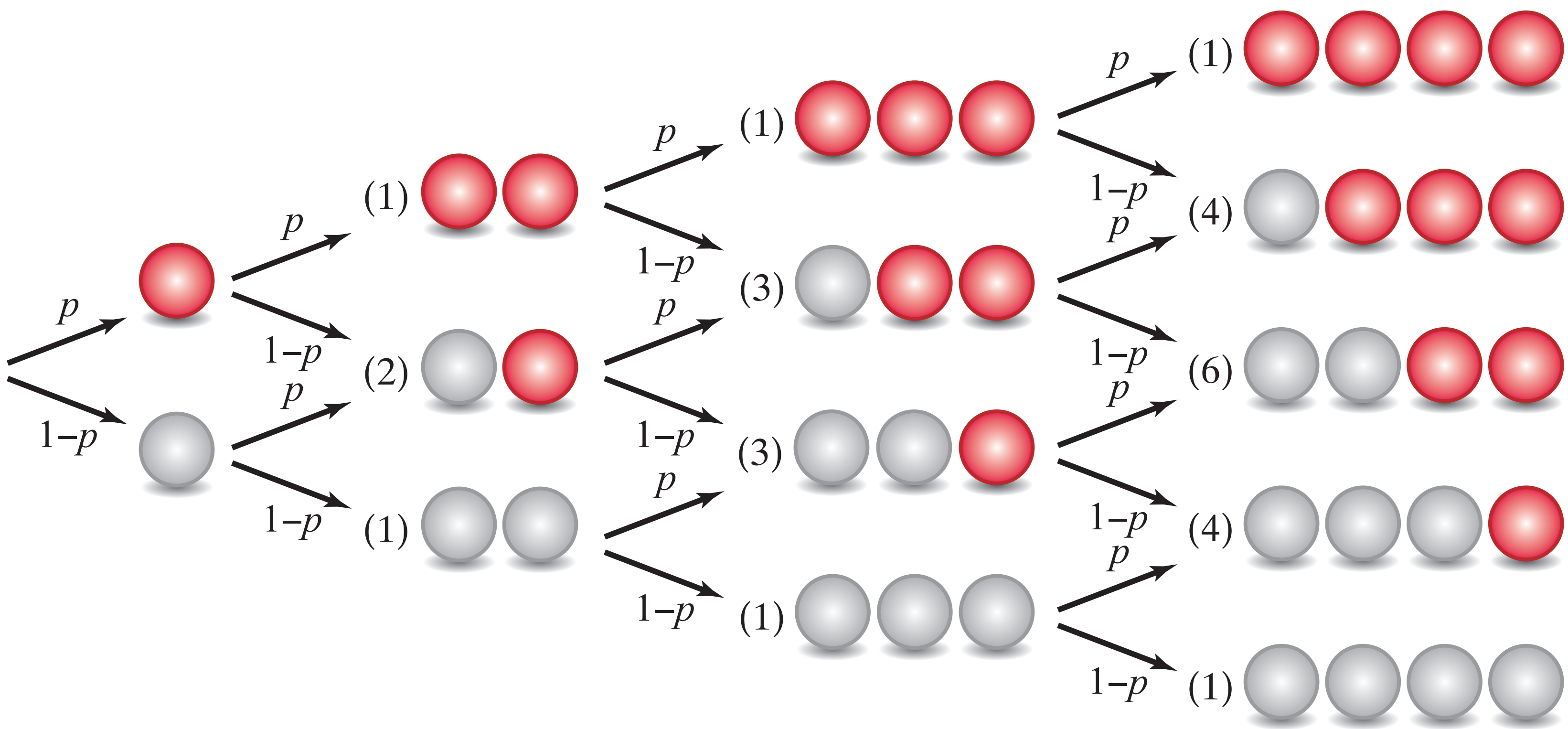
- Probability to extract n red balls over N trials, given the fraction p of red balls in a basket
- Each trial is called Bernoulli process

- Red:  $p = 3/10$
- White:  $1 - p = 7/10$

$$p = 3/10$$

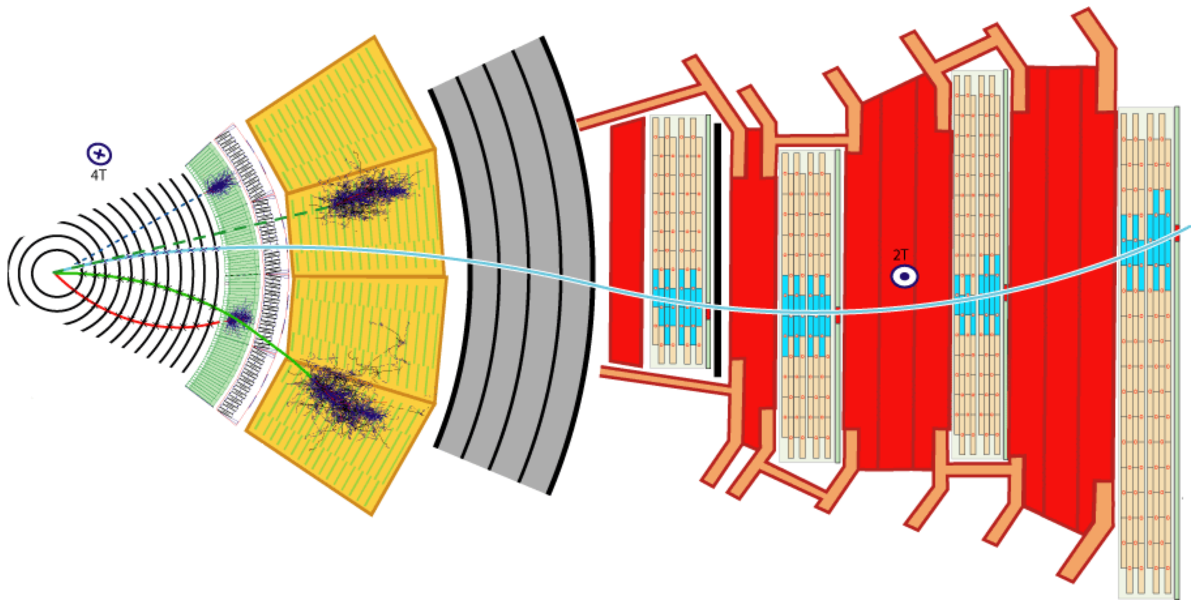


Binomial distribution



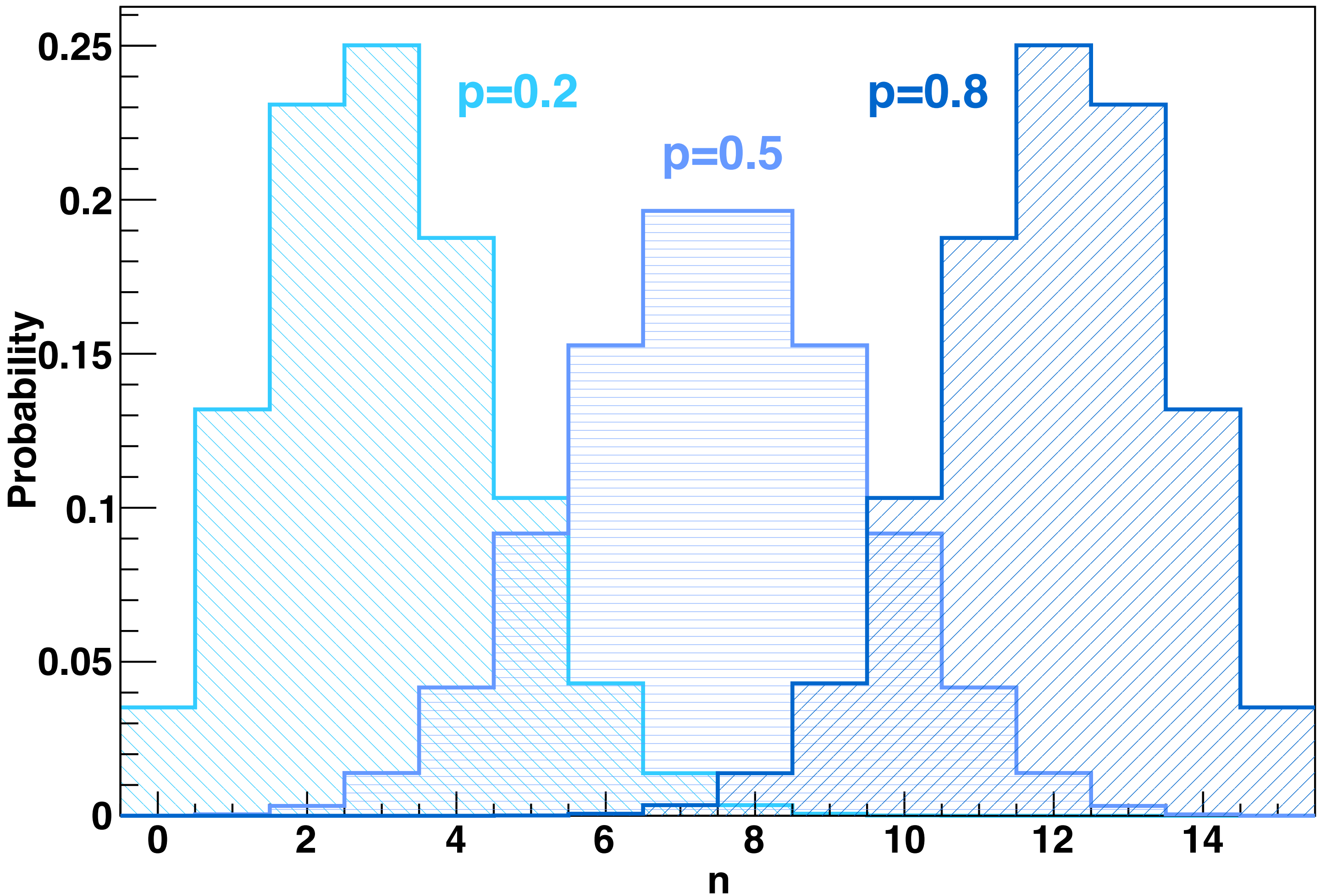
Binomial distribution

- Typical application in physics: detector efficiency ($\varepsilon = p$)



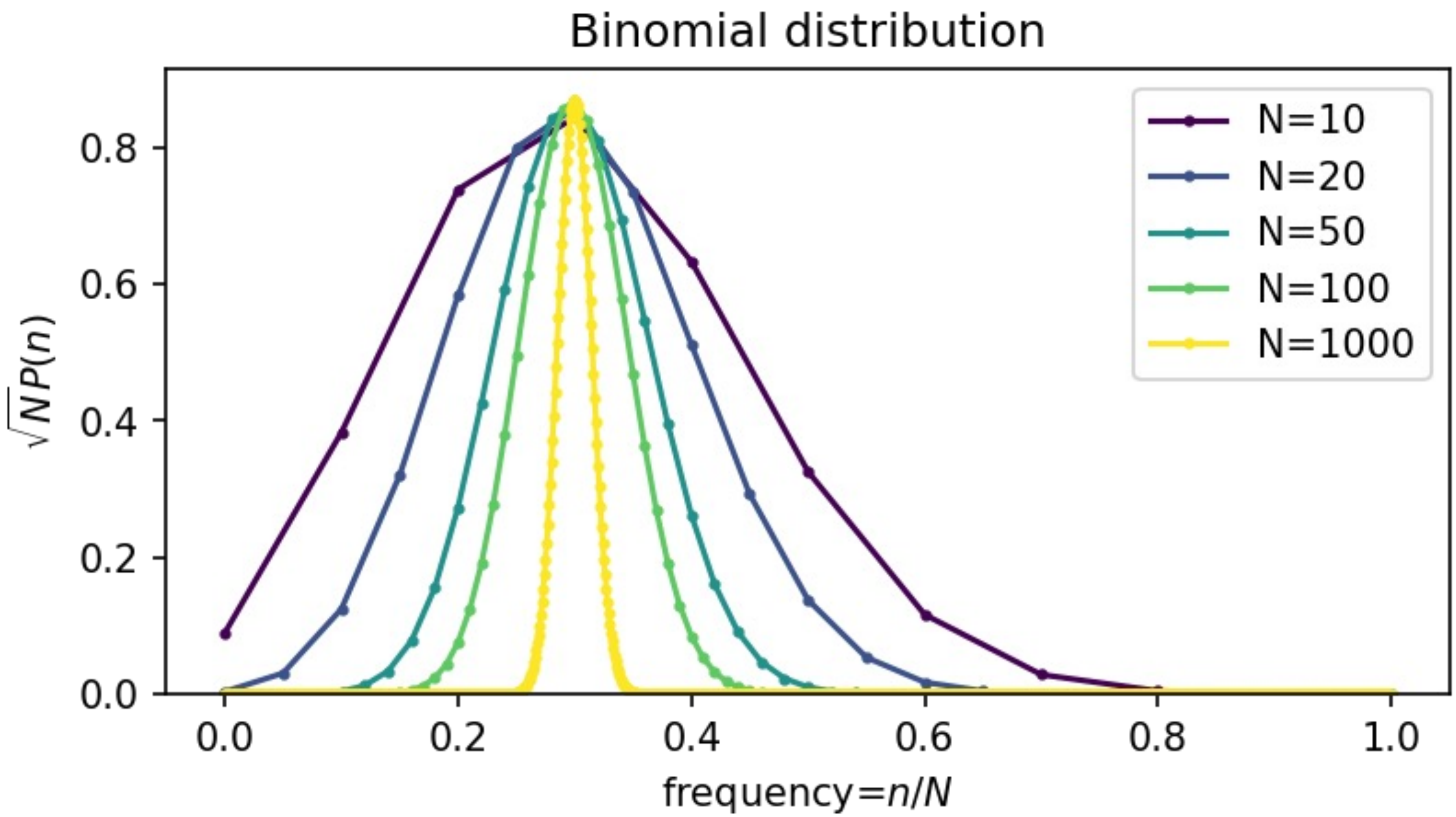
$$P(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Binomial distribution



Law of large numbers

- $\forall \varepsilon \quad \lim_{N\to\infty} P\left(\left|\frac{n}{N} - p\right| < \varepsilon\right) = 1$



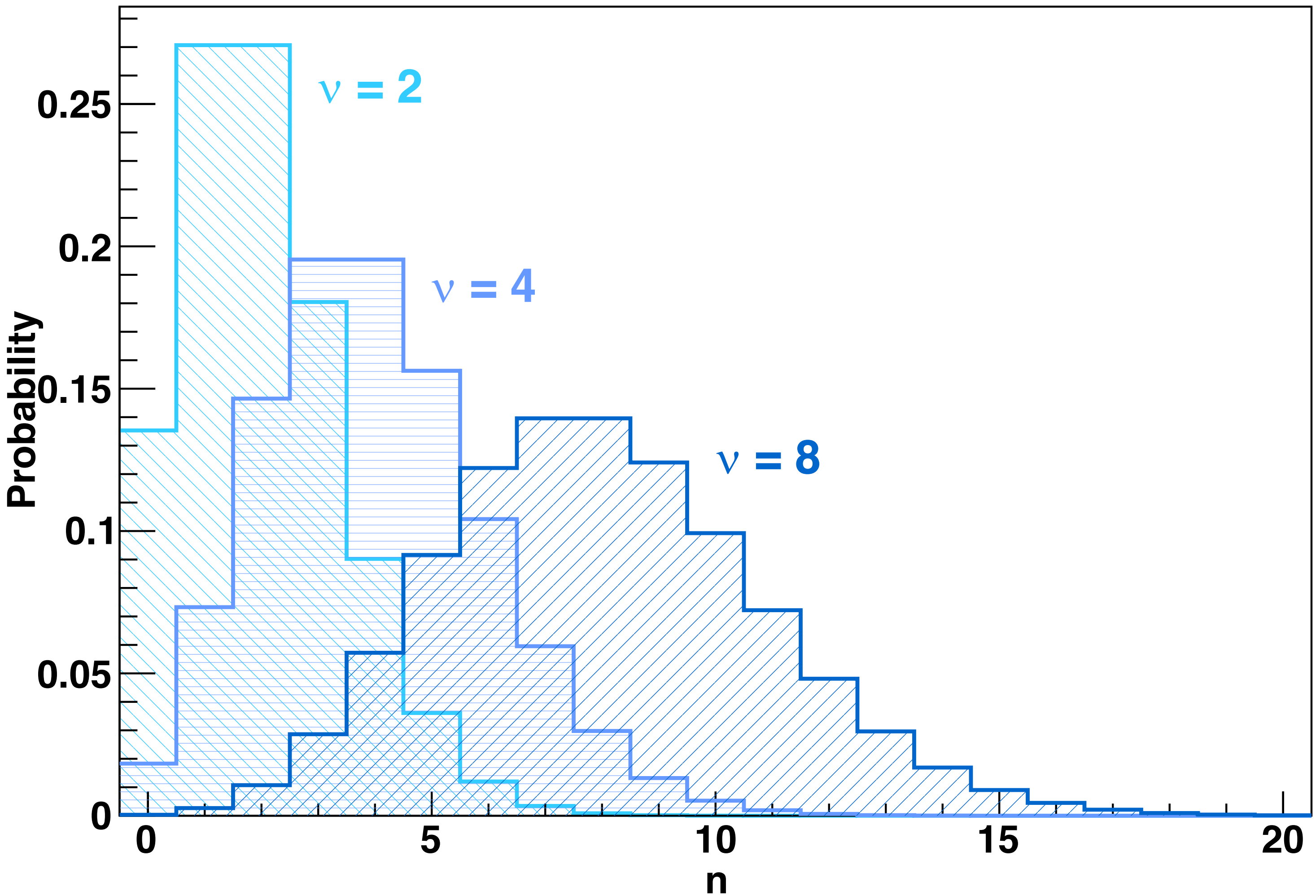
Poisson distribution

- Limit of Binomial for $N \to \infty$, $\nu = Np = \text{const.}$
- Distribution of the number of occurrences of random event uniformly distributed in a measurement range whose rate is known
  - E.g.: number of rain drops in a given area and in a given time interval, number of cosmic rays crossing a detector in a given time interval

Siméon-Denis Poisson (1781-1840)

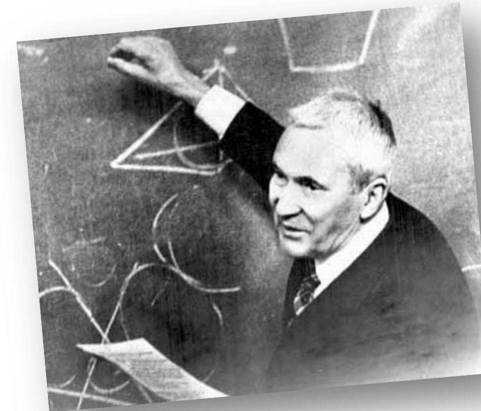
$$P(n;\nu) = \frac{\nu^n}{n!}e^{-\nu}$$

Poisson distribution



Kolmogorov approach

- Axiomatic probability definition
  - Terminology: Ω = sample space, F = event space, P = probability measure
  - Let (Ω, $F \subseteq 2^\Omega$, P) be a measure space that satisfy:

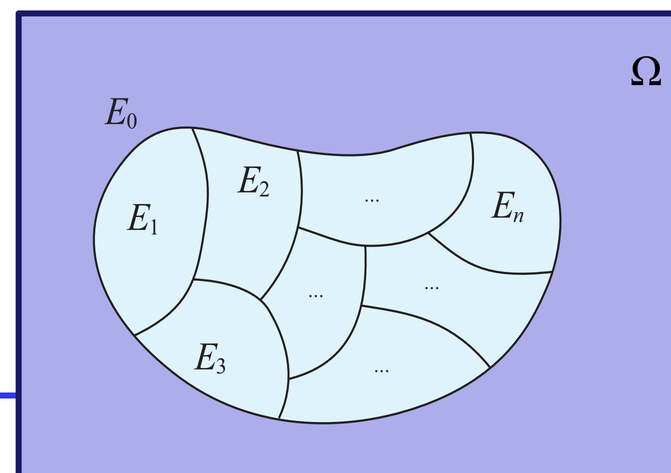


Andrej Nikolaevič Kolmogorov (1903-1987)

  - 1 $P(E) \geq 0 \quad \forall E \in F$
  - 2 $P(\Omega) = 1$ (normalization)
  - 3 $\forall (E_1, \cdots, E_n) \in F^n : E_i \cap E_j = 0$

$$P\left(\bigcup_{i=1,\cdots,n} E_i\right) = \sum_{i=1,\cdots,n} P(E_i)$$

- The same formalism applies to either frequentist and Bayesian probability



Subjective (Bayesian) probability

- Expresses one's degree of belief that a claim is true
  - How strong? How much would you bet?
  - Applicable to all unknown events/claims, not only repeatable experiments
  - Each individual may have a different opinion/prejudice

- Quantitative rules exist about how subjective probability should be modified after **learning** about some observation/evidence
  - Consistent with Bayes theorem (→ will be introduced in next slides)
  - Prior probability → Posterior probability (following observation)
  - The more information we receive, the more Bayesian probability is insensitive on prior subjective prejudice (unless in pathological cases…)

Probability distributions

- Given a discrete random variable, we can assign a probability to each individual value:

$$P(x) = P(\{x\})$$

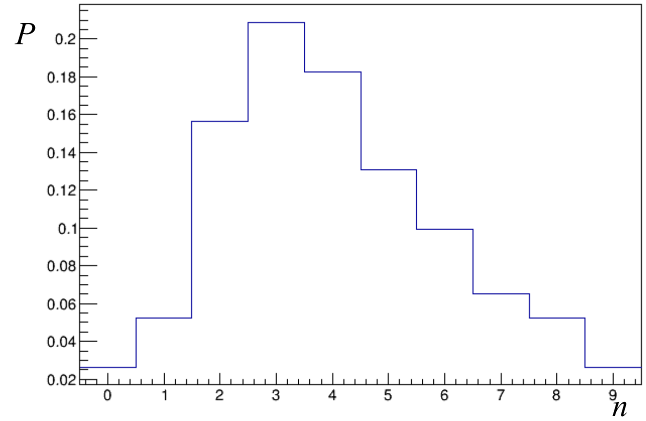
- In case of a continuous variable, the probability assigned to an individual value may be zero
- A probability density better quantifies the probability content (unlike $P(\{x\}) = 0$!):

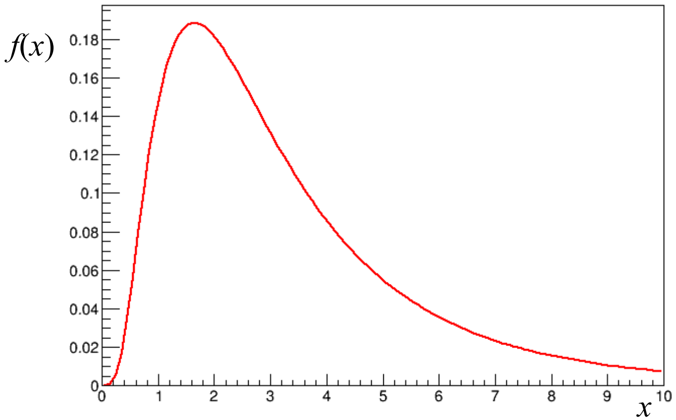
$$\frac{\mathrm{d}P(x)}{\mathrm{d}x} = f(x)$$

- Discrete and continuous distributions can be combined using Dirac's delta functions.
- E.g.:

$$\frac{\mathrm{d}P}{\mathrm{d}x} = \frac{1}{2}\delta(x) + \frac{1}{2}f(x)$$

50% prob. to have zero ($P(\{0\}) = 0.5$), 50% distributed according to $f(x)$

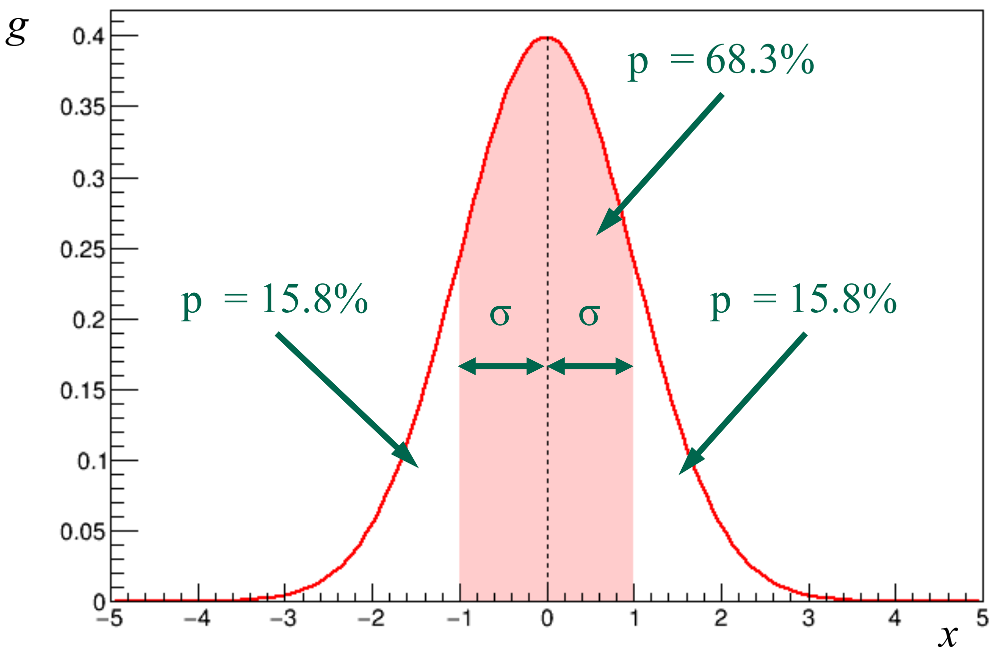




Gaussian distribution

- Many random variables in real experiments follow a Gaussian distribution
- Central limit theorem: approximate sum of multiple random contributions, regardless of the individual distributions
- Frequently used to model detector resolution

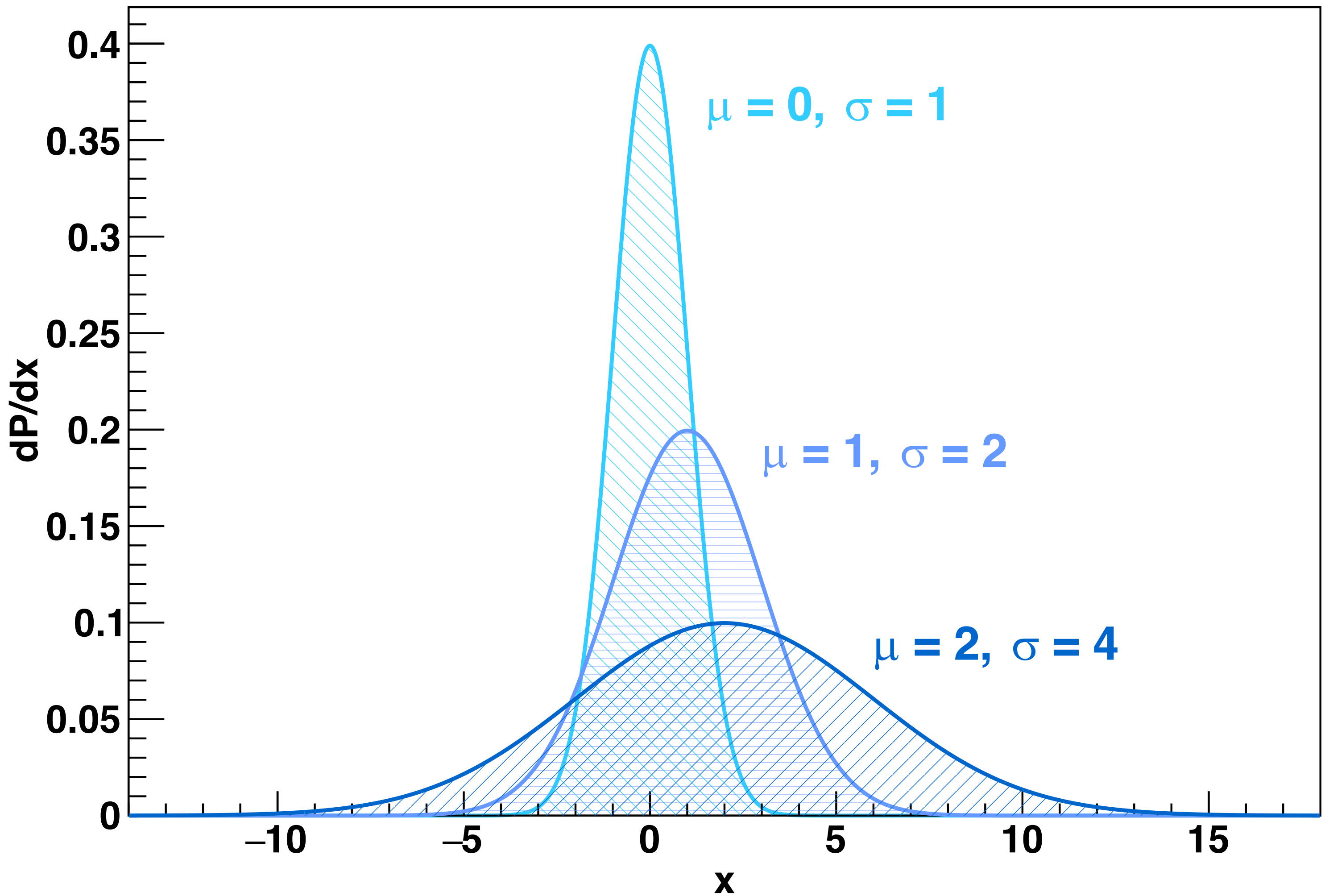
Carl Friedrich Gauss (1777-1855)



$$g(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$n\sigma$	Prob.
1	0.683
2	0.954
3	0.997
4	$1 - 6.3\times10^{-5}$
5	$1 - 5.7\times10^{-7}$

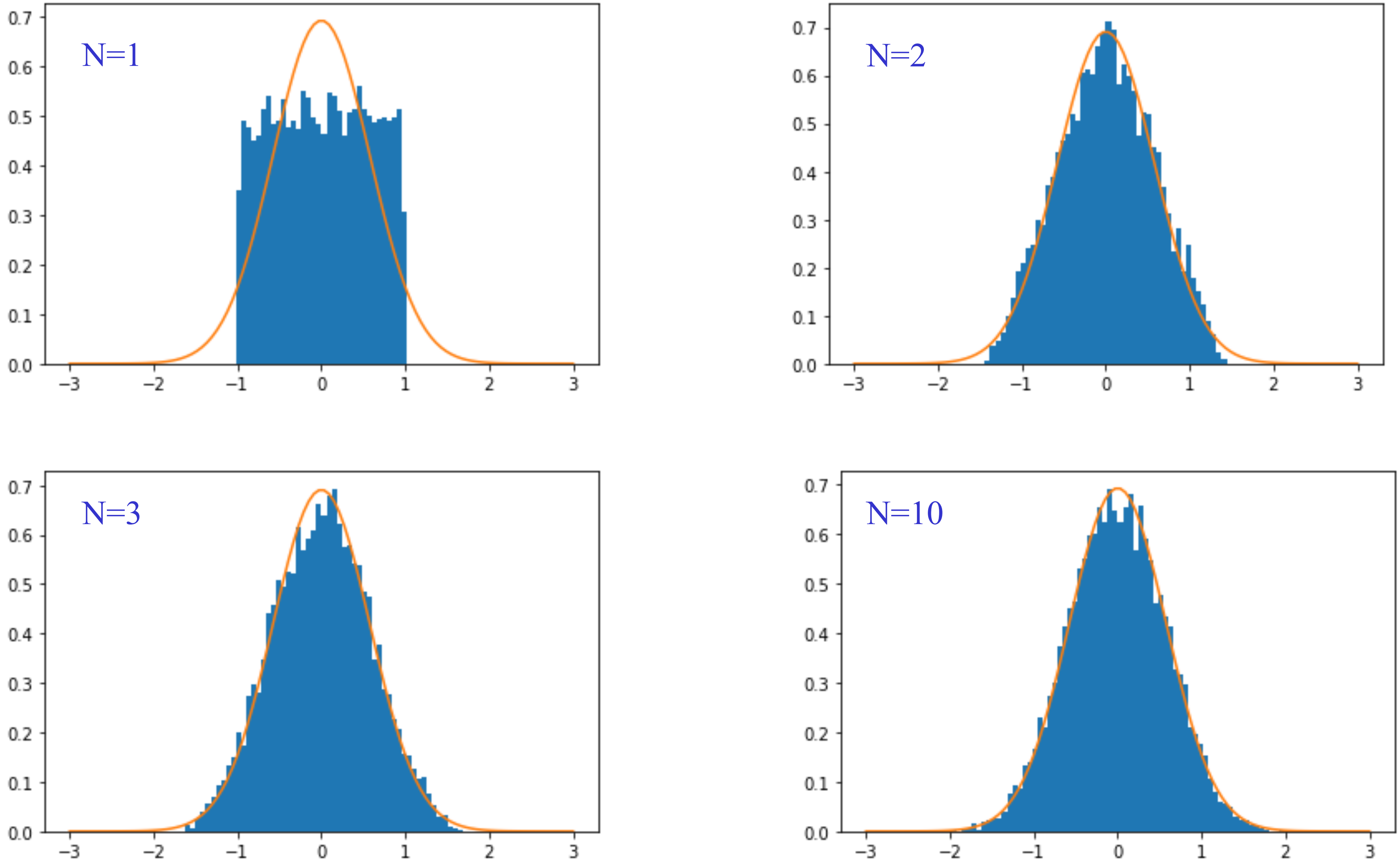
Gaussian distribution



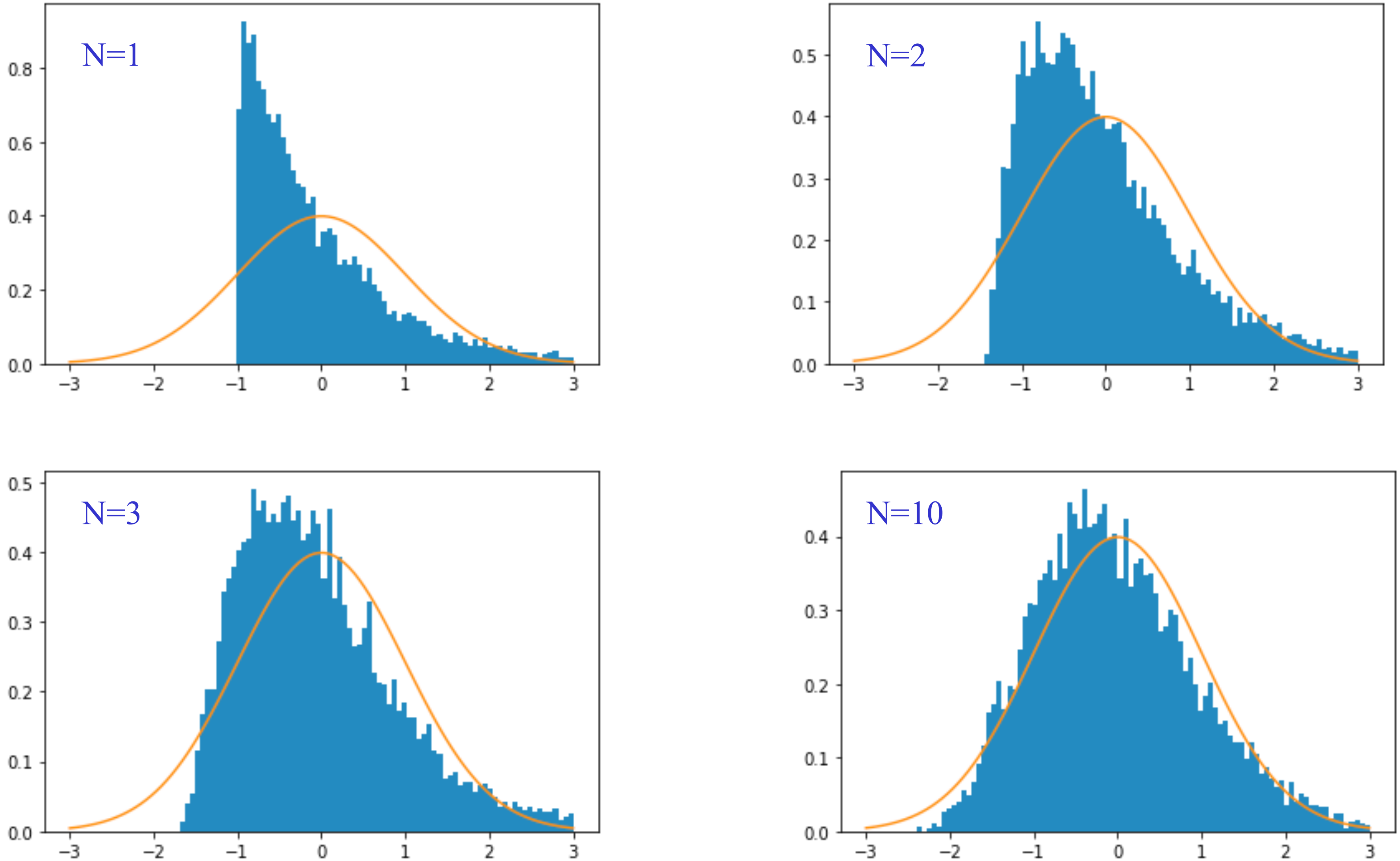
Central limit theorem

- The distribution of the sum or average of N random variables all having the same distribution with finite variance tends to a Gaussian for large N

Central limit theorem

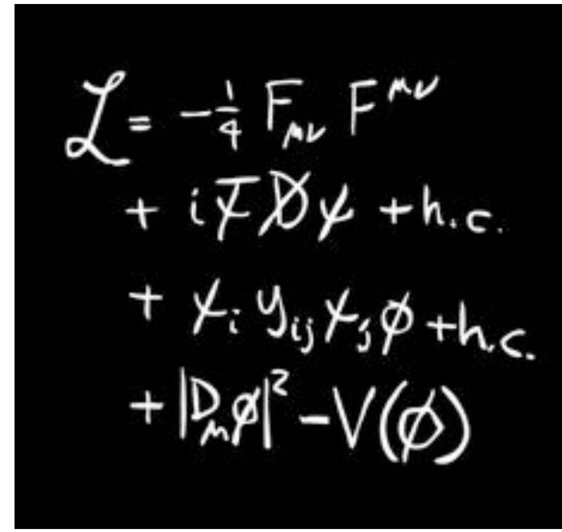


Central limit theorem



Theory inputs

- Distributions of measured quantities in data:
  - are predicted by a theory model,
  - depend on some theory parameters,
  - e.g.: particle mass, cross section, etc.
- Given our data sample, we want to:
  - measure theory parameters,
    - e.g.: $m_{\mathrm{t}} = 173.49 \pm 1.07\ \mathrm{GeV}$
  - answer questions about the nature of data
    - Is there a Higgs boson? ➔ Yes! (strong evidence? Quantify!)
    - Is there a Dark Matter? ➔ No evidence, so far…
    - If not, what is the range of theory parameters compatible with the observed data? What parameter range can we exclude?



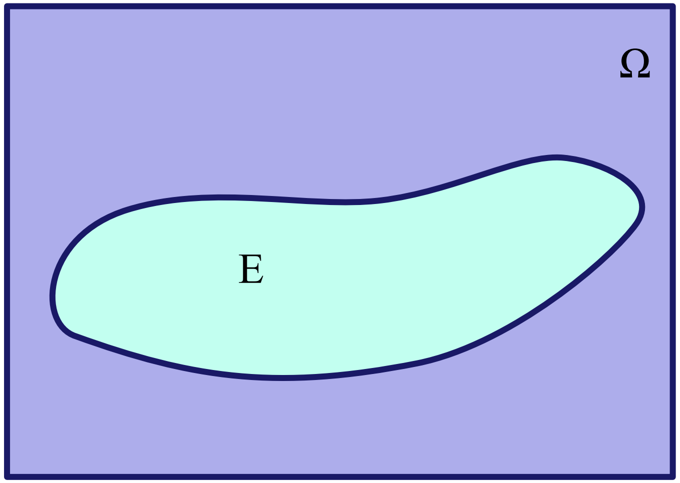
PDFs in more dimensions

- In more dimensions (n random variables), PDF can be defined as:

$$\frac{\mathrm{d}^n P}{\mathrm{d}x_1 \cdots \mathrm{d}x_n} = f(x_1, \cdots, x_n)$$

- The probability associated to an event E is obtained by integrating the PDF over the corresponding set in the sample space

$$P(E) = \int_E f(x_1, \cdots, x_n)\mathrm{d}^n x$$



Mean and variance of random vars.

- Given a random variable x with distribution $f(x)$ we can define:
- Mean or expected value:

$$\begin{cases} \mathrm{E}[x] = \langle x \rangle = \int x f(x) \mathrm{d}x \\ \mathrm{E}[g(x)] = \langle g(x) \rangle = \int g(x) f(x) \mathrm{d}x \end{cases}$$

- Variance:

$$\mathrm{Var}[x] = \langle (x - \langle x \rangle)^2 \rangle = \overbrace{\langle x^2 \rangle}^{\text{r.m.s.: root mean square}} - \langle x \rangle^2$$

- Standard deviation:

$$\sigma_x = \sqrt{\mathrm{Var}[x]} = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$$

- Covariance and correlation coefficient of two variables x and y:

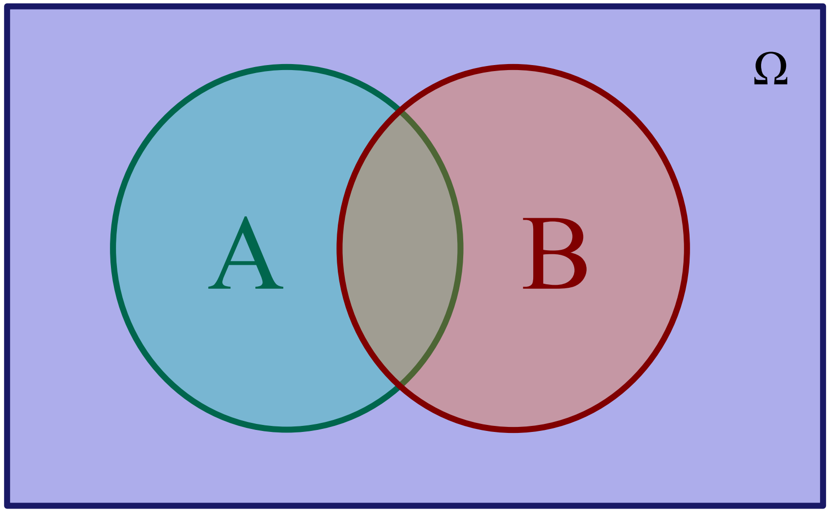
$$\mathrm{cov}(x, y) = \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle$$

$$\rho_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

Note: integration extends over x and y in two dimensions when computing an average value!

Conditional probability

- Probability of A, given B: $P(A \mid B)$, i.e.: probability that an event known to belong to set B also belongs to set A:
  - $P(A \mid B) = P(A \cap B) / P(B)$
  - Notice that:
    $P(A \mid \Omega) = P(A \cap \Omega) / P(\Omega)$



- Event A is said to be independent of B if the probability of A given B is equal to the probability of A:
  - $P(A \mid B) = P(A)$
- If A is independent of B then $P(A \cap B) = P(A)\, P(B)$
- → If A is independent on B, B is independent on A

Independent variables

$$\frac{\mathrm{d}^2 P}{\mathrm{d}x\mathrm{d}y} = f(x, y)$$

- 1D projections: (marginal distributions)

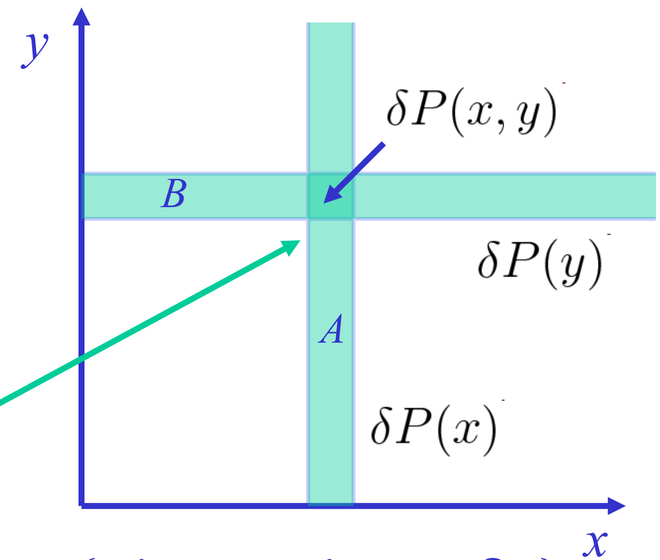
$$\begin{cases} f_x(x) = \int f(x,y)\mathrm{d}y \\ f_y(y) = \int f(x,y)\mathrm{d}x \end{cases}$$

- x and y are independent if:

$$f(x, y) = f_x(x) f_y(y)$$

- We saw that A and B are independent events if:

$$P(A \cap B) = P(A)P(B)$$



- Where $A = \{x' : x < x' < x + \delta x\}$, $B = \{y' : y < y' < y + \delta y\}$

The Bayes theorem

$$\underbrace{P(A|B) = \frac{P(A \cap B)}{P(B)} \quad P(B|A) = \frac{P(A \cap B)}{P(A)}}$$

$$P(A|B)P(B) = P(B|A)P(A)$$

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

Thomas Bayes (1702-1761)

- $P(A)$ = prior probability
- $P(A|B)$ = posterior probability

Bayesian posterior probability

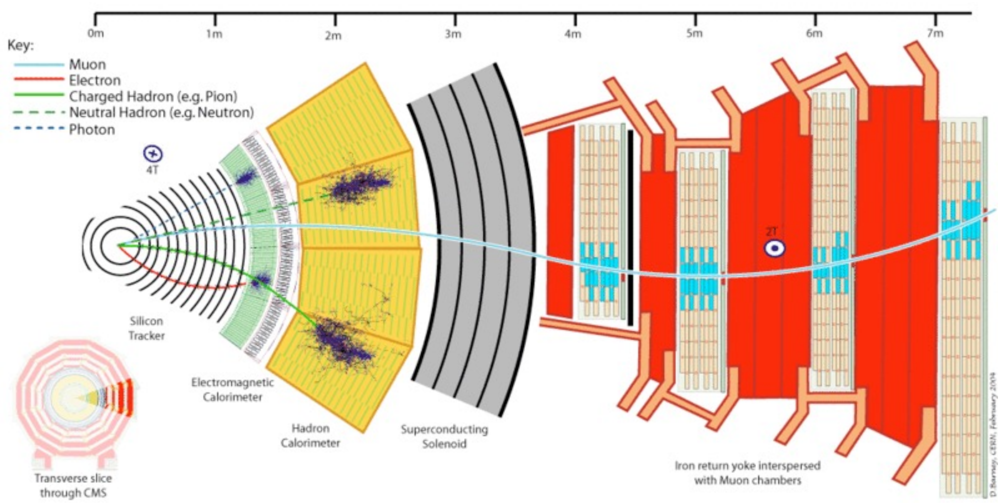
- Bayes theorem allows to determine probability about hypotheses or claims H that not related random variables, given an observation or evidence E:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- $P(H)$ = prior probability
- $P(H \mid E)$ = posterior probability, given E
- The Bayes rule allows to define a rational way to modify one's prior belief once some observation is known

Example (frequentist): muon fake rate

- A detector identifies muons with high efficiently, $\varepsilon = 95\%$
- A small fraction $\delta = 5\%$ of pions are incorrectly identified as muons ("fakes")
- If a particle is identified as a muon, what is the probability it is really a muon?
  - The answer also depends on the composition of the sample!
  - i.e.: the fraction of muons and pions in the overall sample



This example is usually presented as an epidemiology case.

Naïve answers about fake positive probability are often wrong!

Fakes and Bayes theorem

- Using Bayes theorem:
  - $P(\mu|+) = P(+|\mu)\, P(\mu) \,/\, P(+)$

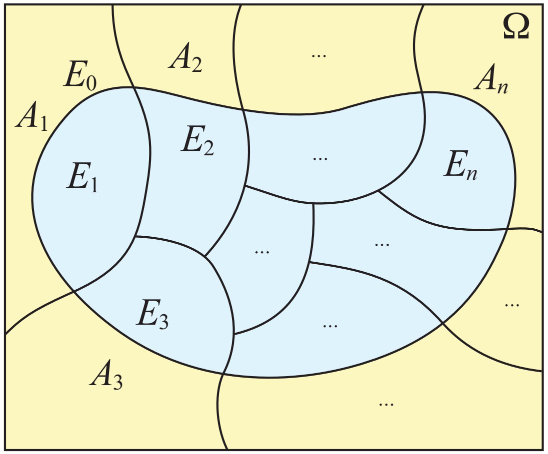
+ denotes a positive id

- Where our inputs are:
  - $P(+|\mu) = \varepsilon = 0.95$, $P(+|\pi) = \delta = 0.05$
- We can decompose $P(+)$ as:
  - $P(+) = P(+|\mu)\, P(\mu) + P(+|\pi)\, P(\pi)$

normalization term

- Putting all together:
  - $P(\mu|+) = \varepsilon\, P(\mu) \,/\, (\varepsilon\, P(\mu) + \delta\, P(\pi))$

Law of total probability

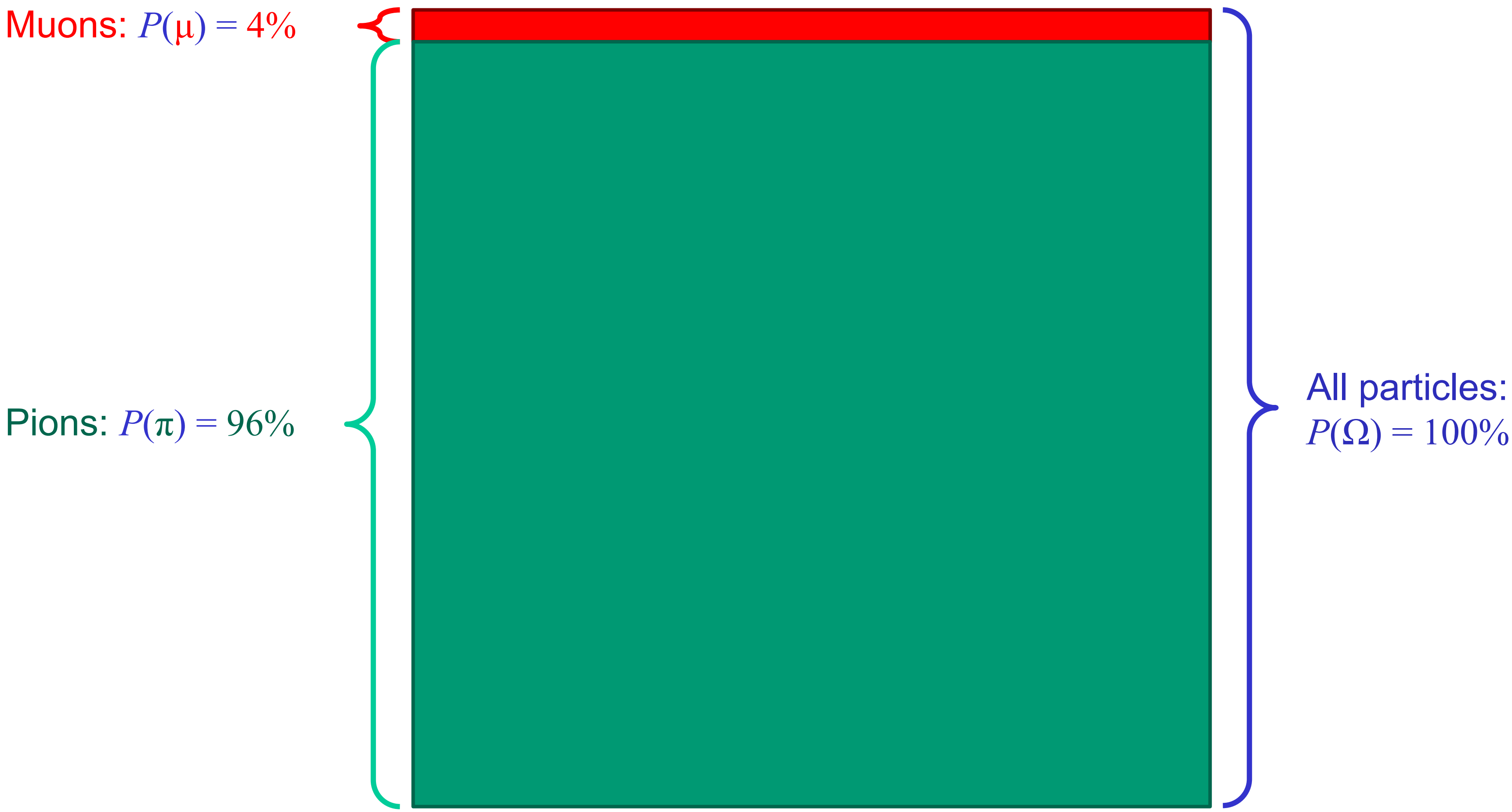


$$P(E_0) = \sum_{i=1}^{n} P(E_0|A_i)P(A_i)$$

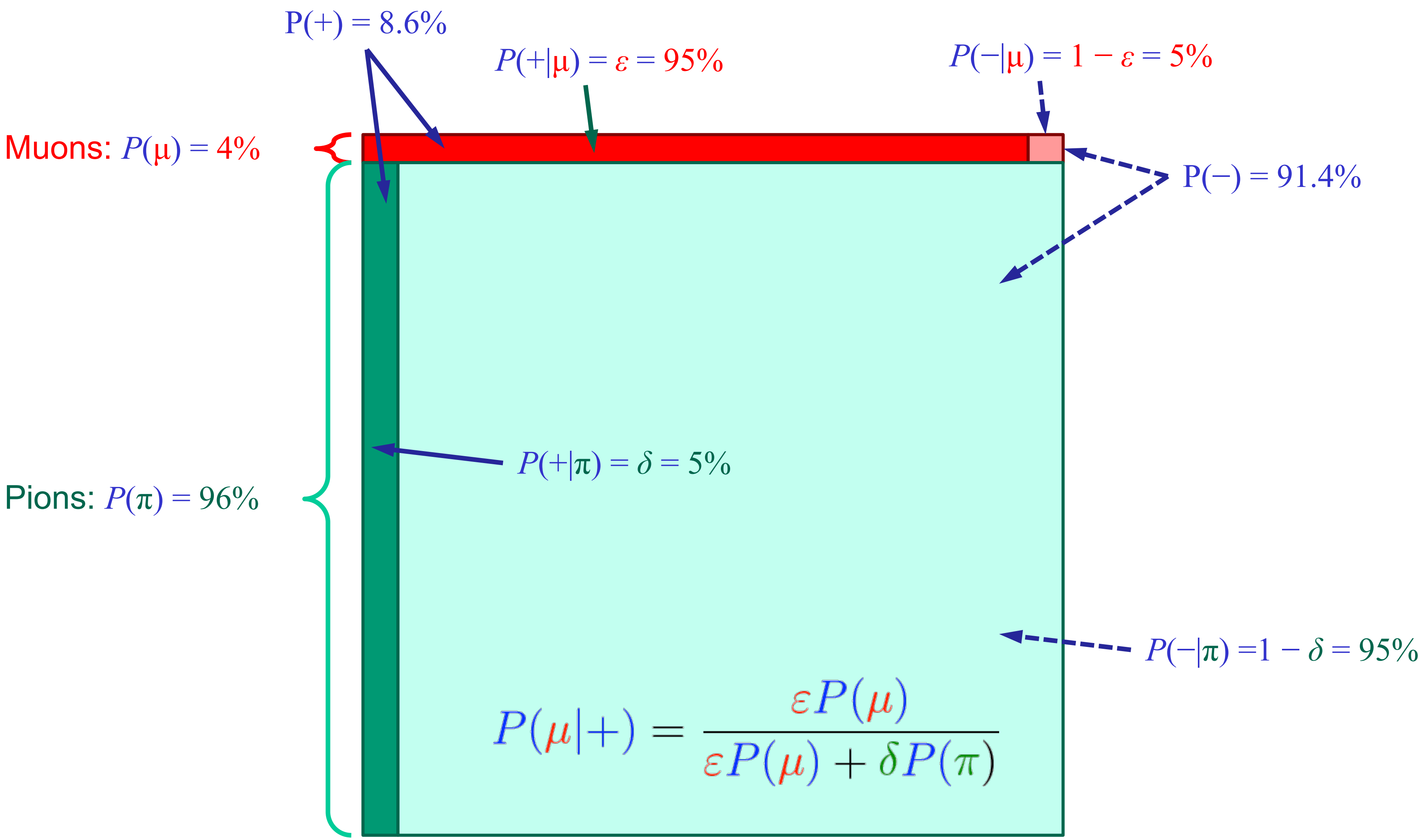
$E_0 = \text{'+'}$, $A_i = \mu, \pi$

- Assume we have a sample made of $P(\mu)=4\%$ muons and $P(\pi)=96\%$ pions, we have:
  - $P(\mu|+) = 0.95 \times 0.04 \,/\, (0.95 \times 0.04 + 0.05 \times 0.96) \cong 0.44$
- Even if the selection efficiency is very high, the low sample purity makes $P(\mu|+)$ lower than 50%.

Before any muon id. information



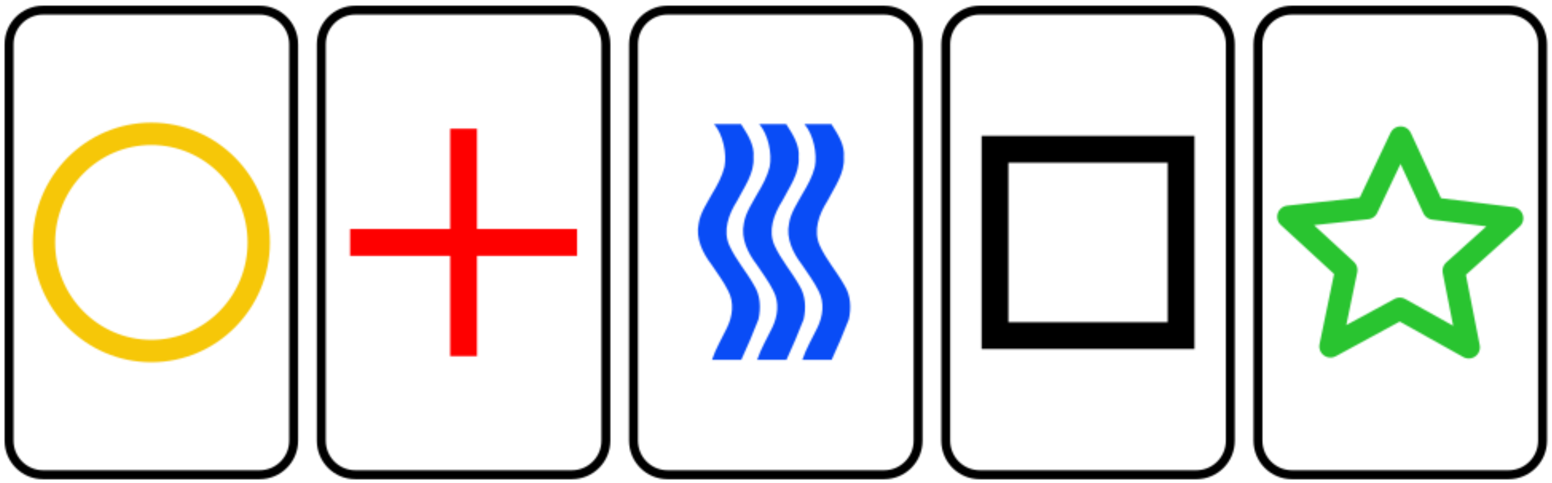
After the muon id. measurement



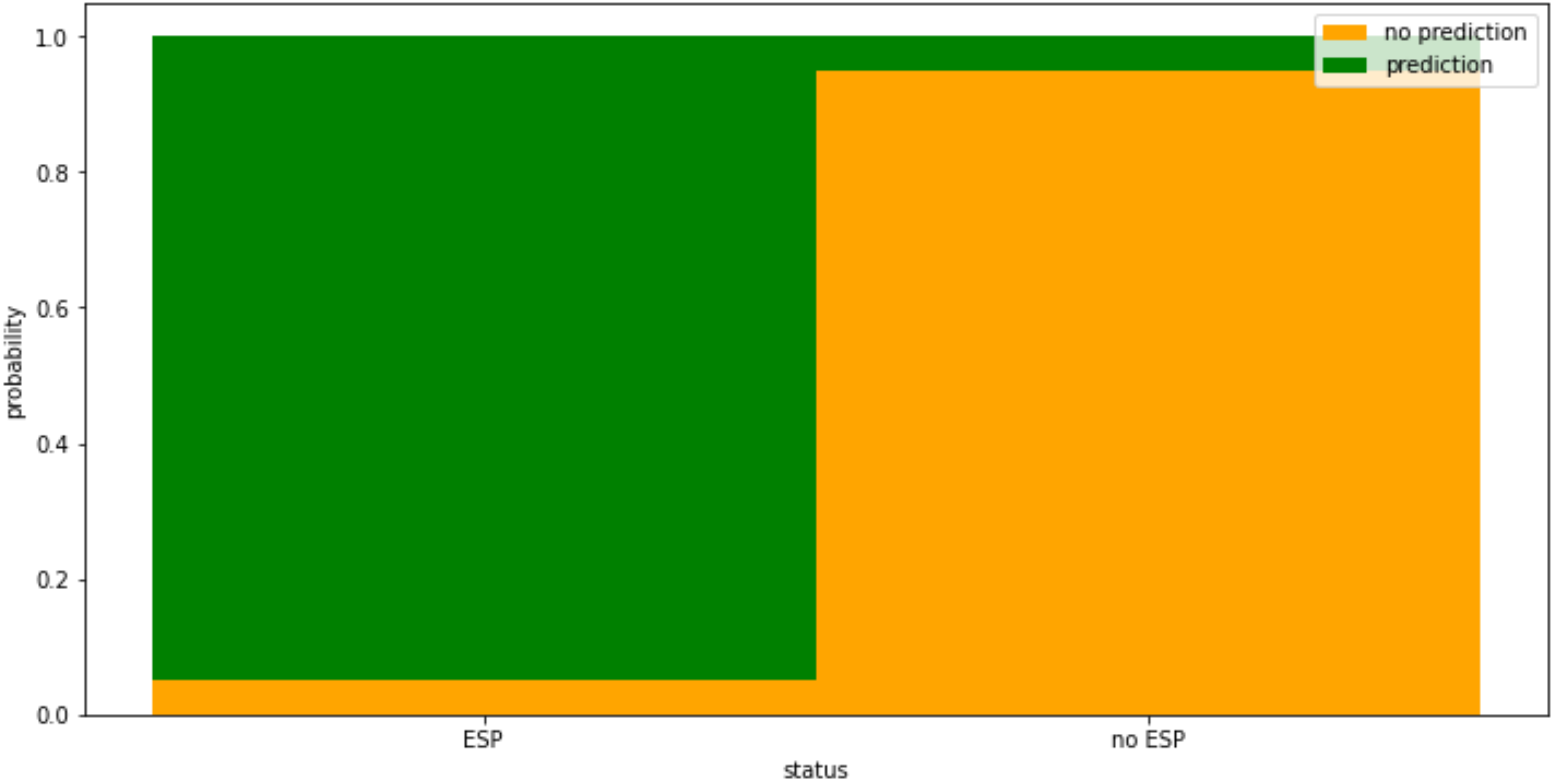
$$P(\mu|+) = \frac{\varepsilon P(\mu)}{\varepsilon P(\mu) + \delta P(\pi)}$$

The same approach to unknowns

- ESP: extra-sensory perception: Supernatural (?) prediction of a set of cards

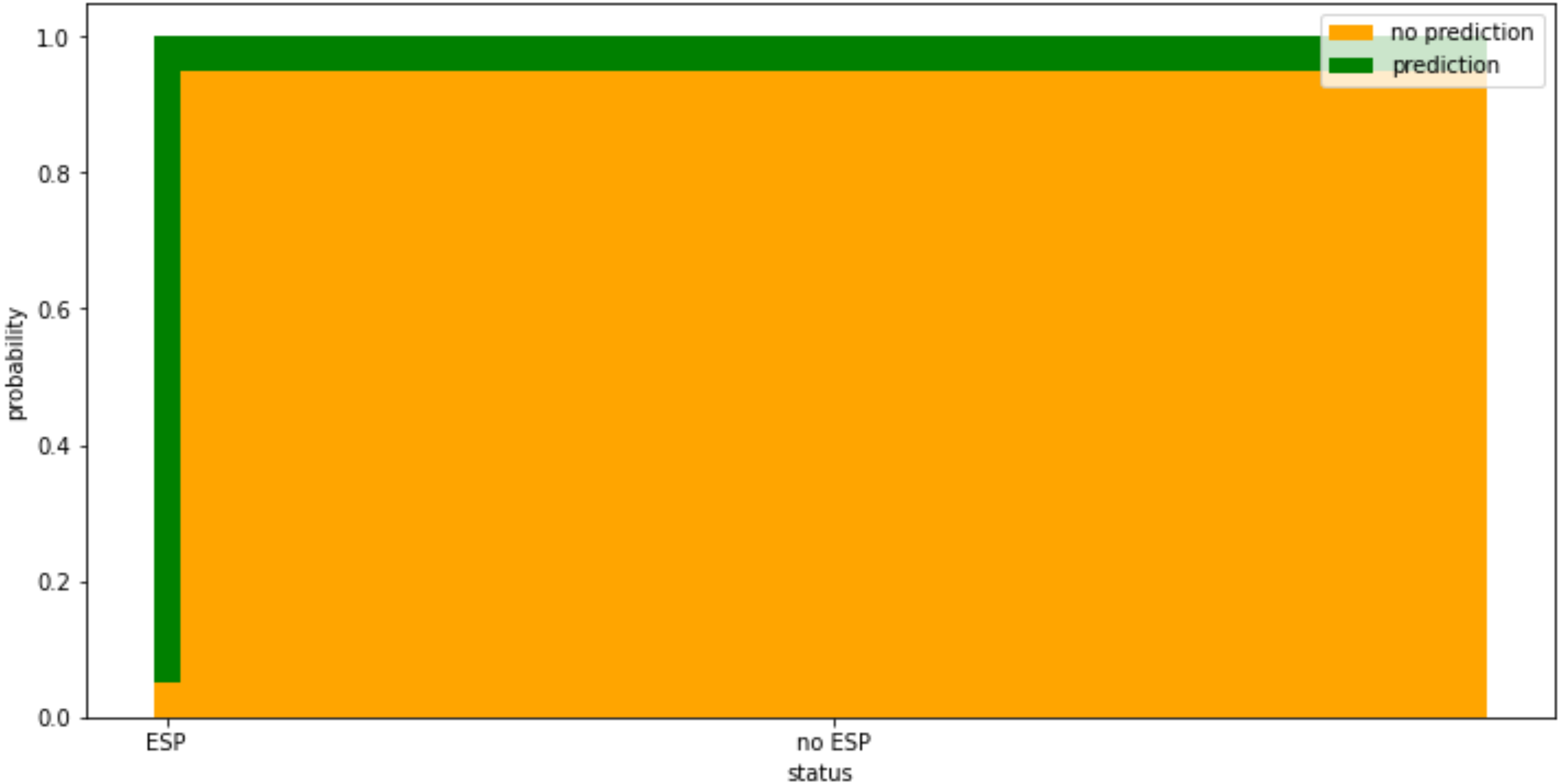


50%-50% prior

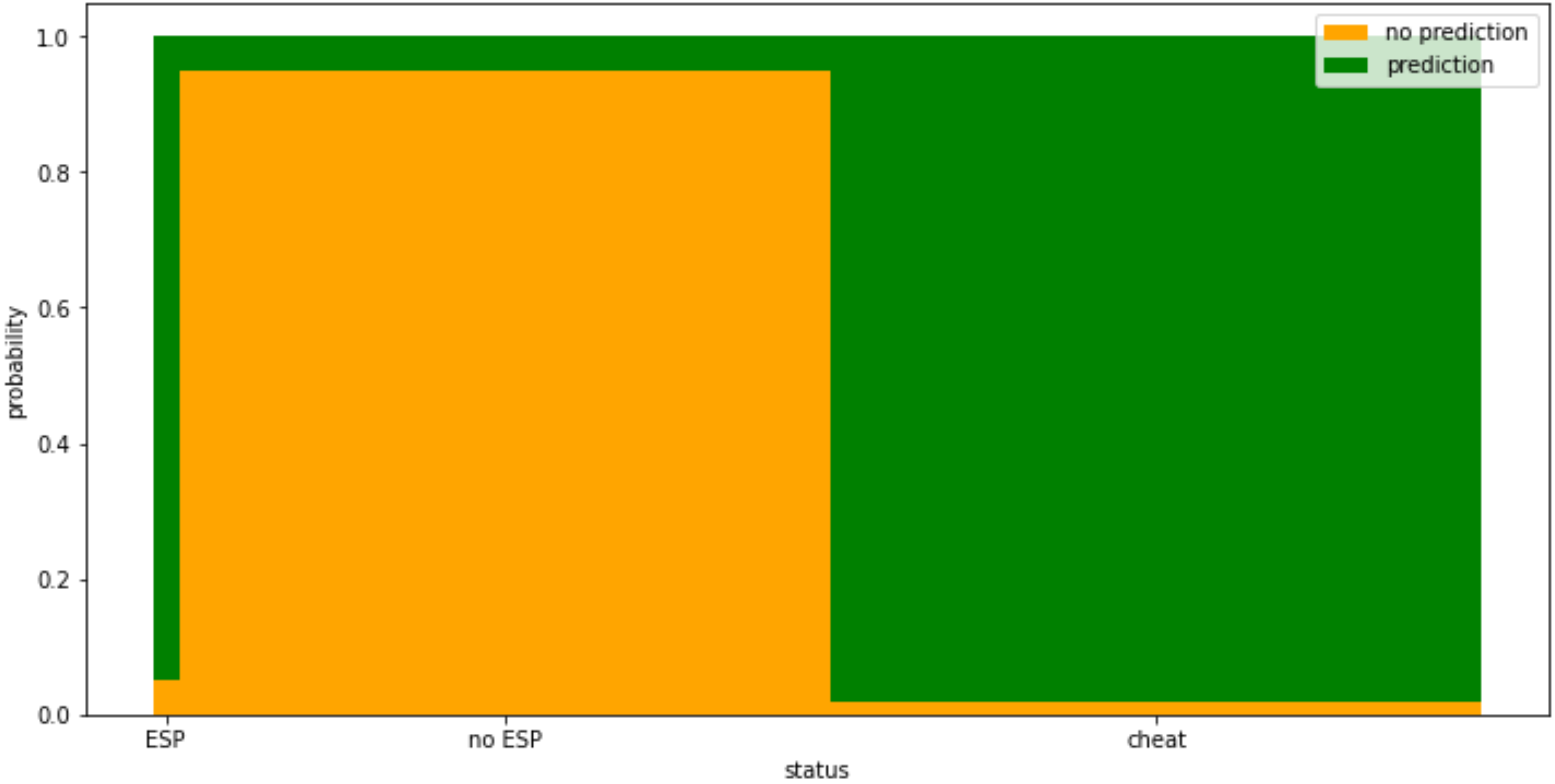


no prediction
prediction
1.0
0.8
0.6
0.4
0.2
0.0
probability
ESP
no ESP
status

Skeptical prior



Suspicious prior



Conspiracy theories

- Note that claiming that all scientific evidences are fakes is the approach of conspiracy theorists
- The main difference is that the prior chosen by conspiracy theorists ignore most of the evidence in favour of fantasy inventions
- Scientific reasoning is closely related to the Bayesian approach, provided that evidences are not discarded on purpose

The likelihood function

- In many cases, the outcome of our experiment can be modeled as a set of random variables $x_1, \ldots, x_n$ whose distribution takes into account:
  - intrinsic sample randomness (quantum physics is intrinsically random),
  - detector effects (resolution, efficiency, …).
- Theory and detector effects can be described according to some parameters $\theta_1, ..., \theta_m$, whose values are, in most of the cases, unknown
- The overall PDF, evaluated at our observation $x_1, \ldots, x_n$, is called likelihood function:

$$L = f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)$$

- In case our sample consists of N independent measurements (collision events) the likelihood function can be written as:

$$L = \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)$$

Bayes rule and likelihood function

- Given a set of measurements $x_1, \ldots, x_n$, Bayesian posterior PDF of the unknown parameters□$\theta_1, \ldots, \theta_m$ can be determined as:

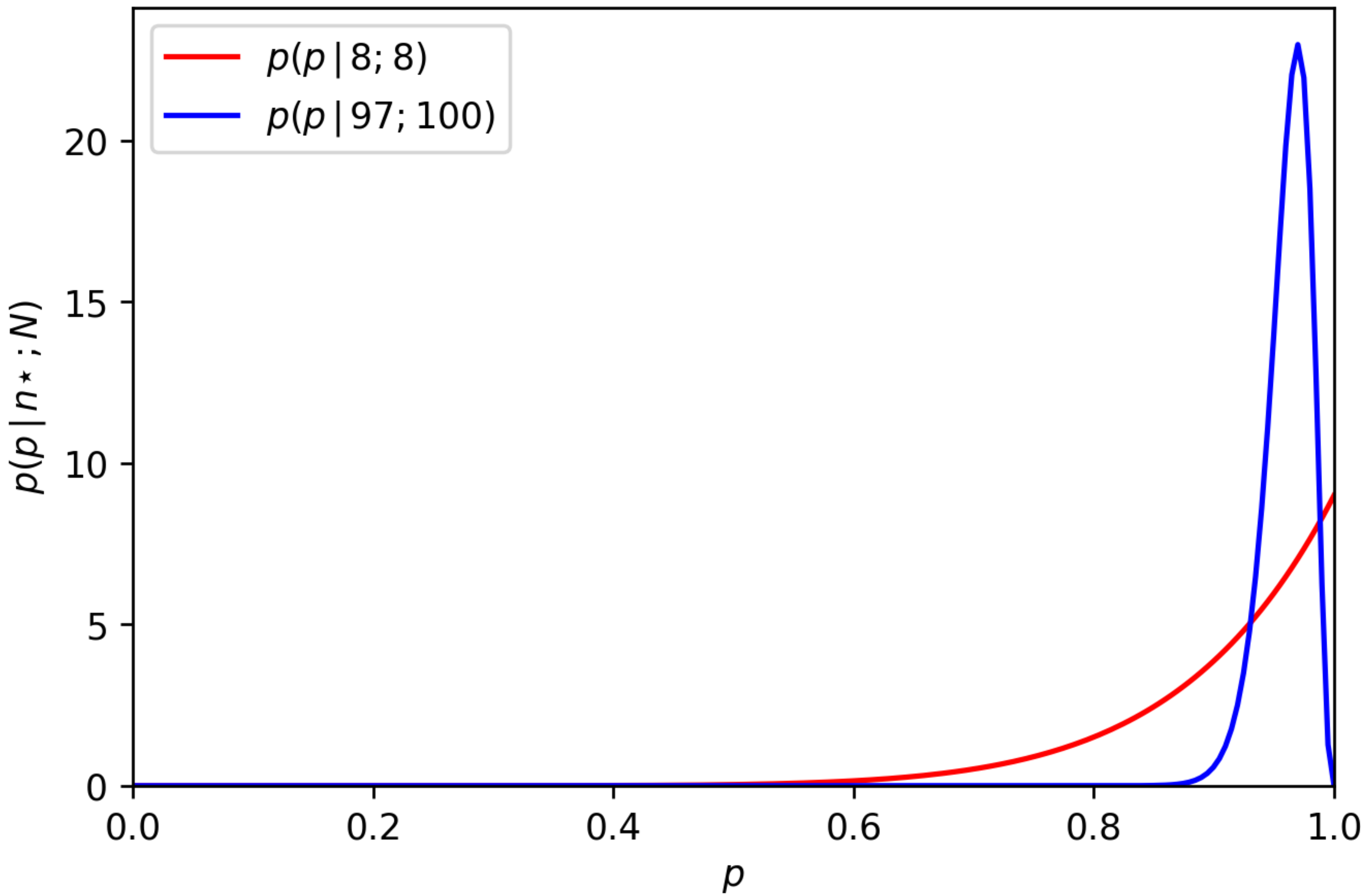
$$P(\theta_1, \cdots, \theta_m | x_1, \cdots, x_n) = \frac{L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)\pi(\theta_1, \cdots, \theta_m)}{\int L(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)\pi(\theta_1, \cdots, \theta_m)\mathrm{d}^m\theta}$$

- Where $\pi(\theta_1, \ldots, \theta_m)$ is the subjective prior probability
- The denominator $\int L(x, \theta)\, \pi(\theta)\, \mathrm{d}^m\theta$ is a normalization factor
- The observation of $x_1, \ldots, x_n$ modifies the prior knowledge of the unknown parameters $\theta_1, \ldots, \theta_m$
- If $\pi(\theta_1, \ldots, \theta_m)$ is sufficiently smooth and L is sharply peaked around the true values $\theta_1, \ldots, \theta_m$, the resulting posterior will not be strongly dependent on the prior's choice

Feedbacks and reliability

- The number of positive/negative feedbacks of an online seller can be assumed to to follow a binomial distribution
- If a seller has 100% of positive feedbacks (8/8) and another one has 97% (97/100), which one is more reliable?
- The posterior a special case of the so-called Beta distribution, prop to: $p^n(1-p)^{n-N}$:

$$p(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)}$$


$p(p\,|\,8;8)$
$p(p\,|\,97;100)$
$p(p\,|\,n_\star;N)$
p
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0

Repeated use of Bayes theorem

- Bayes theorem can be applied sequentially for repeated independent observations (posterior PDF = learning from experiments)

$P_0 = \text{Prior}$

Prior

observation 1

$P_1 \propto P_0 \times L_1$

Conditioned posterior 1

observation 2

$P_2 \propto P_1 \times L_2 \propto P_0 \times L_1 \times L_2$

Conditioned posterior 2

observation 3

$P_3 \propto P_0 \times L_1 \times L_2 \times L_3$

Conditioned posterior 3

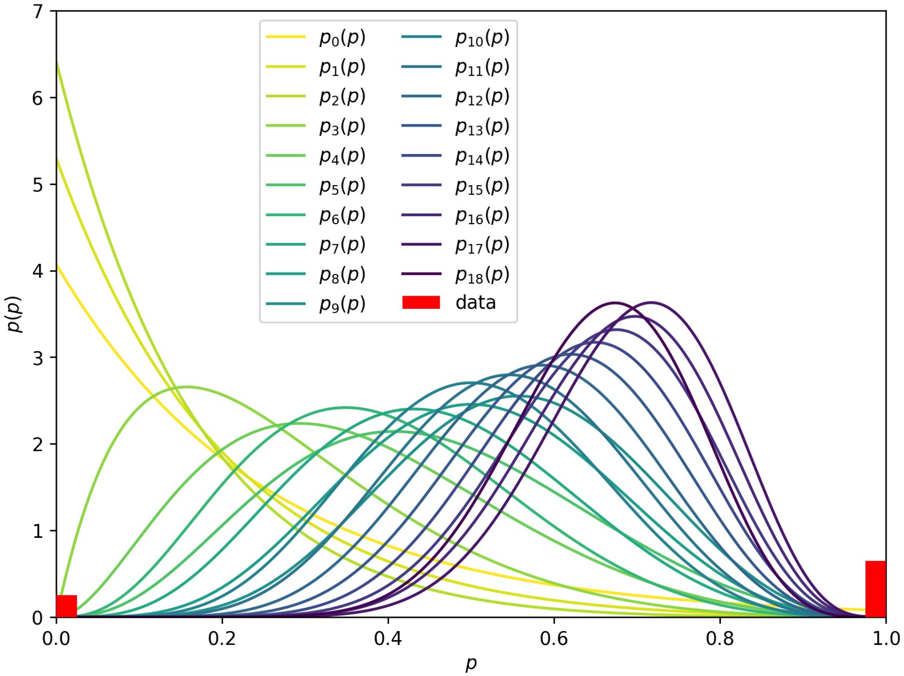
Note that applying Bayes theorem directly from prior to (obs1 + obs2) leads to the same result:

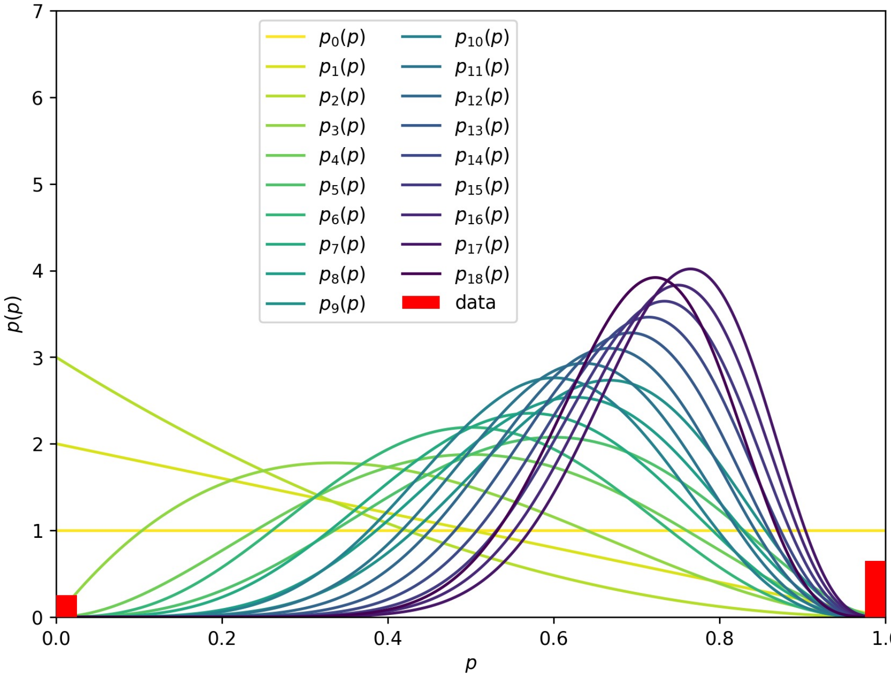
$$P_{1+2} = P_0 \times L_{1+2} = P_0 \times L_1 \times L_2 = P_2$$

Composite likelihood = product of individual likelihoods (for independent observations)

Bayesian inference as learning

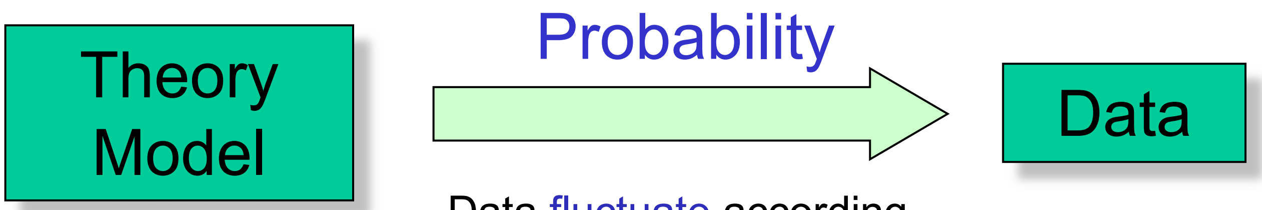
- Inference of a Binomial parameter as repeated application of Bayes rule for many Bernoulli extractions:
- $1 \rightarrow p_{i+1} = p_i \times p$
- $0 \rightarrow p_{i+1} = p_i \times (1 - p)$



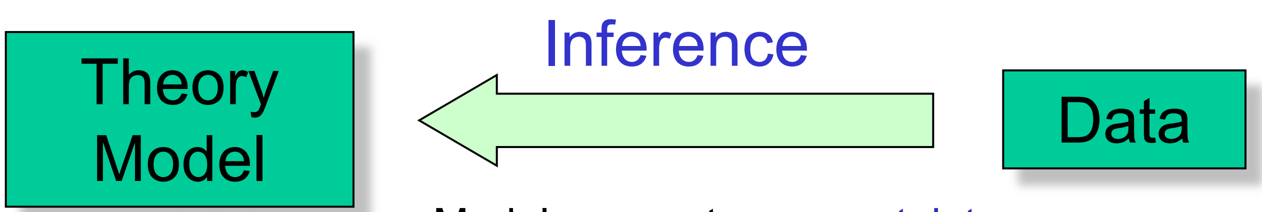


Inference

- Determinig information about unknown parameters using probability theory



Data fluctuate according to process randomness



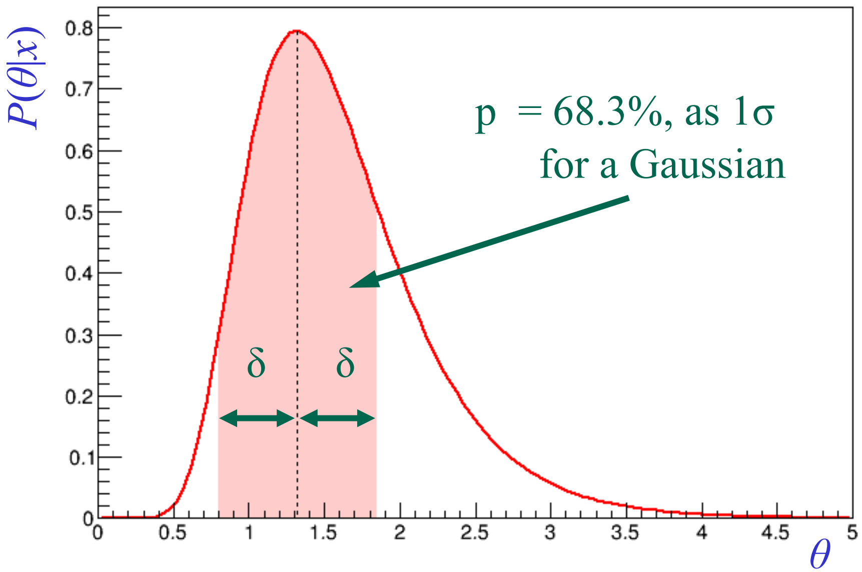
Model parameters uncertainty due to fluctuations of the data sample

Bayesian inference

- The posterior PDF provides all the information about the unknown parameters (let's assume here it's just a single parameter θ for simplicity)

$$P(\theta|x) = \frac{L(x;\theta)\pi(\theta)}{\int L(x;\theta)\pi(\theta)\mathrm{d}\theta}$$

- Given $P(\theta\,|x)$, we can determine:
  - The most probable value (best estimate)
  - Intervals corresponding to a specified probability
- Notice that if $\pi(\theta)$ is a constant, the most probable value of θ correspond to the maximum of the likelihood function



Bayesian inference of a Poissonian

- Posterior PDF, assuming the prior to be $\pi(s)$:

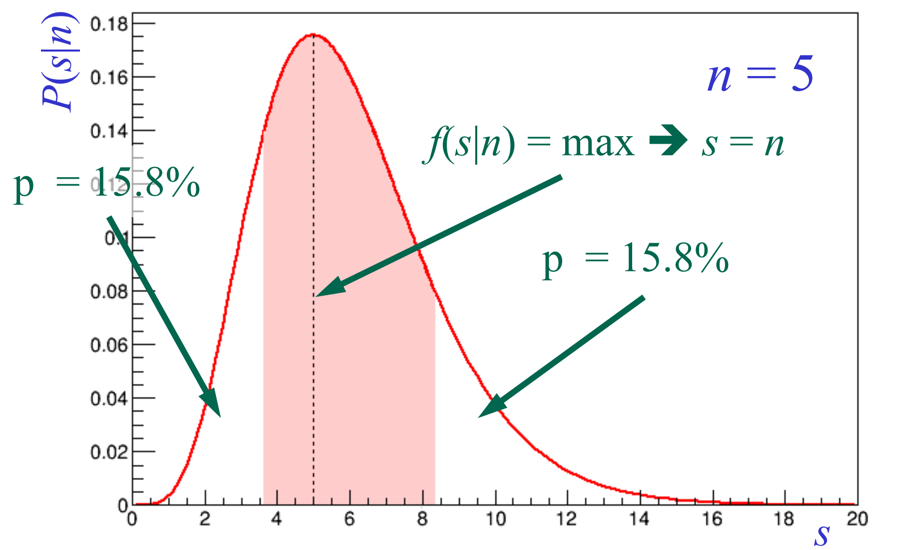
$$P(s|n) = \frac{\frac{s^n e^{-s}}{n!}\pi(s)}{\int_0^\infty \frac{s'^n e^{-s'}}{n!}\pi(s')\mathrm{d}s'}$$

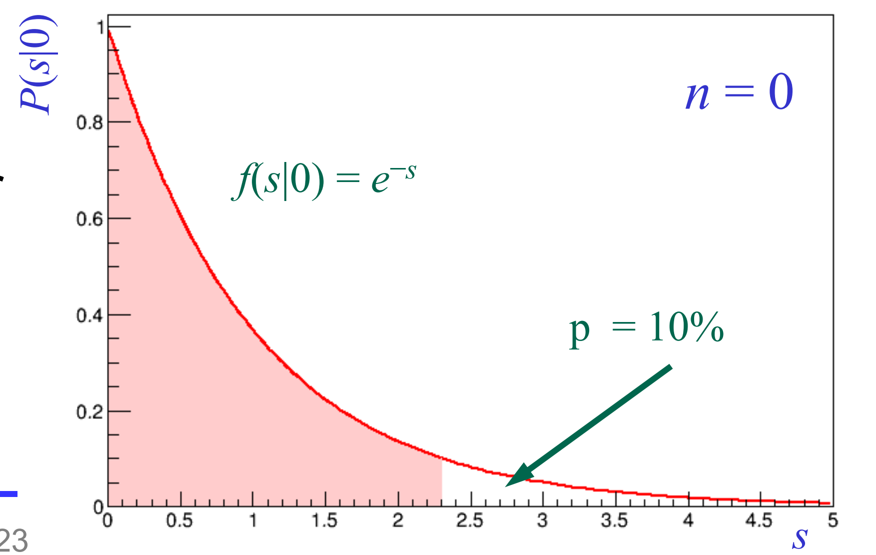
- If is $\pi(s)$ is uniform:

$$\boxed{P(s|n) = \frac{s^n e^{-s}}{n!}}$$

- Note: $\langle s \rangle = n+1\,, \quad \mathrm{Var}[s] = n+1$
- For $n = 0$, one may quote an upper limit at 90% or 95% CL:
- $s < 2.303$ (90% CL)
- $s < 2.996$ (95% CL)

zero observed events





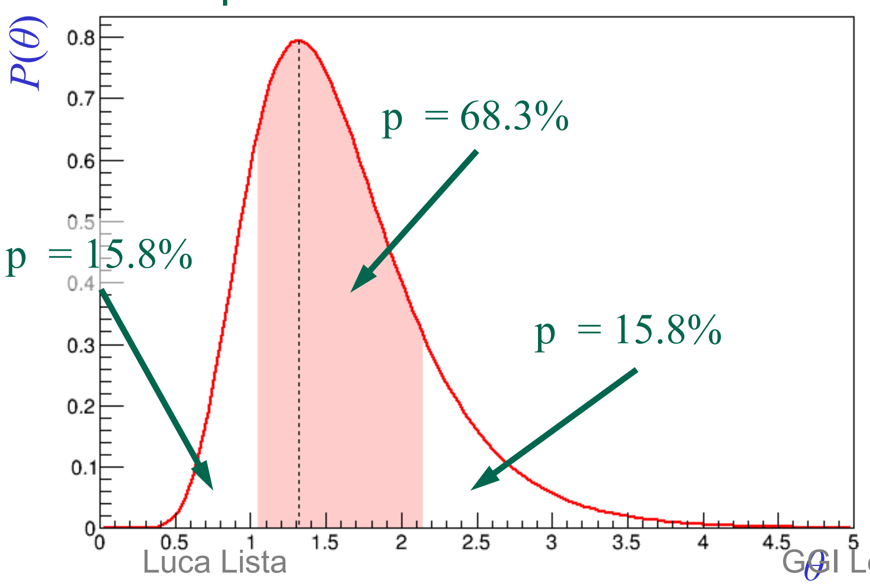
Choice of 68% prob. intervals

- Different interval choices are possible, corresponding to the same probability level (usually 68%, as 1σ for a Gaussian)
  - Equal areas in the right and left tails
  - Symmetric interval
  - Shortest interval
  - …

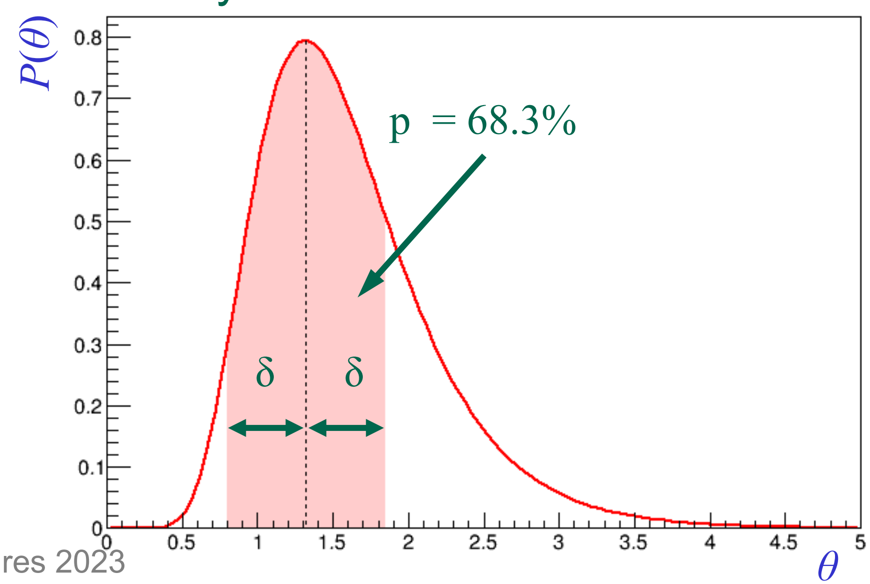
All equivalent for a symmetric distribution (e.g. Gaussian)

- Reported as $\theta = \hat{\theta} \pm \delta$ (sym.) or $\theta = \hat{\theta}^{+\delta_1}_{-\delta_2}$ (asym.)

Equal tails interval

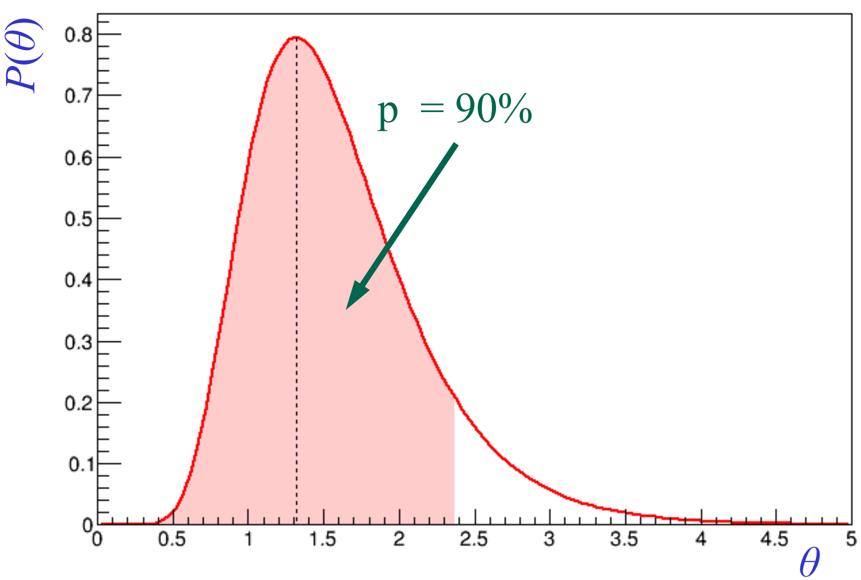


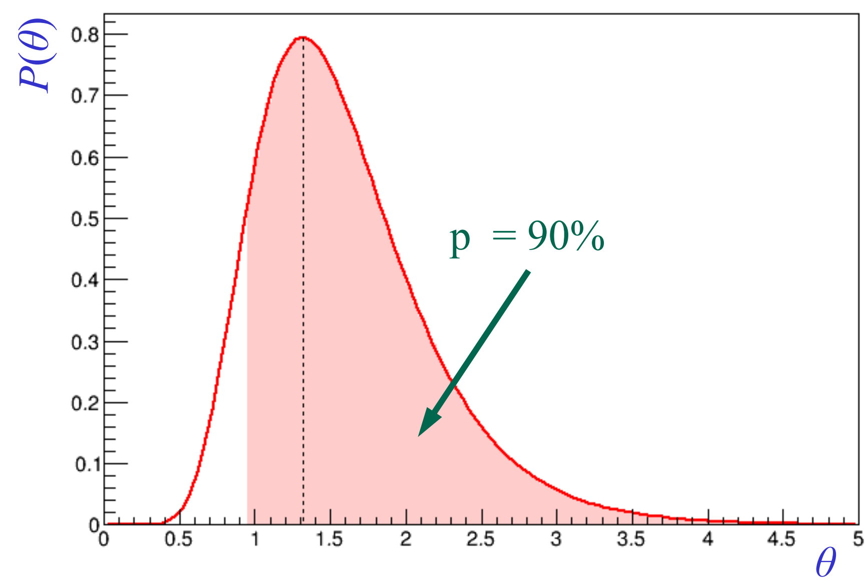
Symmetric interval



Upper and lower limits

- A fully asymmetric interval choice is obtained setting one extreme of the interval to the lowest or highest allowed range
- The other extreme indicates an upper or lower limits to the “allowed” range
- For upper or lower limits, usually a probability of 90% or 95% is preferred to the usual 68% adopted for central intervals
- Reported as: $\theta < \theta^{up}$ (90% CL) or $\theta > \theta^{lo}$ (90% CL)





Error propagation: Bayesian inference

- Applying a parameter transformation, say $\eta = H(\theta)$, results in a transformed central value and transformed uncertainty interval
- The error propagation can be done transforming the posterior PDF, then computing the interval on the transformed PDF:

$$f'(\eta) = \int \delta(\eta - H(\theta)) f(\theta) \mathrm{d}\theta$$

- Transformations for cases with more than one variable proceed in a similar way:

$\eta = H(\theta_1, \theta_2)$:

$$f'(\eta) = \int \delta(\eta - H(\theta_1, \theta_2)) f(\theta_1, \theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2$$

$\eta_1 = H_1(\theta_1, \theta_2)$, $\eta_1 = H_1(\theta_1, \theta_2)$:

$$f'(\eta_1, \eta_2) = \int \delta(\eta_1 - H_1(\theta_1, \theta_2)) \delta(\eta_2 - H_2(\theta_1, \theta_2)) f(\theta_1, \theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2$$

Choosing the prior PDF

- If the prior PDF is uniform in a choice of variable, it won't be uniform when applying coordinate transformation
- Given a prior PDF in a random variable, there is always a transformation that makes the PDF uniform
- The problem is: chose one metric where the PDF is uniform
- Harold Jeffreys' prior: chose the prior form that is invariant under parameter transformation

$$\pi(\theta) \propto \sqrt{\det I(\vec{\theta})} \qquad I_{ij}(\vec{\theta}) = \left\langle \frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial \theta_i} \frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial \theta_j} \right\rangle$$

- Some commonly used cases:
  - Poissonian mean: $\pi(\mu) \propto 1/\sqrt{\mu}$
  - Poissonian mean with background b: $\pi(\mu) \propto 1/\sqrt{\mu + b}$
  - Gaussian mean: $\pi(\mu) \propto 1$
  - Gaussian standard deviation: $\pi(\mu) \propto 1/\sigma$
  - Binomial parameter: $\pi(\mu) \propto 1/\sqrt{\varepsilon(1-\varepsilon)}$

Note: the previous simple Poissonian example was obtained with $\pi(\mu) = \text{const.}$!

- Problematic with PDF in more than one dimension!

Lifetime estimate

$$L(t_1, \ldots, t_N|\tau) = \frac{1}{\tau^N}\prod_{i=1}^{N} e^{-t_i/\tau}$$

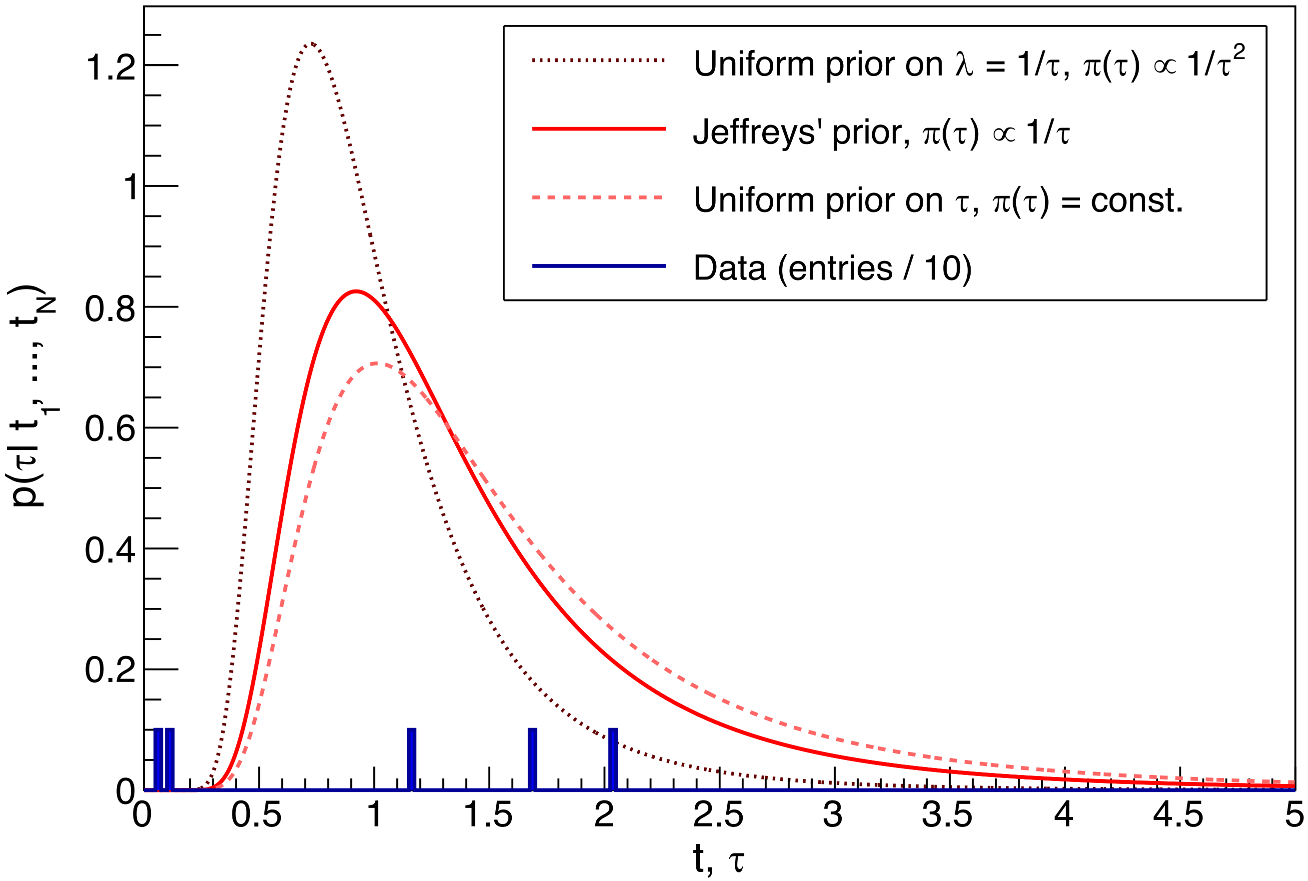
- The posterior for τ is:

$$p(\tau|t_1, \ldots, t_N) = \frac{\pi(\tau)e^{-\sum_{i=1}^N t_i/\tau}/\tau^N}{\int \pi(\tau')e^{-\sum_{i=1}^N t_i/\tau'}/\tau'^N \mathrm{d}\tau'}$$

- The prior can chosen in a different way:
  - Uniform in τ, $\pi(\tau) = \mathrm{const.}$
  - Uniform in $\lambda = 1/\tau$, $\pi(\tau) = 1/\tau^2$
  - Jeffrey's prior, $\pi(\tau) = 1/\tau$
- All choices give as posterior a gamma distribution with different parameters ($k = N$, $N + 2$, $N + 1$)
  - $p(\tau|t_1, \ldots, t_N) = C\,\tau^k e^{-\sum_{i=1}^N t_i/\tau}$
- The maximum of the PDF is at $\tau = \sum_{i=1}^N t_i\,/k$ which is equal to $N\bar{t}/k$, that, for large N, tends to $\bar{t}$, regardless of the prior choice.

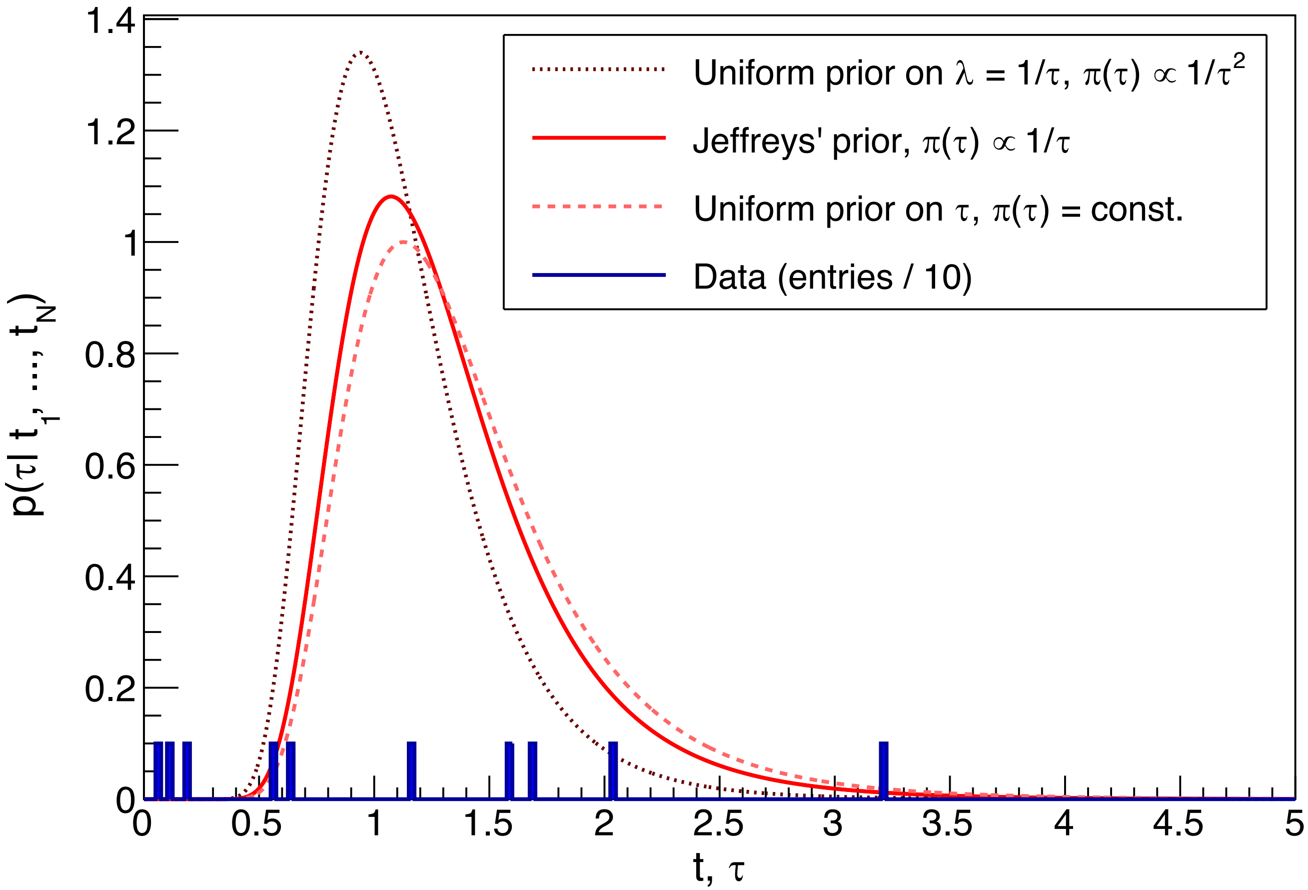
Prior dependence: lifetime

$n = 5$



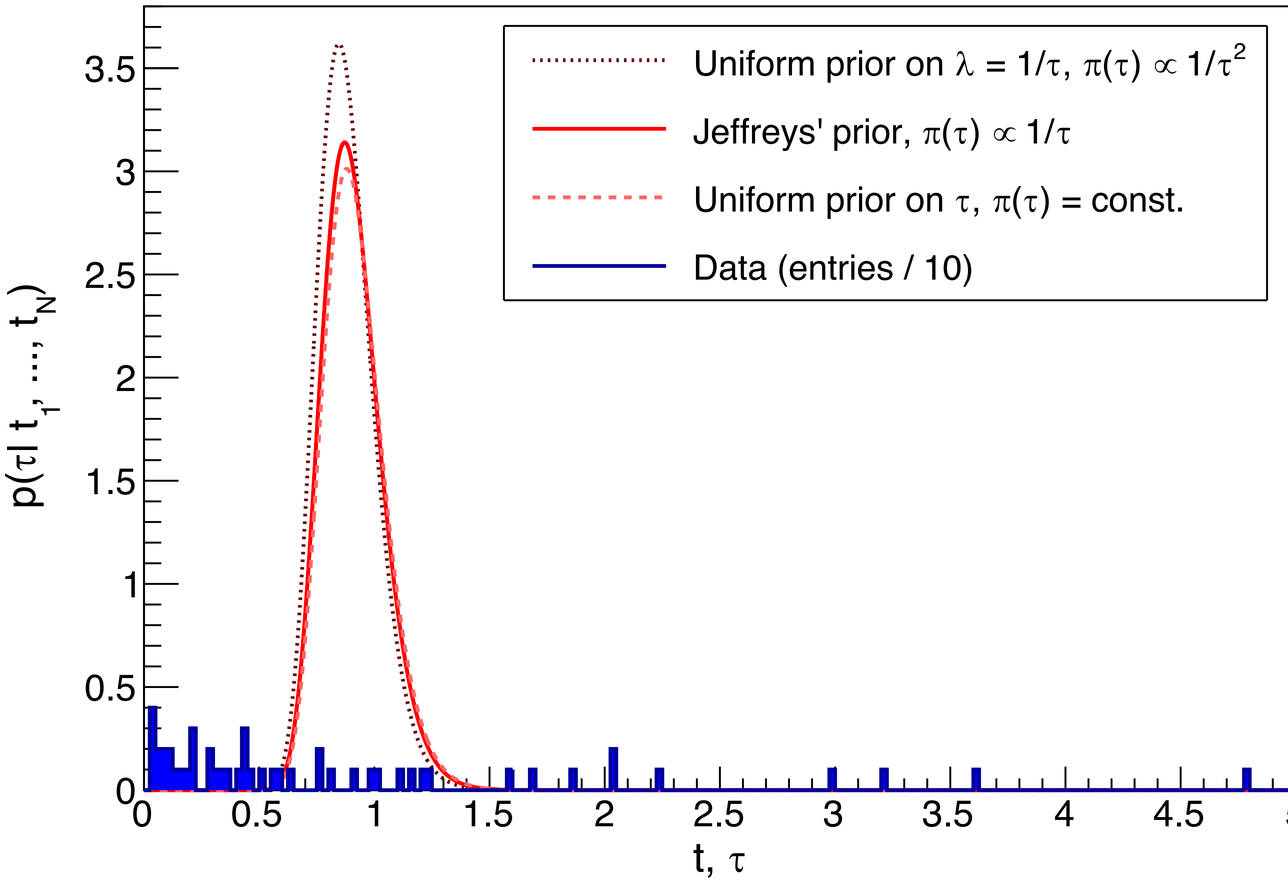
Prior dependence: lifetime

$n = 10$



Prior dependence: lifetime

$n = 50$

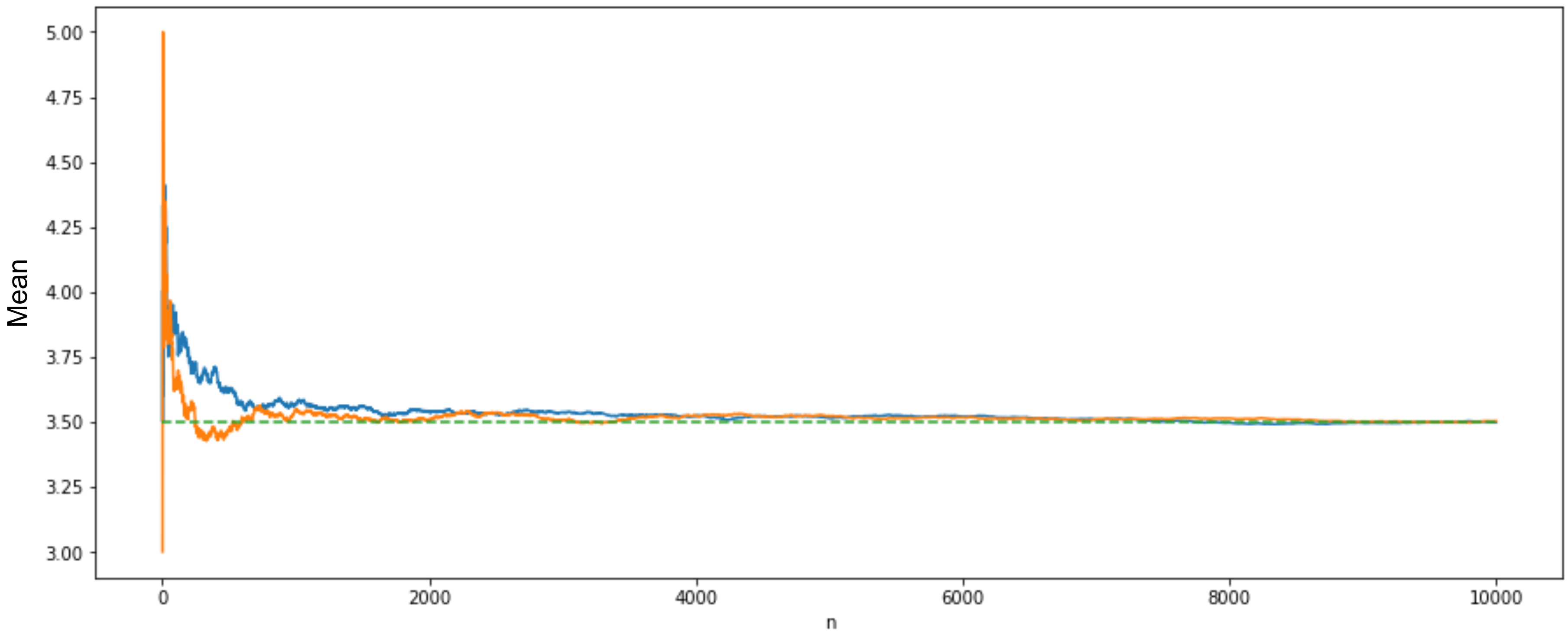


Lecture 2

- Law of large numbers and frequentist probability
- Inference with the frequentist approach: coverage
- The maximum likelihood method
- Extended likelihood functions
- Binned and unbinned fits
- Properties of frequentist estimators
- Neyman's confidence interval
- Binomial confidence intervals according to Clopper and Pearson
- Error estimates with the maximum likelihood method
- Issues with asymmetric uncertainties
- Two-dimensional uncertainty contours
- Binned fits: the minimum chi-squared method
- Goodness of the fit with chi-squared test
- Baker-Cousins binned likelihood ratio fits
- Combination of measurements and the BLUE method

Law of large numbers

- The mean of a large number of random extractions following the same distribution tends to the expected value
- This laws assumes an underlying probability model



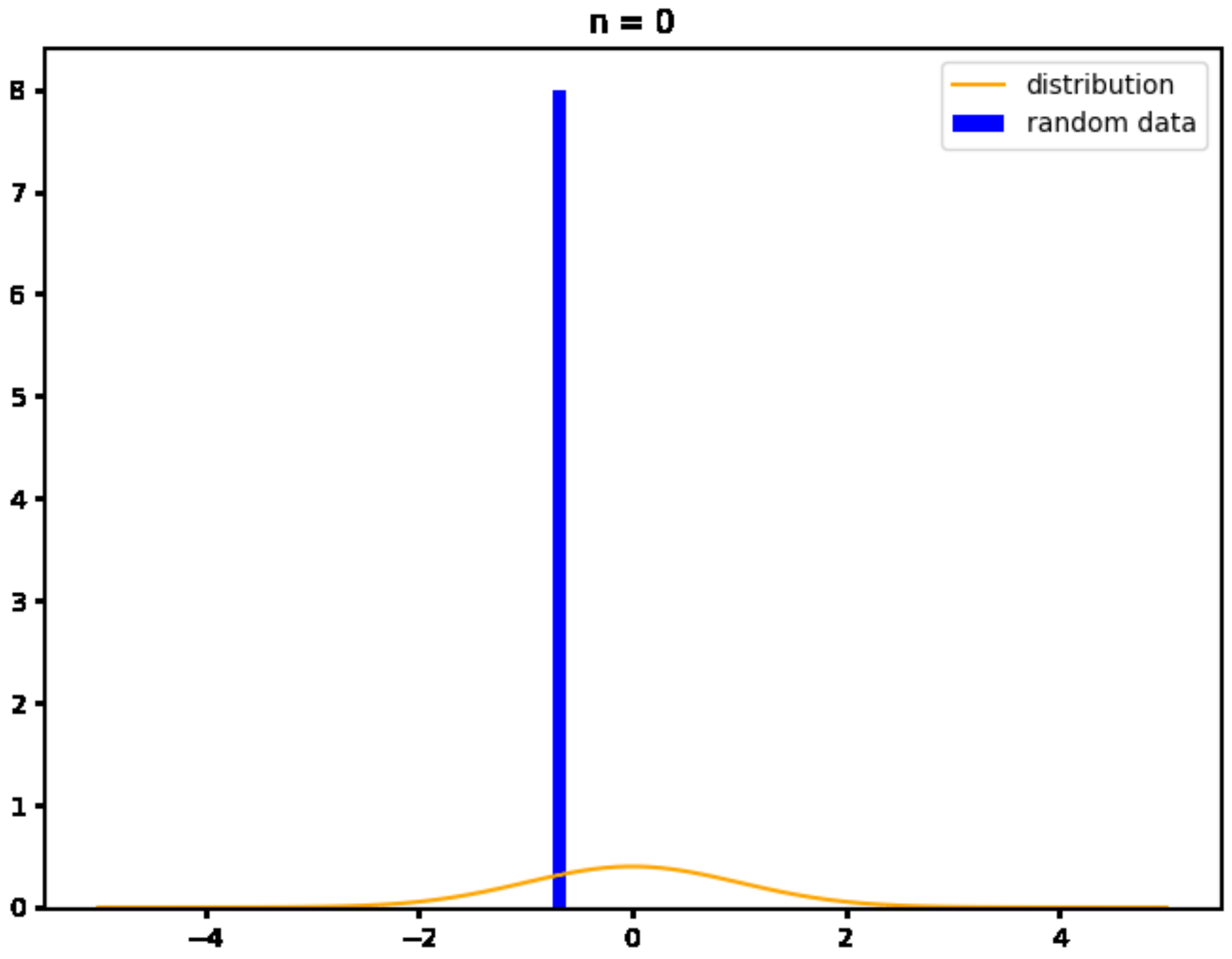
Frequentist probability

- Probability P = frequency of occurrence of an event in the limit of very large number ($N \to \infty$) of repeated trials

$$\text{Probability: } P = \lim_{N \to \infty} \frac{\text{Number of favorable cases}}{N = \text{Number of trials}}$$

- Exactly realizable only with an infinite number of trials
  - Conceptually may be unpleasant
  - Pragmatically acceptable by physicists
- Only applicable to repeatable experiments

Large numbers and probability



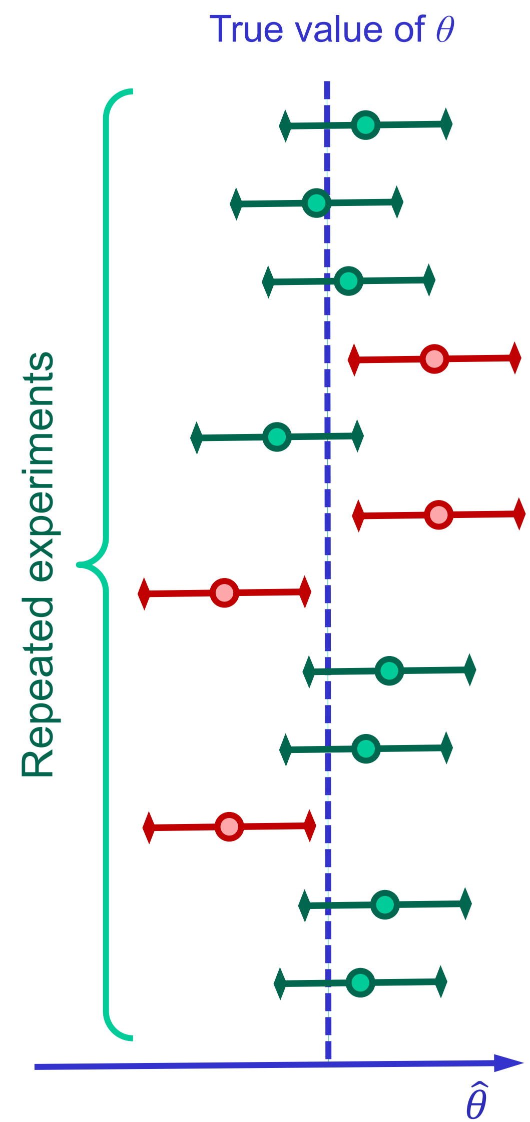
Frequentist inference

- Assigning a probability level of an unknown parameter makes no sense in the frequentist approach
  - Parameters are not random variables!
- A frequentist inference procedure determines a central value and an uncertainty interval that depend on the observed measurements
- The central value and interval extremes are random variables
- No subjective element is introduced in the determination

- The function that returns the central value given an observed measurement is called estimator
- Different estimator choices are possible, the most frequently adopted is the maximum likelihood estimator because of its statistical properties discussed in the following

Frequentist coverage

- Repeating the experiment will result each time in a different data sample
- For each data sample, the estimator returns a different central value $\hat{\theta}$
- An uncertainty interval $[\hat{\theta} - \delta, \hat{\theta} + \delta]$ can be associated to the estimator's value $\hat{\theta}$
- Some of the confidence intervals contain the fixed and unknown true value of θ, corresponding to a fraction equal to 68% of the times, in the limit of very large number of experiments (coverage)

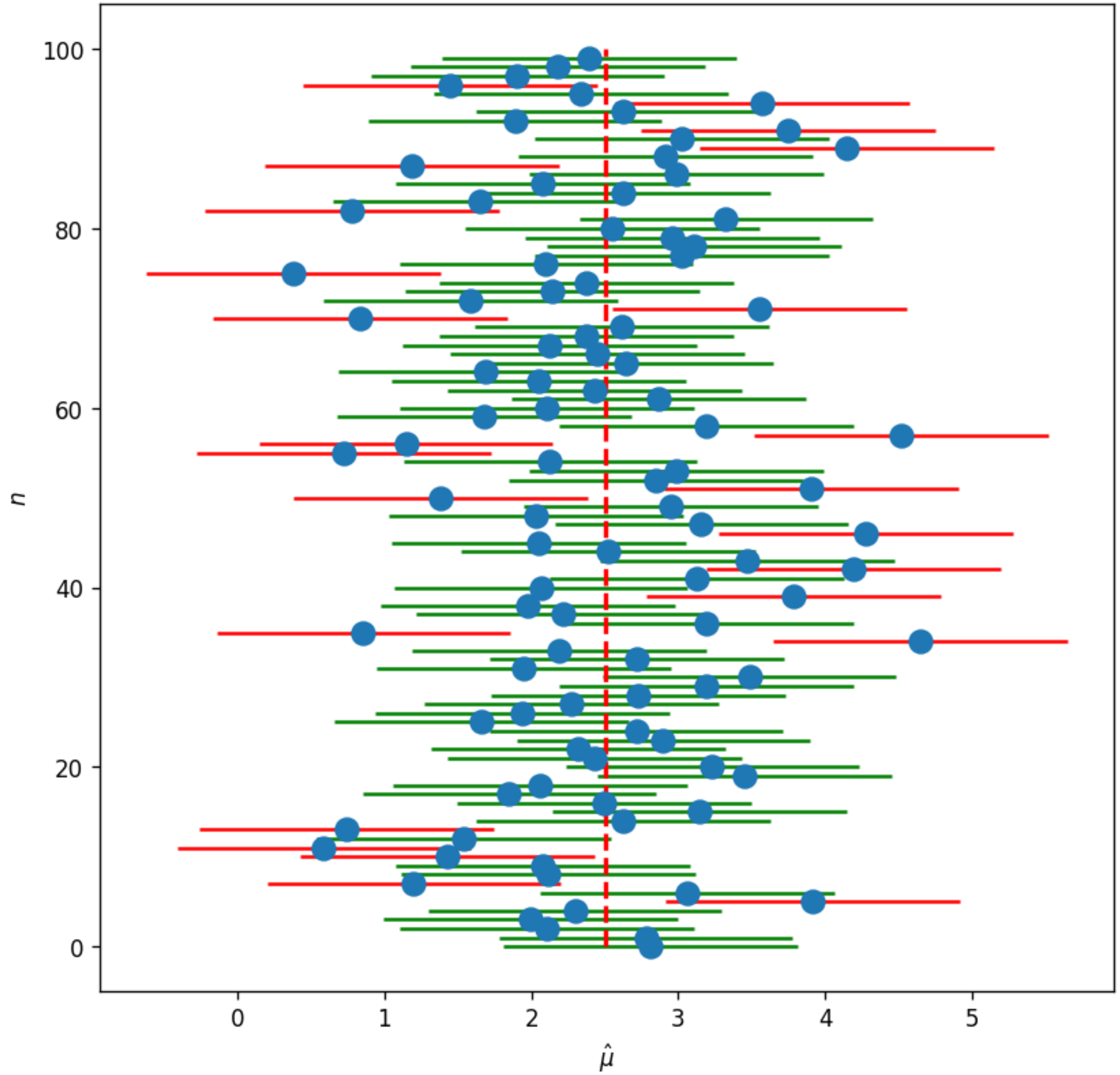


Frequentist inference:

- An estimator is a function of a given set of measurements that provides an approximate value of a parameter of interest which appears in our PDF model ("best fit")
- Simplest example:
  - Assume a Gaussian PDF with a *known* σ and an *unknown* μ
  - A single experiment provides a measurement x
  - We estimate μ as $\hat{\mu} = x$
  - The distribution of $\hat{\mu}$ (repeating the experiment many times) is the original Gaussian
  - 68.3% of the experiments (in the limit of large number of repetitions) will provide an estimate within: $\mu - \sigma < \hat{\mu} < \mu + \sigma$
- We can quote:

$$\boxed{\mu = x \pm \sigma}$$

Coverage for frequentist inference



The maximum-likelihood method

- The maximum-likelihood estimator is the most adopted parameter estimator
- The "best fit" parameters correspond to the set of values that maximizes the likelihood function
  - Good statistical properties (→ next slides)
- The maximization can be performed analytically only in the simplest cases, and numerically for most of realistic cases

- Minuit is historically the most widely used minimization engine in High Energy Physics
  - F. James, 1970's; rewritten in C++ and released under CERN's ROOT framework

Gaussian case

- If we have n independent measurements all modeled with (or approximated to) the same Gaussian PDF, we have:

$$-2\ln L = \underbrace{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2}}_{\text{(example of a } \chi^2 \text{ variable)}} + n(\ln 2\pi + 2\ln\sigma)$$

- An analytical minimization of $-2\ln L$ w.r.t μ (assuming σ^2 is known) gives the arithmetic mean as ML estimate of μ:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- If σ^2 is also unknown, the ML estimate of σ^2 is:

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

- The above estimate can be demonstrated to have an unpleasant feature, called *bias* (→ next slide)

Some estimator properties

- **Consistency**: for large number of measurements the estimator $\hat{\theta}$ should converge, in probability, to the true value θ.
  - ML estimators are consistent

- **Bias**: the bias of a parameter is the average value of its deviation from the true value

$$\mathrm{b}(\theta) = \left\langle \hat{\theta} - \theta \right\rangle = \left\langle \hat{\theta} \right\rangle - \theta$$

  - ML estimators may have a bias, but the bias decreases with large number of measurements (if the fit model is correct…!)
  - E.g.: in the case of the estimate of a Gaussian's σ^2, the unbiased estimate is the well known:

$$\hat{\sigma^2}_{\mathrm{unbias.}} = \frac{n}{n-1}\hat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

ML method underestimates the variance σ^2

Efficiency of an estimator

- The variance of any consistent estimator is subject a lower bound (Cramér-Rao bound):

$$\mathbb{V}[\hat{\theta}] \geq \frac{\left(1 + \frac{\partial b(\theta)}{\partial \theta}\right)^2}{\mathbb{E}\left[\left(\frac{\partial^2 lnL(\theta)}{\partial \theta^2}\right)^2\right]} = \frac{\left(1 + \underbrace{\frac{\partial b(\theta)}{\partial \theta}}_{\text{bias of } \theta}\right)^2}{\underbrace{\mathbb{E}\left[-\frac{\partial^2 lnL(\theta)}{\partial \theta^2}\right]}_{\text{Fisher information}}} = V_{CR}$$

- **Efficiency** can be defined as the ratio of Cramér-Rao bound and the estimator's variance:

$$\varepsilon(\hat{\theta}) = \frac{V_{CR}}{\mathbb{V}\mathrm{ar}[\hat{\theta}]}$$

- Efficiency for ML estimators tends to 1 for large number of measurements

$$\lim_{n\to\infty} \mathbb{V}[\hat{\theta}] = -\frac{1}{\mathbb{E}\left[\left(\frac{\partial^2 lnL(\theta)}{\partial \theta^2}\right)^2\right]} \cong -\frac{1}{\left.\frac{\partial^2 lnL(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}}}$$

  – I.e.: ML estimates have, asymptotically, the smallest possible variance

Approx. maximum likelihood errors

- A parabolic approximation of $-2\ln L$ around the minimum is equivalent to a Gaussian approximation
  - Sufficiently accurate in many but not all cases

$$-2\ln L = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} + \text{const.}$$

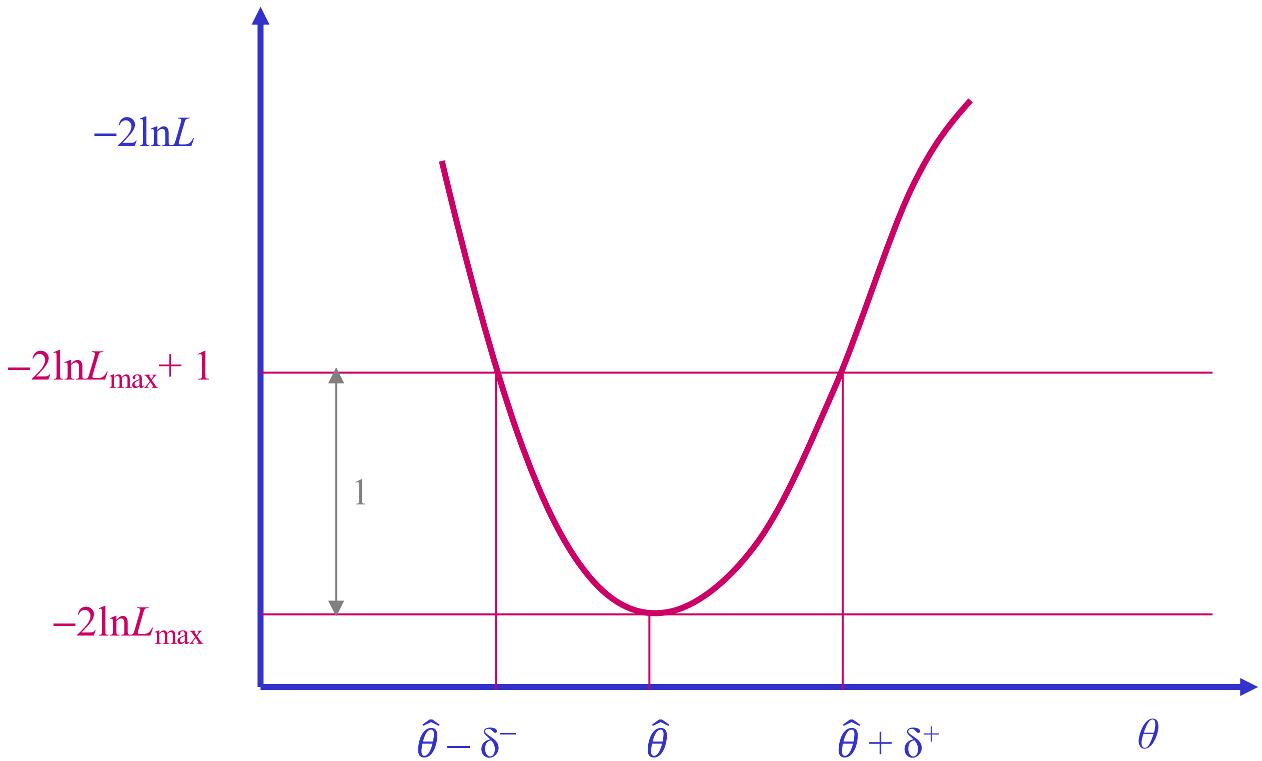
- Estimate of the covariance matrix from 2nd order partial derivatives w.r.t. fit parameters at the minimum:

$$V_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\theta_k = \hat{\theta}_k}$$

- Implemented in Minuit as MIGRAD/HESSE function

Asymmetric errors

- Another approximation alternative to the parabolic one may be to evaluate the excursion range of $-2\ln L$.
- Error ($n\sigma$) determined by the range around the maximum for which $-2\ln L$ increases by $+1$ ($+n^2$ for $n\sigma$ intervals)



- Errors can be asymmetric
- For a Gaussian PDF the result is identical to the 2nd order derivative matrix
- Implemented in Minuit as MINOS function

Measure a particle's lifetime

- The probability distribution for a single measurement is:

$$p(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

- And the likelihood function is

$$L(t_1, \ldots, t_n; \tau) = \frac{1}{\tau^n}\left(\prod_{i=1}^{n} e^{-t_i/\tau}\right)$$

- Minimization gives:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

- Which is clearly unbiased
- The distribution of $\hat{\tau}$ is a gamma distribution scale parameter τ and shape parameter 1:

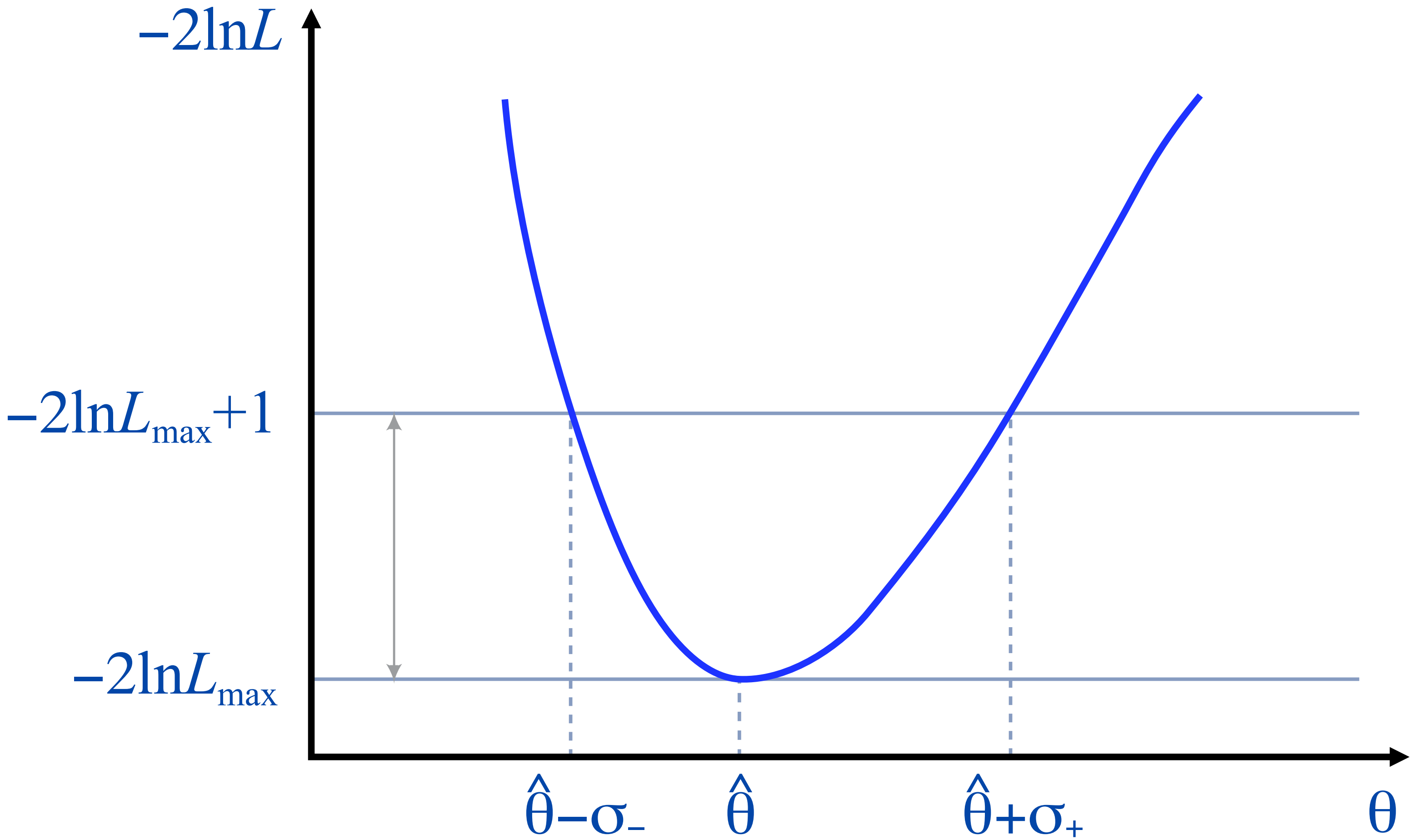
$$p(\hat{\tau}) = \hat{\tau}^{2n} e^{-\frac{\hat{\tau}}{\tau}} / (\tau^{2n}(2n-1)!)$$

- With uncertainty given by the square root of the variance equal to:

$$\sigma_{\hat{\tau}} = \tau/\sqrt{n}$$

- This is also equal to the Cramer-Rao bound

ML uncertainty estimate



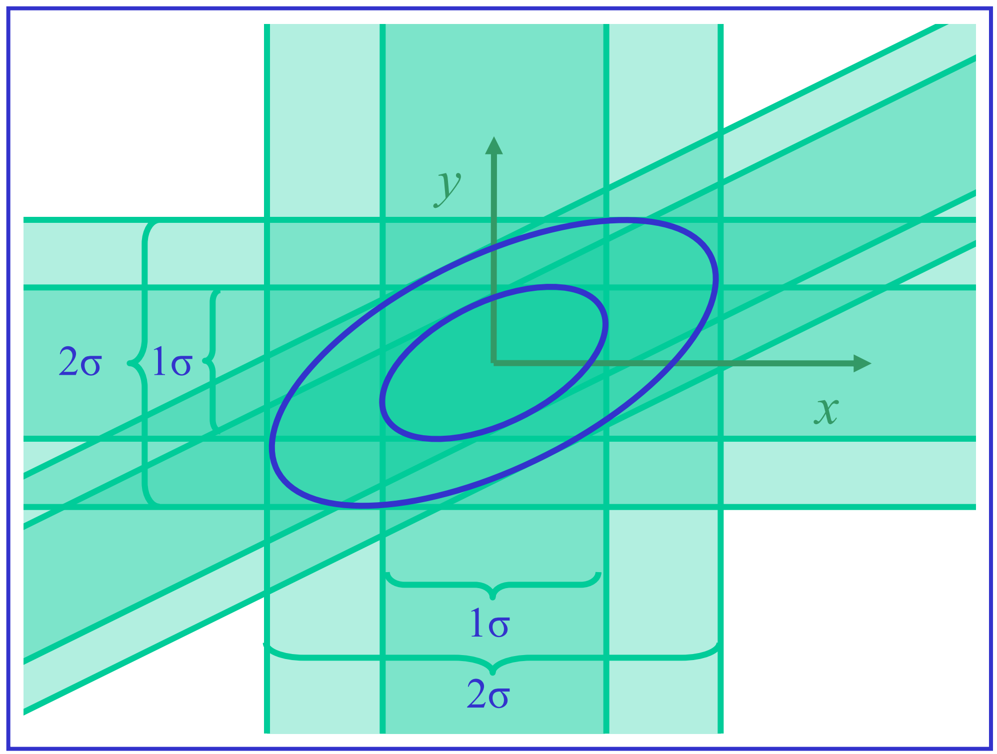
2D intervals

- In more dimensions one can determine 1σ and 2σ contours
- Note: different probability content in 2D compared to one dimension
- 68% and 95% contours are usually preferable

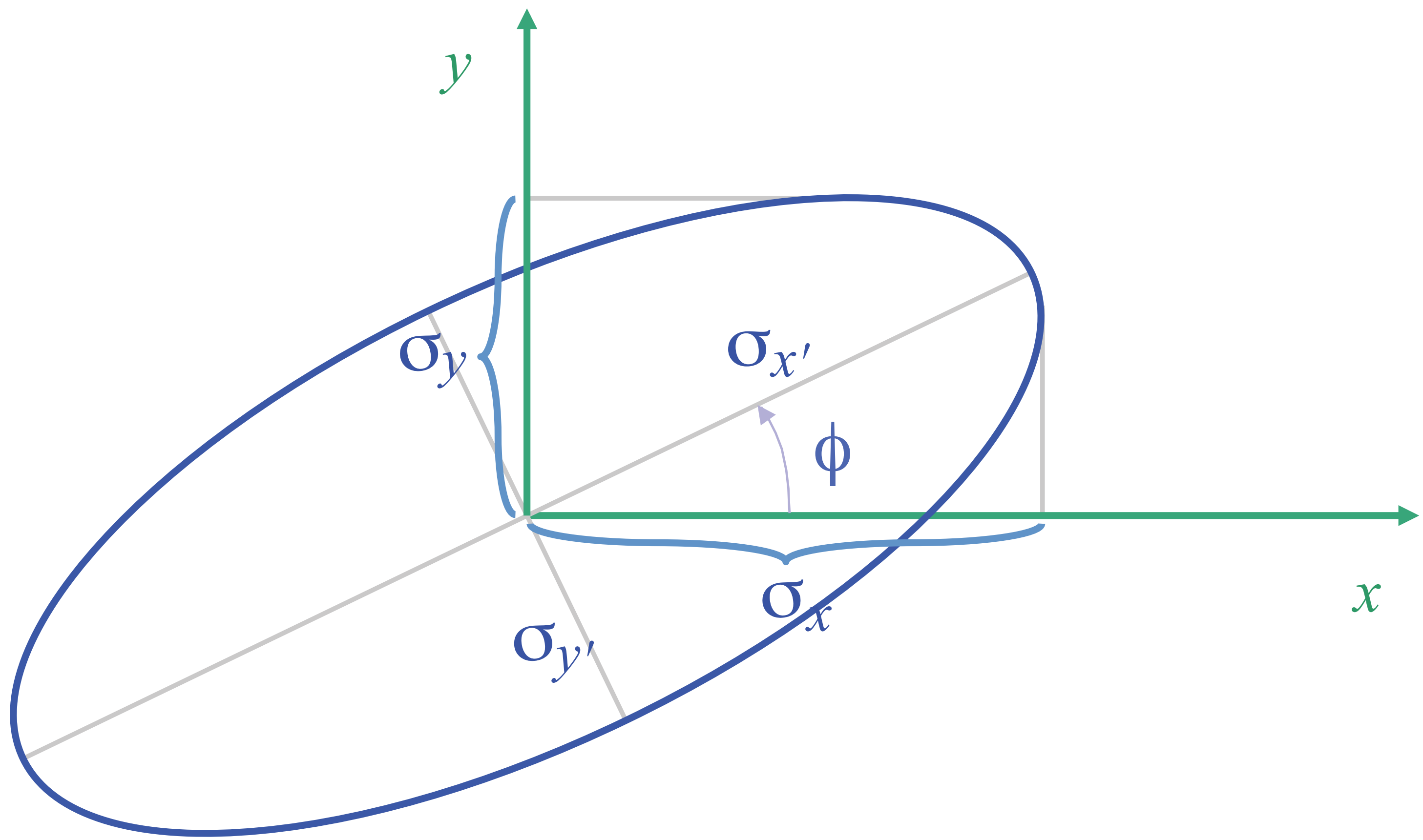
$$P_{1D}(n\sigma) = \sqrt{\frac{2}{\pi}} \int_0^n e^{-\frac{x^2}{2}} \mathrm{d}x = \mathrm{erf}\left(\frac{n}{\sqrt{2}}\right)$$

$$P_{2D}(n\sigma) = \int_0^n e^{-\frac{r^2}{2}} r\mathrm{d}r = 1 - e^{-\frac{n^2}{2}}$$

Width	P_{1D}	P_{2D}
1σ	0.6827	0.3934
2σ	0.9545	0.8647
3σ	0.9973	0.9889
1.515σ		0.6827
2.486σ		0.9545
3.439σ		0.9973



In more dimensions



Extended likelihood function

- Given a sample of N measurements of the variables $(x_1, \ldots, x_n)$, the likelihood function expresses the probability density of the sample, as a function of the unknown parameters:

$$L = \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)$$

- If the size N of the sample is also a random variable, the extended likelihood function is usually also used:

$$L = P(N; \theta_1, \cdots, \theta_m) \prod_{i=1}^{N} f(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m)$$

- Where $P(N; \theta_1, \ldots, \theta_m)$ is in practice always a Poisson distribution whose expected rate is a function of the unknown parameters

- In many cases it is convenient to use $-\ln L$ or $-2\ln L$: $\prod_i \to \sum_i$

Extended likelihood function

- For Poissonian signal and background processes:

$$L(x_i; s, b, \theta) = \frac{(s+b)^n e^{-(s+b)}}{n!} \prod_{i=1}^{n} (f_s P_s(x_i;\theta) + f_b P_b(x_i;\theta))$$

$$\left.\begin{aligned} f_s &= \frac{s}{s+b} \\ f_b &= \frac{b}{s+b} \end{aligned}\right\} \rightarrow\ = \frac{e^{-(s+b)}}{n!} \prod_{i=1}^{n} (sP_s(x_i;\theta) + bP_b(x_i;\theta))$$

- We can fit simultaneously s, b and θ minimizing:

constant!

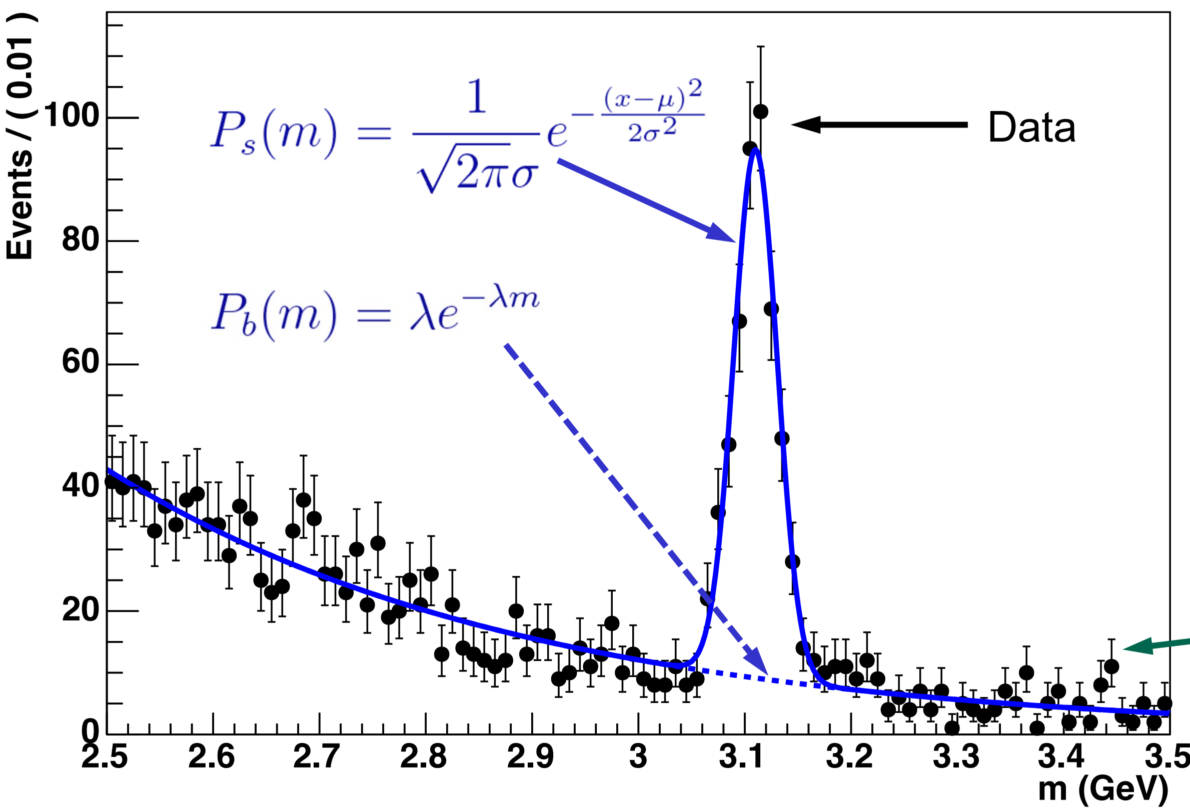
$$-\ln L = s + b - \sum_{i=1}^{n} \ln(sP_s(x_i;\theta) + bP_b(x_i;\theta)) + \ln n!$$

- Sometimes s is replaced by $\mu\, s_0$, where s_0 is the theory estimate and μ is called signal strength

Example of ML fit

- $P_s(m)$: Gaussian peak
- $P_b(m)$: exponential shape

Exponential decay parameter λ, Guassian mean μ and standard deviation σ can be fit together with sig. and bkg. yields s and b.



The additional parameters, beyond the parameters of interest (s in this case), used to model background, resolution, etc. are examples of nuisance parameters

In the plot, data are accumulated into **bins** of a given width

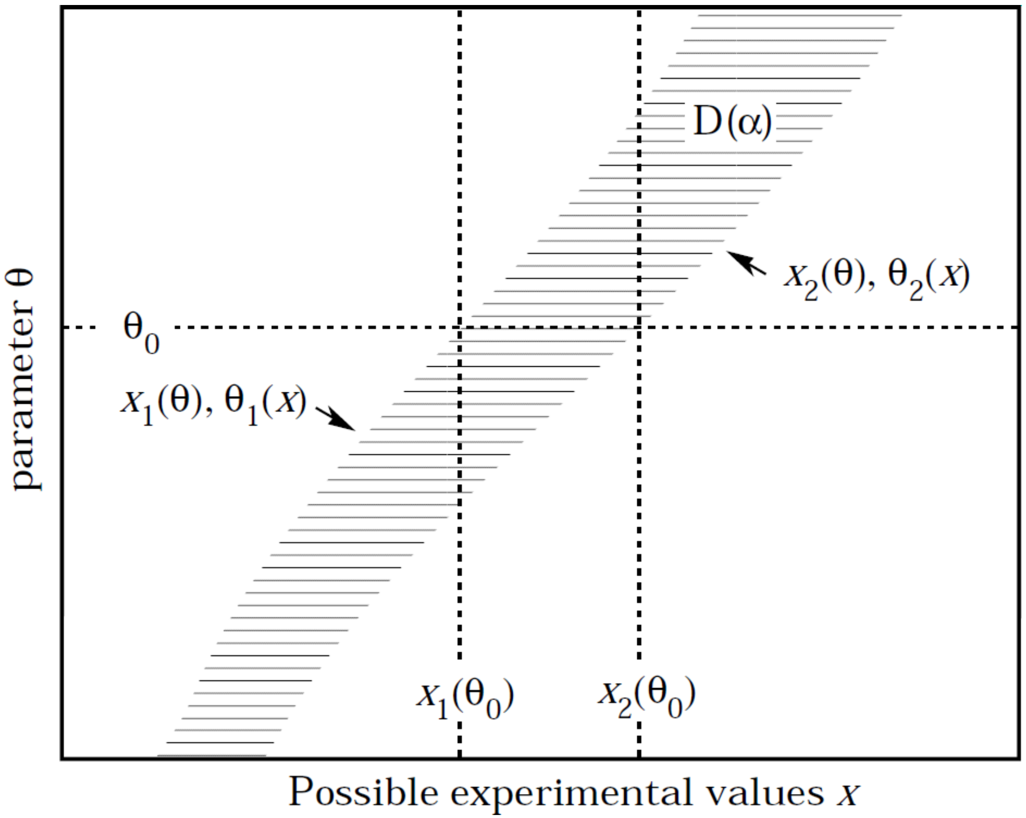
Error bars usually represent uncertainty on each bin count (in this case: Poissonian)

Neyman's confidence intervals

Procedure to determine frequentist confidence intervals

Plot from PDG statistics review

- Scan the allowed range of an unknown parameter θ
- Given a value of θ compute the interval $[x_1, x_2]$ that contain x with a probability $1-\alpha$ equal to 68% (or 90%, 95%)
- **Choice of interval needed!**
- Invert the confidence belt: for an observed value of x, find the interval $[\theta_1, \theta_2]$
- A fraction of the experiments equal to $1-\alpha$ will measure x such that the corresponding $[\theta_1, \theta_2]$ contains ("covers") the true value of θ ("coverage")
- **Note**: the random variables are $[\theta_1, \theta_2]$, not θ !

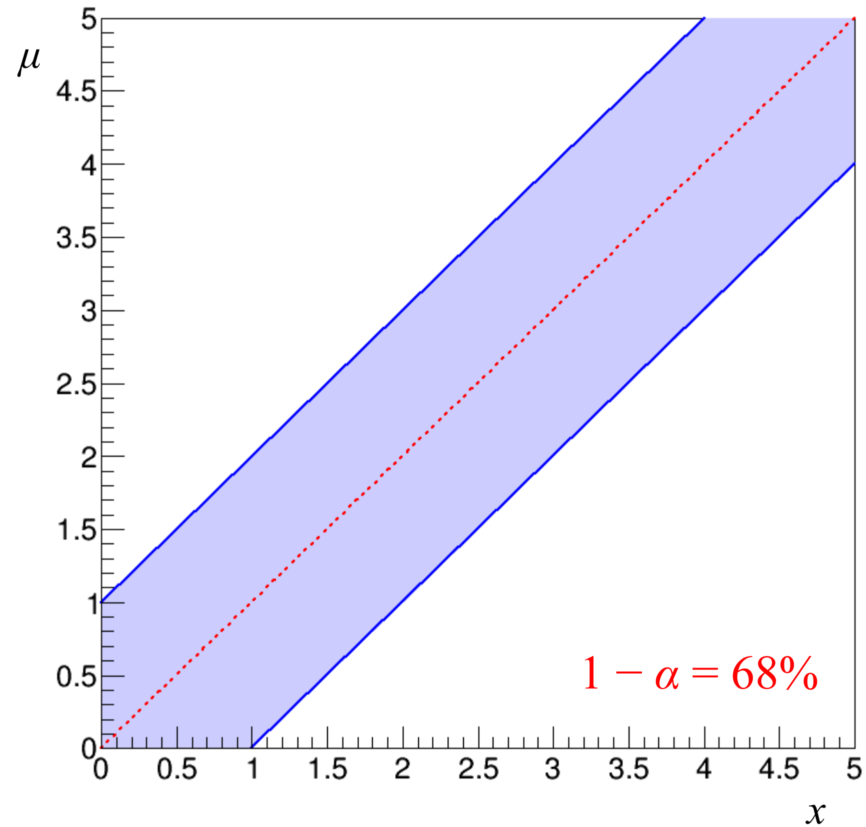


α = significance level

Simplest example: Gaussian case

- Assume a Gaussian distribution with unknown average μ and known $\sigma = 1$
- The belt inversion is trivial and gives the expected result:
  Central value $\hat{\mu} = x$,
  $$[\mu_1, \mu_2] = [x - \sigma, x + \sigma]$$
- So we can quote:

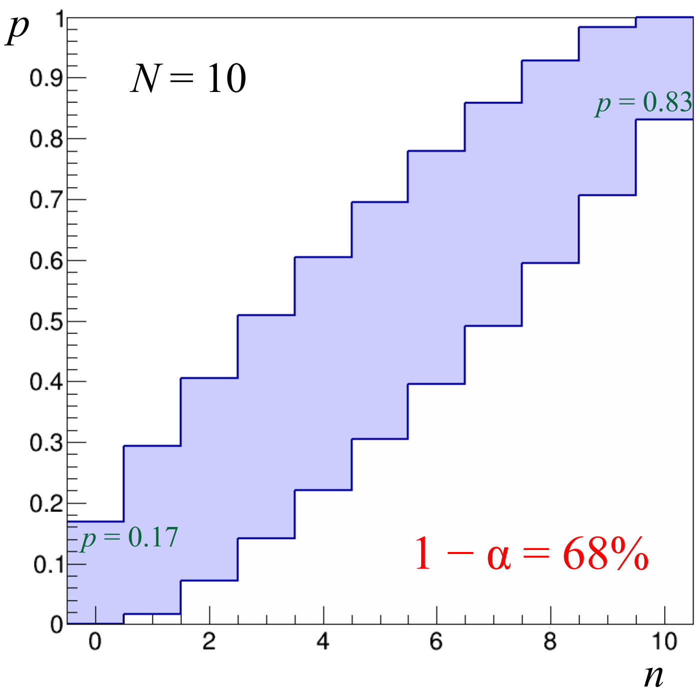
$$\boxed{\mu = x \pm \sigma}$$



Binomial intervals

- The Neyman's belt construction may only guarantee approximate coverage in case of discrete variables
- For a Binomial distribution: find the interval $\{n_{\min}, \ldots, n_{\max}\}$ such that:

$$\sum_{n=n_{\min}}^{n=n_{\max}} \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \geq 1-\alpha$$

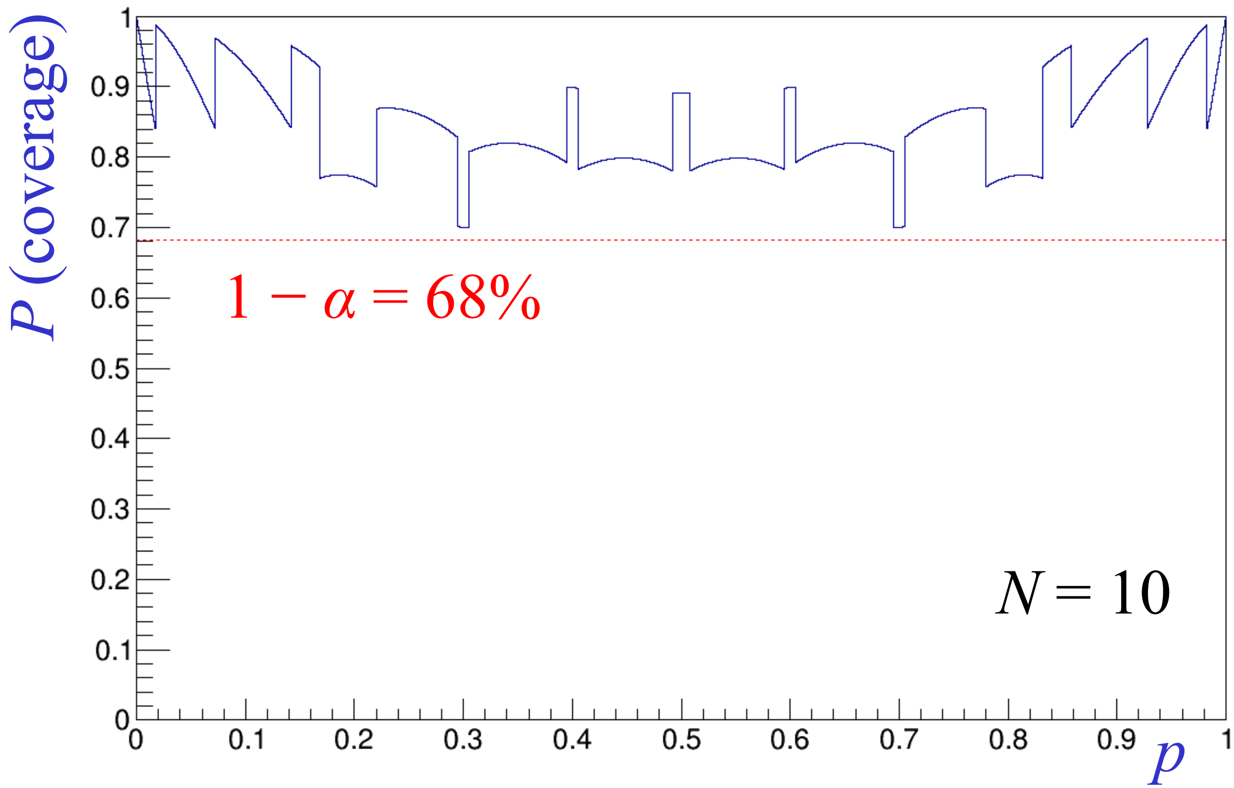


- Clopper and Pearson (1934) solved the belt inversion problem for central intervals
- For an observed $n = k$, find lowest p^{lo} and highest p^{up} such that:
- $P(n \leq k \mid N, p^{\mathrm{lo}}) = \alpha/2, \quad P(n \geq k \mid N, p^{\mathrm{up}}) = \alpha/2$
- E.g.: $n = N = 10$, $P(N|N) = p^N = \alpha/2$, hence: $p^{\mathrm{lo}} = \sqrt[10]{\alpha/2} = 0.83$ (68% CL), 0.74 (90% CL)
- A frequently used approximation, which fails for $n = 0$, N is:

$$\hat{p} = \frac{n}{N}, \; \sigma_{\hat{p}} \simeq \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

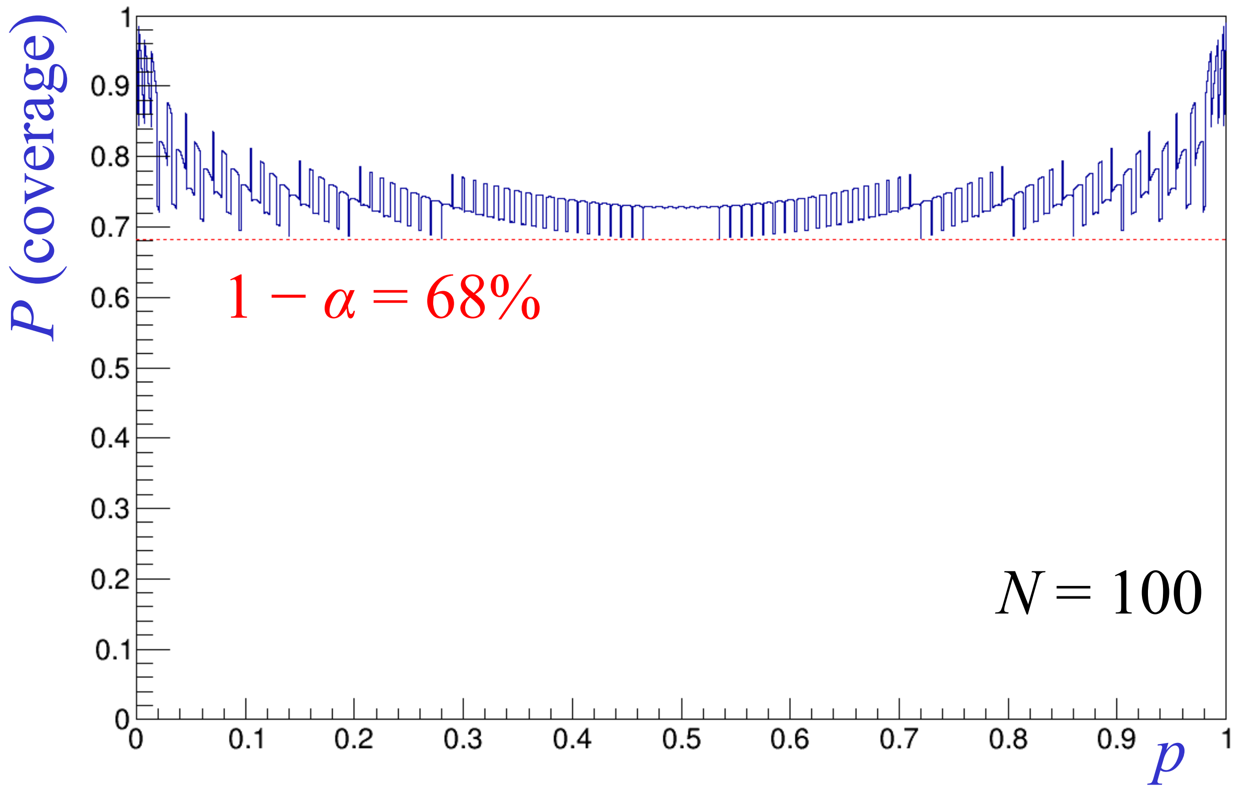
Clopper-Pearson coverage (I)

- CP intervals are often defined as “exact” in literature
- Exact coverage is often impossible to achieve for discrete variables



Clopper-Pearson coverage (II)

- For larger N the "ripple" gets closer to the nominal 68% coverage



Coverage for discrete variables

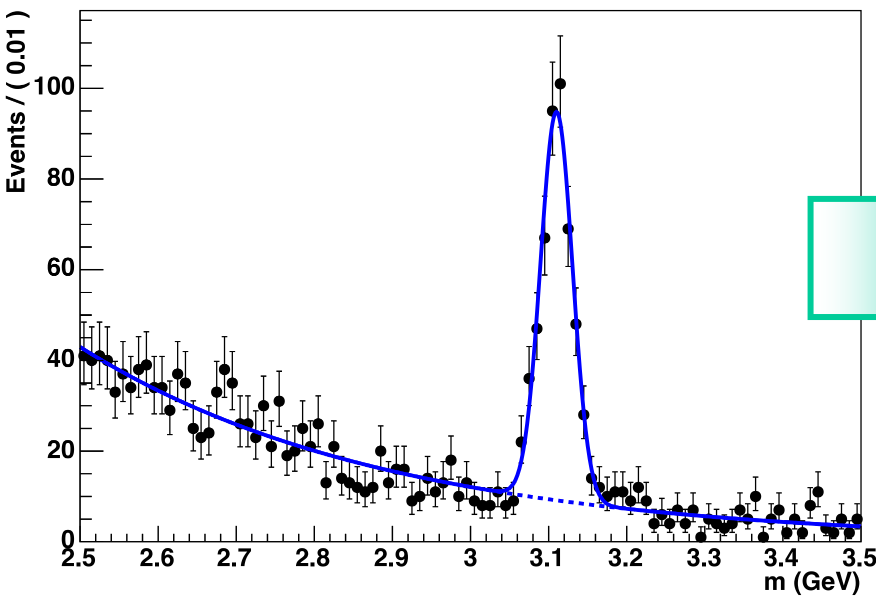
- Note: if the true value is $p = 0$ (similarly for $p = 1$), the observed value is always $n = 0$, therefore the confidence interval is $[0, pup[$
- The true value is therefore contained in the confidence interval with 100% probability instead of 68% (or 90%, or whatever)
- This is against the definition of frequentist coverage, but it is unavoidable for discrete variable
- This feature is typical of cases with low number of counts, for instance, for Poissonian counting experiments

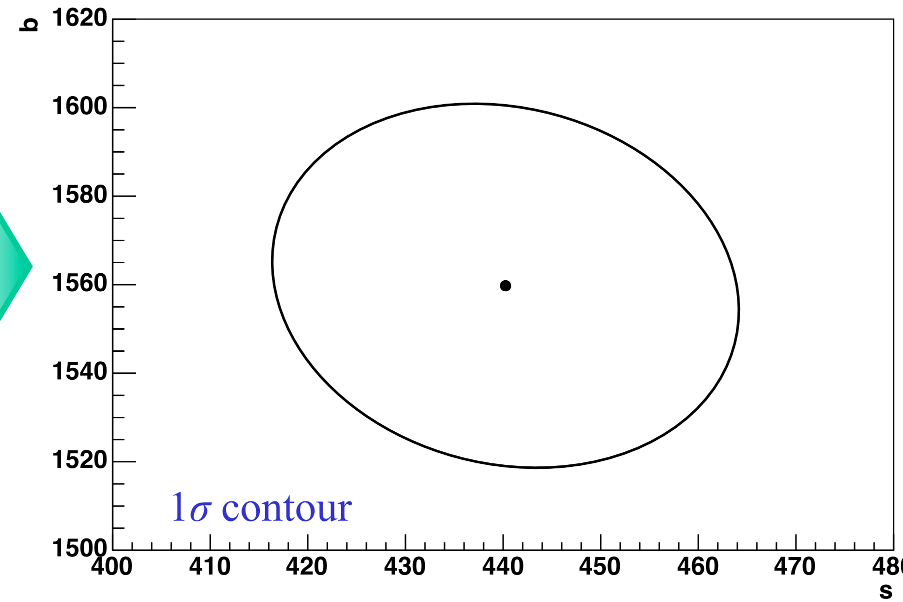
Example of 2D contour

- From previous fit example:
  - $P_s(m)$: Gaussian peak
  - $P_b(m)$: exponential shape

Exponential decay parameter, Gaussian mean and standard deviation are fit together with s and b yields.

The contour shows for this case a mild correlation between s and b





Error propagation

- Assume we estimate from a fit the parameter set: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and we know their covariance matrix Θ_{ij}
- We want to determine a new set of parameters that are functions of $\boldsymbol{\theta}$: $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$.
- For small uncertainties, a linear approximation maybe sufficient
- A Taylor expansion around the central values of $\boldsymbol{\theta}$ gives, using the error matrix Θ_{ij}:

$$H_{ij} = \sum_{k,l} \frac{\partial \eta_i}{\partial \theta_k} \frac{\partial \eta_j}{\partial \theta_l} \Theta_{kl}$$

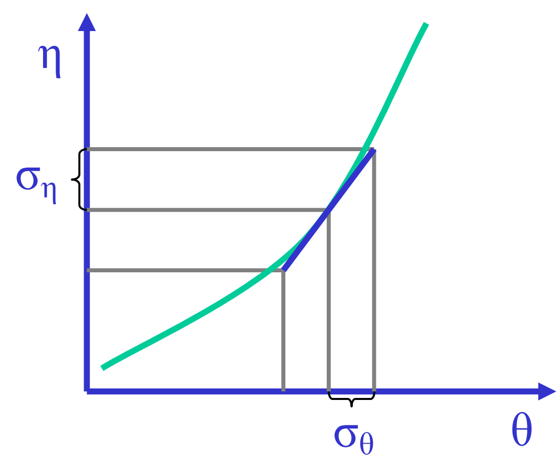
- Few examples in case of no correlation:

$$\sigma_{x+y} = \sigma_{x-y} = \sqrt{\sigma_x^2 + \sigma_y^2}$$

$$\frac{\sigma_{xy}}{xy} = \frac{\sigma_{x/y}}{x/y} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2}$$

$$\sigma_{x^2} = 2x\sigma_x$$

$$\sigma_{\ln x} = \frac{\sigma_x}{\sqrt{x}}$$



Binned likelihood

- Sometimes data are available as binned histogram
  - Most often each bin obeys Poissonian statistics (event counting)
- The likelihood function is the product of Poisson PDFs corresponding to each bin having entries n_i
- The expected number of entries n_i depends on some unknown parameters: $\mu_i = \mu_i(\theta_1, \ldots, \theta_m)$
- The function to minimize is the following $-2 \ln L$:

$$
\begin{aligned}
-2\ln L &= -2\ln \prod_{i=1}^{n_{\text{bins}}} \text{Poiss}(n_i; \mu_i(\theta_1, \cdots, \theta_m)) \\
&= -2\ln \prod_{i=1}^{n_{\text{bins}}} \frac{e^{-\mu_i(\theta_1, \cdots, \theta_m)} \mu_i(\theta_1, \cdots, \theta_m)^{n_i}}{n_i!}
\end{aligned}
$$

- The expected number of entries μ_i is often approximated by a continuous function $\mu(x)$ evaluated at the center x_i of the bin
- Alternatively, μ_i can be a combination of other histograms (“templates”)
  - E.g.: sum of different simulated processes with floating yields as fit parameters

Binned fits: minimum χ^2

- Bin entries can be approximated by Gaussian variables for sufficiently large number of entries with standard deviation equal to n_i (Neyman's χ^2)
- Maximizing L is equivalent to minimize:

$$\chi^2 = \sum_{i=1}^{n_{\mathrm{bins}}} \frac{(n_i - \mu(x_i; \theta_1, \cdots, \theta_m))^2}{n_i}$$

- Sometimes, the denominator n_i is replaced (Pearson's χ^2) by:

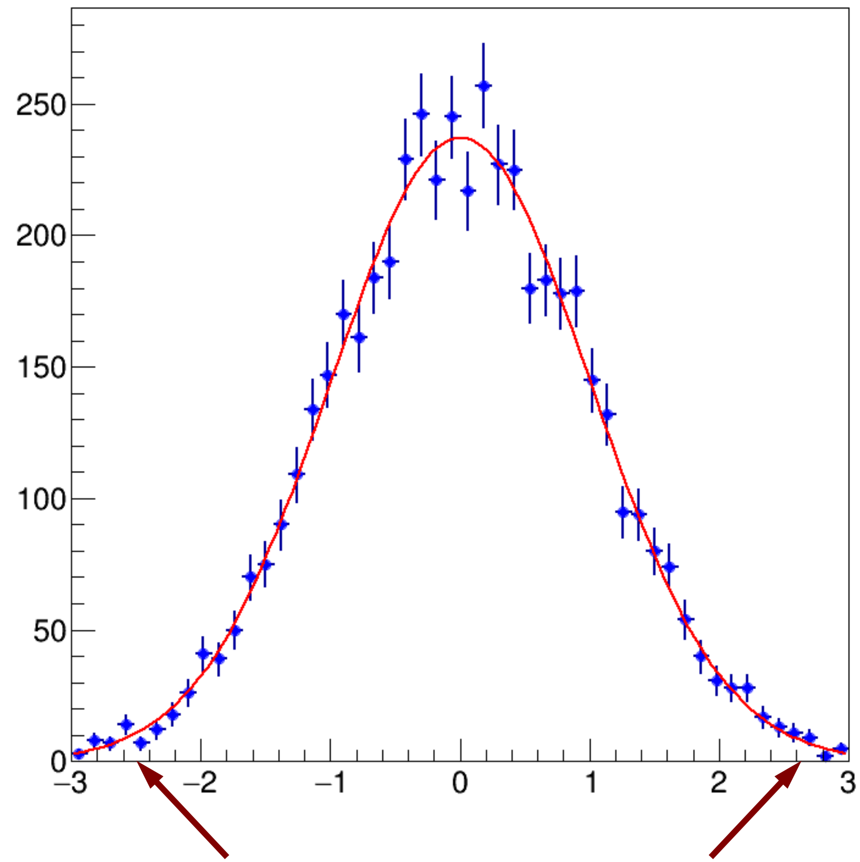
$$\mu_i = \mu\,(x_i;\, \theta_1, \ldots, \theta_m)$$

  in order to avoid cases with zero or small n_i
- Analytic solution exists for linear and other simple problems
  - E.g.: linear fit model
- Most of the cases are treated numerically, as for unbinned ML fits

Binned fit example

- Binned fits are convenient w.r.t. unbinned fits because the number of input variables decreases from the number of entries to the number of bins
  - Usually simpler and faster numerically
  - Unbinned fits become unpractical for very large number of entries
- A fraction of the information is lost, hence a possible loss of precision may occur for small number of entries
- Treat correctly bins with smalll number of entries!

Gaussian fit (determine yield, μ and σ)



Bins with small number of entries!

Fit quality (χ^2 test)

- The maximum value of the likelihood function obtained from the fit doesn't usually give information about the goodness of the fit
- The χ^2 of a fit with a Gaussian underlying model is distributed according to a known PDF

$$P(\chi^2; n) = \frac{2^{-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} \chi^{n-2} e^{-\frac{\chi^2}{2}}$$

n is the number of degrees of freedom (n. of bins − n. of params.)

  - The cumulative distribution of $P(\chi^2; n)$ follows a uniform distribution between 0 and 1 (p-value)
  - If the model deviates from the assumed distribution, the distribution of the p-value will be more peaked around zero
- Note! p-values are not the "*probability of the fit hypothesis*"
  - This would be a Bayesian probability, with a different meaning, and should be computed in a different way

Binned likelihood ratio

- A better alternative to the (Gaussian-inspired, Neyman and Pearson's) χ^2 has been proposed by Baker and Cousins using the following likelihood ratio:

$$
\begin{aligned}
\chi^2_\lambda &= -2\ln\prod_i \frac{L(n_i;\mu_i)}{L(n_i;n_i)} = -2\ln\prod_i \frac{e^{-\mu_i}\mu_i^{n_i}}{\cancel{n_i!}}\frac{\cancel{n_i!}}{e^{-n_i}n_i^{n_i}} \\
&= 2\sum_i\left[\mu_i(\theta_i,\cdots,\theta_m) - n_i + n_i\ln\left(\frac{n_i}{\mu_i(\theta_1,\cdots,\theta_m)}\right)\right]
\end{aligned}
$$

- Same minimum value as from Poisson likelihood function, since a constant term has been added to the log-likelihood function
- In addition, it provides goodness-of-fit information, and asymptotically obeys chi-squared distribution with $n - m$ degrees of freedom
(Wilks' theorem, see following slides)

S. Baker, R. Cousins NIM 221 (1984) 437

Combining measurements

- Assume two measurements with different uncorrelated (Gaussian) errors: $m_1 \pm \sigma_1, \quad m_2 \pm \sigma_2$

- Build the χ^2: $$\chi^2 = \frac{(m-m_1)^2}{\sigma_1^2} + \frac{(m-m_2)^2}{\sigma_2^2}$$

- Minimize the χ^2: $$0 = \frac{\partial \chi^2}{\partial m} = 2\frac{(m-m_1)}{\sigma_1^2} + 2\frac{(m-m_2)}{\sigma_2^2}$$

- Estimate m as: $$m = \frac{\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{w_1 m_1 + w_2 m_2}{w_1 + w_2}$$ ← Weighted average, $w_i = \sigma_i^{-2}$

- Error estimate: $$\frac{1}{\sigma_m^2} = -\frac{\partial^2 \ln L}{\partial m^2} = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial m^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}$$

Combining correlated measurements

- Correlation coefficient $\rho \neq 0$:

$$m_1 \pm \sigma_1, \quad m_2 \pm \sigma_2$$

- Build χ^2 including correlation terms:

$$\chi^2 = \begin{pmatrix} m - m_1 & m - m_2 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} m - m_1 \\ m - m_2 \end{pmatrix}$$

- The χ^2 minimization gives:

$$m = \frac{m_1(\sigma_2^2 - \rho\sigma_1\sigma_2) + m_2(\sigma_1^2 - \rho\sigma_1\sigma_2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}$$

$$\sigma_m^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho)^2}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}$$

a.k.a "**BLUE**":
Best **L**inear **U**nbiased **E**stimator

Correlated errors

- The "common error" σ_C is defined as: $\sigma_C^2 = \rho\sigma_1\sigma_2$
- Using error propagation, this also implies that:

$$\sigma^2_{m_1-m_2} = \left(\frac{\partial(m_1-m_2)}{\partial m_1}\right)^2\sigma_1^2 + \left(\frac{\partial(m_1-m_2)}{\partial m_2}\right)^2\sigma_2^2 + 2\left(\frac{\partial(m_1-m_2)}{\partial m_1}\right)\left(\frac{\partial(m_1-m_2)}{\partial m_2}\right)\rho\sigma_1\sigma_2$$

$$\Downarrow$$

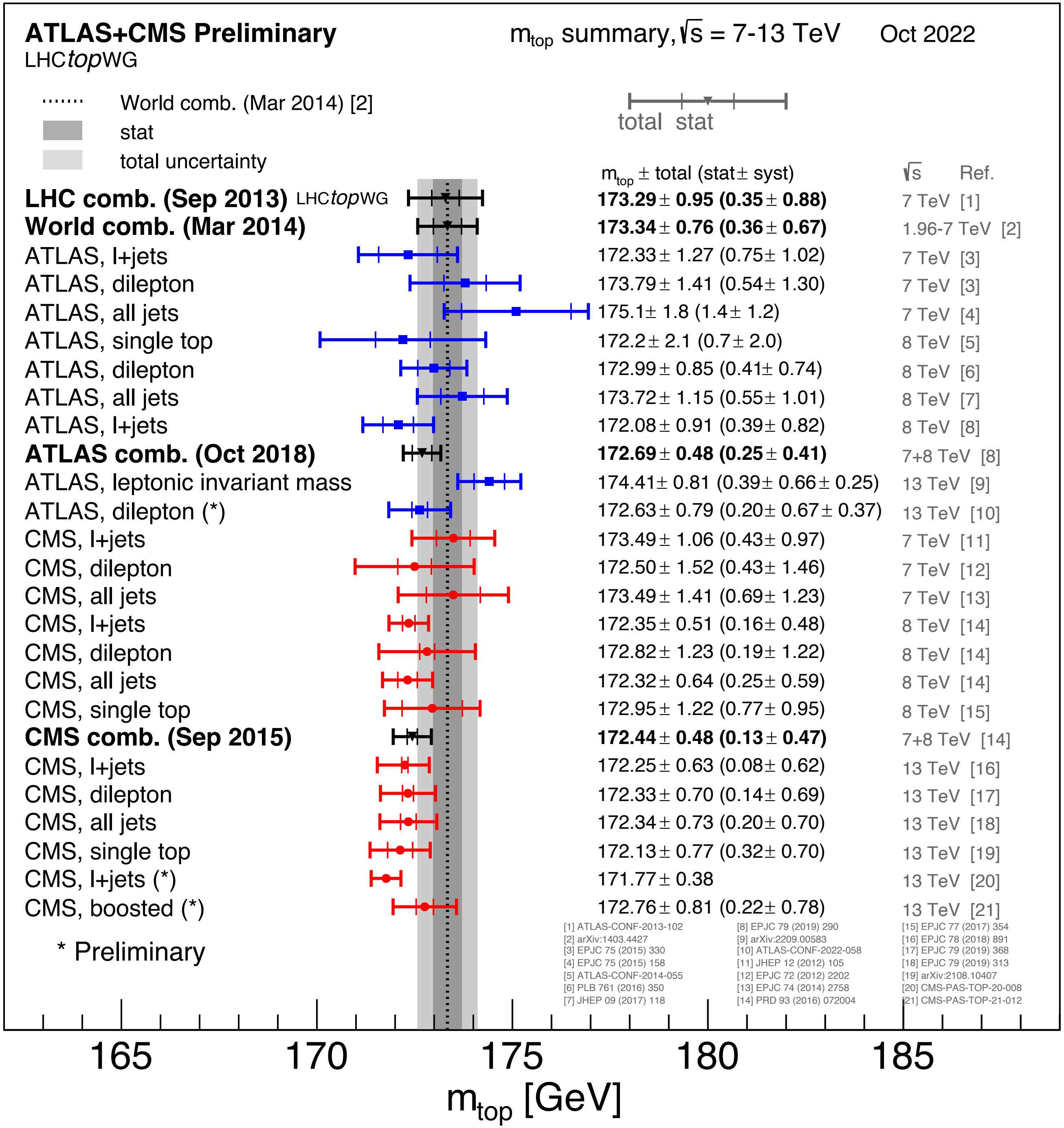
$$\sigma^2_{m_1-m_2} = (\sigma_1^2-\sigma_C^2) + (\sigma_2^2-\sigma_C^2)$$

- The previous formulas can be written as a weighted average:

$$m = \frac{\frac{m_1}{\sigma_1^2-\sigma_C^2}+\frac{m_2}{\sigma_2^2-\sigma_C^2}}{\frac{1}{\sigma_1^2-\sigma_C^2}+\frac{1}{\sigma_2^2-\sigma_C^2}} = \frac{w_1m_1+w_2m_2}{w_1+w_2}$$

Note: weights may be negative!

$$\sigma_m^2 = \frac{1}{\frac{1}{\sigma_1^2-\sigma_C^2}+\frac{1}{\sigma_2^2-\sigma_C^2}} + \sigma_C^2$$



ATLAS+CMS Preliminary
LHCtopWG
m_top summary, √s = 7-13 TeV
Oct 2022
World comb. (Mar 2014) [2]
stat
total uncertainty
total stat
m_top ± total (stat ± syst)
√s Ref.
LHC comb. (Sep 2013) LHCtopWG 173.29 ± 0.95 (0.35 ± 0.88) 7 TeV [1]
World comb. (Mar 2014) 173.34 ± 0.76 (0.36 ± 0.67) 1.96-7 TeV [2]
ATLAS, l+jets 172.33 ± 1.27 (0.75 ± 1.02) 7 TeV [3]
ATLAS, dilepton 173.79 ± 1.41 (0.54 ± 1.30) 7 TeV [3]
ATLAS, all jets 175.1 ± 1.8 (1.4 ± 1.2) 7 TeV [4]
ATLAS, single top 172.2 ± 2.1 (0.7 ± 2.0) 8 TeV [5]
ATLAS, dilepton 172.99 ± 0.85 (0.41 ± 0.74) 8 TeV [6]
ATLAS, all jets 173.72 ± 1.15 (0.55 ± 1.01) 8 TeV [7]
ATLAS, l+jets 172.08 ± 0.91 (0.39 ± 0.82) 8 TeV [8]
ATLAS comb. (Oct 2018) 172.69 ± 0.48 (0.25 ± 0.41) 7+8 TeV [8]
ATLAS, leptonic invariant mass 174.41 ± 0.81 (0.39 ± 0.66 ± 0.25) 13 TeV [9]
ATLAS, dilepton (*) 172.63 ± 0.79 (0.20 ± 0.67 ± 0.37) 13 TeV [10]
CMS, l+jets 173.49 ± 1.06 (0.43 ± 0.97) 7 TeV [11]
CMS, dilepton 172.50 ± 1.52 (0.43 ± 1.46) 7 TeV [12]
CMS, all jets 173.49 ± 1.41 (0.69 ± 1.23) 7 TeV [13]
CMS, l+jets 172.35 ± 0.51 (0.16 ± 0.48) 8 TeV [14]
CMS, dilepton 172.82 ± 1.23 (0.19 ± 1.22) 8 TeV [14]
CMS, all jets 172.32 ± 0.64 (0.25 ± 0.59) 8 TeV [14]
CMS, single top 172.95 ± 1.22 (0.77 ± 0.95) 8 TeV [15]
CMS comb. (Sep 2015) 172.44 ± 0.48 (0.13 ± 0.47) 7+8 TeV [14]
CMS, l+jets 172.25 ± 0.63 (0.08 ± 0.62) 13 TeV [16]
CMS, dilepton 172.33 ± 0.70 (0.14 ± 0.69) 13 TeV [17]
CMS, all jets 172.34 ± 0.73 (0.20 ± 0.70) 13 TeV [18]
CMS, single top 172.13 ± 0.77 (0.32 ± 0.70) 13 TeV [19]
CMS, l+jets (*) 171.77 ± 0.38 13 TeV [20]
CMS, boosted (*) 172.76 ± 0.81 (0.22 ± 0.78) 13 TeV [21]
* Preliminary
[1] ATLAS-CONF-2013-102
[2] arXiv:1403.4427
[3] EPJC 75 (2015) 330
[4] EPJC 75 (2015) 158
[5] ATLAS-CONF-2014-055
[6] PLB 761 (2016) 350
[7] JHEP 09 (2017) 118
[8] EPJC 79 (2019) 290
[9] arXiv:2209.00583
[10] ATLAS-CONF-2022-058
[11] JHEP 12 (2012) 105
[12] EPJC 72 (2012) 2202
[13] EPJC 74 (2014) 2758
[14] PRD 93 (2016) 072004
[15] EPJC 77 (2017) 354
[16] EPJC 78 (2018) 891
[17] EPJC 79 (2019) 368
[18] EPJC 79 (2019) 313
[19] arXiv:2108.10407
[20] CMS-PAS-TOP-20-008
[21] CMS-PAS-TOP-21-012
165
170
175
180
185
m_top [GeV]

JE
on
and

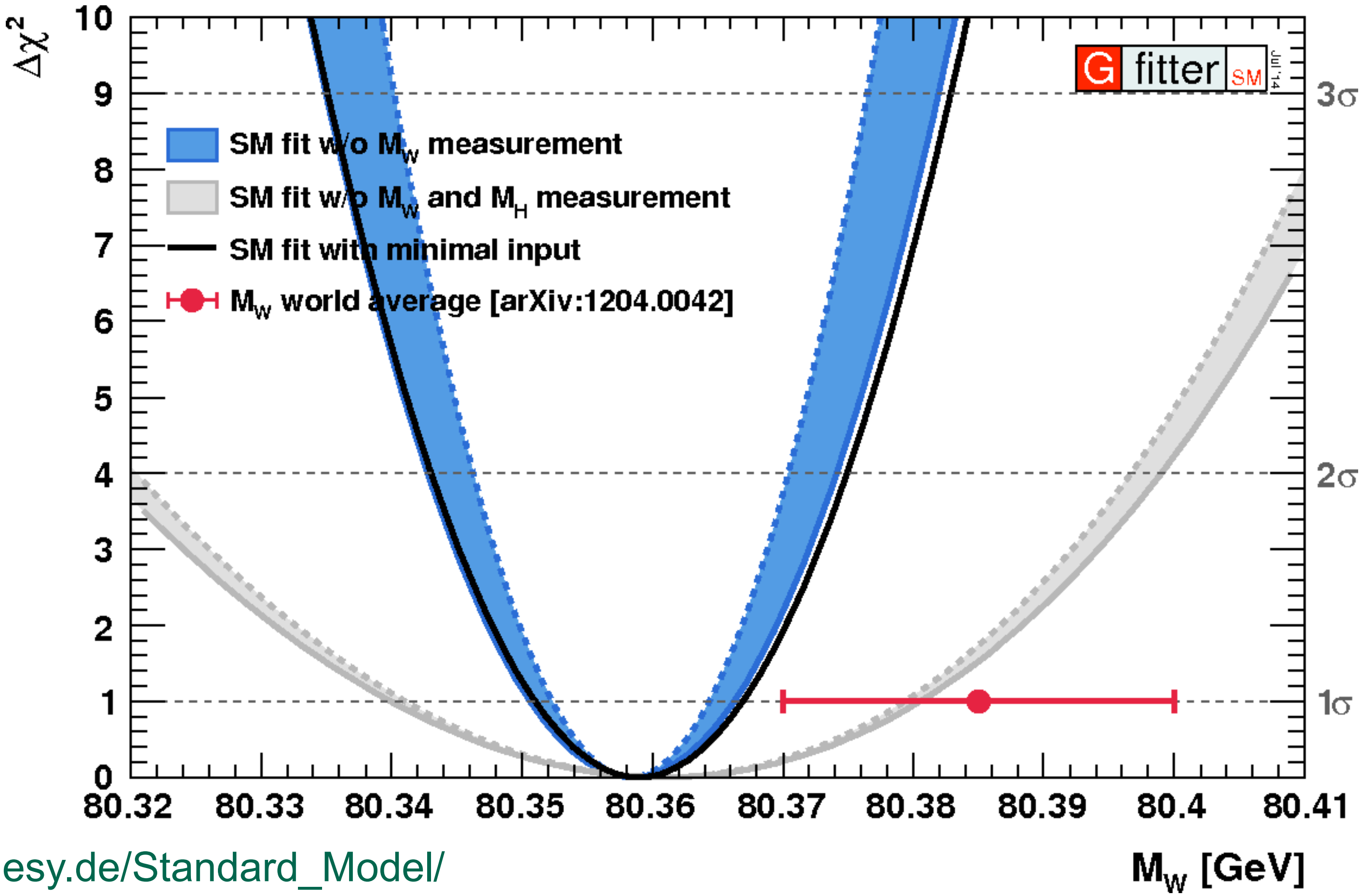
Generalization of χ^2 to n dimensions

- We have n measurements, $(m_1, \ldots, m_n)$ with a $n \times n$ covariance matrix (C_{ij})
- Expected values for $m_1, \ldots, m_n$, $M_1, \ldots, M_n$ may depend on some theory parameter(s) θ
- The following χ^2 can be minimized to have an estimate of the parameter(s) θ:

$$
\begin{aligned}
\chi^2 &= \sum_{i,j=1}^{n} (m_i - M_i(\theta)) C_{ij}^{-1} (m_j - M_j(\theta)) \\
&= \begin{pmatrix} m_1 - M_1, & \cdots, & m_n - M_n \end{pmatrix} \begin{pmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & \ddots & \vdots \\ C_{n1} & \cdots & C_{nn} \end{pmatrix}^{-1} \begin{pmatrix} m_1 - M_1 \\ \cdots \\ m_n - M_n \end{pmatrix} \\
&= (\boldsymbol{m} - \boldsymbol{M}(\theta))^T \boldsymbol{C}^{-1} (\boldsymbol{m} - \boldsymbol{M}(\theta))
\end{aligned}
$$

Global electroweak fit

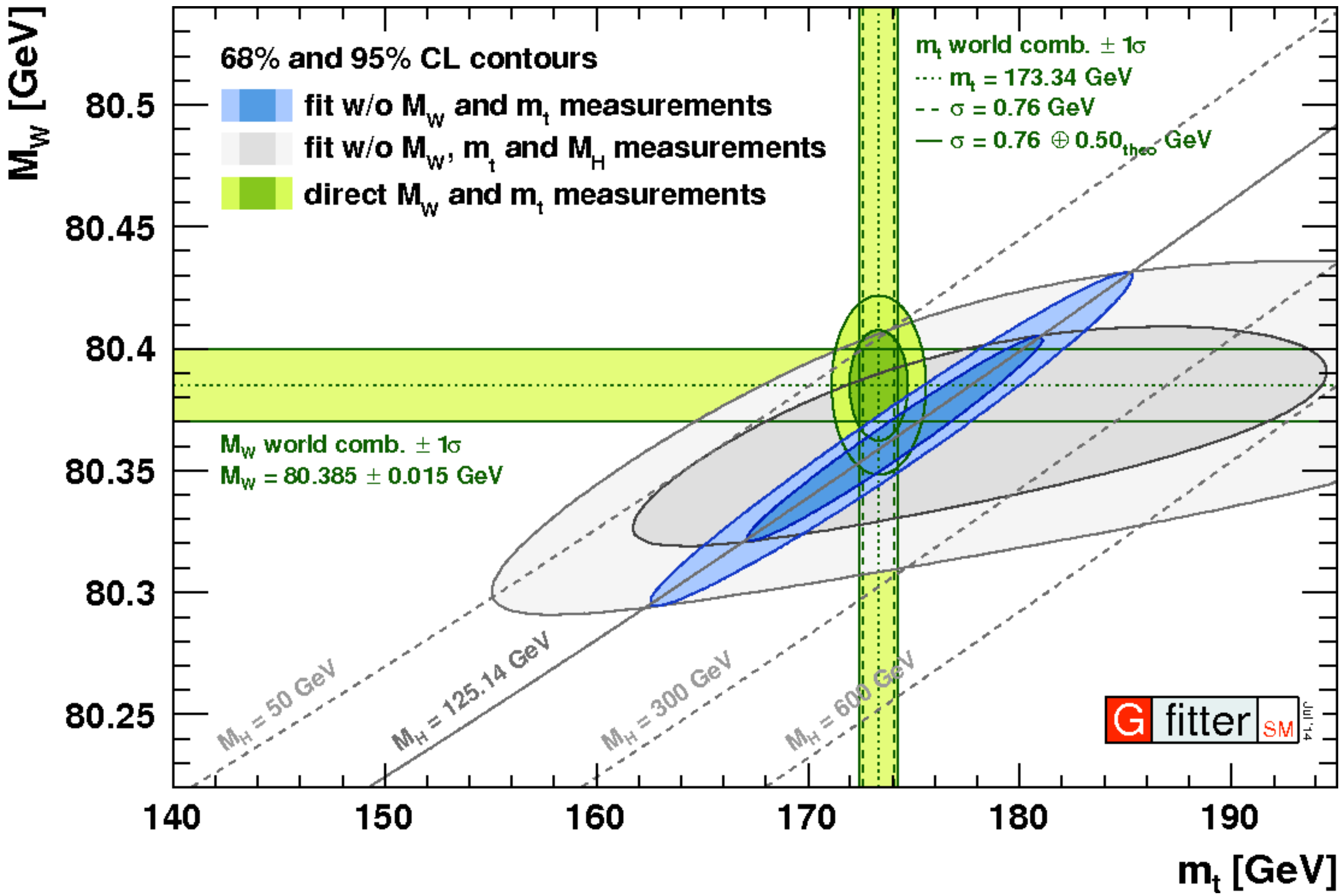
- A Global χ^2 fit to electroweak measurements predicts the W mass allowing a comparison with direct measurements



Details on:
http://gfitter.desy.de/Standard_Model/

More on electroweak fit

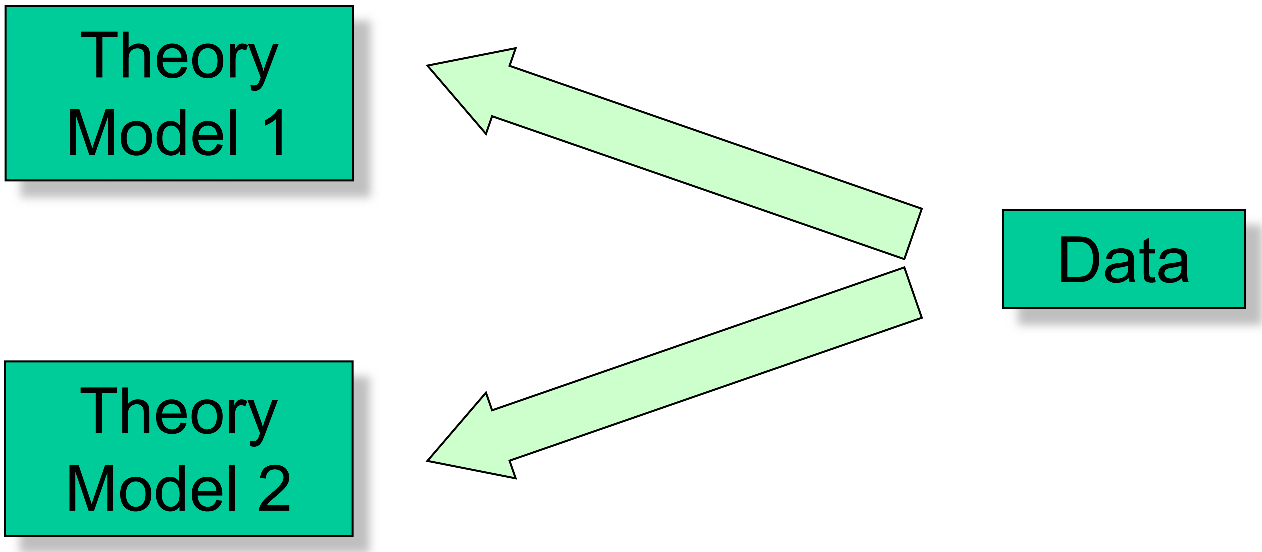
- W mass vs top-quark mass from global electroweak fit



Lecture 3

- Hypothesis testing
- The ROC curve
- The Neyman-Pearson lemma
- Multivariate discrimination
- Projective likelihood ratio
- Fisher discriminant
- Introduction to machine learning
- Artificial Neural Networks
- Boosted decision trees
- The bias-variance tradeoff
- Issues with machine learning: underfitting, overfitting

Hypothesis testing



Which hypothesis is the most consistent with the experimental data?

Bayesian approach

- Bayesian probability gives meaning to the probability that an hypothesis is true:

$$P(H_1|x) = \frac{P(x|H_1)\pi(H_1)}{P(x)}$$

- The evaluation of $P(x)$ requires the decomposition over all possible hypotheses:

$$P(x) = P(x|H_0)\pi(H_0) + P(x|H_1)\pi(H_1) + \cdots$$

- The ratio of probabilities for two hypothesis does not depend on $P(x)$, and can be computed without considering all possible hypotheses:

$$\frac{P(H_1|x)}{P(H_0|x)} = \frac{P(x|H_1)\pi(H_1)}{P(x|H_0)\pi(H_0)}$$

Bayes factors

- It is possible to introduce Bayes factor:

$$\frac{P(H_1|x)}{P(H_0|x)} = B_{1/0}(x)\frac{\pi(H_1)}{\pi(H_0)}$$

- In other words, this defines the posterior odds as a function of prior odds:

$$O_{1/0}(x) \ = B_{1/0}(x)\, o_{1/0}$$

- In word:

posterior odds = Bayes factor times prior odds

- Bayes factor can be used to measure how favoured is one hypothesis against another, and, in the simplest cases, it is equal to the likelihood ratio
- Typical range values are:
  - 1-3: very weak evidence
  - 3-20: positive evidence
  - 20-150: strong evidence
  - >150: very strong evidence

In presence of other parameters

$$P(H_1, \theta_1 | x) = \frac{P(x|H_1, \theta_1)\pi(H_1, \theta_1)}{P(x)}$$

$$P(H_1|x) = \frac{\int P(x|H_1, \theta_1)\pi(H_1, \theta_1)\mathrm{d}\theta_1}{P(x)} = \frac{\pi(H_1)\ \int P(x|H_1, \theta_1)\pi(\theta_1|H_1)\mathrm{d}\theta_1}{P(x)}$$

- Because one has:

$$\pi(H_0, \theta_0) = \pi(\theta_0|H_0)\pi(H_0)$$
$$\pi(H_1, \theta_1) = \pi(\theta_1|H_1)\pi(H_1)$$

- The Bayes factor should defined as:

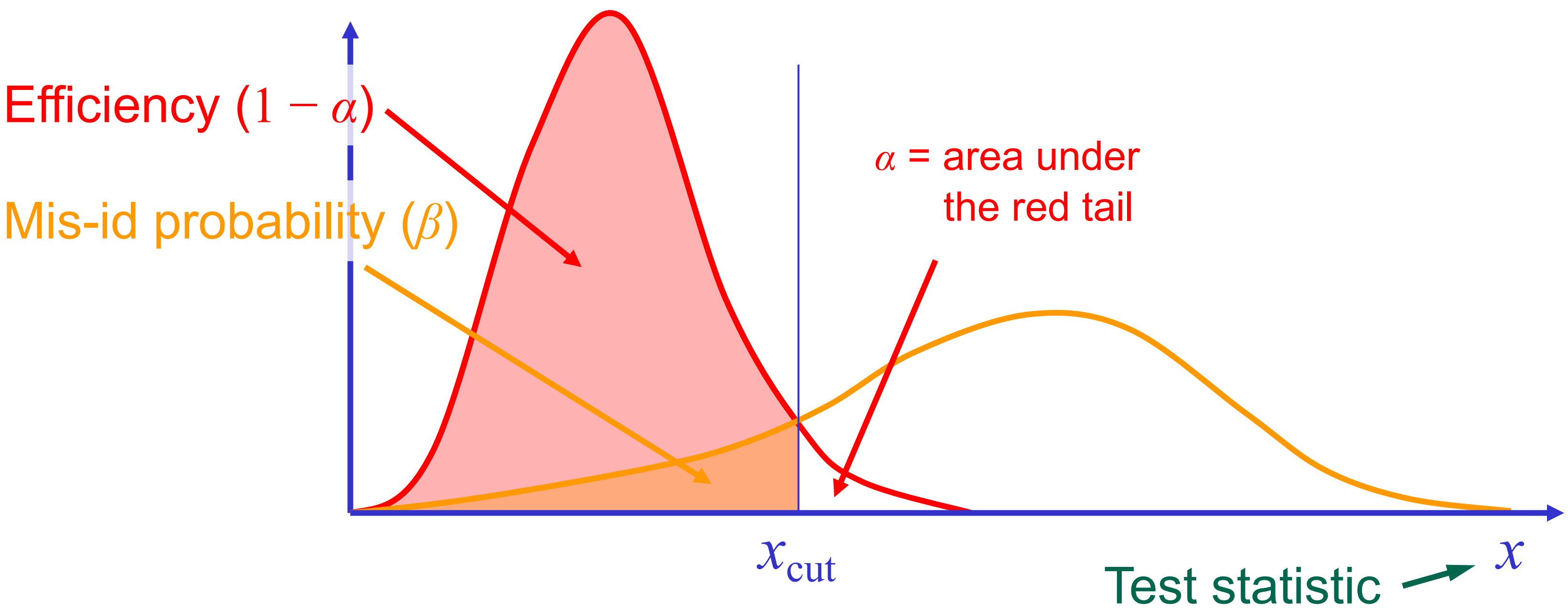
$$B_{1/0}(x) = \frac{P(x|H_1)}{P(x|H_0)}$$

- I.e.: it is the ratio of marginal likelihood for x:

$$B_{1/0}(x) = \frac{P(x|H_1)}{P(x|H_0)} = \frac{\int P(x|H_1, \theta_1)\pi(\theta_1|H_1)\mathrm{d}\theta_1}{\int P(x|H_0, \theta_0)\pi(\theta_0|H_0)\mathrm{d}\theta_0}$$

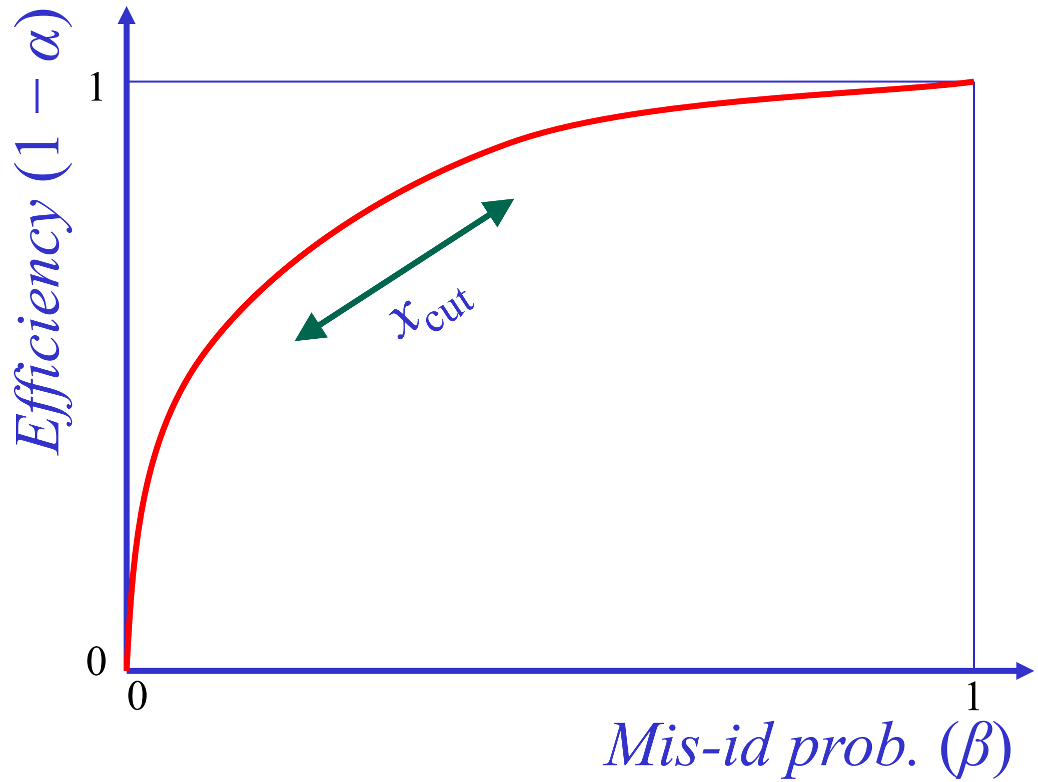
Simplest case: cut analysis

- Selection ("cut") on one (or more) variable(s):
  - If $x \leq x_{\mathrm{cut}}$ $\Rightarrow$ signal
  - Else, if $x > x_{\mathrm{cut}}$ $\Rightarrow$ background



Efficiency *vs* mis-id

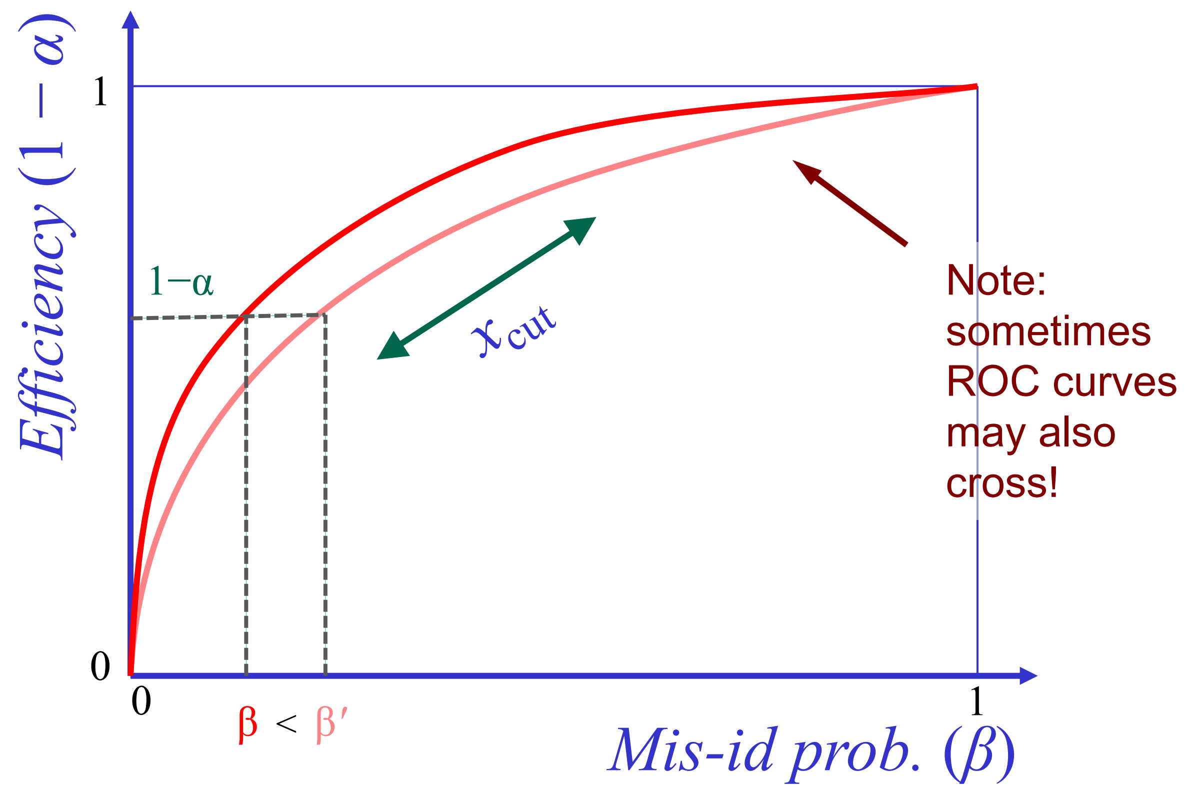
- Varying the applied cut on the test statistic both the efficiency and mis-id probability change



Sometimes also referred to as **ROC curve** (*Receiver Operating Characteristic*)

Performance comparison

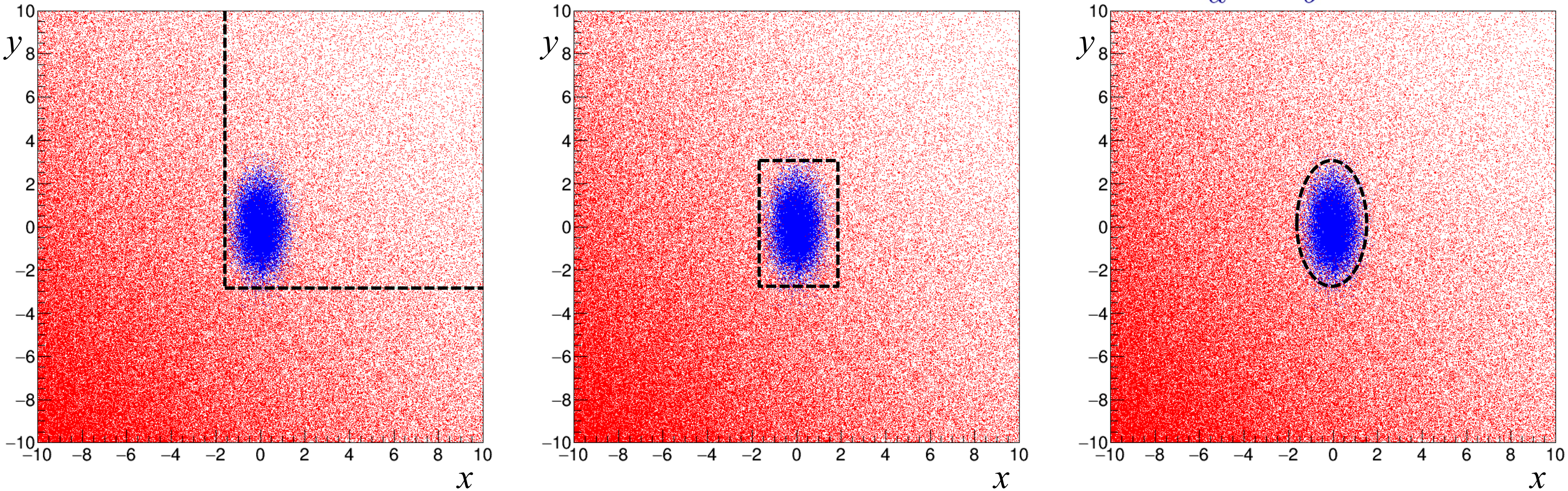
- One test is preferable to another if, for the same level of efficiency ($1-\alpha$), it has lower mis-id probability (β)



Multidimensional case

- In case of multiple discriminating variables (multivariate), the choice of the optimal selection is not always straightforward

$x > x_{\mathrm{cut}}$ and $y > y_{\mathrm{cut}}$ $\quad$ $x_1 < x < x_2$ and $y_1 < y < y_2$ $\quad$ $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} < k^2$



$$P_s(x, y) = \mathrm{Gauss}(x; 0, \sigma_x) \times \mathrm{Gauss}(y; 0, \sigma_y)\ , \quad P_b(x, y) = \alpha\, e^{-\alpha x} \times \beta\, e^{-\beta y}$$

- In many cases it’s convenient to find a single variable (test statistic) that ‘summarizes’ all the sample information

Terminology

- Statisticians' terminology is sometimes not very natural for physics applications, but it has become popular among physicists as well:

- H_0 = **null hypothesis**
  - Ex. 1: *"a sample contains only background"*
  - Ex. 2: *"a particle is a pion"*
- H_1 = **alternative hypothesis**
  - Ex. 1: *"a sample contains background + signal"*
  - Ex. 2: *"a particle is a muon"*

- **Test statistic**: a variable computed from our sample that discriminates between the two hypotheses H_0 and H_1. Usually a 'summary' of the information available in the sample
- ***α* = significance level**: probability to reject H_1 if $H_0$is assumed to be true (error of first kind, false positive)
  - $\alpha = 1$ – misidentification probability
- ***β* = misidentification probability**, i.e.: probability to reject H_0 if H_1 is assumed to be true (error of second kind, false negative)
  - $1 - \beta$ = **power of the test =** selection efficiency
- ***p*-value**: probability, assuming H_0, of observing a result at least as extreme as the observed test statistic

The Neyman-Pearson lemma

- For a fixed significance level (α) or signal efficiency ($1-\alpha$), a selection based on the likelihood ratio gives the lowest possible mis-id probability (β):

$$\lambda(x) = \frac{L(x|H_1)}{L(x|H_0)} > k_\alpha$$

- The likelihood function can't always be determined exactly
- If we can't determine the exact likelihood function, we can choose other discriminators as test statistics that approximates the exact likelihood
- Neural Networks, Boosted Decision Trees and other machine-learning algorithms are example of discriminators that may closely approximate the performances of the exact likelihood ratio approaching the Neyman-Pearson limit

Multivariate discrimination

- In general, when we consider algorithms that provide a test statistic for samples with multiple variables we talk about multivariate discriminators
  - Simple mathematical algorithms exist, as well as complex implementation based on extensive CPU computations
- In general, the algorithms are 'trained' using input samples whose nature is known (training samples)
  - I.e.: where H_0 or H_1 is know to be true
  - Example: use data samples simulated with computer algorithms (Monte Carlo)
- Some of the most common problems:
  - The size of samples is finite, hence the true distributions for the considered hypotheses can't be determined exactly
  - The distribution of the input samples does not reproduce exactly the true distribution of real data (e.g.: systematic uncertainties)

Projective likelihood ratio

- The likelihood function is approximated by the product of projective PDF in each variable

$$\lambda(x) = \frac{L(x_1, \cdots, x_n | H_1)}{L(x_1, \cdots, x_n | H_0)} \simeq \underbrace{\frac{\prod_{i=1}^{n} f_i(x_i | H_1)}{\prod_{i=1}^{n} f_i(x_i | H_0)}}_{x_1,\ \ldots,\ x_n \text{ approximately considered independent variables}}$$

- Exact only in case of independent variables
- The approximation may be improved if the variables are first rotated in order to eliminate correlation (principal component analysis)
  - Find eigenvectors of the covariance matrix
  - **Note**: uncorrelated variables are not necessarily independent

Fisher discriminant

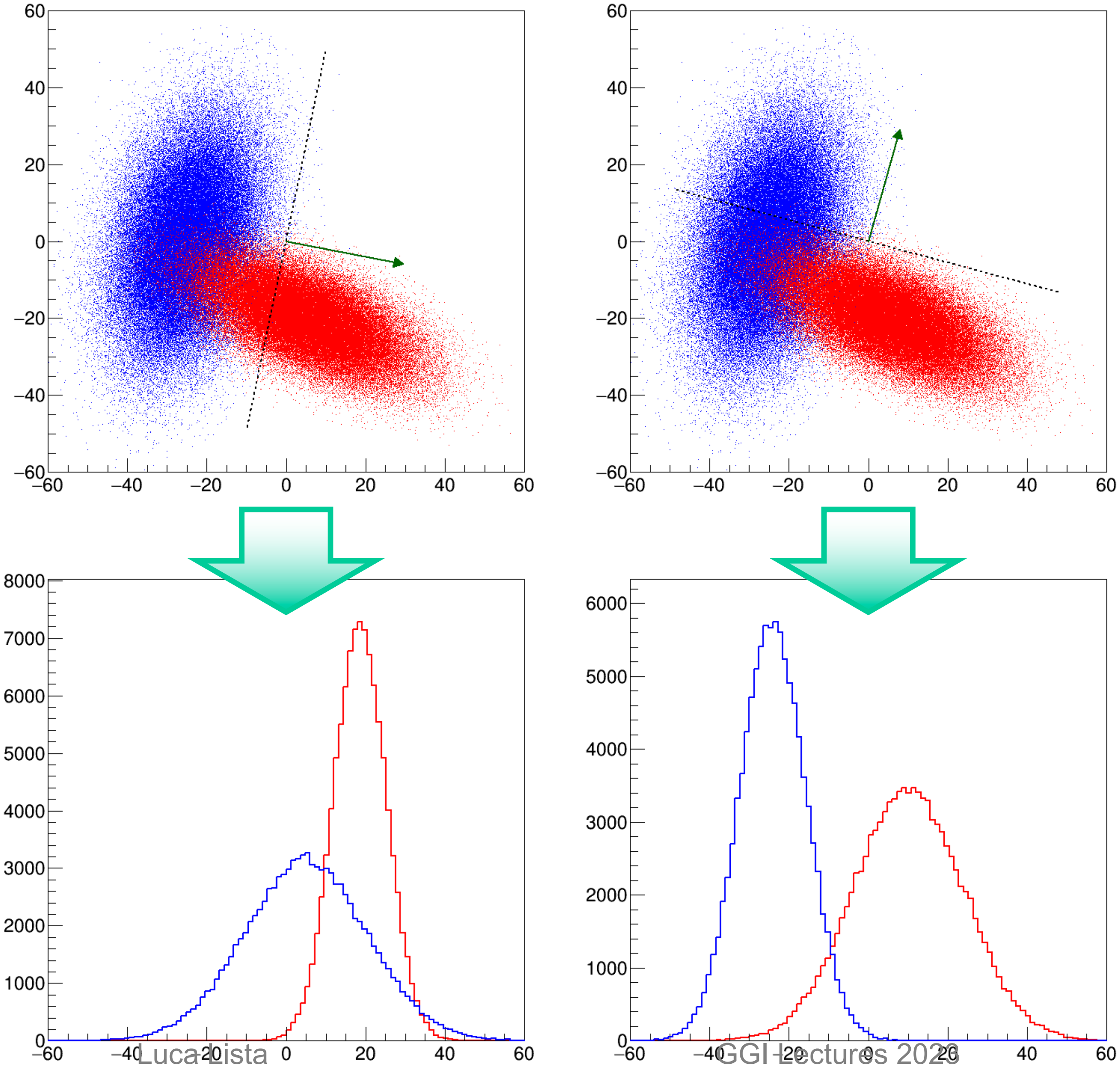
- Linear combination of input variables that maximizes the distance of the means of the two classes while minimizing the variance projected along a direction $\mathbf{w}$:

$$J(\mathbf{w}) = \frac{|\mu_0 - \mu_1|^2}{\sigma_0^2 + \sigma_1^2} = \frac{\mathbf{w}^{\mathrm{T}} \cdot (\mathbf{m}_0 - \mathbf{m}_1)}{\mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)\mathbf{w}}$$

Sir Ronald Aylmer Fisher (1890-1962)

- The selection is achieved by requiring $J(\mathbf{w}) > J_{\mathrm{cut}}$, which determines an hyperplane perpendicular to $\mathbf{w}$ that separates the two samples
- The maximization problem can be solved analytically using linear algebra

Fisher discriminant

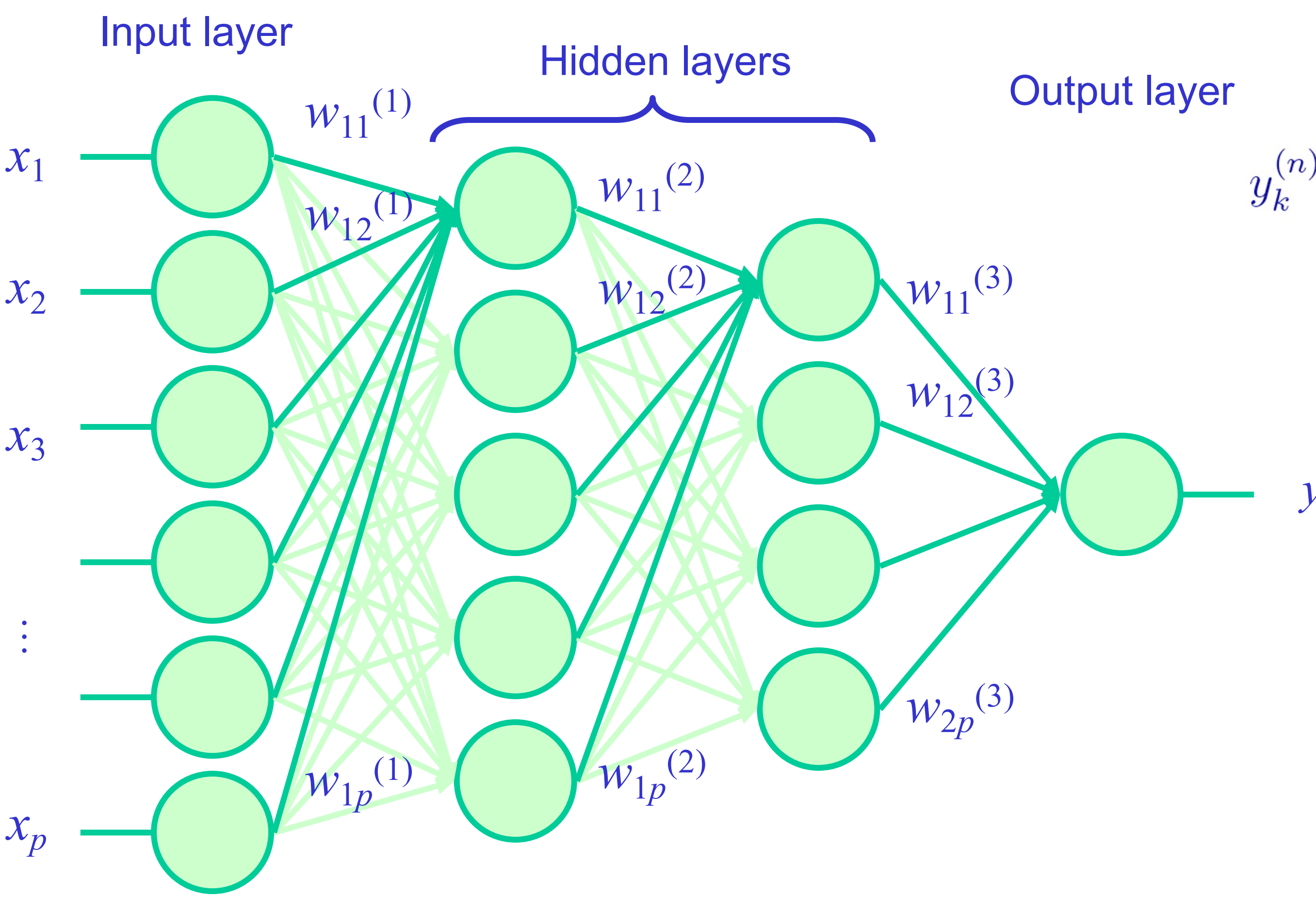


Projections along different directions achieve different overlap level

Maximum separation achieved by maximizing Fisher discriminant

Artificial Neural Networks

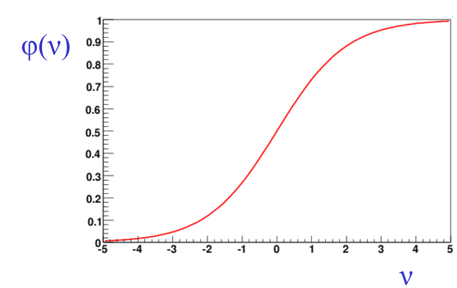
- Artificial simplified model of how neuron cells work: multilayer perceptron



$$y_k^{(n)}(\vec{x}) = \varphi\left(\sum_{j=1}^{p} w_{kj}^{(n)} x_j\right)$$

$\varphi(\nu)$ = Activation function

$$\varphi(\nu) = \frac{1}{1+e^{-\lambda\nu}}$$



Network training

- Find the optimal set of network parameters $w_{ij}^{(n)}$ that minimize the "loss function" defined on a set of N training events:

$$L(w) = \sum_{i=1}^{N} (y_i^{true} - y(\vec{x}_i))^2$$

- Variation of loss function are also used, like cross-entropy, adopted more for binomial models with parameter $p_i = y_i^{true}$:

$$L(w) = -\log\left(\prod_{i=1}^{N} y(\vec{x}_i)^{y_i^{true}} (1 - y(\vec{x}_i))^{1-y_i^{true}}\right) =$$

$$= -\sum_{i=1}^{N} \left[y_i^{true} \log y(\vec{x}_i) - \left(1 - y_i^{true}\right) \log(1 - y(\vec{x}_i))\right]$$

- Where $y_i^{\mathrm{true}} = 1$ for signal (H_1), 0 for background (H_0)
- Usually achieved with stochastic gradient descent: weights are modified for each training event (back propagation):

$$w_{ij} \rightarrow w_{ij} - \eta \frac{\partial L(w)}{\partial w_{ij}}$$

ANN and function modeling

- Artificial neural network with a single hidden layer may approximate any analytical function within a given approximation if the number of neurons is sufficiently high
- Demonstration in:
  - H. N. Mhaskar, Neural Computation, Vol. 8, No. 1, Pages 164-177 (1996), *Neural Networks for Optimal Approximation of Smooth and Analytic Functions*:

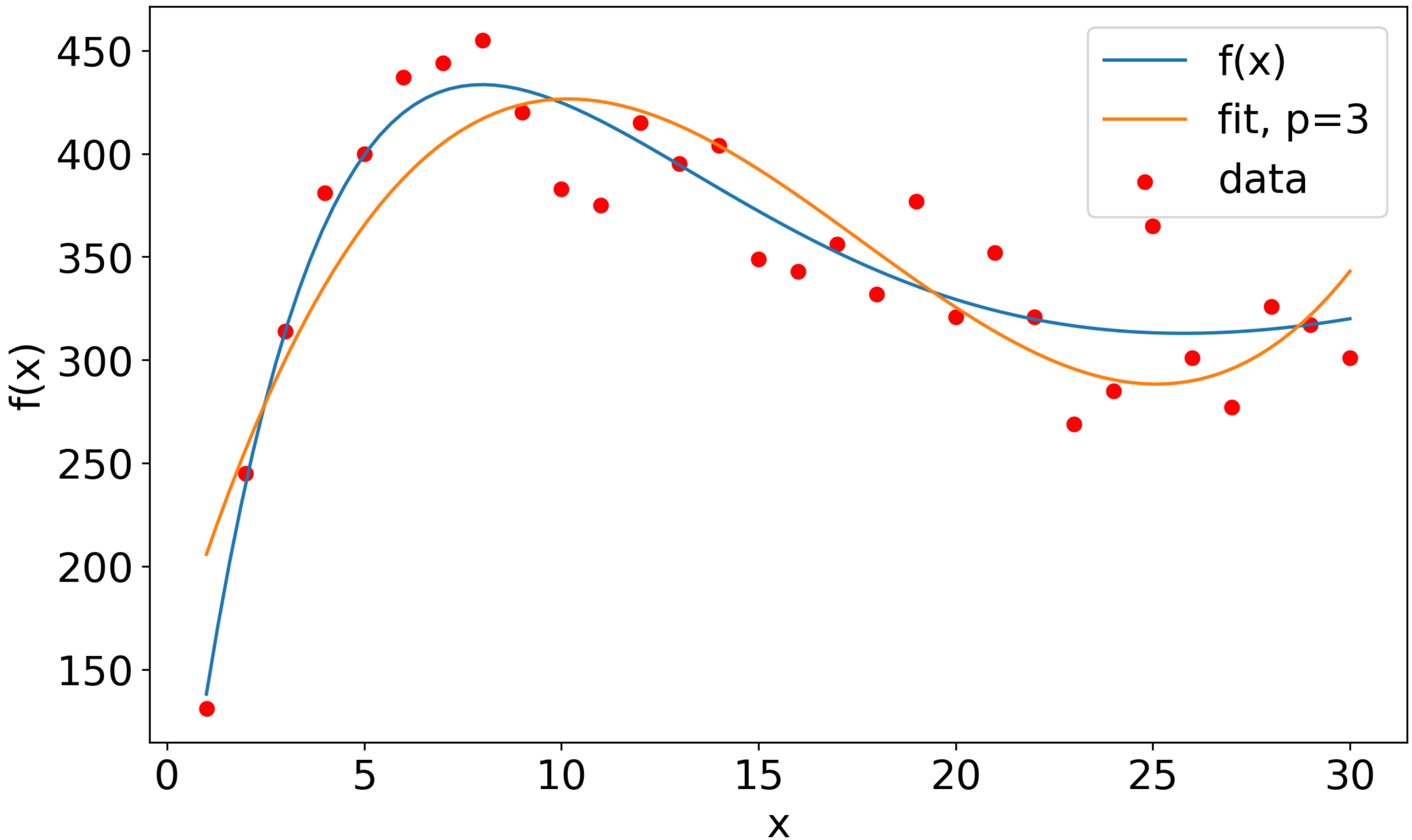
  "We prove that neural networks with a single hidden layer are capable of providing an optimal order of approximation for functions assumed to possess a given number of derivatives, if the activation function evaluated by each principal element satisfies certain technical conditions"

- Anyway, the finite number of layer may lead to reaching this goal only approximately
- Deep learning: networks with several hidden layers
  - Can manage complex variables combinations, e.g.: exploiting invariant mass distributions using four-vectors as input!
  - Almost untreatable in the past, a lot of progress has been done recently, with better training algorithms and more easily available CPU power

Fit seen as machine learning

- Let us take a polynomial fit of a points whose true model is not known
- The degree of the polynomial, i.e.: the number of parameter, can be chosen arbitrarily and determines the complexity of the model, and how it can flexibly adapt to the data
- By varying the number of parameter, one may improve the agreement of the fit curve to data at the cost of introducing a larger variance

Fit seen as machine learning

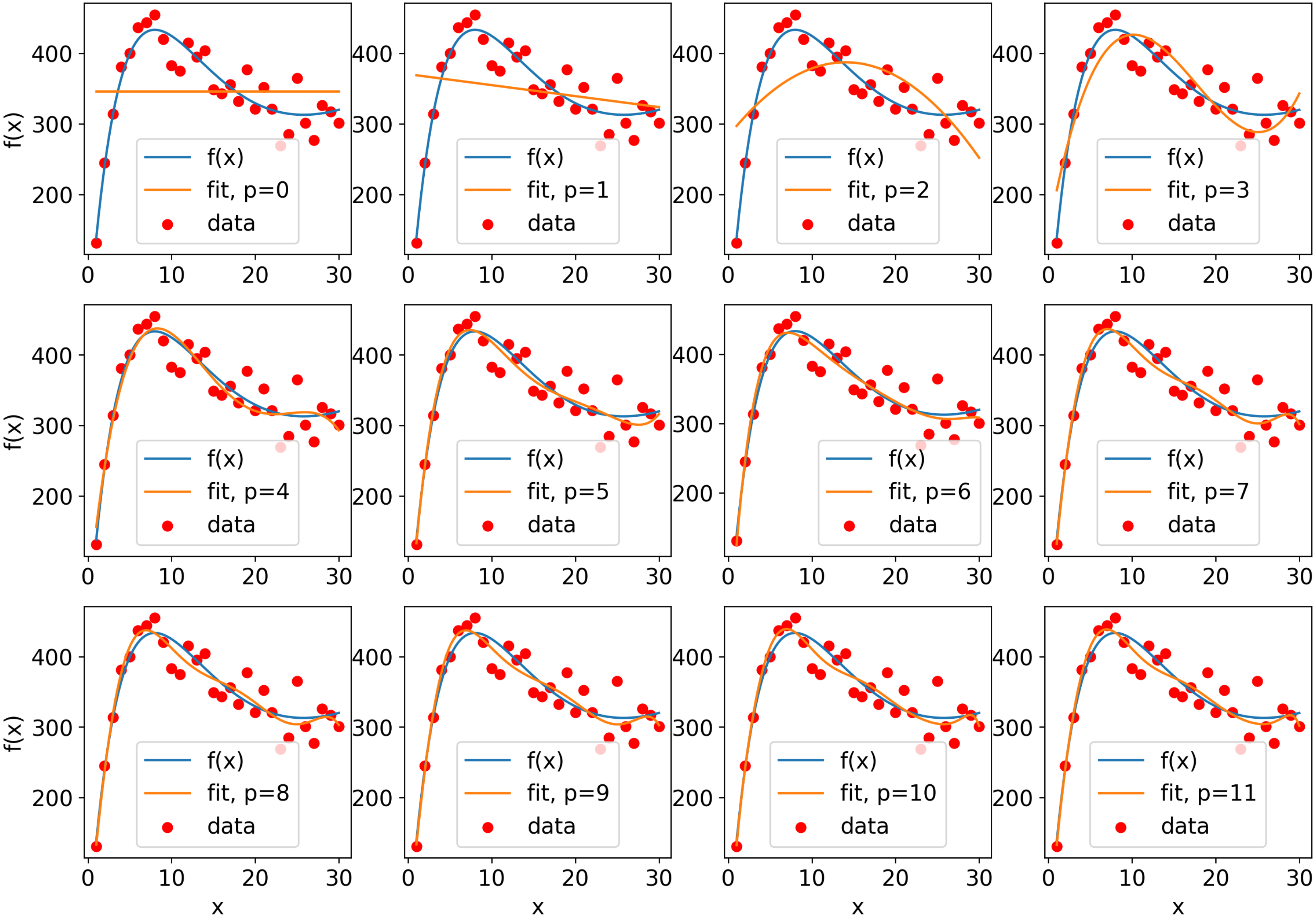


Bias variance decomposition

- Mean squared error:

$$\varepsilon^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^* - \hat{f}(x_i)\right)^2$$

- By increasing the number of parameters, the mean squared error decreases for the fit sample, but may increase for independently extracted samples
- The fit follows the fluctuations in data more closely than the original distribution, that is not known (overfitting)



f(x)
x
f(x)
fit, p=0
fit, p=1
fit, p=2
fit, p=3
fit, p=4
fit, p=5
fit, p=6
fit, p=7
fit, p=8
fit, p=9
fit, p=10
fit, p=11
data
200
300
400
0
10
20
30

Bias variance decomposition

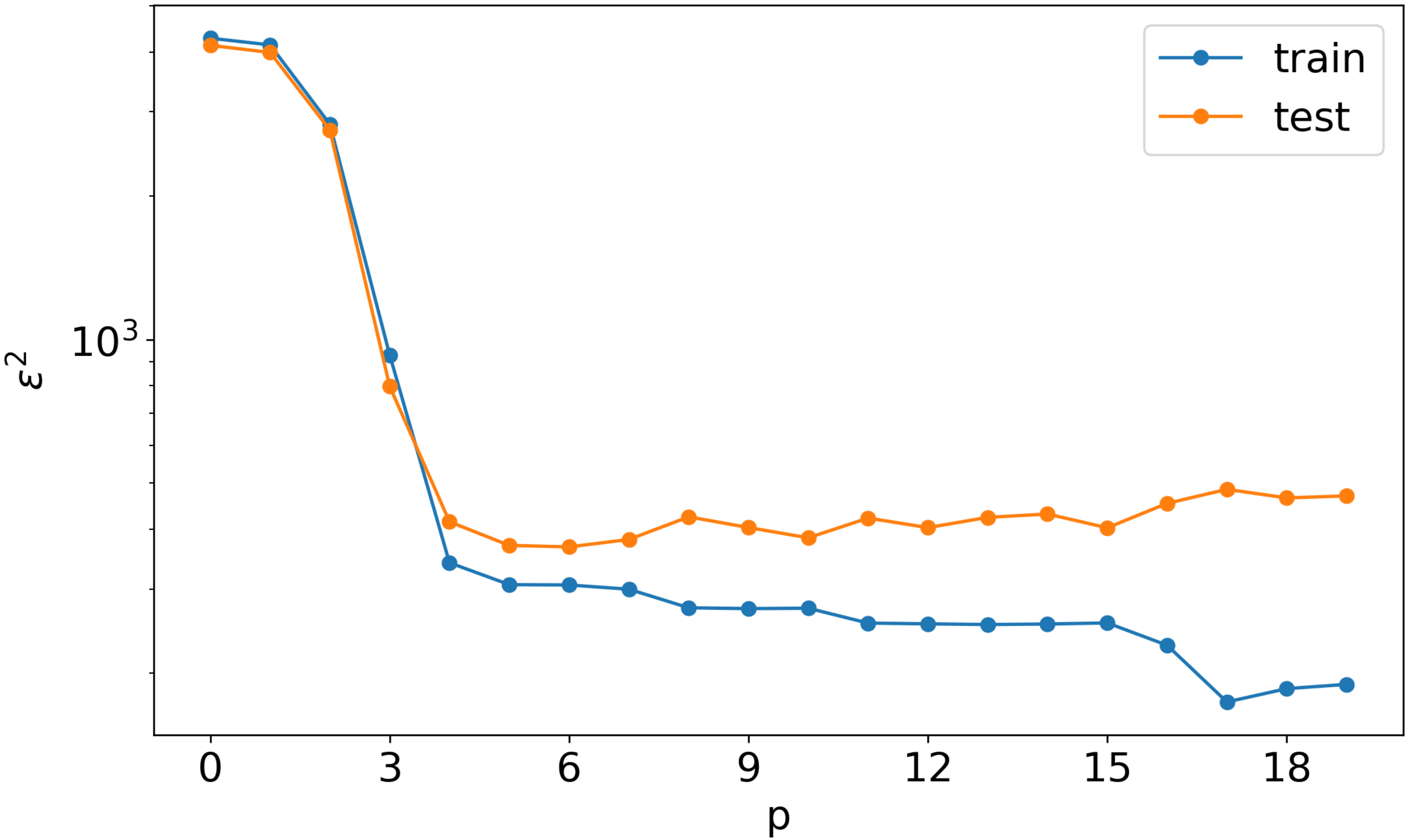
- Expected mean squared error:

$$\langle \varepsilon^2 \rangle = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i^* - \hat{f}(x_i)\right)^2\right]$$

- It is possible to demonstrate that the following decomposition holds, if we use the notation $\hat{\theta} = \hat{f}(x_i)$, $\theta^* = y_i^*$:

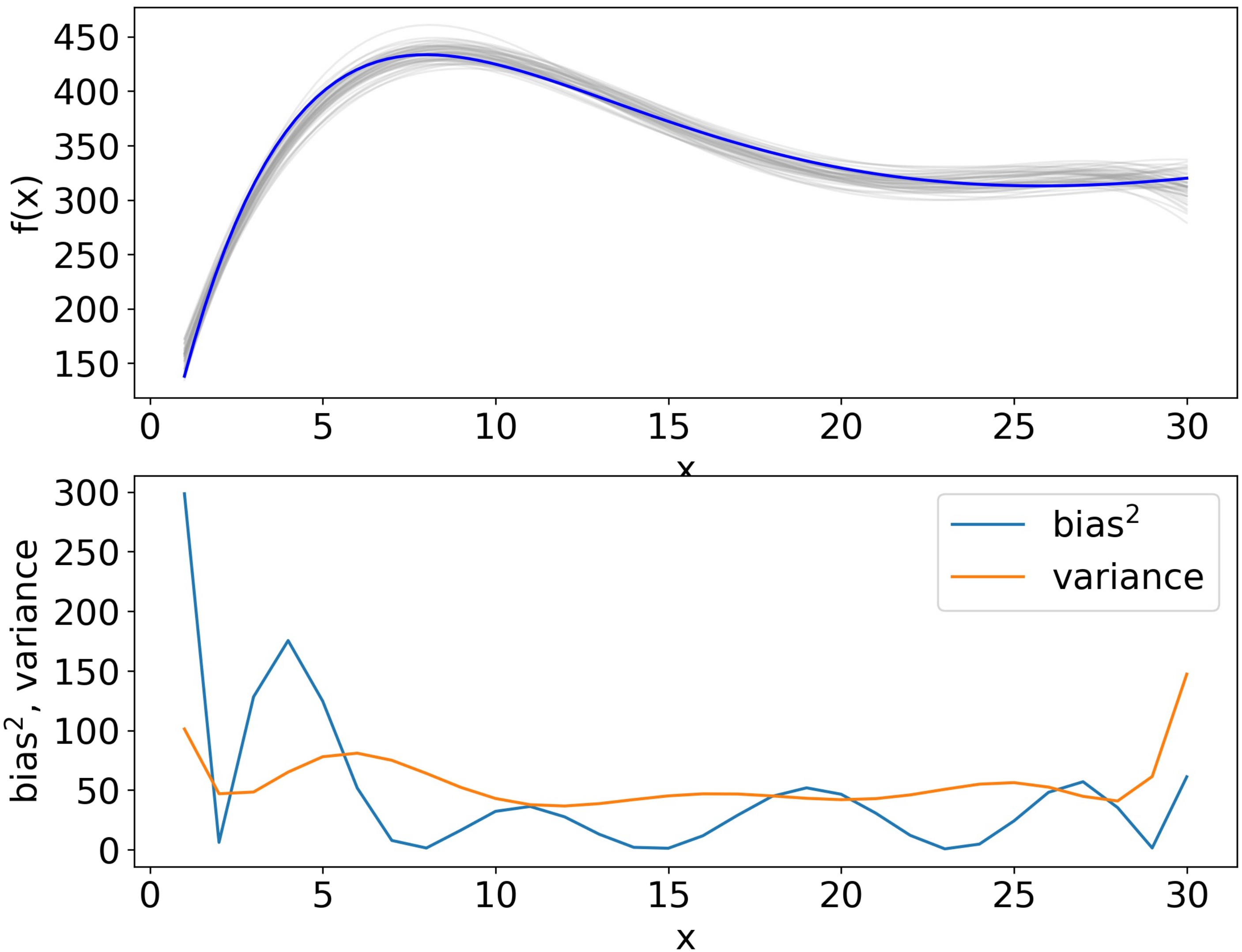
$$\langle \varepsilon^2 \rangle = \mathbb{V}\mathrm{ar}\left[\hat{\theta}\right] + \mathbb{B}\mathrm{ias}\left[\hat{\theta}\right]^2 + \mathbb{V}\mathrm{ar}[\theta^*]$$

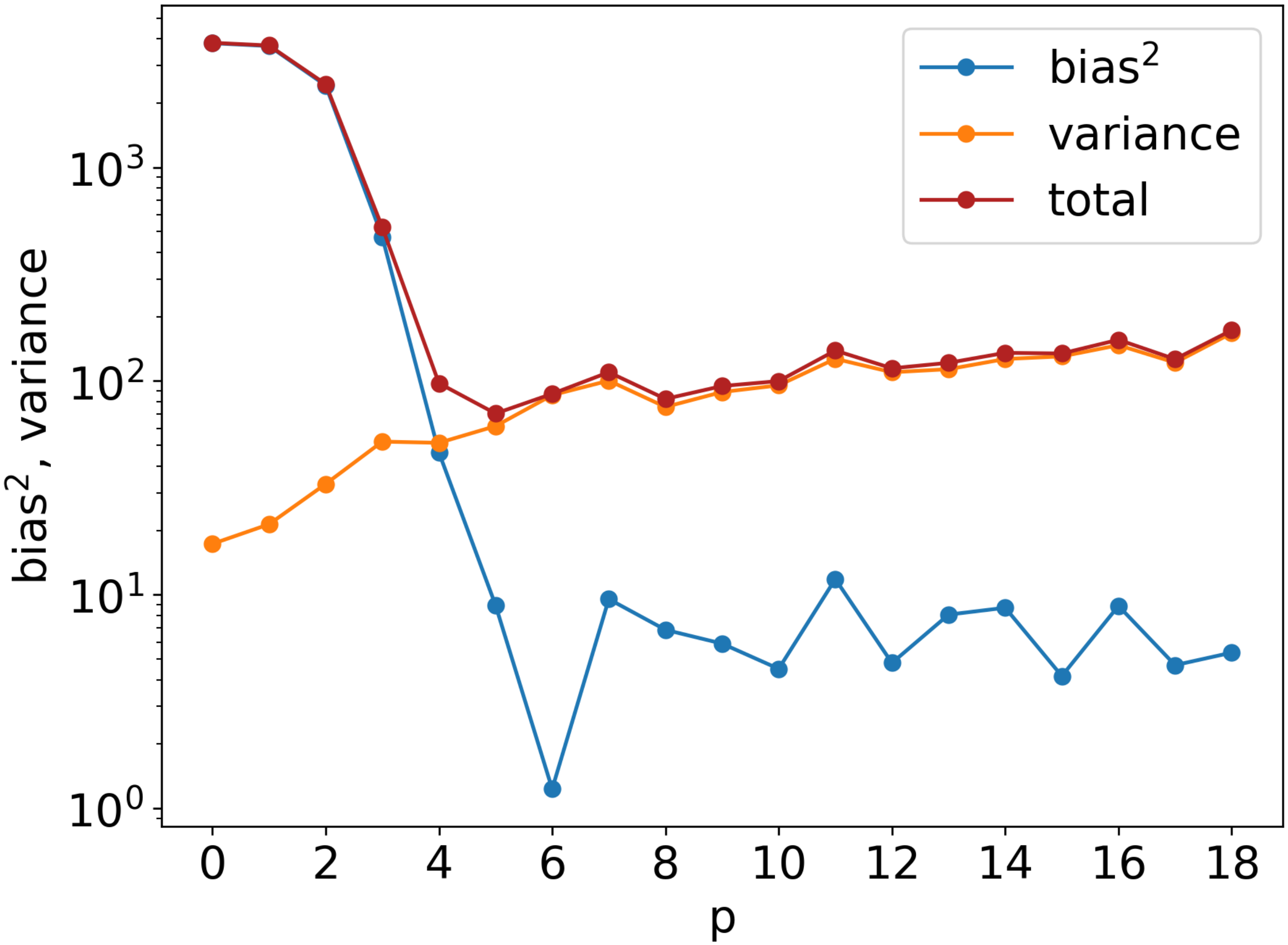
- While the intrinsic noise of the data $\mathbb{V}\mathrm{ar}[\theta^*]$ can't be improved with the fit, $\mathbb{V}\mathrm{ar}\left[\hat{\theta}\right]$ and $\mathbb{B}\mathrm{ias}\left[\hat{\theta}\right]$ depend on the fit model, i.e.: on the number of parameters
- One cannot achieve at the same time optimum variance and optimum bias, and a trade off must be chosen



train
test
10^3
ε^2
0
3
6
9
12
15
18
p

$p = 4$

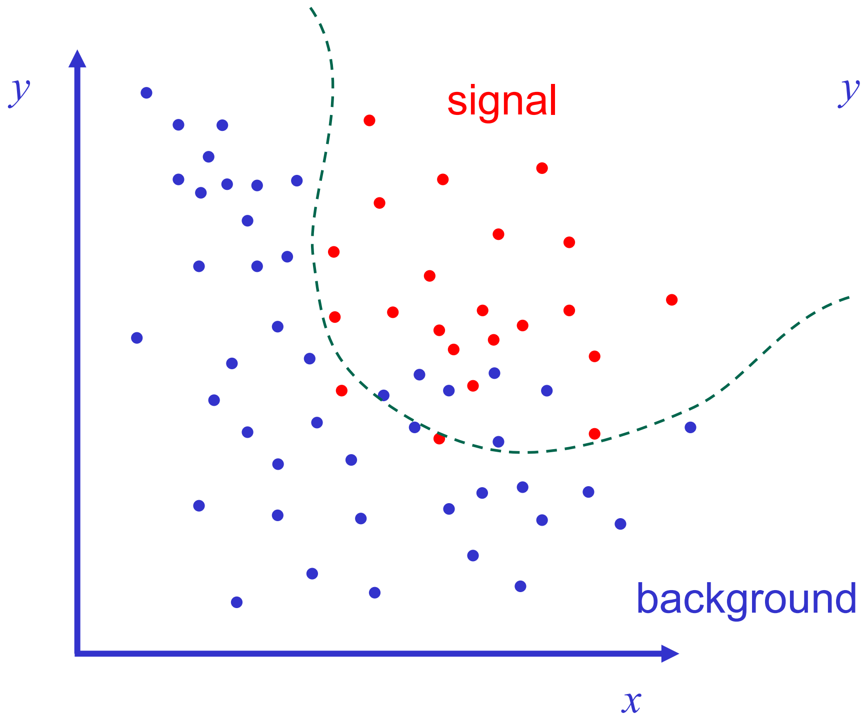

450
400
350
300
250
200
150
f(x)
0
5
10
15
20
25
30
x
300
250
200
150
100
50
0
bias², variance
0
5
10
15
20
25
30
x
bias²
variance

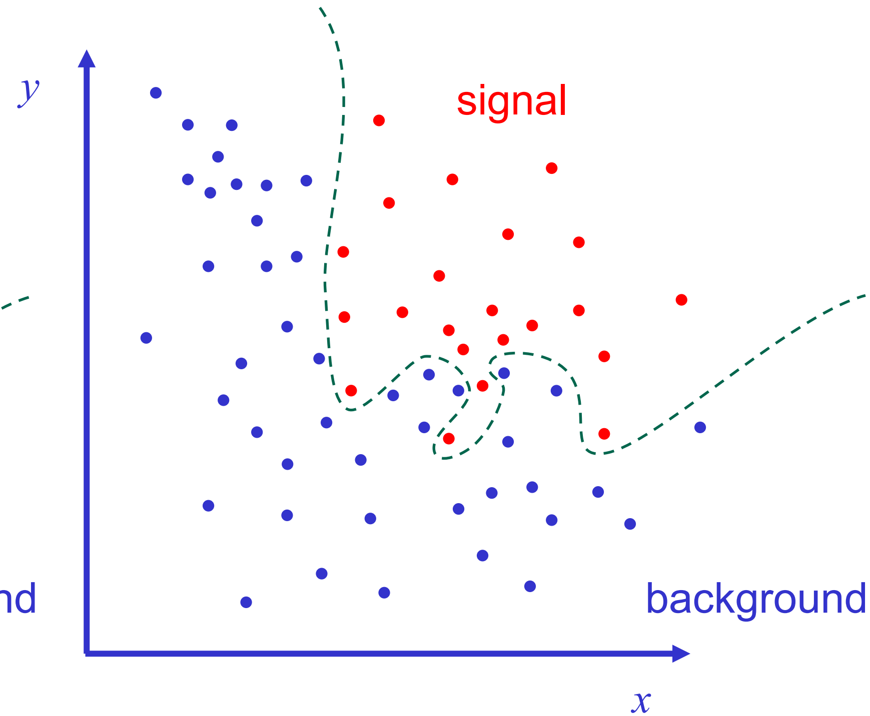


bias^2
variance
total
10^0
10^1
10^2
10^3
bias^2, variance
0
2
4
6
8
10
12
14
16
18
p

Overtraining

- Algorithms may learn too much from the training sample, exploiting features that are only due to random fluctuations
- Check for overtraining comparing the discriminator's distributions for the training sample and for an independent test sample (consistent distributions = no overtraining)





Decision Trees

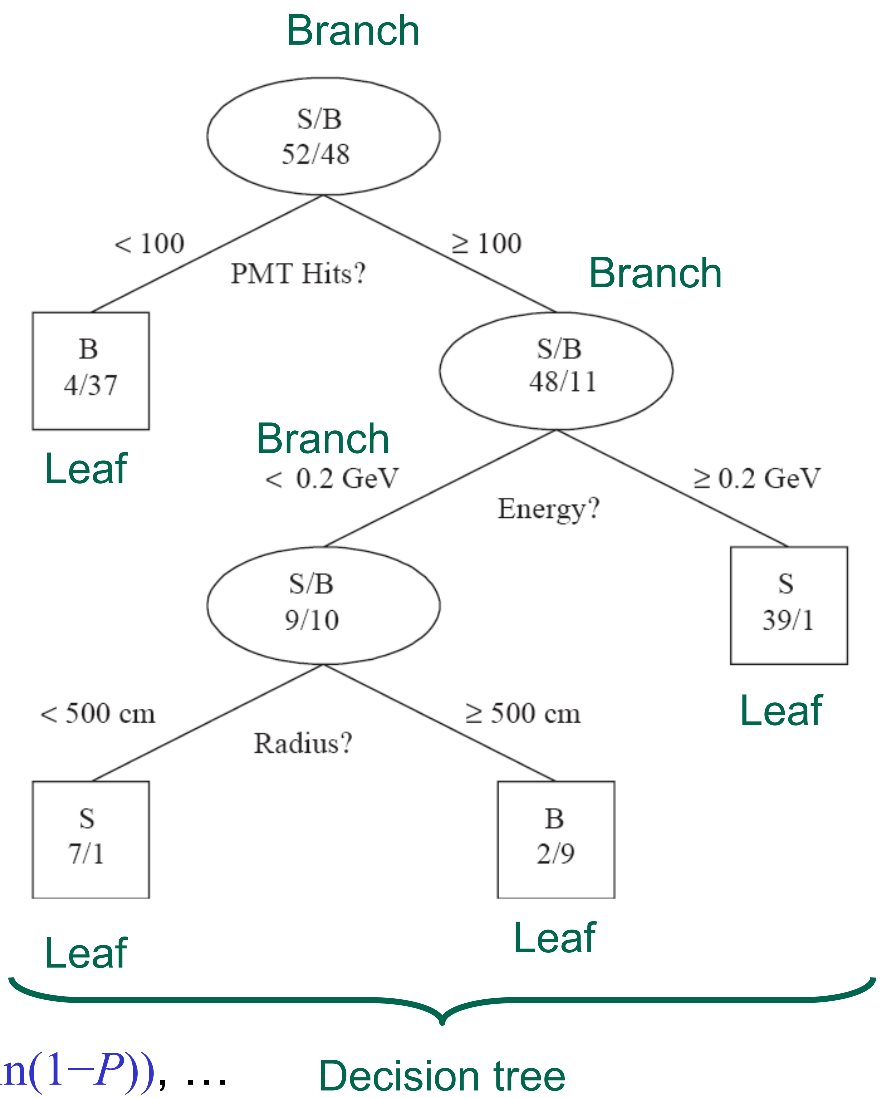
- Cuts are applied sequentially
- Each cut splits the sample into nodes
  - Nodes where signal or background is largely dominant are classified as leafs
    - Alternatively: stop splitting if too few events per note, total number of nodes too high, …
  - One branch = one sequence of cuts
- Cuts can be optimized to achieve the best split level: maximize for each node the gain of Gini index:

$$G = P(1-P)$$

P = node purity

$G = 0$ for nodes with all S or all B events

Gain = $N_{\text{parent}} G_{\text{parent}} - N_{\text{left ch.}} G_{\text{left ch.}} - N_{\text{left ch.}} G_{\text{left ch.}}$



- Alternative metrics exist
  E.g.: cross entropy $= -(P \ln P + (1-P)\ln(1-P))$, …

From trees to forests

- A single tree tends to adapt very closely to data, and performances are usually not very satisfactory
- It may easily provide overtraining if the tree is too deep
- It means: it has large variance, but in general low bias
- Combining many independent trees may reduce the variance
- Random forests are algorithms that combine many different trees and provide as output the combination of individual results

Boosted Decision Trees

- Combine a large number of decision trees (forest) using different weights. Usually: $\mathcal{O}(1000)$ trees used
- More performant and stable than a single optimized tree
- Boosting achieved by iteratively reweighting training sample according to classifier output in previous iteration

1. Reweight events using previous iteration's classifier result
2. Build and optimize a new tree with reweighted events
3. Give a score to each tree
4. The final BDT classifier result is the weighted average over all trees, using the given scores as weights:

$$y(\vec{x}) = \sum_{k=1}^{N_c} w_i C^{(i)}(\vec{x})$$

Adaptive boosting

- Misclassified events are reweight according to the fraction of classification errors of the previous tree:

$$\frac{1-f}{f}\,, f = \frac{N_{\mathrm{misclassified}}}{N_{\mathrm{tot}}}$$

- Also use (log of) misclassification fraction as score for each tree:

$$y(\vec{x}) = \sum_{k=1}^{N_c} \log\left(\frac{1-f^{(i)}}{f^{(i)}}\right) C^{(i)}(\vec{x})$$

- Next iteration will better perform on events poorly classified in the previous iteration
- Further variations and more algorithms available

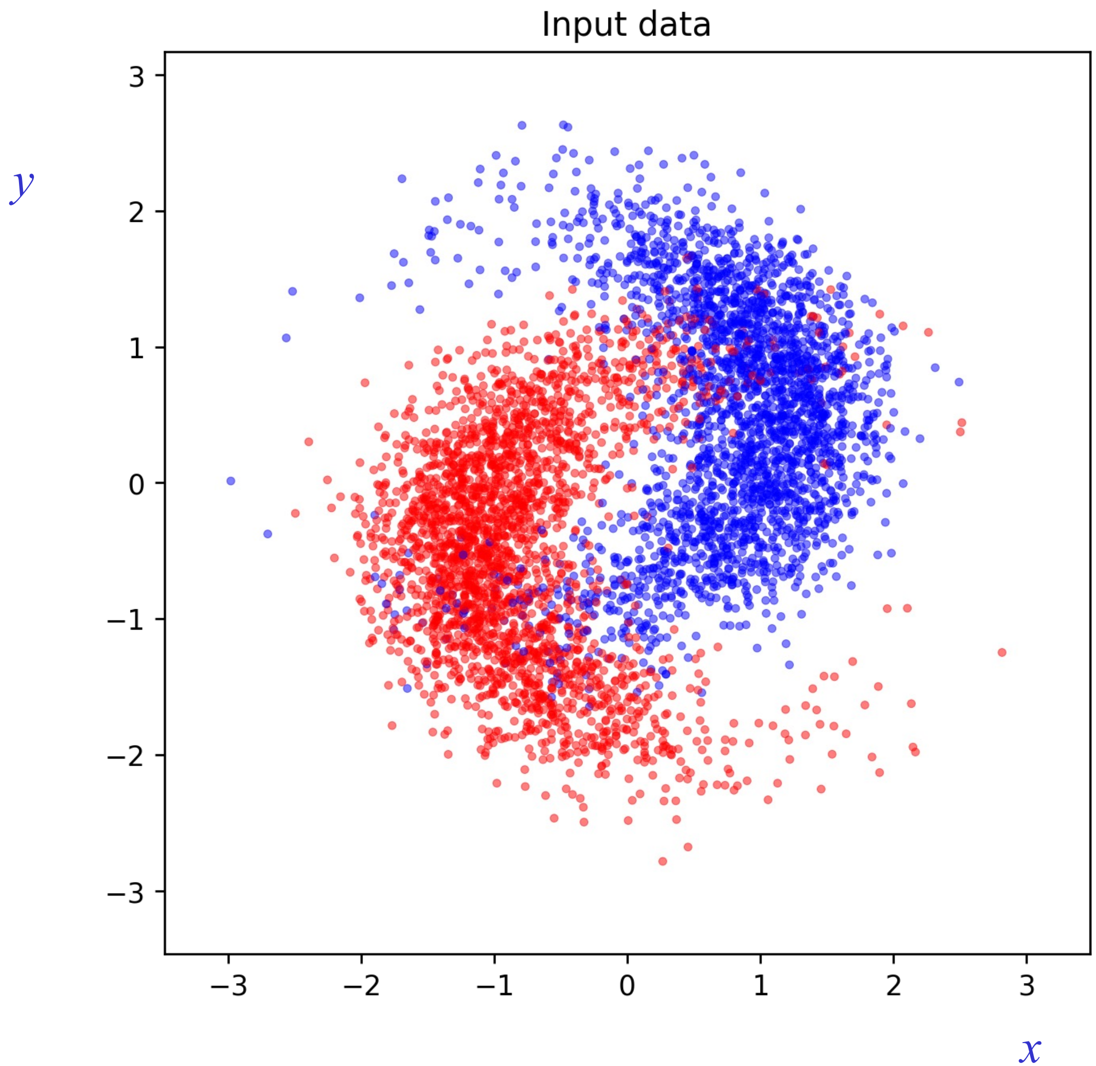
Extreme Gradient Boost

- Introducing a loss function for decision trees allows to use stochastic gradient descent algorithms
- Moreover, it also allows to use decision trees for regression, not only for classification
- The output of a tree can be taken as the index corresponding to one of the possible leaves, for a classification problem: $i = q(\vec{x})$
- For regression problems, we can assume that the output of a tree is a function of the index: $g = w_{q(\vec{x})}$ so that for many trees the output is the combination of their outputs:

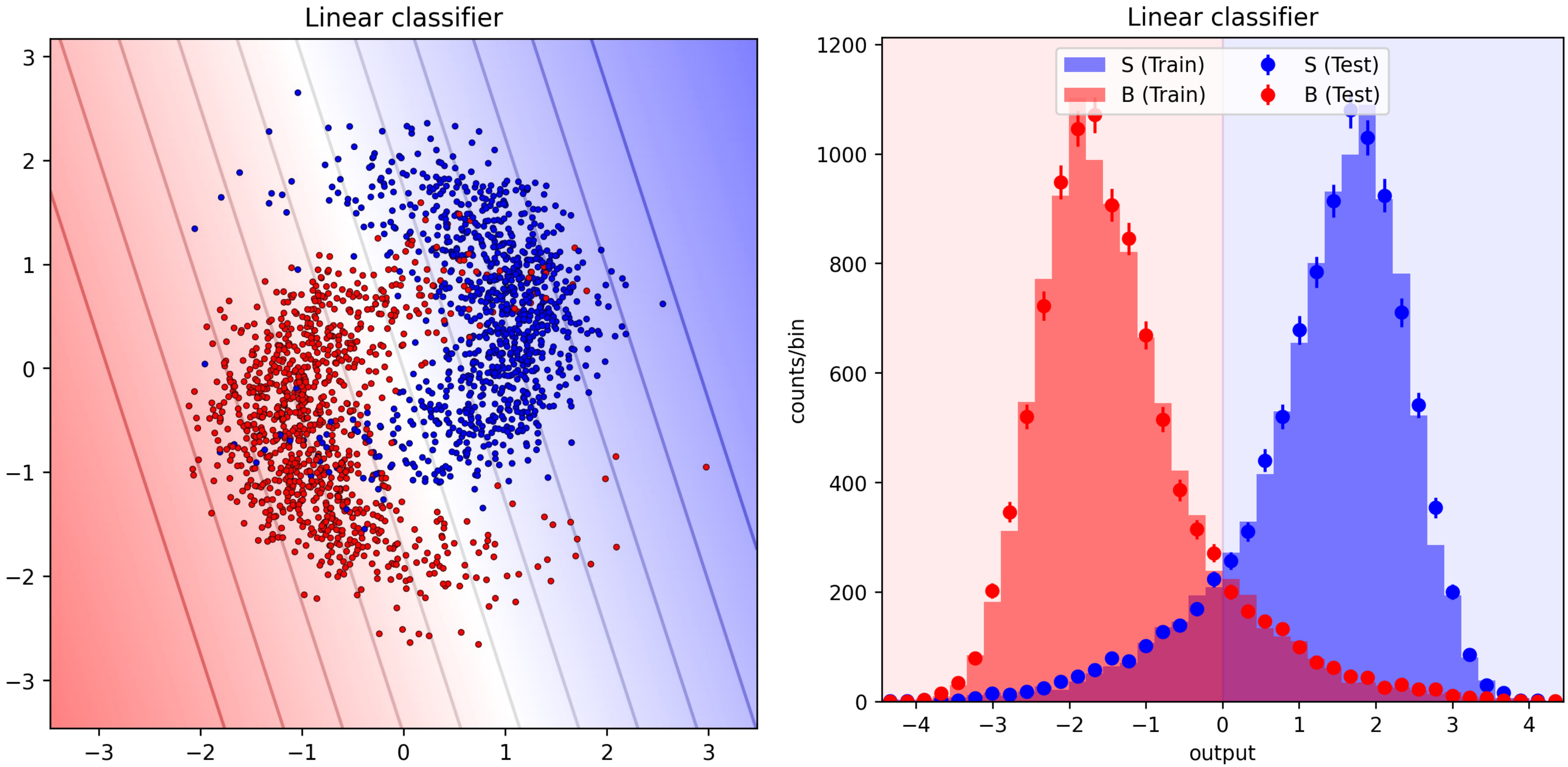
$$\hat{y}(\vec{x}) = \sum_{j=1}^{M} g^{(j)}(\vec{x}) = \sum_{j=1}^{M} w_{q(\vec{x})}^{(j)}$$

- The loss function may be the sum of

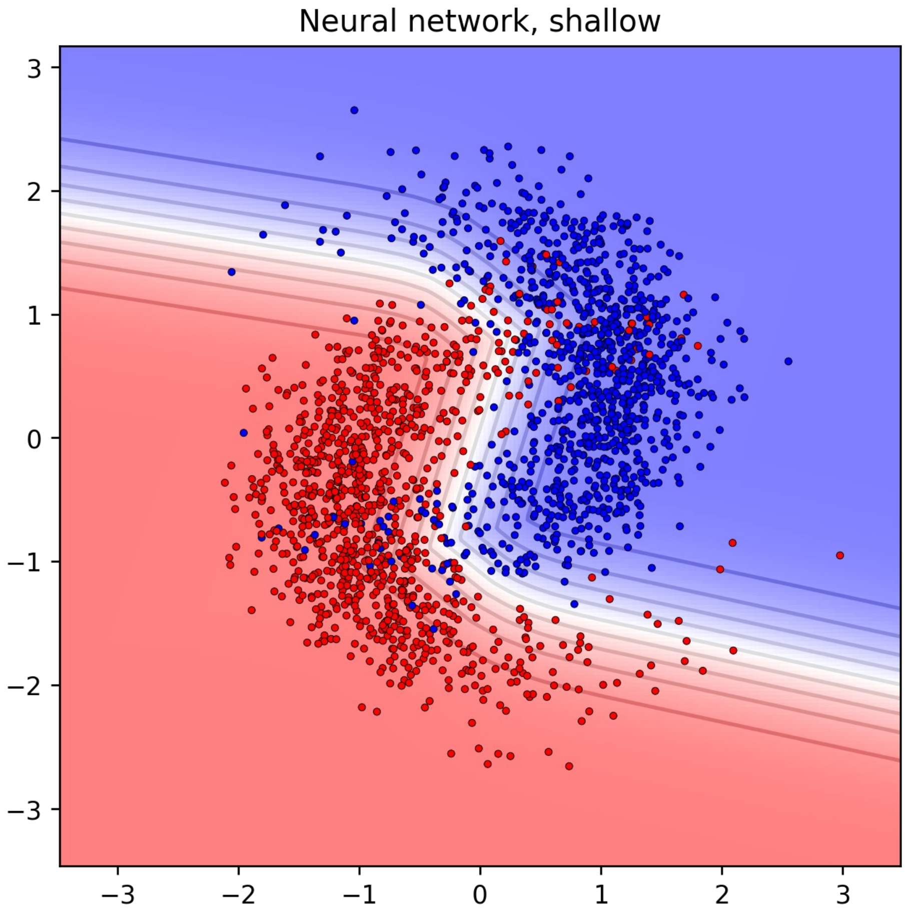
Strongly nonlinear example



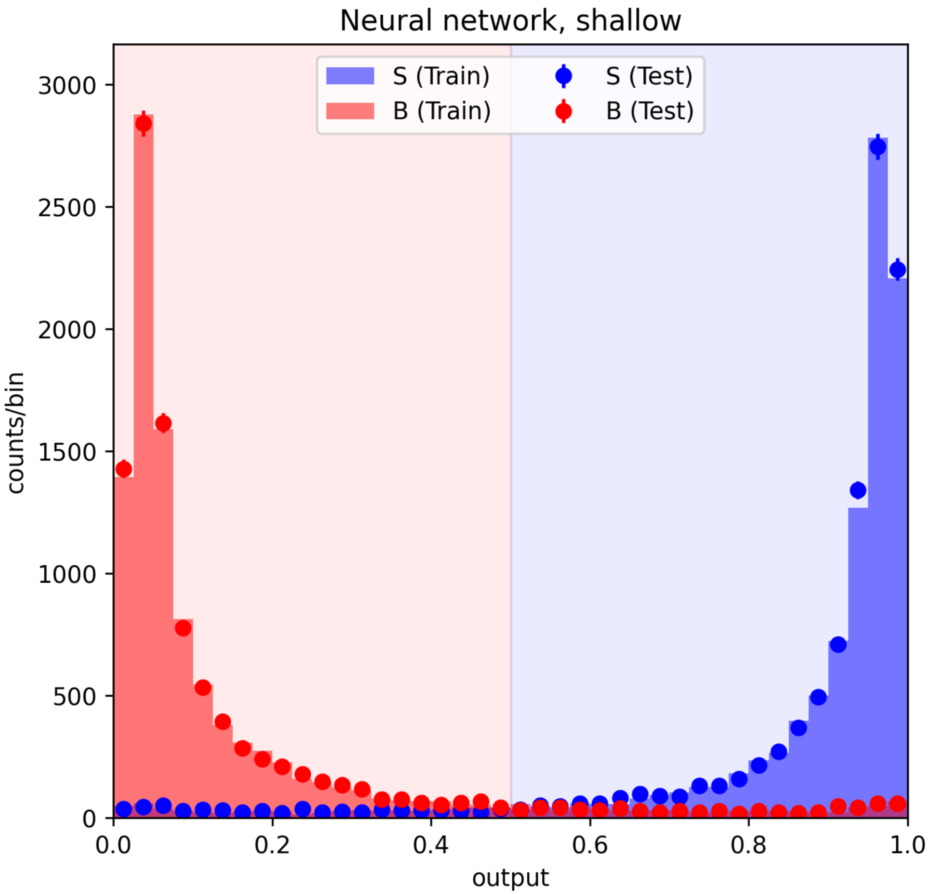
Linear classifier



NN, shallow

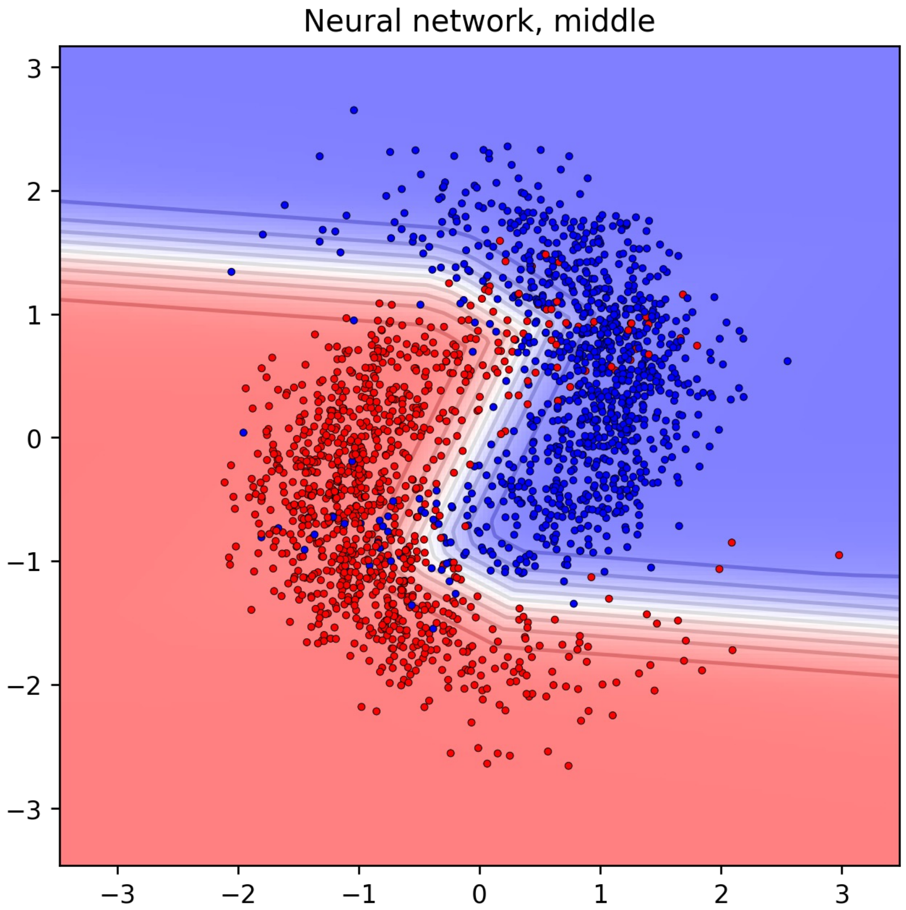


Neural network, shallow

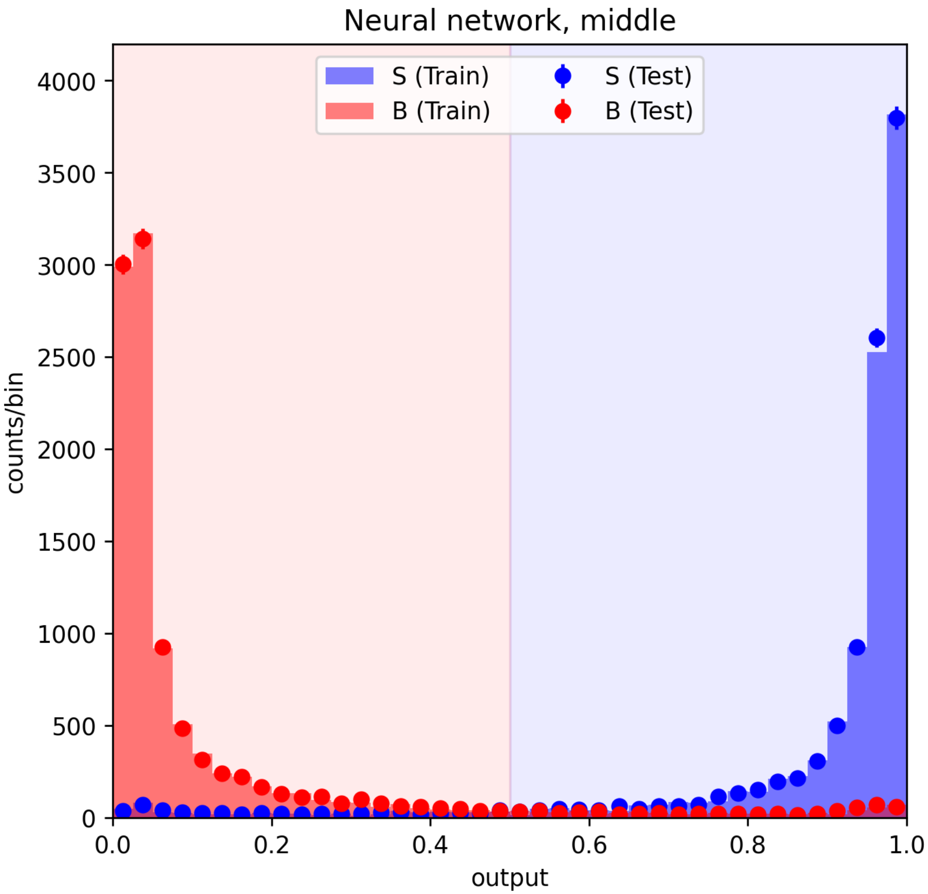


Neural network, shallow
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
output

NN, middle

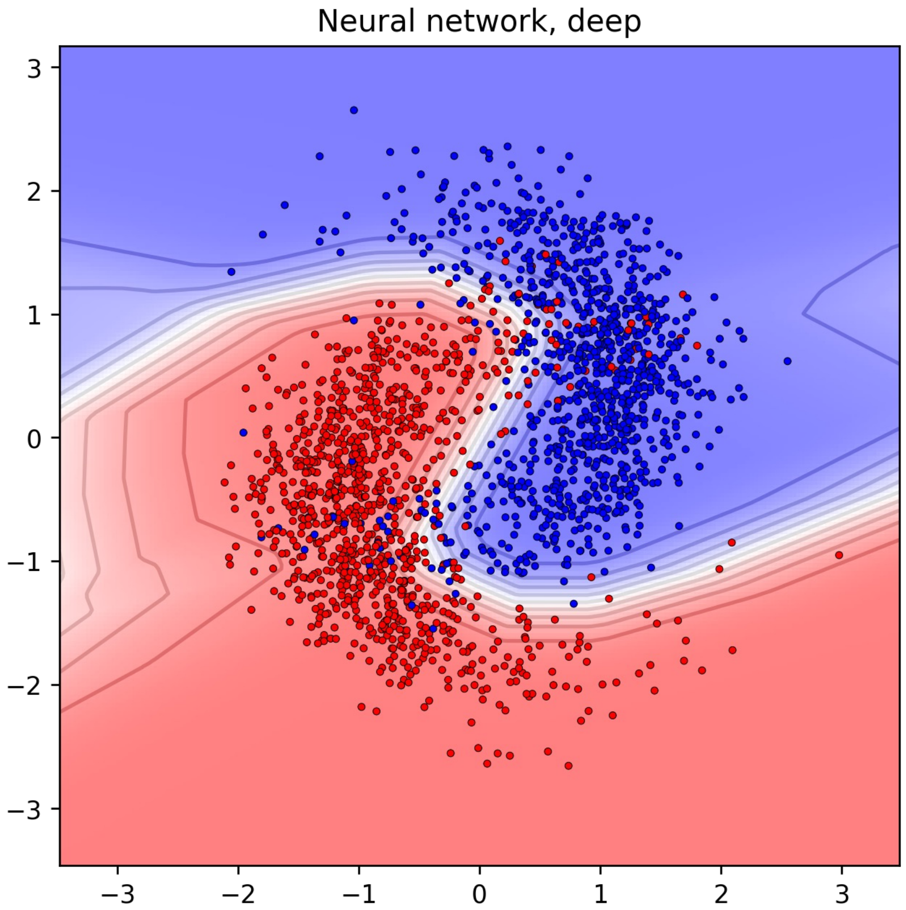


Neural network, middle
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

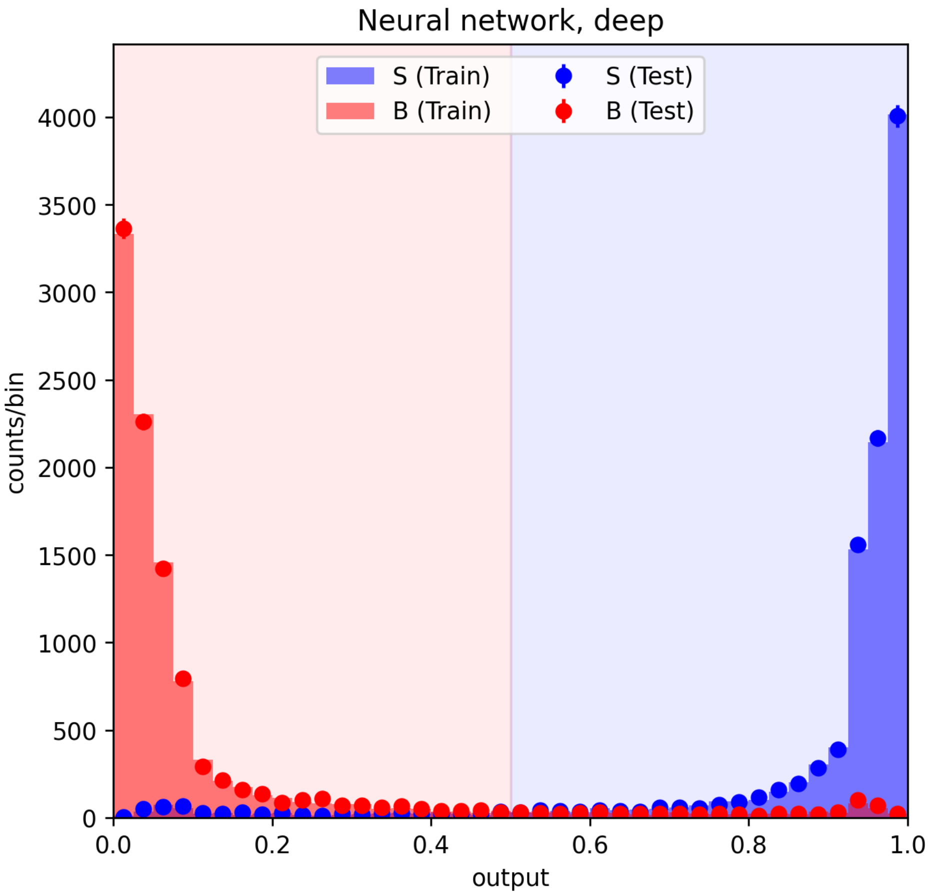


Neural network, middle
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
4000
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
output

NN, deep

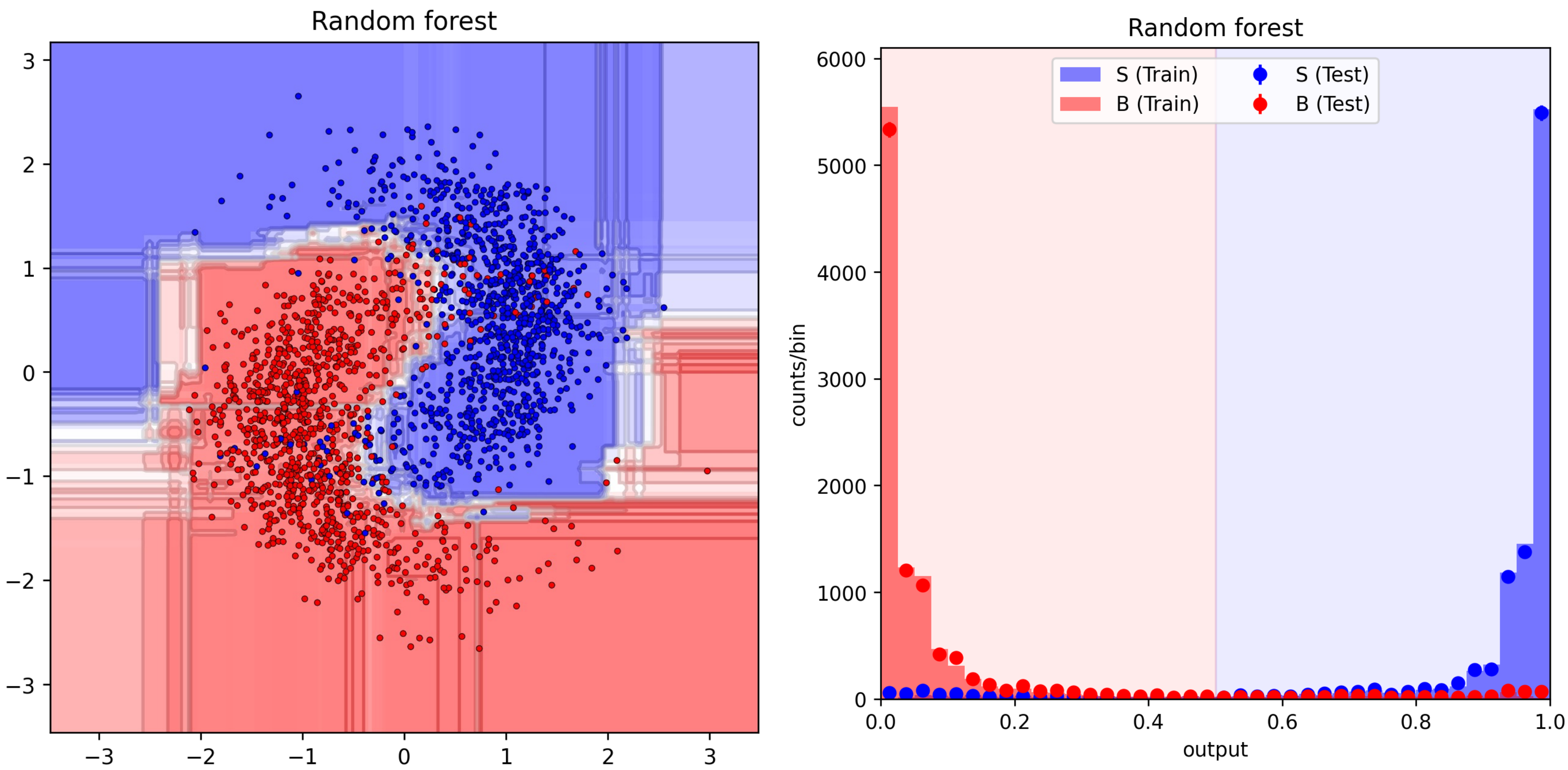


Neural network, deep
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

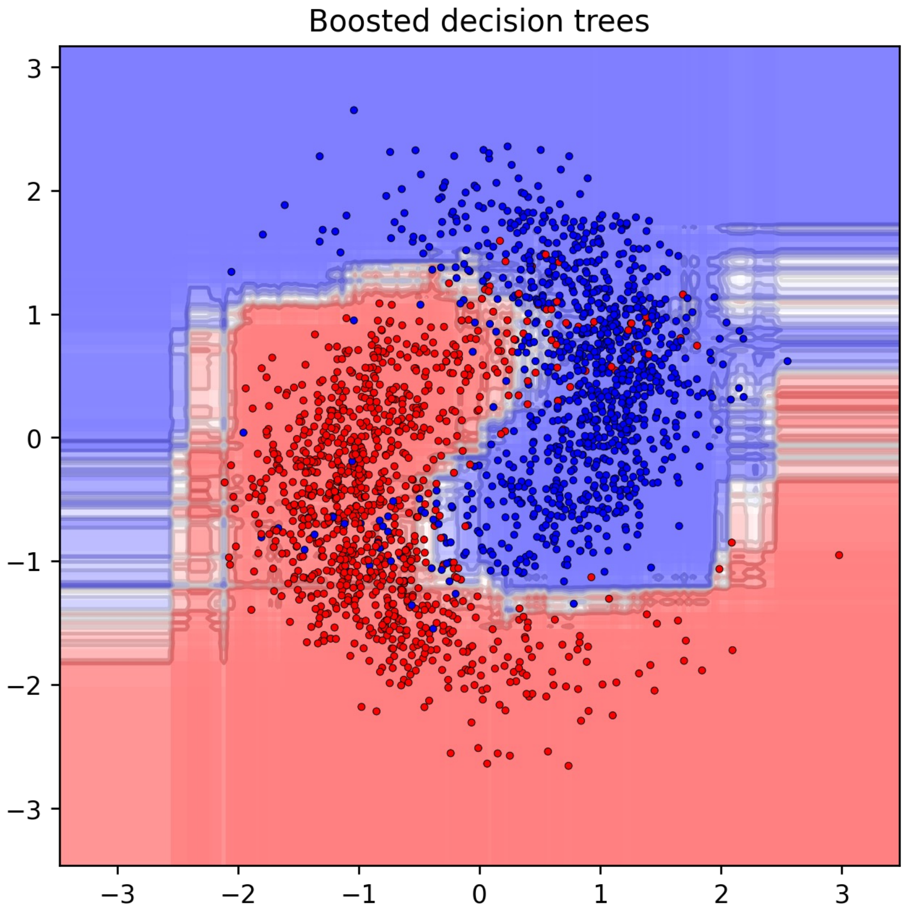


Neural network, deep
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
output
4000
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0

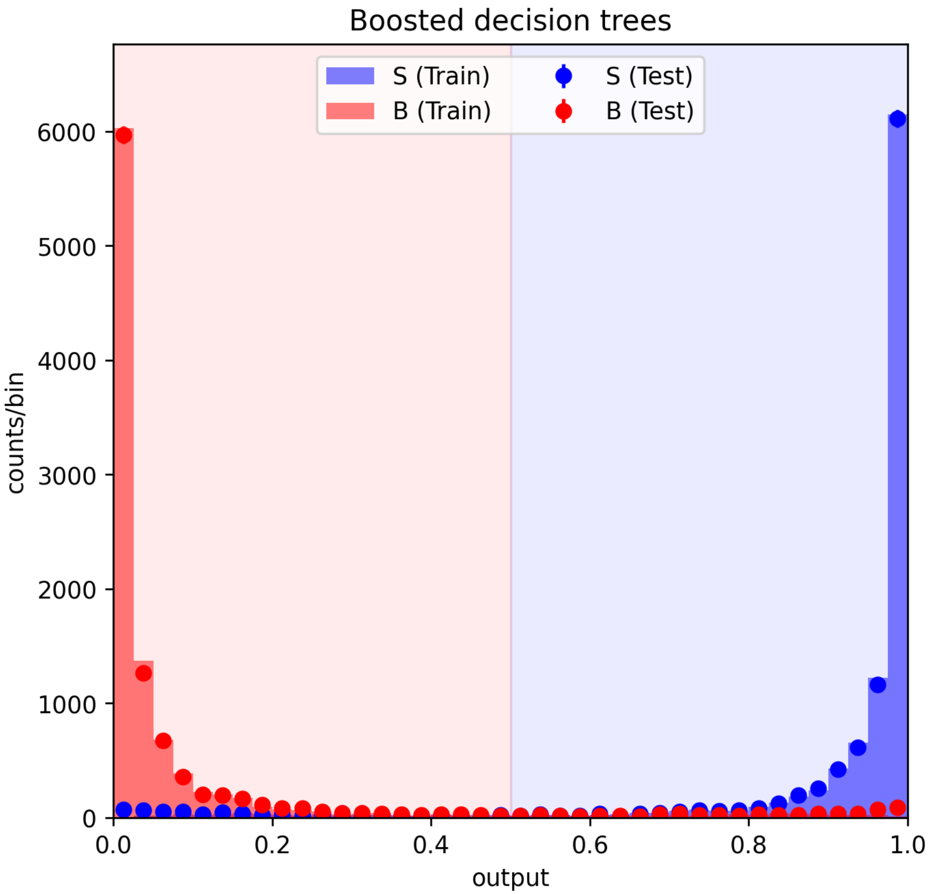
Random forest



XGBoost

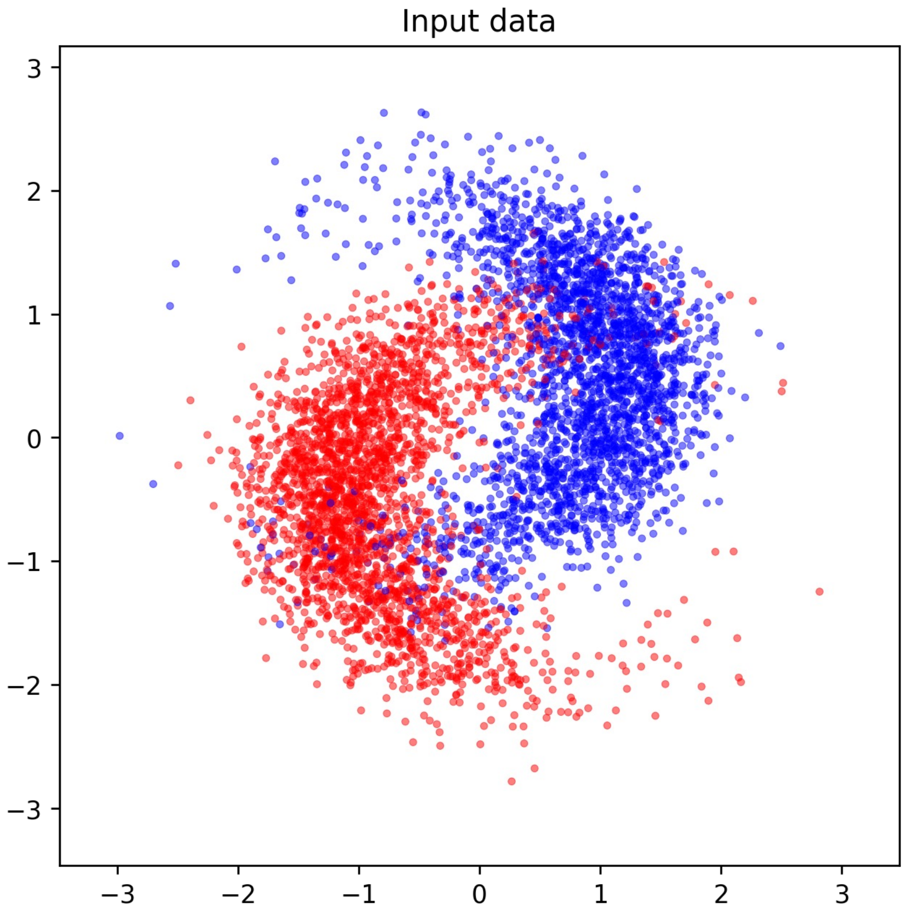


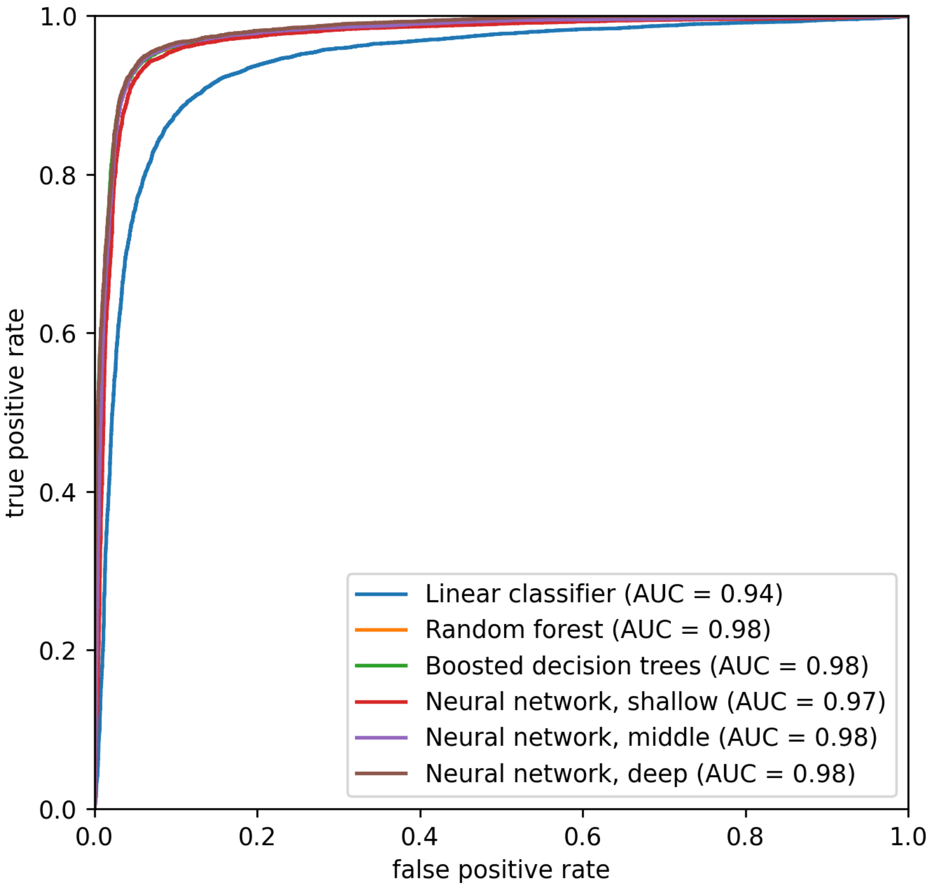
Boosted decision trees
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3



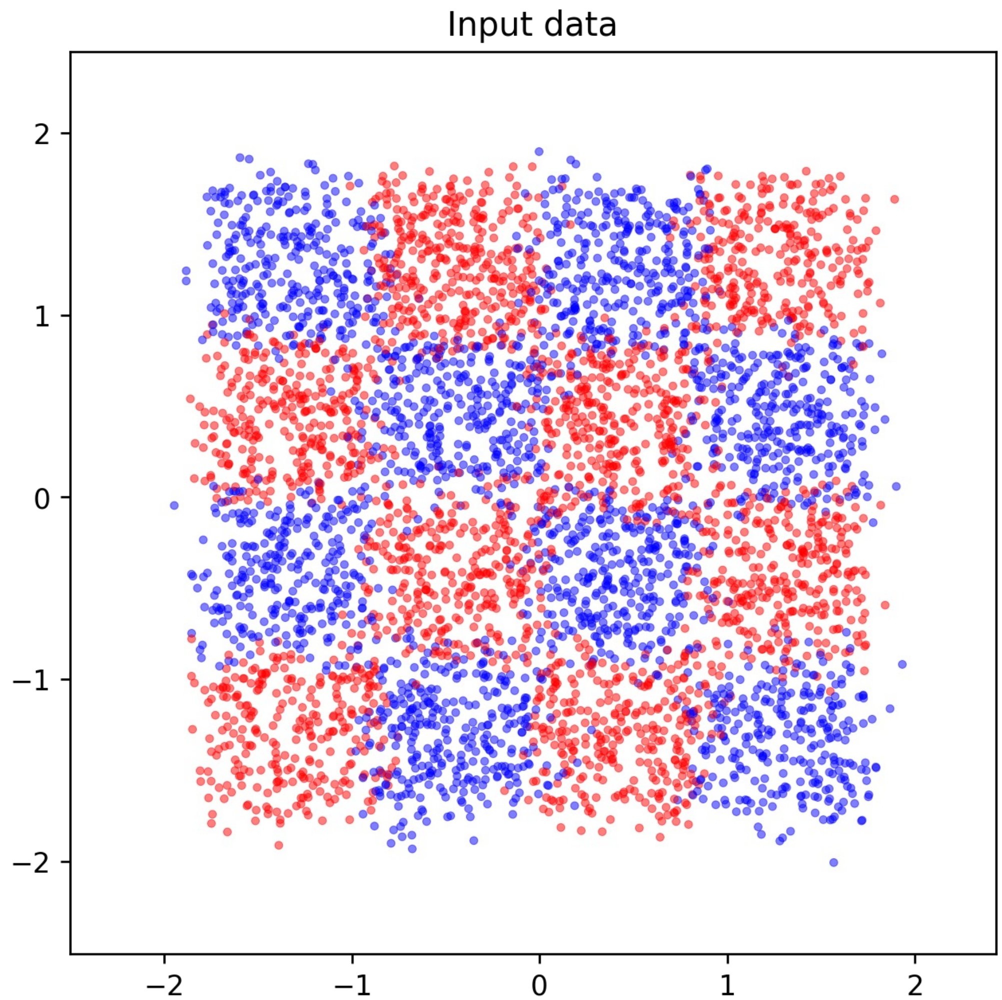
Boosted decision trees
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
output

Linear classifier





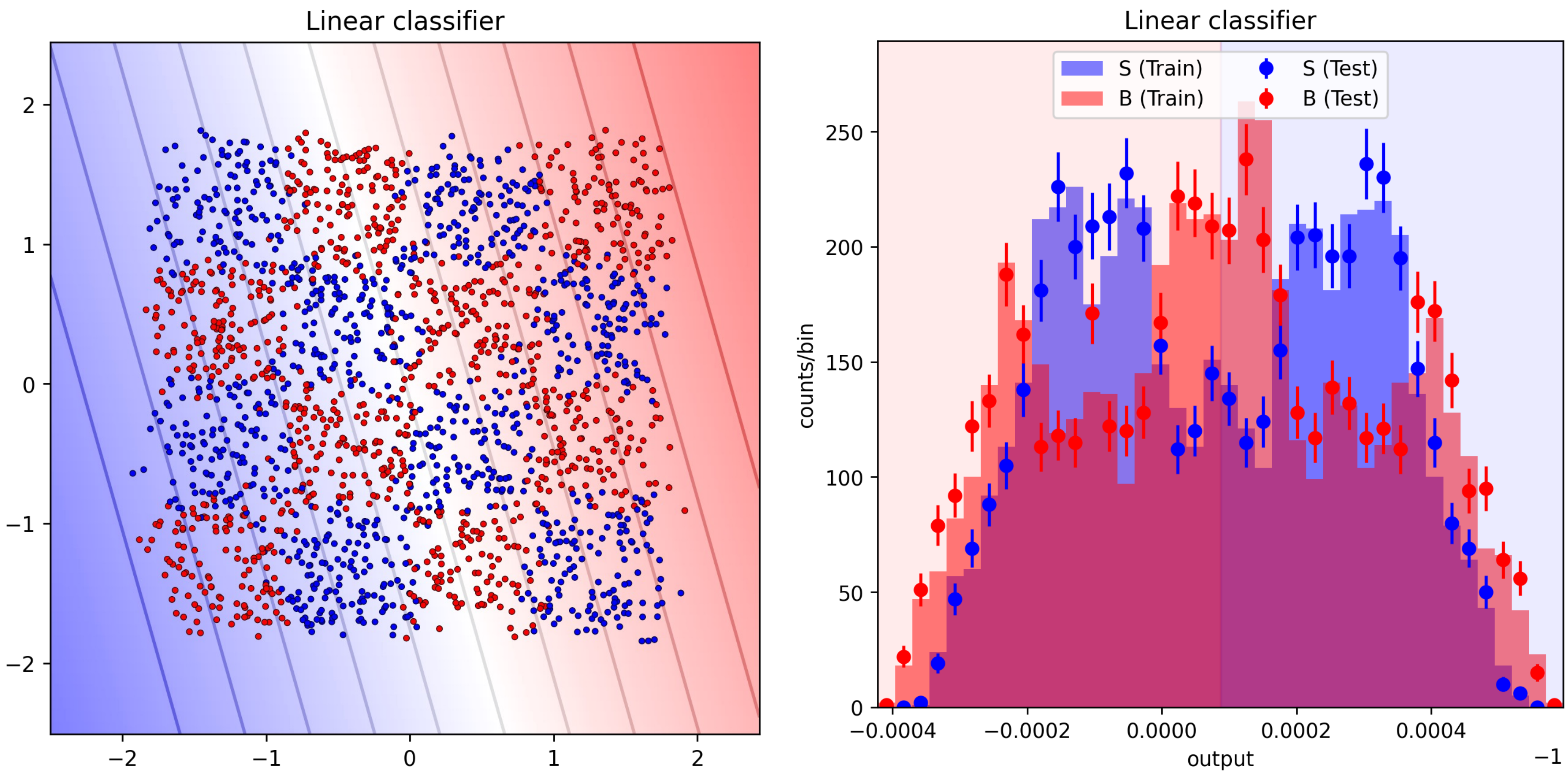
Chess shape



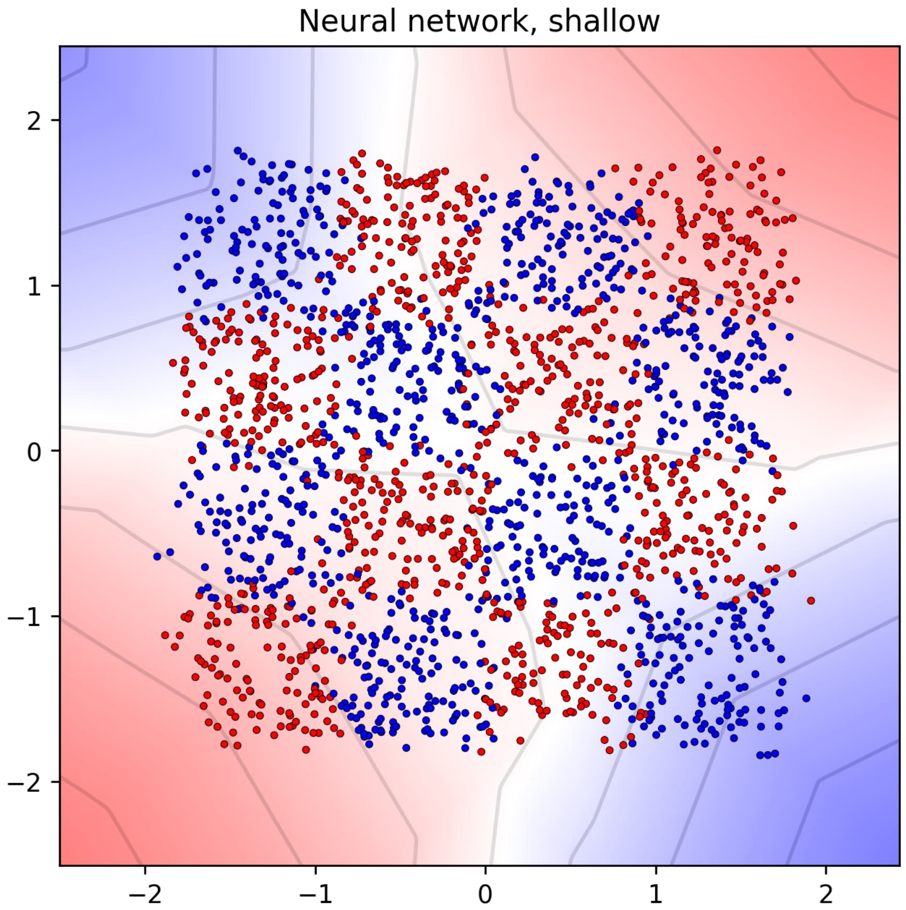
y

x

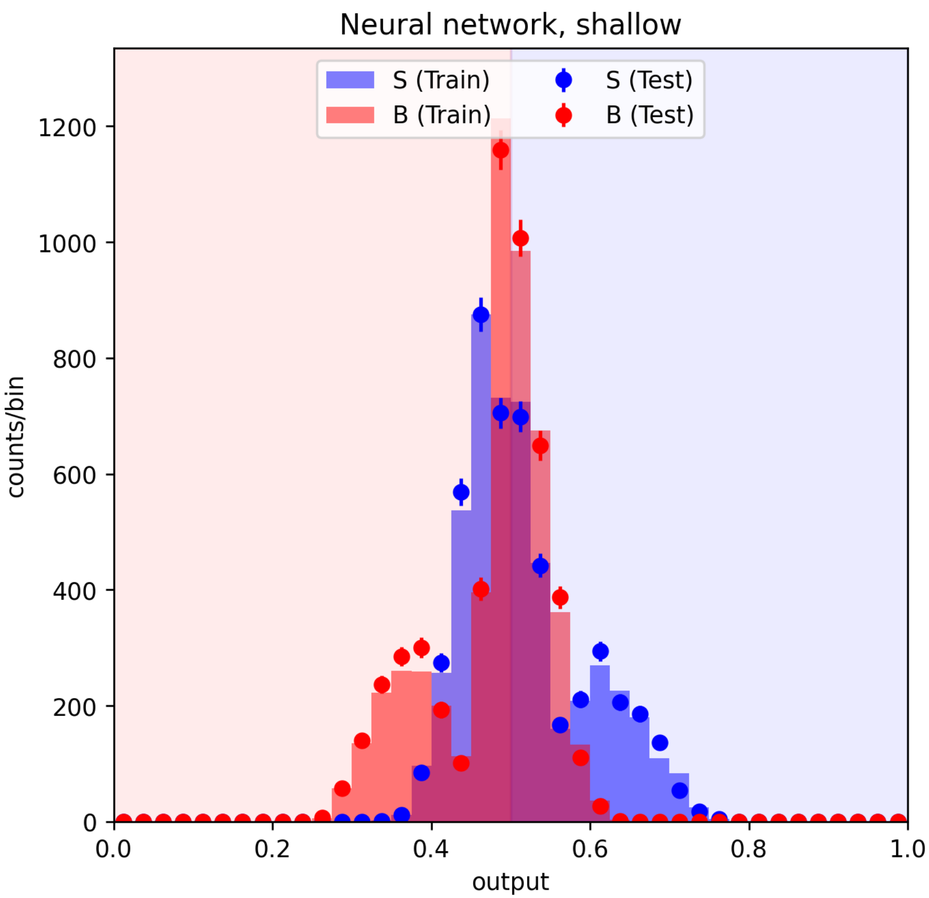
Linear classifier



NN, shallow

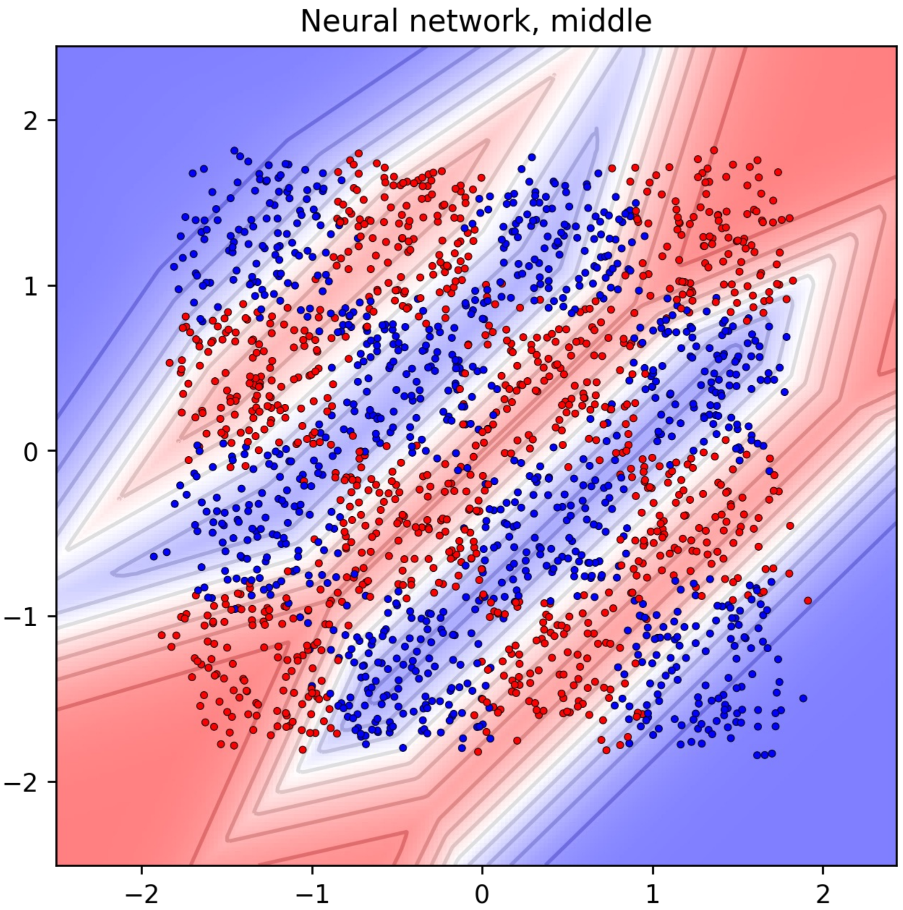


Neural network, shallow
2
1
0
−1
−2
−2
−1
0
1
2

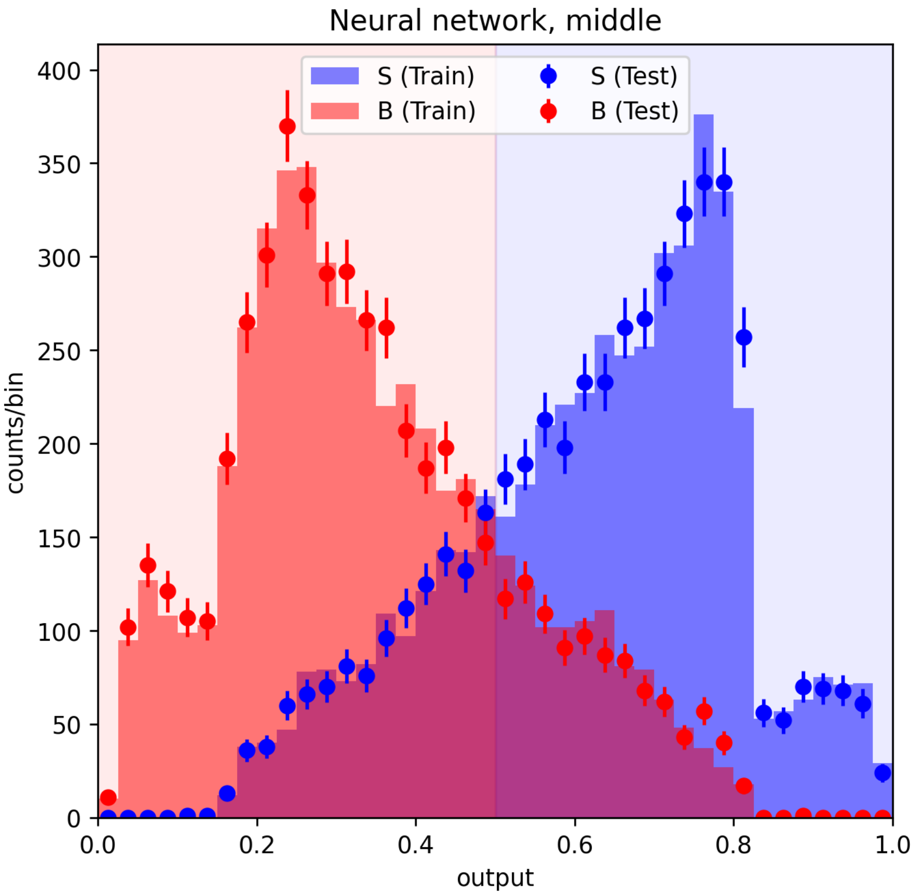


Neural network, shallow
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
output

NN, middle

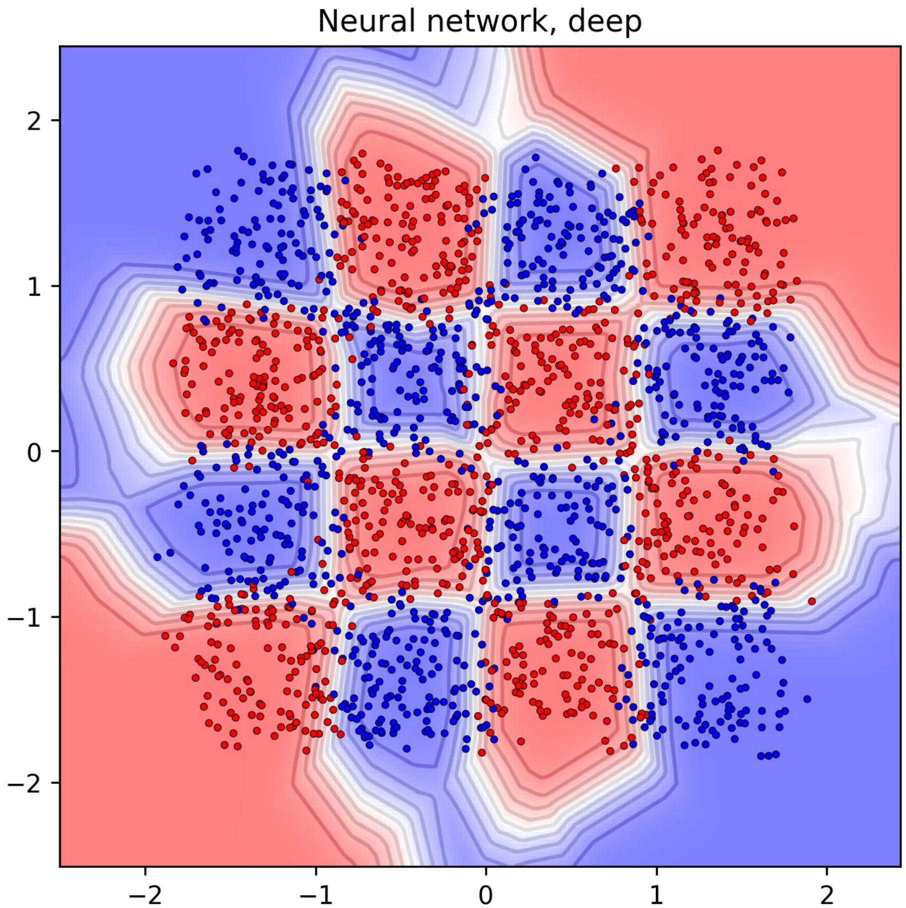


Neural network, middle
2
1
0
−1
−2
−2
−1
0
1
2

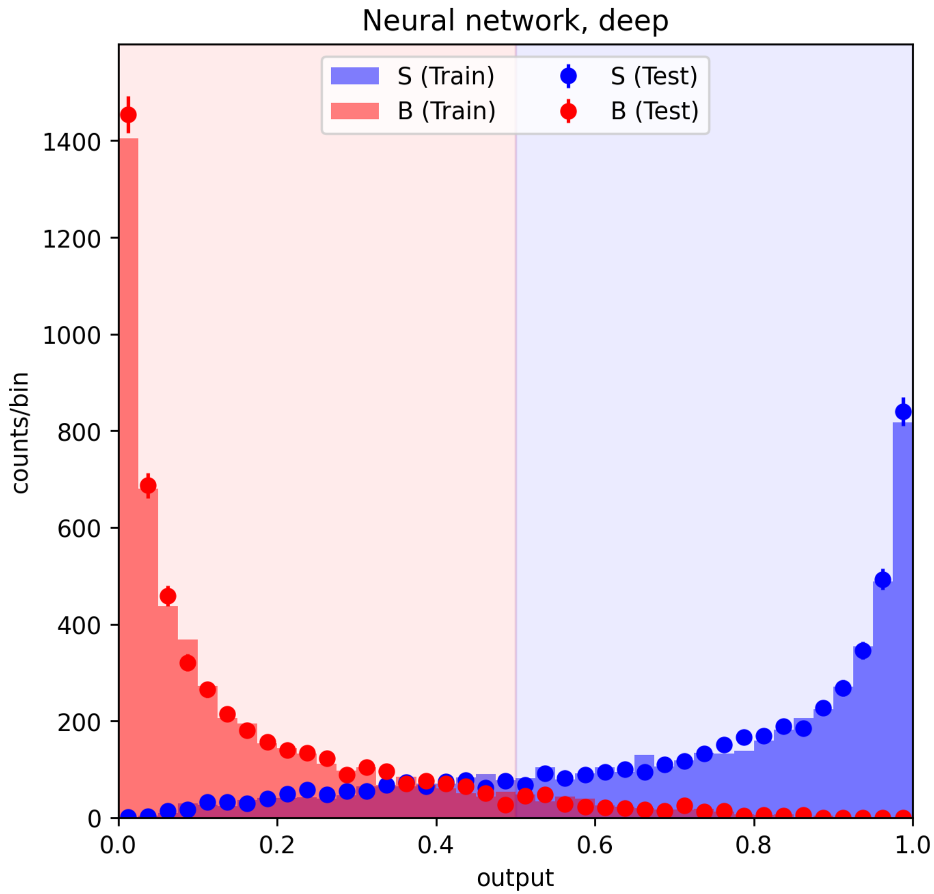


Neural network, middle
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
output
0
50
100
150
200
250
300
350
400
0.0
0.2
0.4
0.6
0.8
1.0

NN, deep

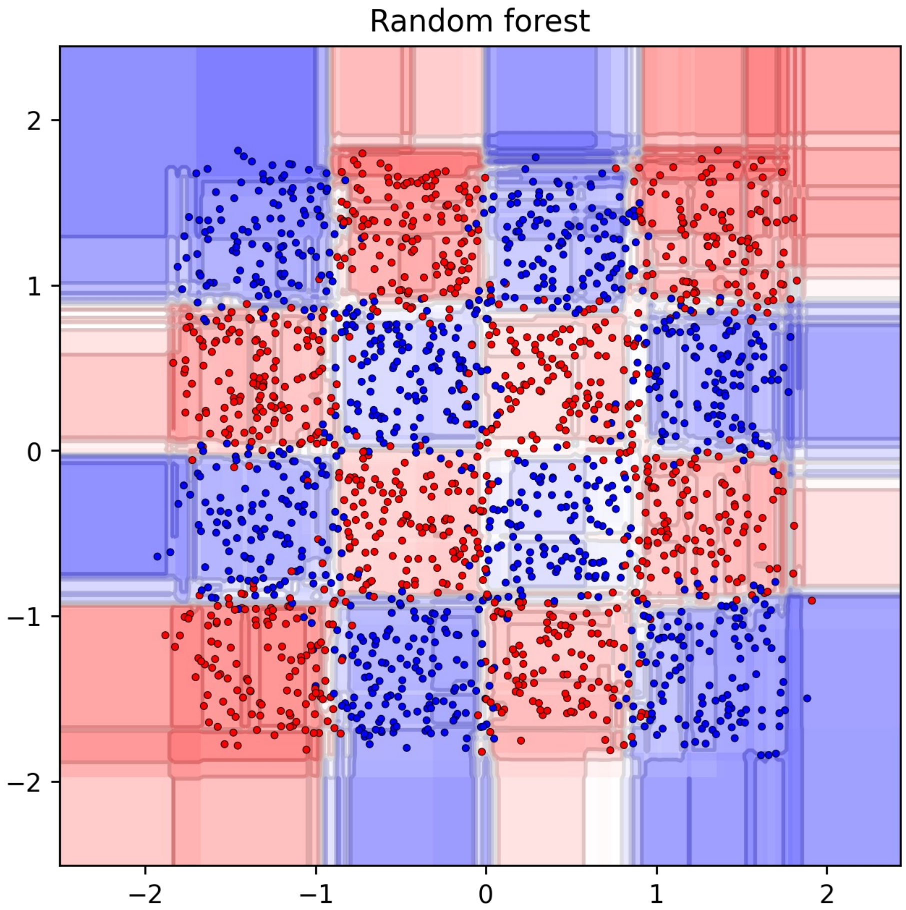


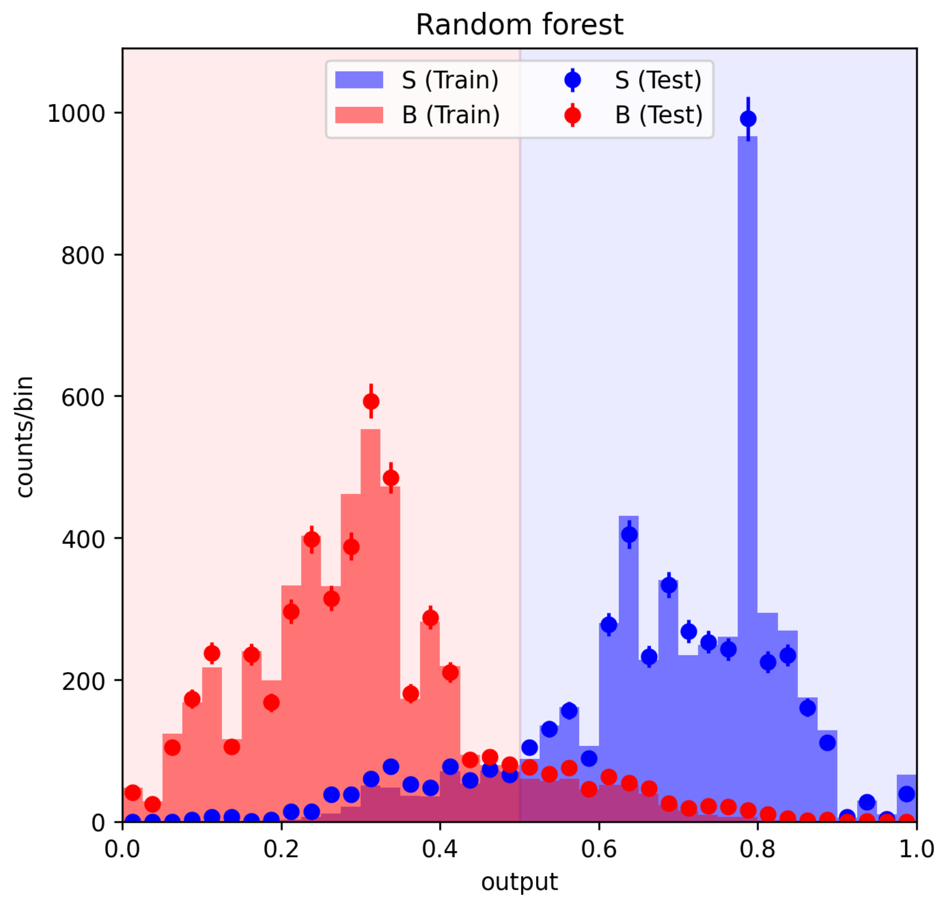
Neural network, deep
2
1
0
−1
−2
−2
−1
0
1
2



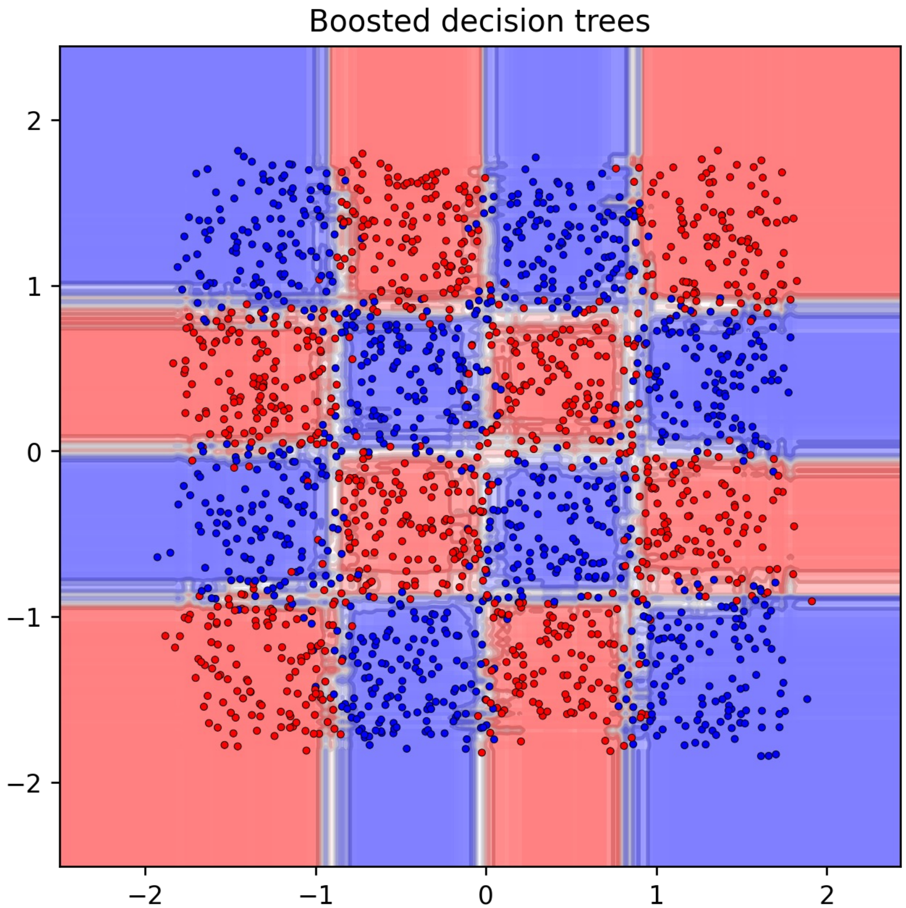
Neural network, deep
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
1400
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
output

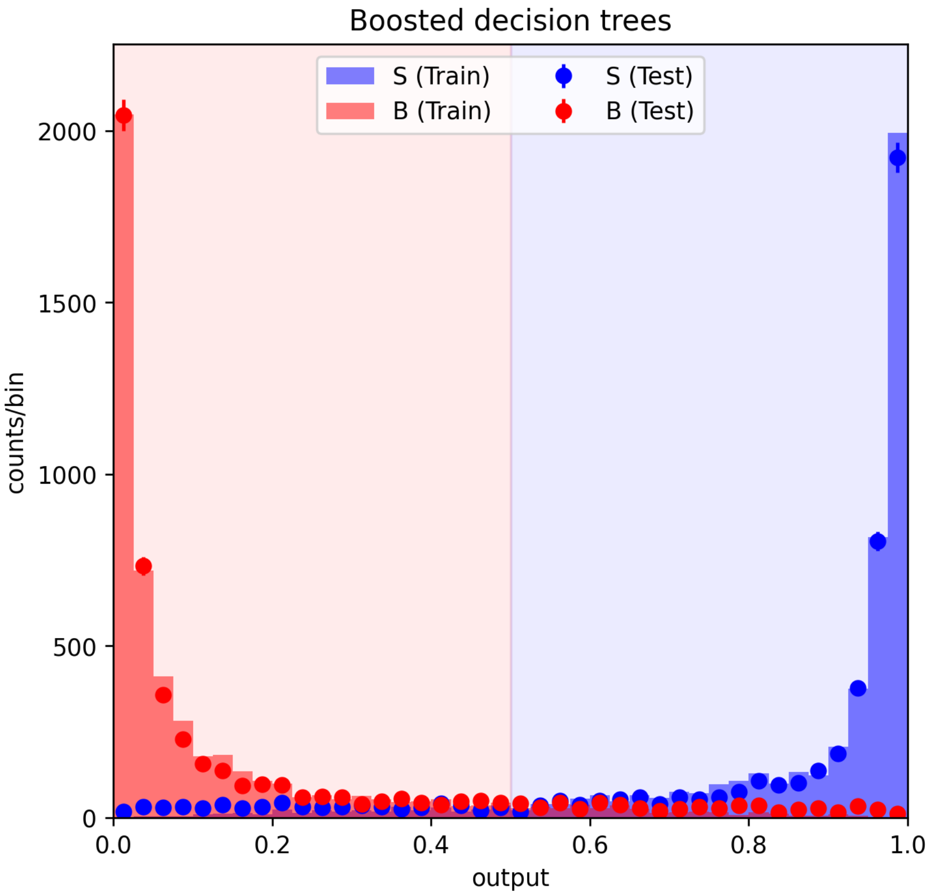
Random forest



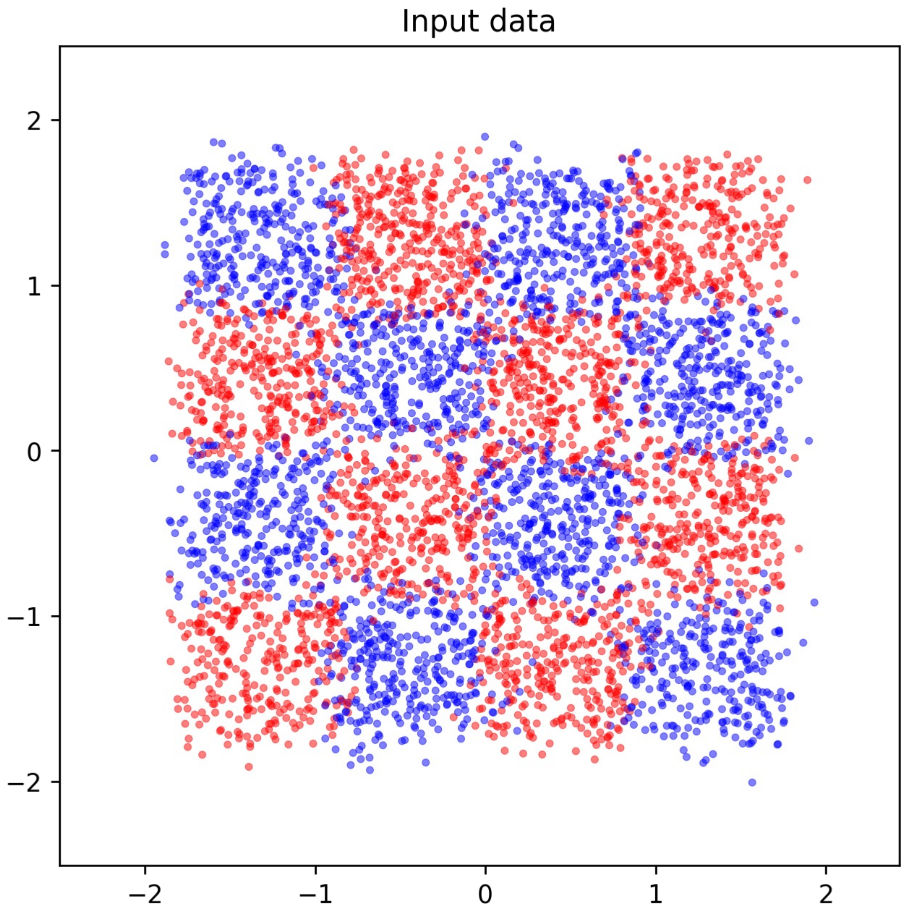


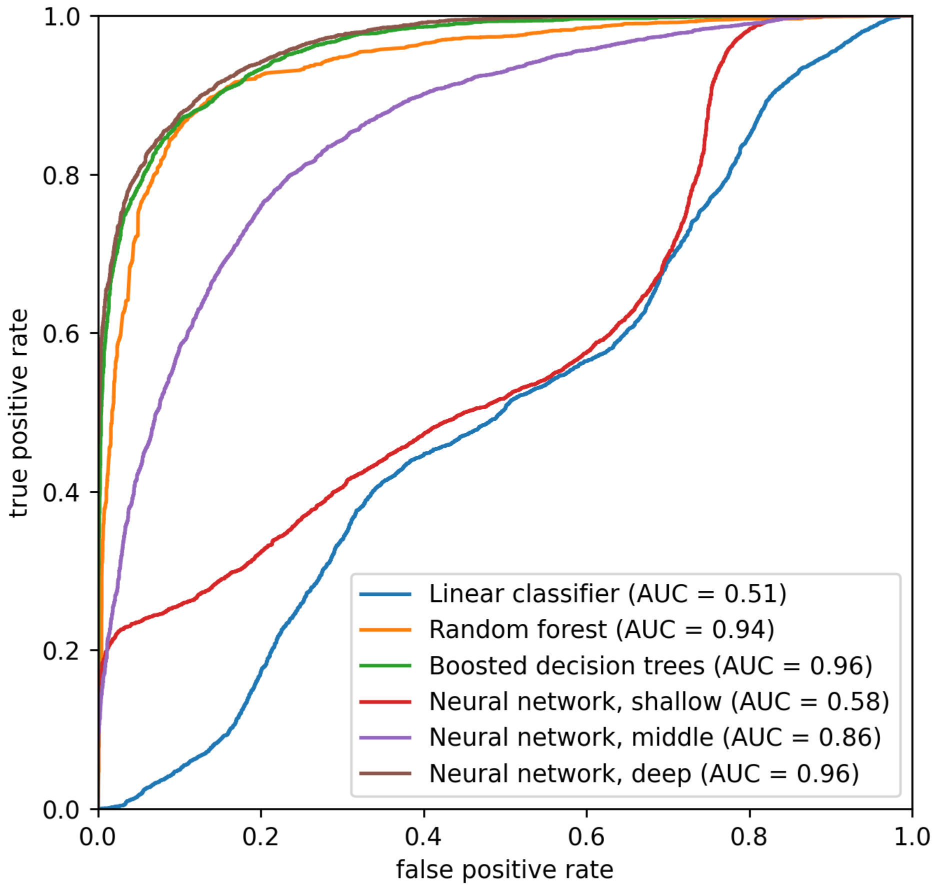
XGBoost


Boosted decision trees
2
1
0
−1
−2
−2
−1
0
1
2

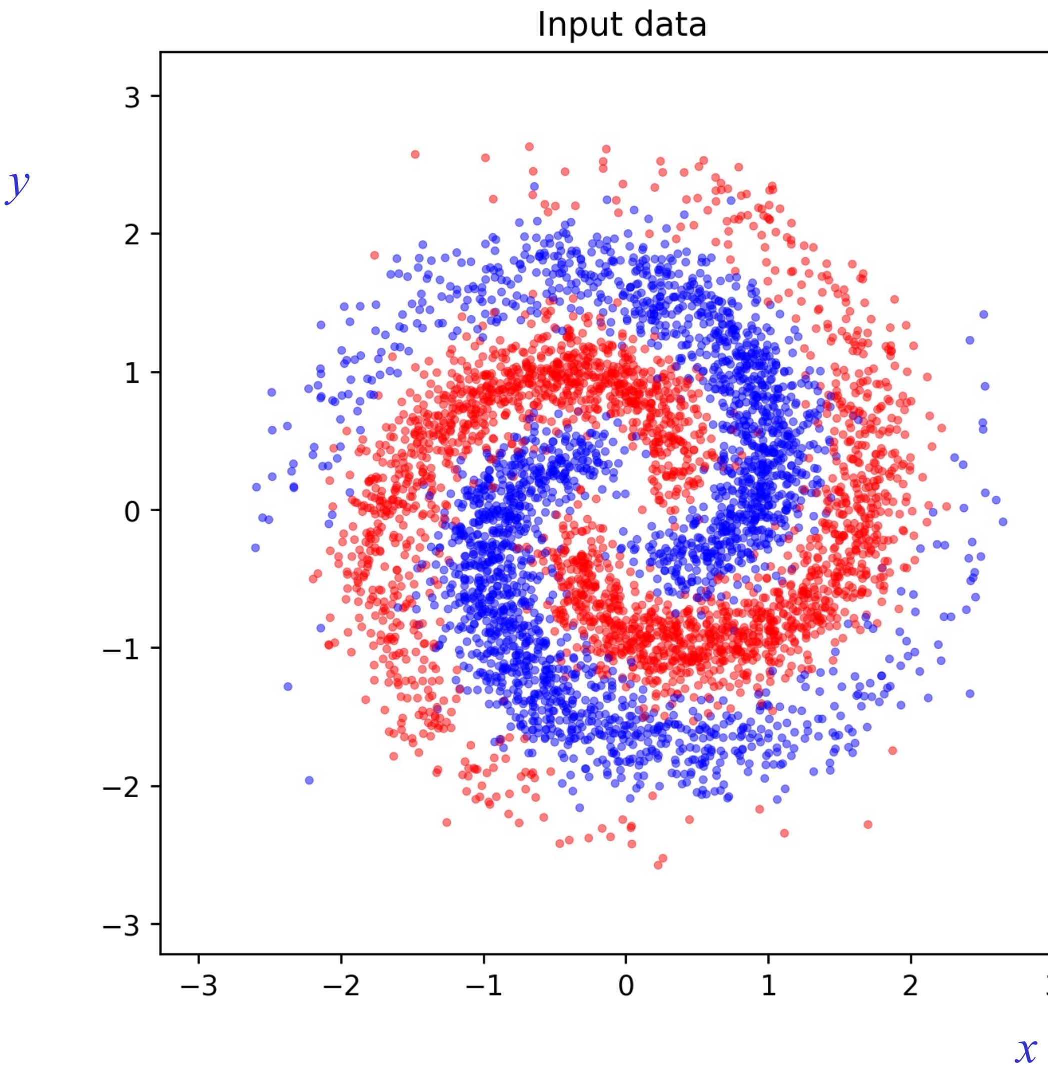

Boosted decision trees
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
output

Intertwained spirals

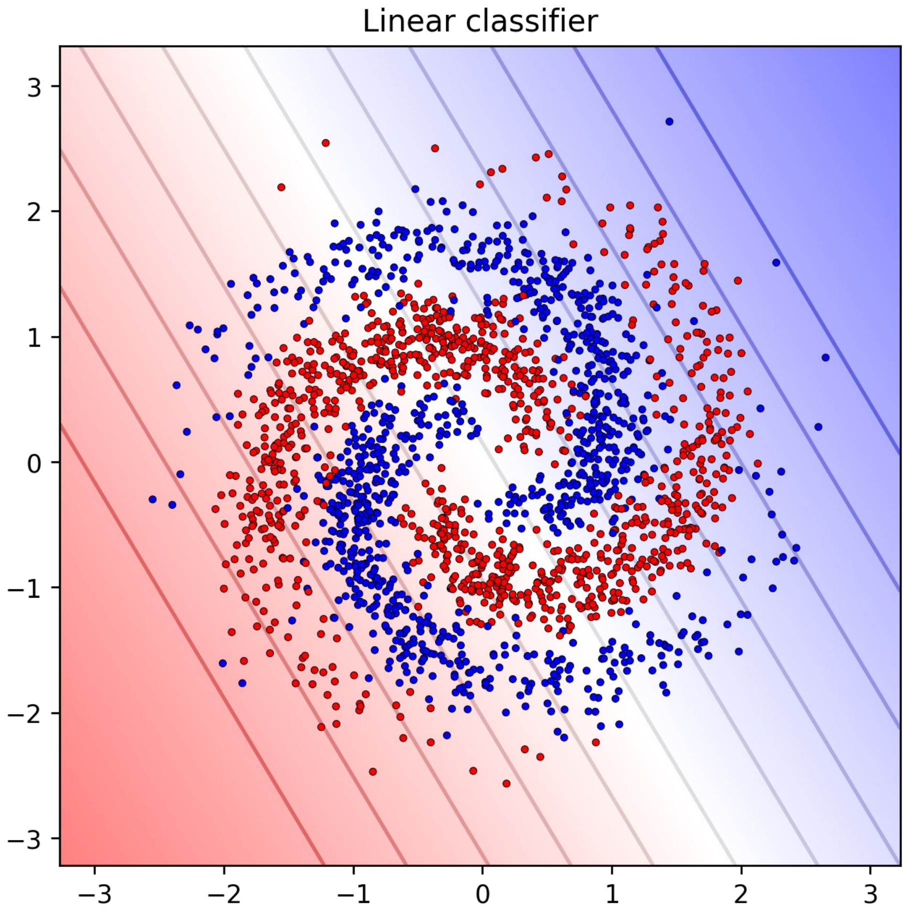


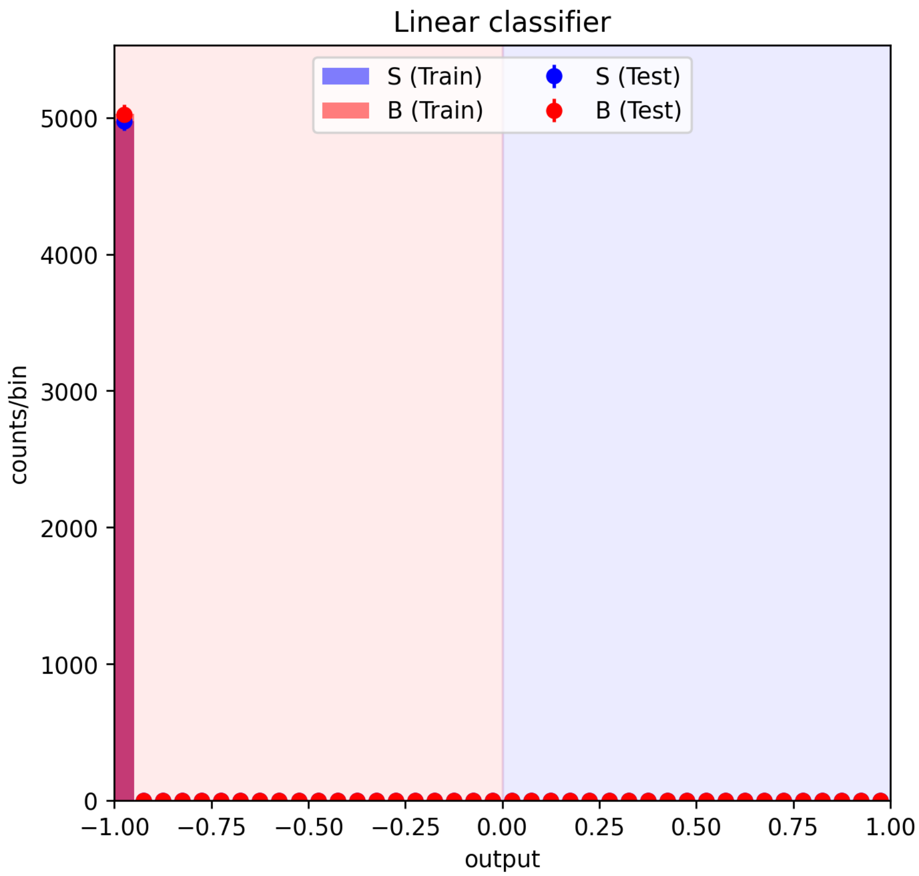


Intertwained spirals

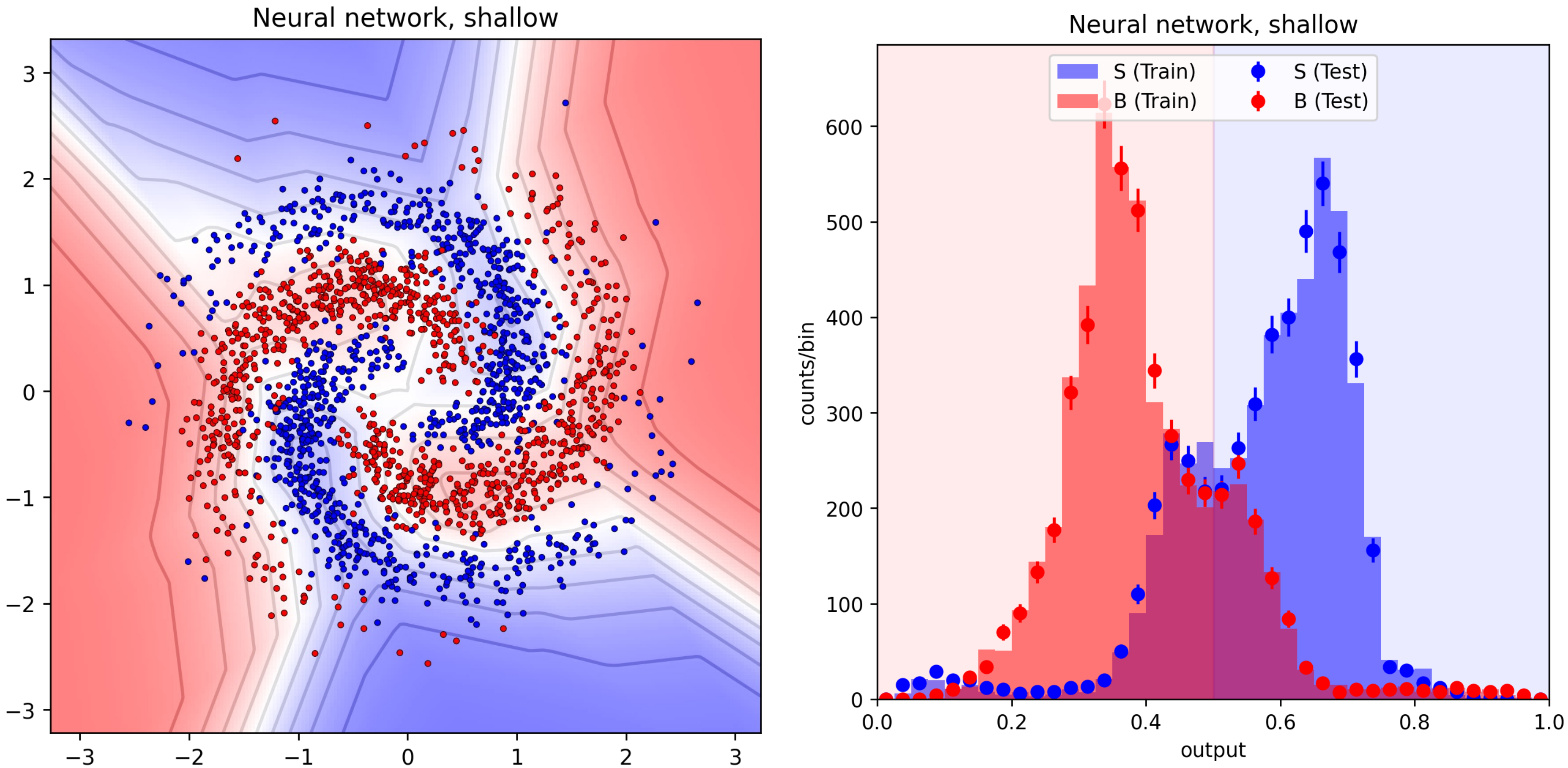


Linear classifier



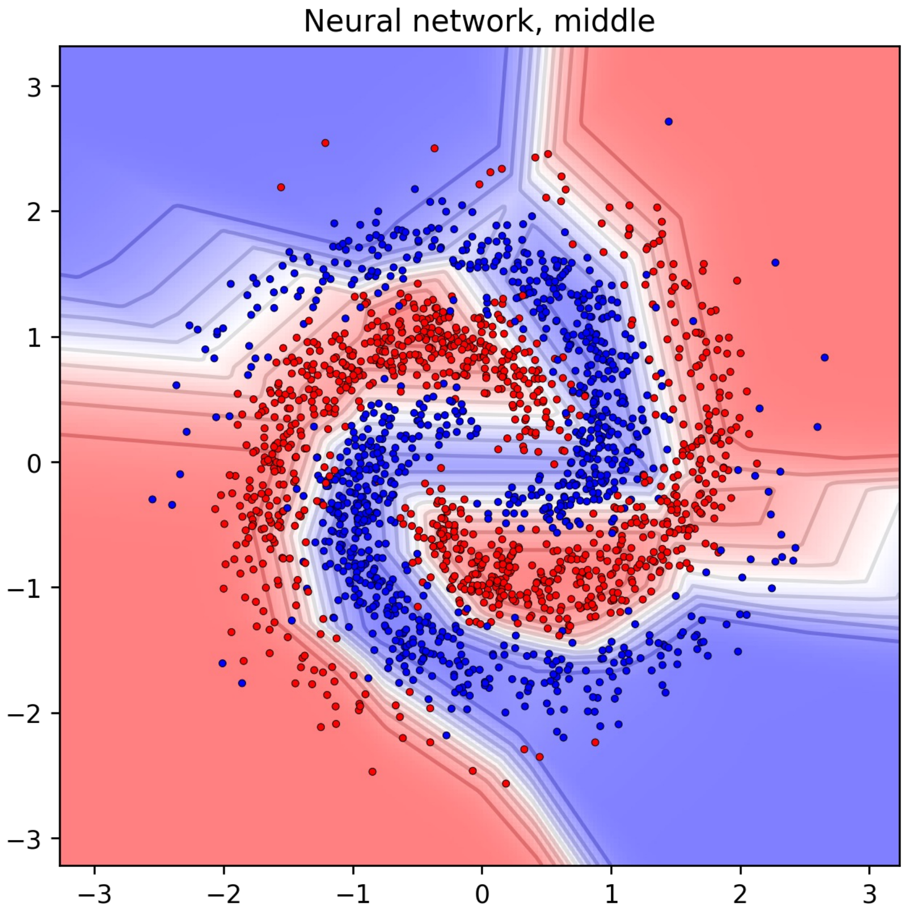


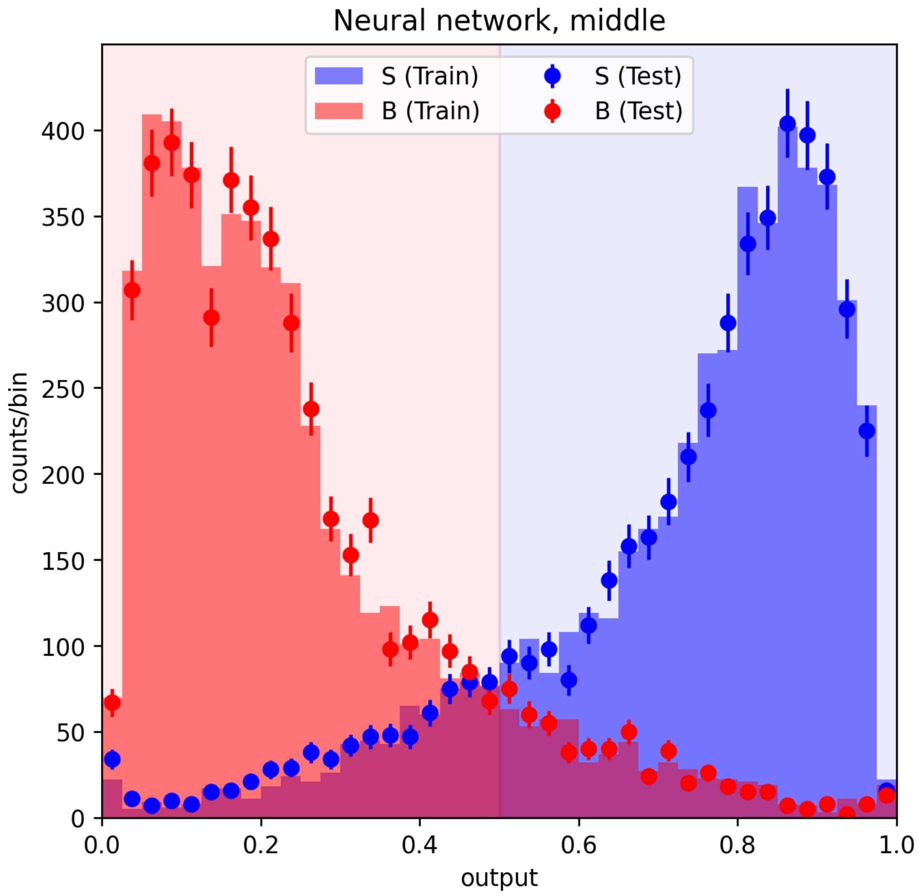
NN, shallow



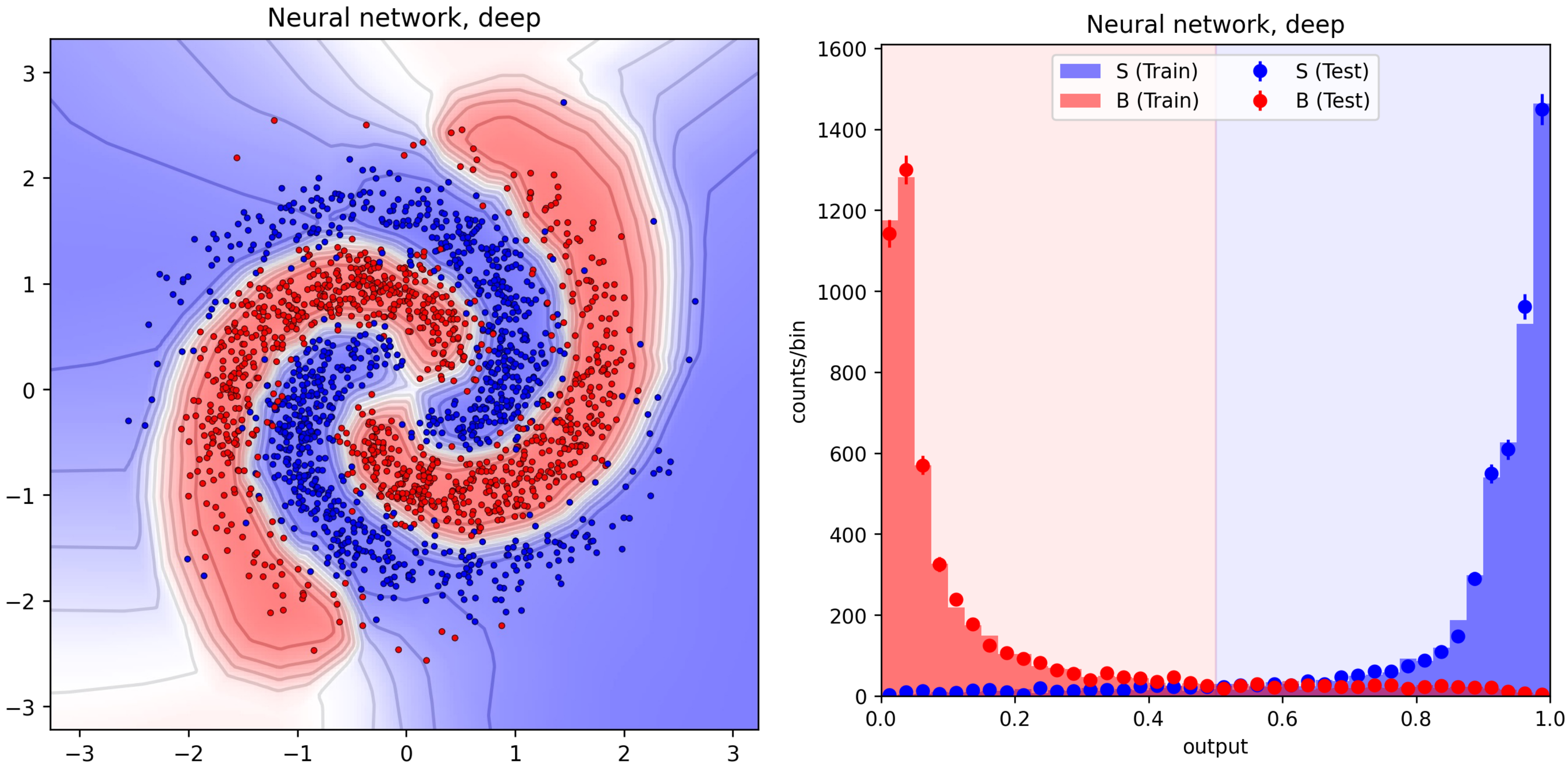
Neural network, shallow
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Neural network, shallow
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
600
500
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0
output

NN, middle

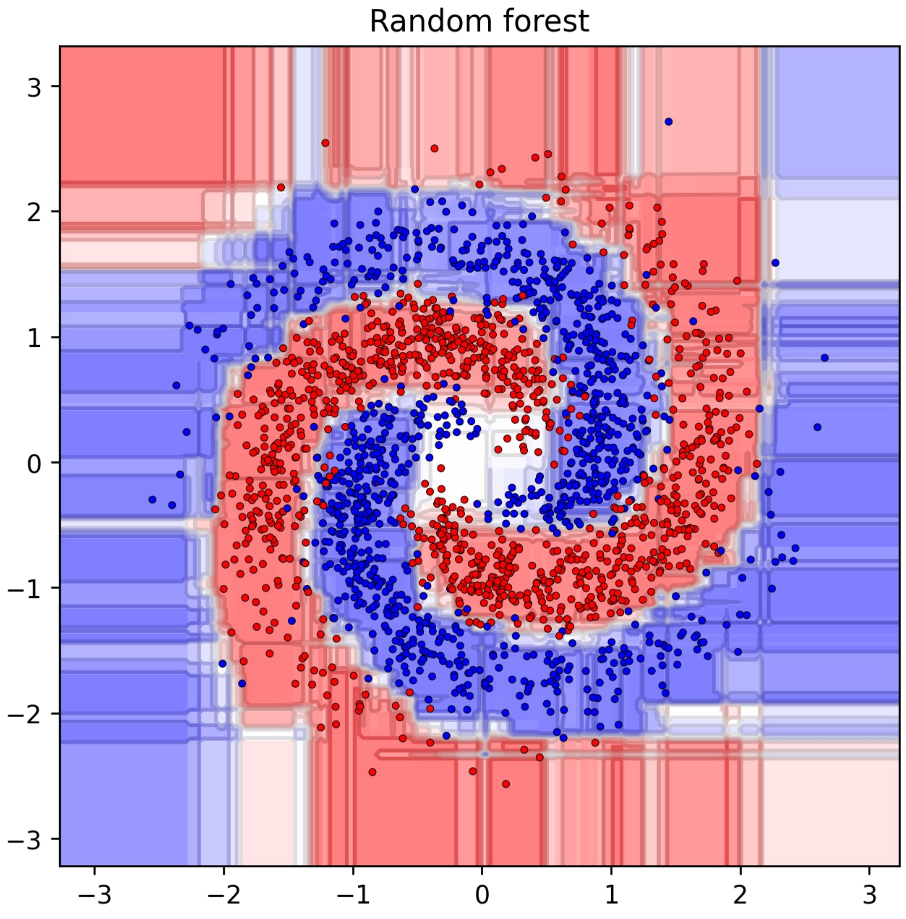

Neural network, middle
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

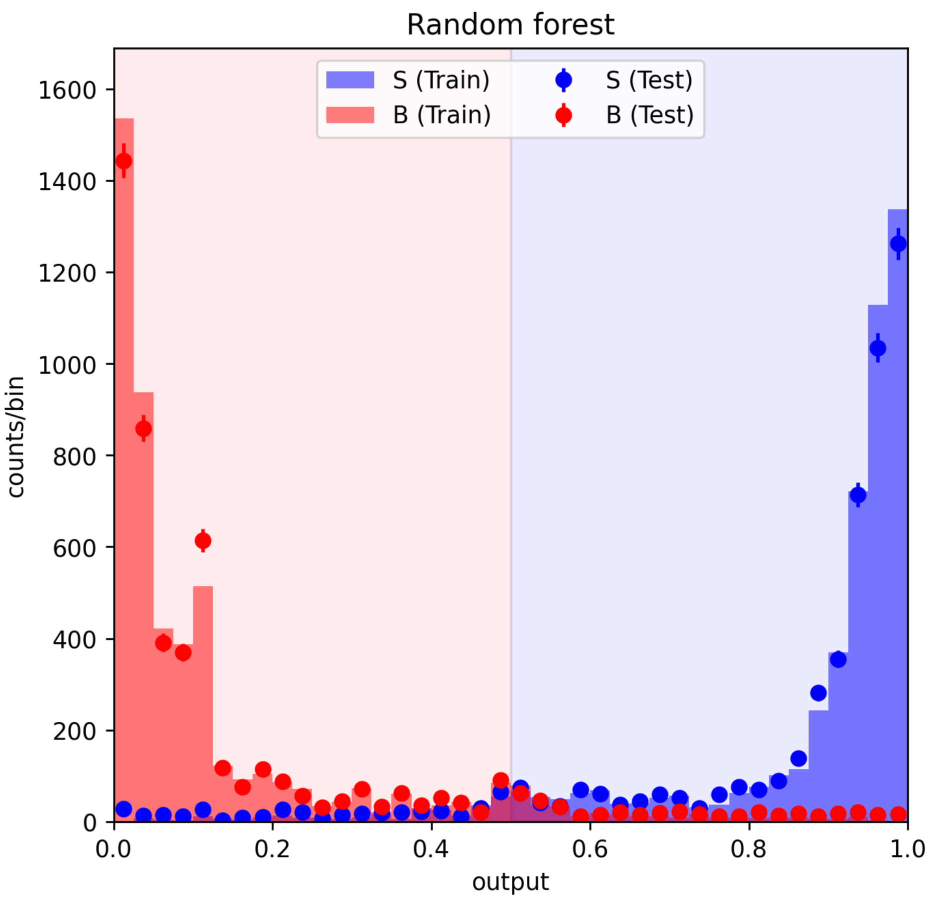

Neural network, middle
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
output
0
50
100
150
200
250
300
350
400
0.0
0.2
0.4
0.6
0.8
1.0

NN, deep

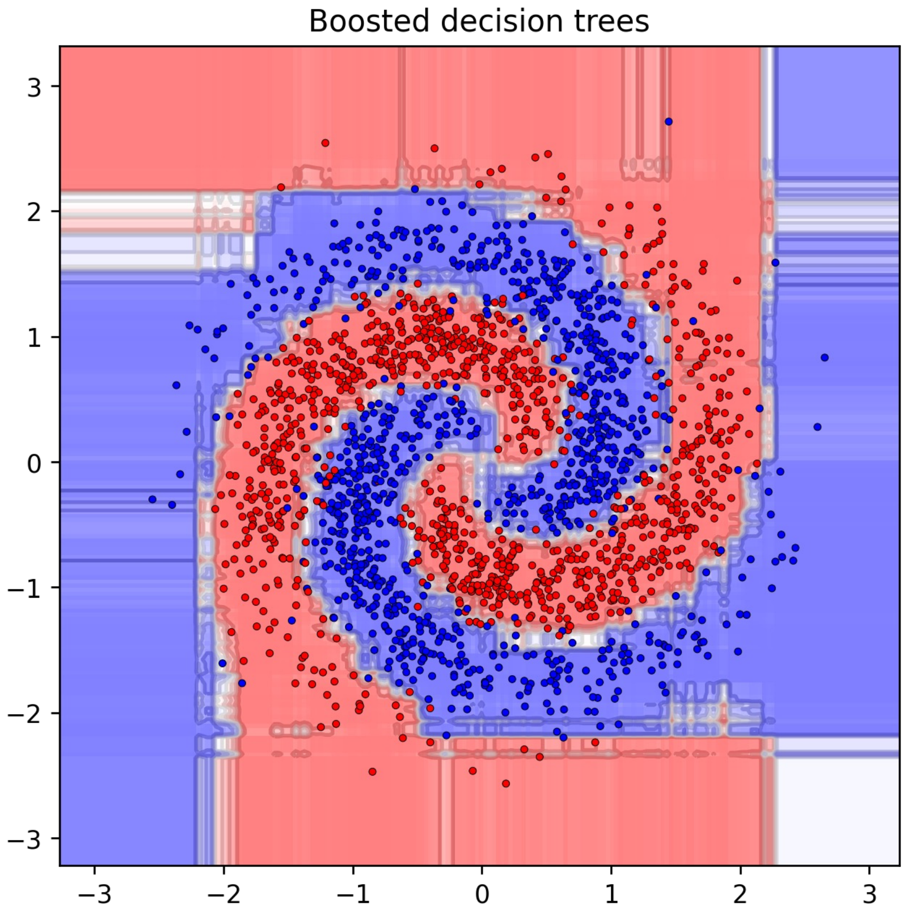

Neural network, deep
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Neural network, deep
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
1600
1400
1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
output

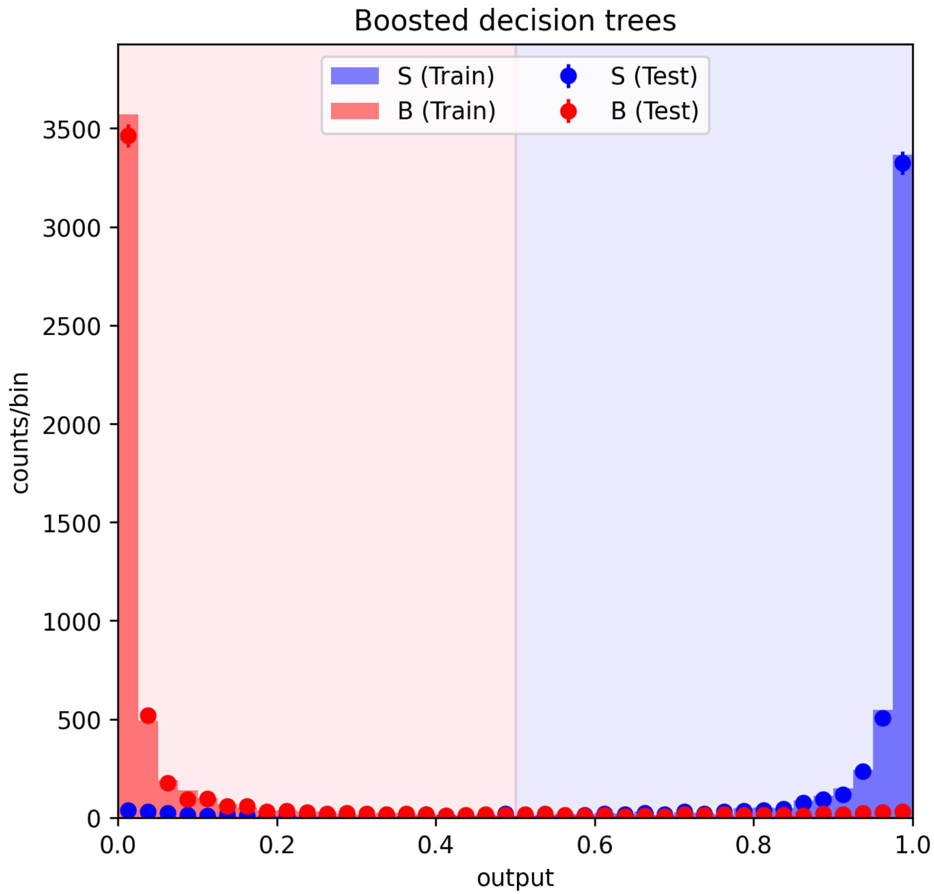
Random forest



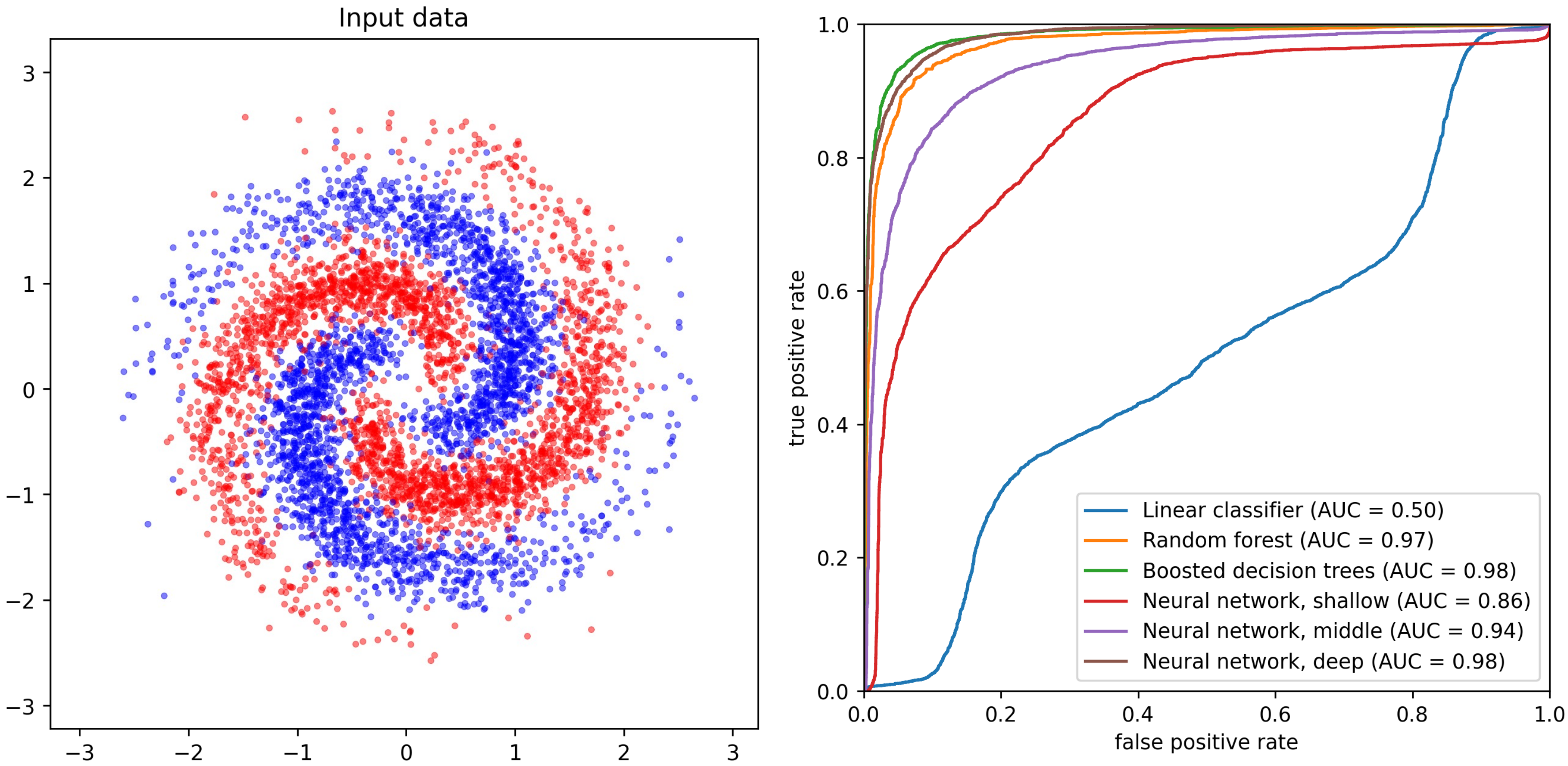


XGBoost


Boosted decision trees
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


Boosted decision trees
S (Train)
B (Train)
S (Test)
B (Test)
counts/bin
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
output

Linear classifier



More machine learning algorithms

- Convolutional neural networks
- Unsupervised learning
  - Clustering Anomaly detection
- Autoencoders
- Generative Networks

Tensor flow playground

- https://playground.tensorflow.org/

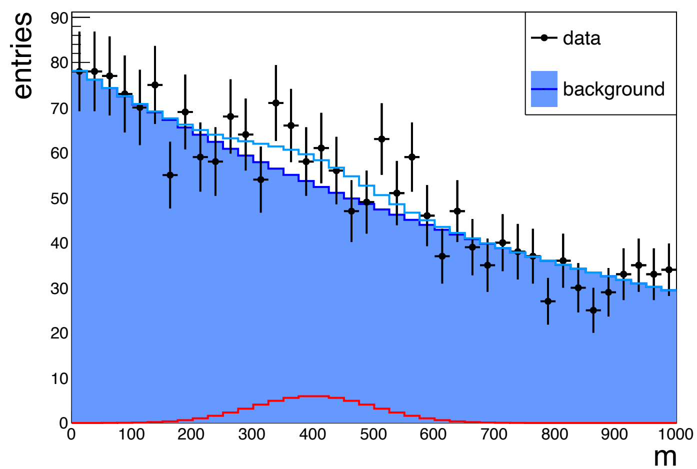
Lecture 4

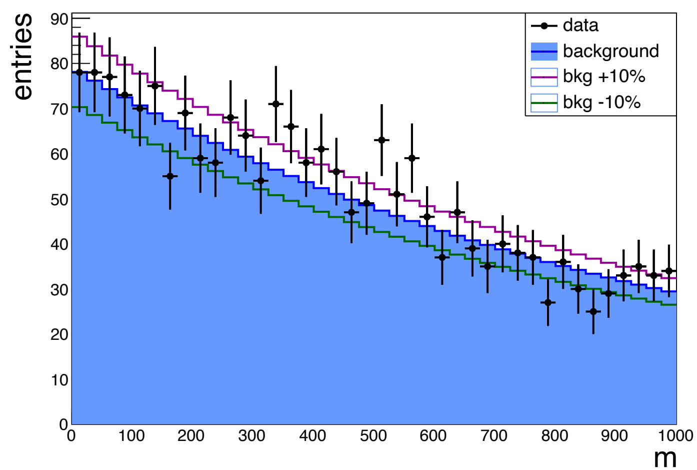
- Discoveries and significance level
- Upper limits
- The flip-flopping issue and the Feldman-Cousins approach
- Bayesian upper limits for event-counting problems
- Modified frequentist approach: the CLs method
- Treatment of nuisance parameters and systematic uncertainties
- The profile likelihood method
- Applications of Wilks' theorem
- Different test statistics
- Asimptotic approximations and the Asimov dataset
- The look-elsewhere effect
- Understanding the "Brazilian" exclusion plots

Claiming a discovery

- We want to test our data sample against two hypotheses about the theoretical underlying model:
  - H_0: the data are described by a model that contains background only
  - H_1: the data are described by a model that contains signal plus background
- Our discrimination is based on a test statistic λ whose distribution is known under the two hypotheses
  - Let's assume λ tends to have (conventionally) large values if H_1 is true and small values if H_0 is true
  - This convention is consistent with λ being the likelihood ratio $L(x|H_1)/L(x|H_0)$
- Under the frequentist approach, compute the p-value as the probability that λ is greater or equal to than the value λ_{obs} we observed

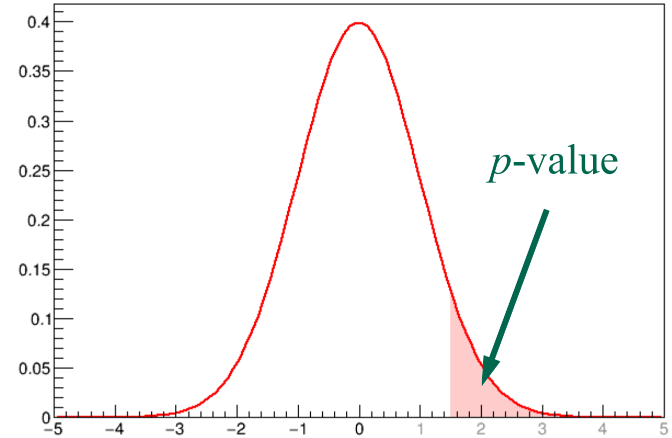
Are data below more consistent with a background fluctuation or with a peaking excess?





Significance

- The p-value is usually converted into an equivalent area of a Gaussian tail:



Z = significance level

$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x = 1 - \Phi(Z)$$

Φ = cumulative of a normal distribution

$$Z = \Phi^{-1}(1-p)$$

- In literature we find, by convention:
  - If the significance is $Z > 3$ ("3σ") one claims "*evidence of*"
    - Probability that background fluctuation will produce a test statistic at least as extreme as the observed value : $p < 1.349 \times 10^{-3}$
  - If the significance is $Z > 5$ ("5σ") one claims "*observation*" (discovery!)
    - $p < 2.87 \times 10^{-7}$
- **Note**: the probability that background produces a large test statistic is not equal to probability of the null hypothesis (background only), which has only a Bayesian sense

Statement by ASA

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

"*The p-value was never intended to be a substitute for scientific reasoning. Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p < 0.05$ era'.*

1. *p-values can indicate how incompatible the data are with a specified statistical model.*
2. *p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*"

Ronald L. Wasserstein, Nicole A. Lazar
The ASA's statement on p-values: context, process, and purpose
DOI:10.1080/00031305.2016.1154108
http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108

Discovery and scientific method

- From Cowan *et al.*, EPJC 71 (2011) 1554:

> *It should be emphasized that in an actual scientific context, rejecting the background-only hypothesis in a statistical sense is only part of discovering a new phenomenon. One's degree of belief that a new process is present will depend in general on other factors as well, such as the plausibility of the new signal hypothesis and the degree to which it can describe the data.*
>
> *Here, however, we only consider the task of determining the p-value of the background-only hypothesis; if it is found below a specified threshold, we regard this as "discovery".*

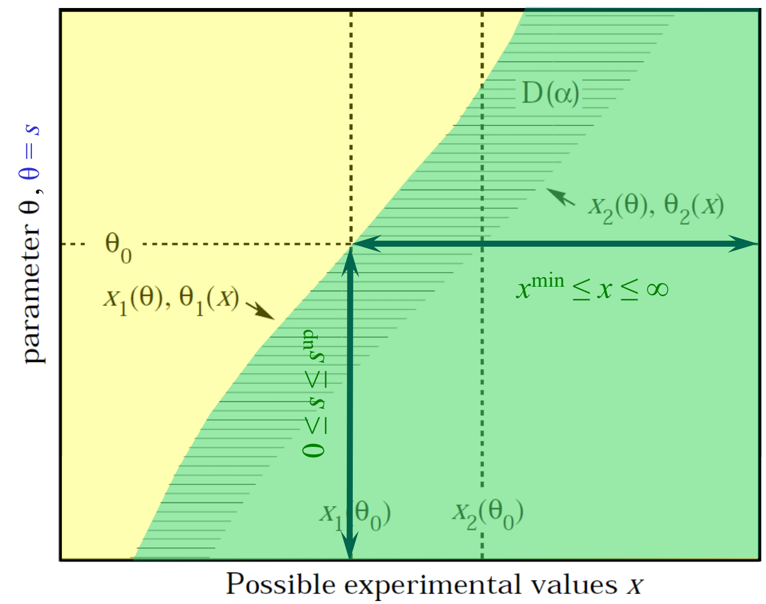
Complementary role of Frequentist and Bayesian approaches ☺

Upper limits

- Measure the amount of excluded region resulting from our (negative) search for a new signal
- Building a fully asymmetric Neyman confidence belt based on the considered test statistic x
- Invert the belt, find the allowed interval:

$$s \in [s_1, s_2] \Rightarrow s \in [0, s^{\mathrm{up}}]$$

- Upper limit = upper extreme of the asymmetric interval $[0, s^{\mathrm{up}}]$
- In case the observable x is discrete (e.g.: the number of events n in a counting experiments), the coverage may not be exact

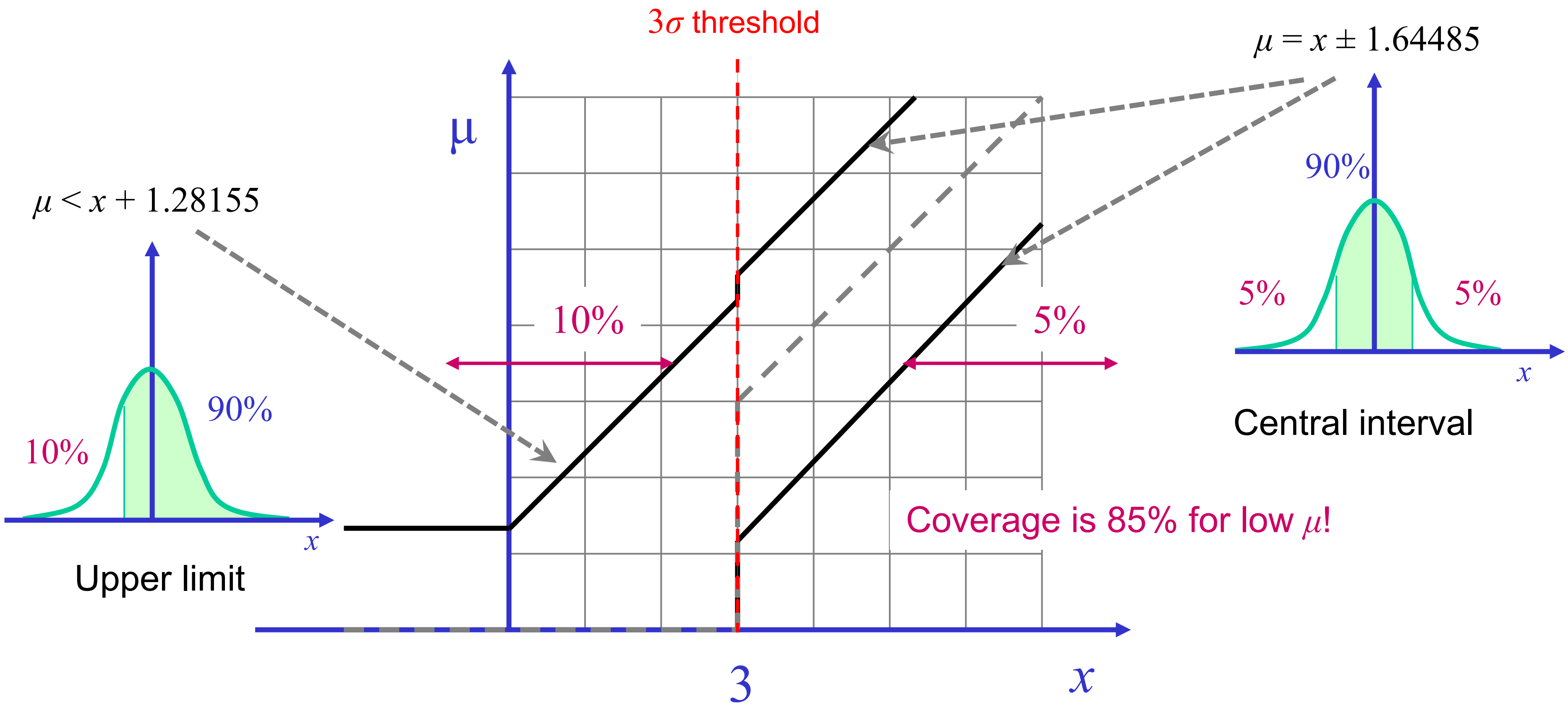


The flip-flopping issue

- When to quote a central value or upper limit?
- A popular choice was:
  - *"Quote a 90% CL upper limit of the measurement if the significance is below 3σ; quote a central value otherwise"*
  - Upper limit ↔ central interval decided according to observed data
- This produces an incorrect coverage!

“Flip-flopping” with a Gaussian PDF

- Assume a Gaussian with a fixed width: $\sigma = 1$



Gary J. Feldman, Robert D. Cousins, Phys.Rev.D57:3873-3889,1998

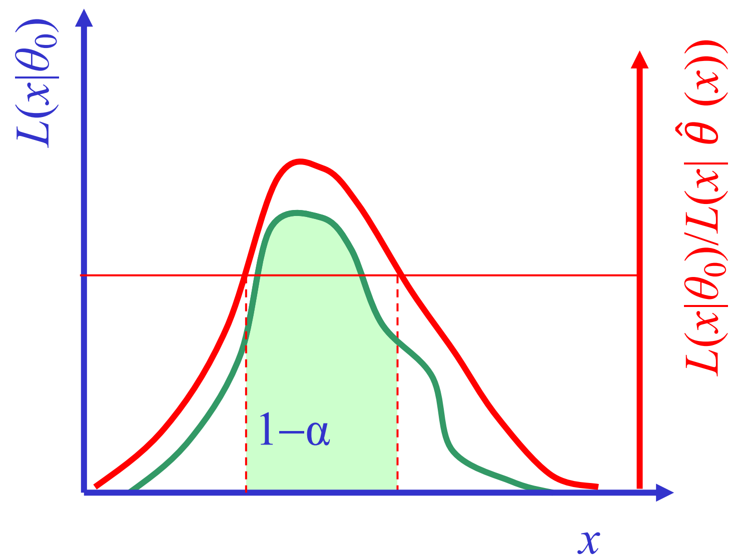
Likelihood ratio & Neyman belt

- Feldman and Cousins proposed a criterion to define the Neyman belt based in a likelihood ratio test:

$$R_\mu = \{x : L(x|\theta_0) \,/\, L(x|\hat{\theta}) > k_\alpha\}$$

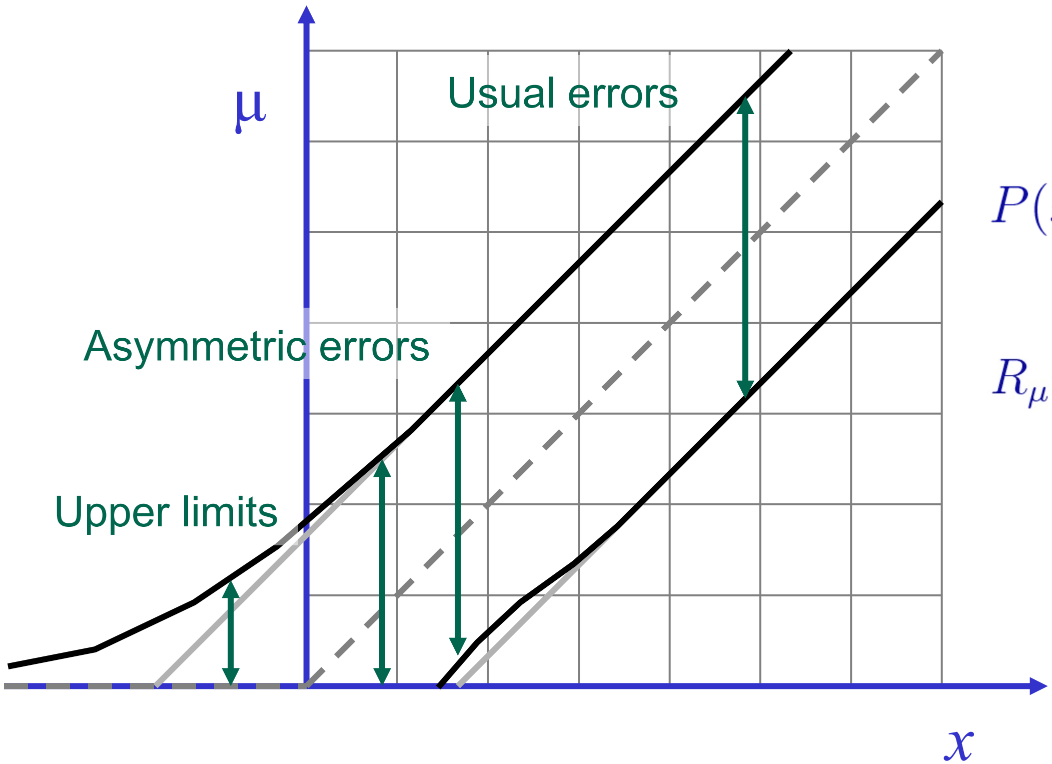
- The value k_α depends on the desired significance level α

- H_0: $\theta = \hat{\theta}$, the best-fit value
- H_1: $\theta = \theta_0$, the specific value considered for the Neyman belt construction



Feldman-Cousins approach

- Application to the Gaussian case:



$$\hat{\mu} = \max(x, 0)$$

$\hat{\mu} = x$ for $x \geq 0$

$$P(x|\hat{\mu}) = \begin{cases} \frac{1}{\sqrt{2\pi}}\,, & x \geq 0\,, \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,, & x < 0\,. \end{cases}$$

$$R_\mu(x) = \frac{P(x|\mu)}{P(x|\hat{\mu})} = \begin{cases} e^{-\frac{(x-\mu)^2}{2}}\,, & x \geq 0\,, \\ e^{-\frac{x\mu - \mu^2}{2}}\,, & x < 0\,. \end{cases}$$

Confidence intervals must be computed numerically, even for this simple Gaussian case!

Upper limits for event counting

- The simplest search for a new signal consists of counting the number of events passing a specified selection
- The number of selected events n is distributed according to a Poissonian distribution
- Expected n for signal + background (H_1): $s + b$
- Expected n for background only (H_0): b

- We measure n events, we want to compare with the two hypotheses H_1 and H_0.
- Simplest case: b is known with negligible uncertainty
  - If not, uncertainty on its estimate must be taken into account

Counting, Bayesian approach

- Let's assume the background b is known with no uncertainty:

$$L(n; s) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

- A uniform prior, $\pi(s) = 1$ simplifies, as usual, the computation:

$$1 - \alpha = \int_0^{s^{\mathrm{up}}} P(s|n)\mathrm{d}s = \frac{\int_0^{s^{\mathrm{up}}} L(n; s)\pi(s)\mathrm{d}s}{\int_0^{\infty} L(n; s)\pi(s)\mathrm{d}s}$$

- Inverting the equation gives the upper limit s^{up}

$$\alpha = e^{-s^{\mathrm{up}}} \frac{\sum_{m=0}^{n} \frac{(s^{\mathrm{up}}+b)^m}{m!}}{\sum_{m=0}^{n} \frac{b^m}{m!}}$$

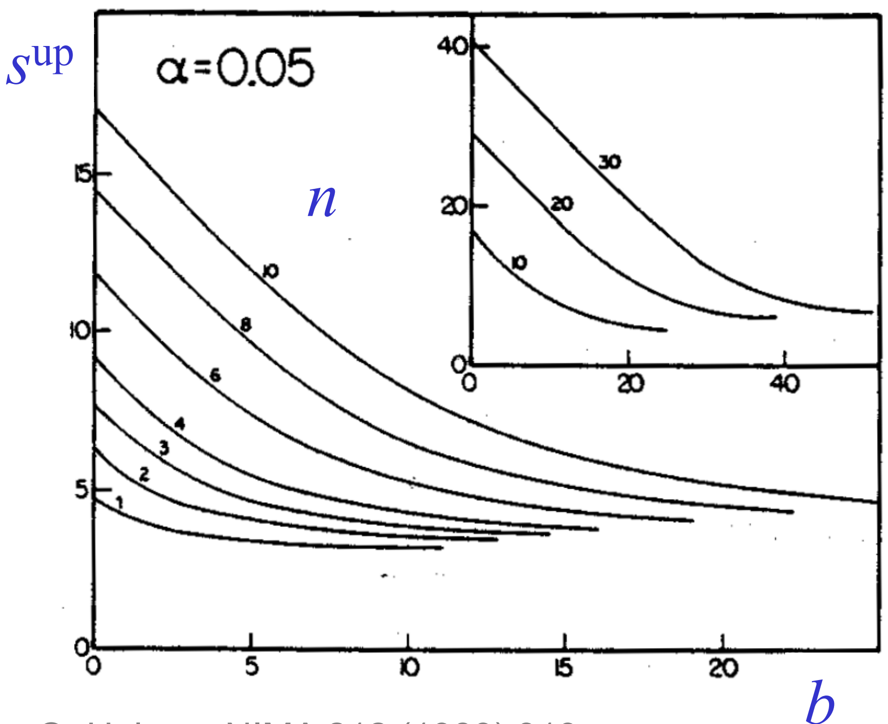
- For $n = 0$ s^{up} does not depend on b:

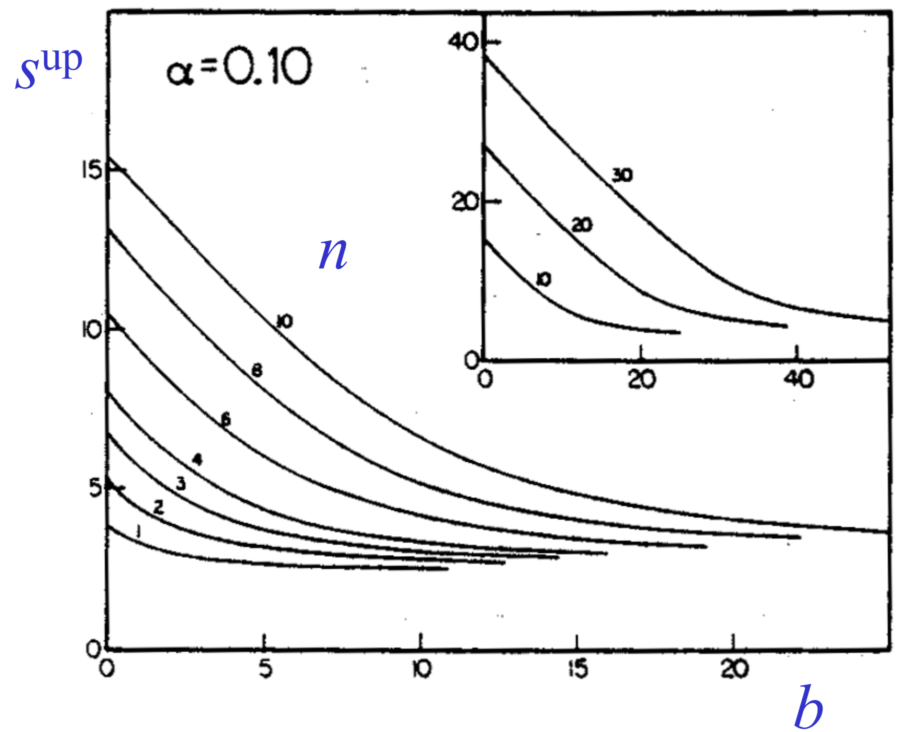
$$\alpha = e^{-s^{\mathrm{up}}}$$

  – $s < 2.303$ (90% CL) $\leftarrow \alpha = 0.1$
  – $s < 2.996$ (95% CL) $\leftarrow \alpha = 0.05$

Counting, Bayesian approach

- Upper limits decrease as b increases and increase as n increases
- For $n = 0$, upper limits are not sensitive on b (given in prev. slide)





O. Helene. NIMA 212 (1983) 319

Frequentist: zero events selected

- Assume we have negligible background ($b = 0$) and we measure zero events ($n = 0$)
- The likelihood function simplifies as:

$$L(n = 0; s) = \mathrm{Poiss}(0; s) = e^{-s}$$

- The (fully asymmetric) Neyman belt inversion is pretty simple:

$$P(n \leq 0; s^{\mathrm{up}}) = \alpha \longrightarrow s^{\mathrm{up}} = -\ln \alpha$$

- The results are by chance identical to the Bayesian computation:

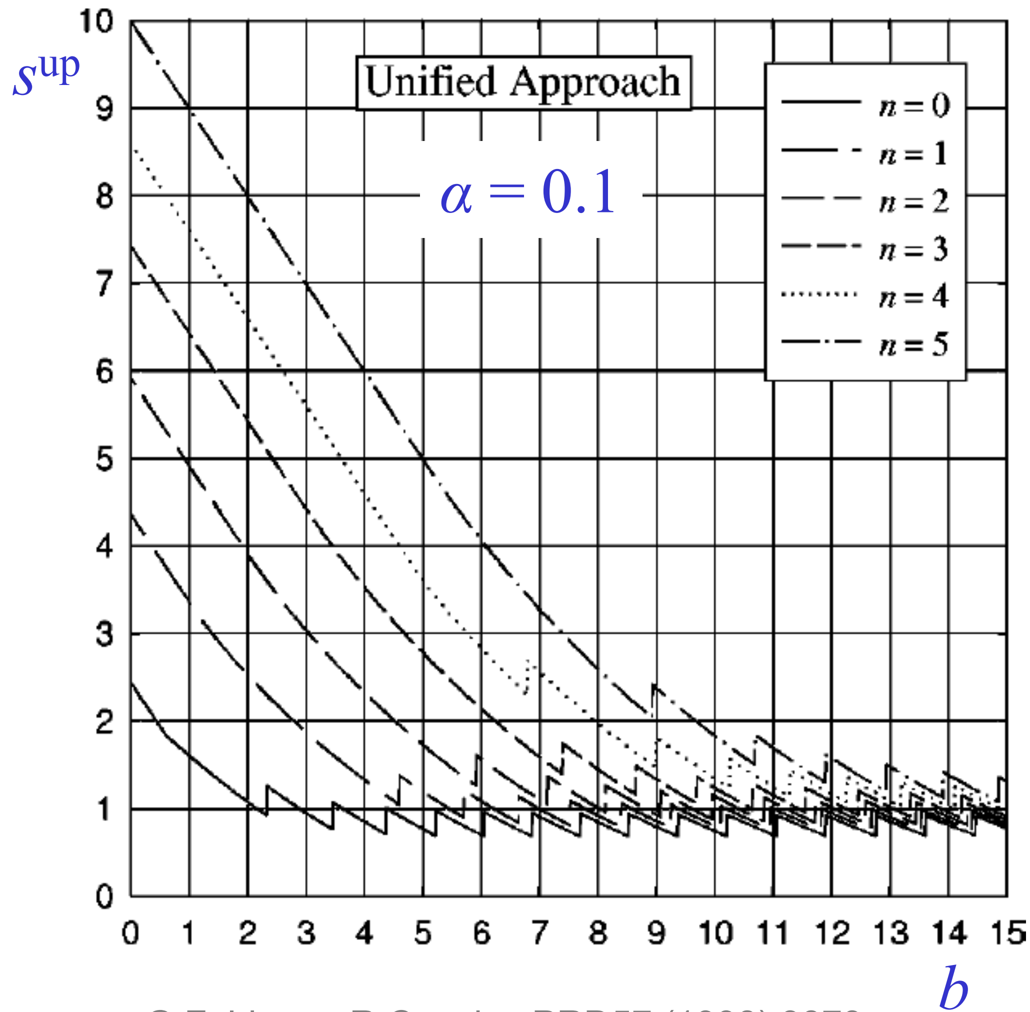
$$s < 2.303 \text{ (90\% CL)} \leftarrow \alpha = 0.1$$

$$s < 2.996 \text{ (95\% CL)} \leftarrow \alpha = 0.05$$

- In spite of the numerical coincidence, the interpretation of frequentist and Bayesian upper limits remain very different!
- **Warning**: this evaluation suffer from the "**flip-flopping**" problem, so the coverage is spoiled if you decide to switch from upper limit to a central value depending on the observed significance!

Conting, Feldman-Cousins

- F&C intervals cure the flip-flopping issue and ensure the correct coverage
  - May overcover for discrete variables
- The "ripple" structure is due to the discrete nature of Poissonian counting
- Note that even for $n = 0$ the upper limit decrease as b increases (apart from ripple effects)
- If two experiment are designed for an expected background of –say– 0.5 and 0.01, the "worse" one has the best expected upper limit



G.Feldman, R.Cousins PRD57 (1998) 3873
C. Giunti, PRD59 (1999), 053001

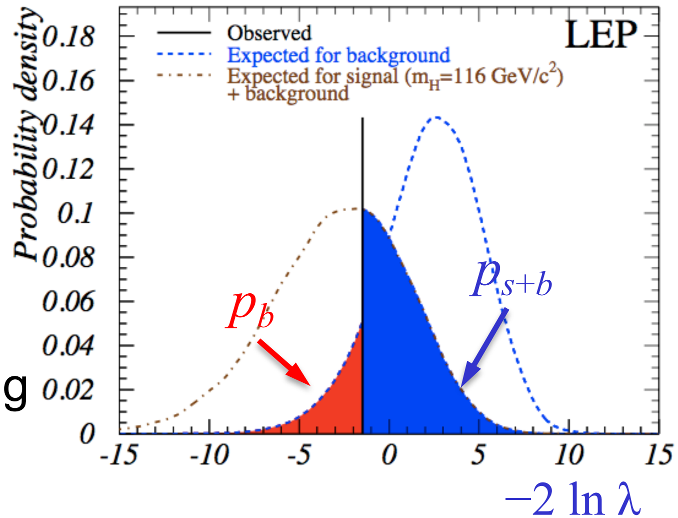
From PDG Review…

- *"The intervals constructed according to the unified procedure [FC] for a Poisson variable n consisting of signal and background have the property that for $n = 0$ observed events, the upper limit decreases for increasing expected background. This is counter-intuitive, since it is known that if $n = 0$ for the experiment in question, then no background was observed, and therefore one may argue that the expected background should not be relevant. The extent to which one should regard this feature as a drawback is a subject of some controversy"*

Modified frequentist approach

- A modified approach was proposed for the first time when combining the limits on the Higgs boson search from the four LEP experiments, ALEPH, DELPHI, L3 and OPAL
- Given a test statistic $\lambda(x)$, determine its distribution for the two hypotheses $H_1(s+b)$ and $H_0(b)$, and compute:

$$\begin{cases} p_{s+b} = P\big(\lambda(x|H_1) \leq \lambda^{\mathrm{obs}}\big) \\ p_b \quad = P\big(\lambda(x|H_0) \geq \lambda^{\mathrm{obs}}\big) \end{cases}$$



- The upper limit is computed, instead of requiring $p_{s+b} \leq \alpha$, on the modified statistic $\mathrm{CL}_s \leq \alpha$:

$$\mathrm{CL}_s = \frac{p_{s+b}}{1 - p_b}$$

- Since $1-p_b \leq 1$, $\mathrm{CL}_s \geq p_{s+b}$, hence upper limits computed with the CL_s method are always **conservative**

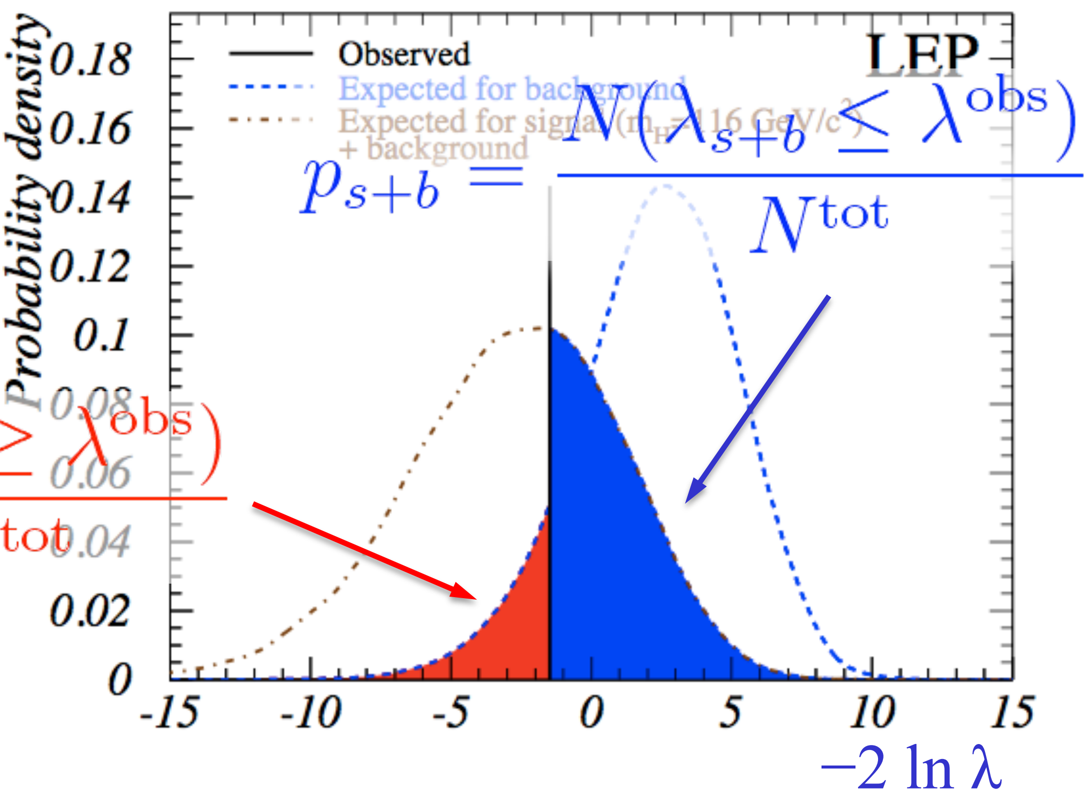
Note: $\lambda \leq \lambda^{\mathrm{obs}}$ implies $-2\ln\lambda \geq \lambda^{\mathrm{obs}}$

CL_s with toy experiments

- In practice, p_b and p_{s+b} are computed in from simulated pseudo-experiments ("toy Monte Carlo")

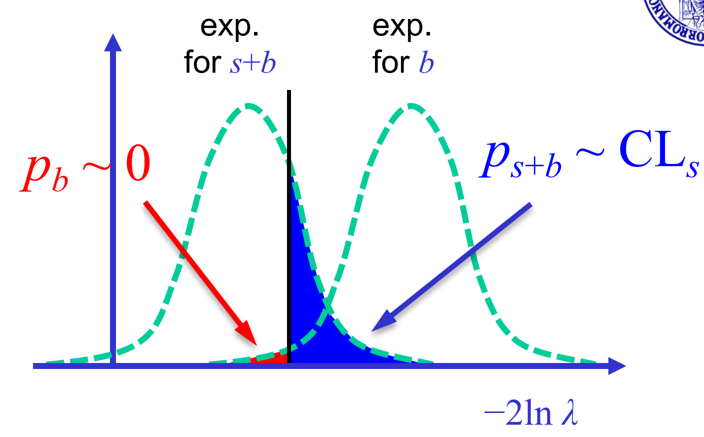
$$\mathrm{CL}_s = \frac{N(\lambda_{s+b} \le \lambda^{\mathrm{obs}})}{N(\lambda_b \le \lambda^{\mathrm{obs}})}$$

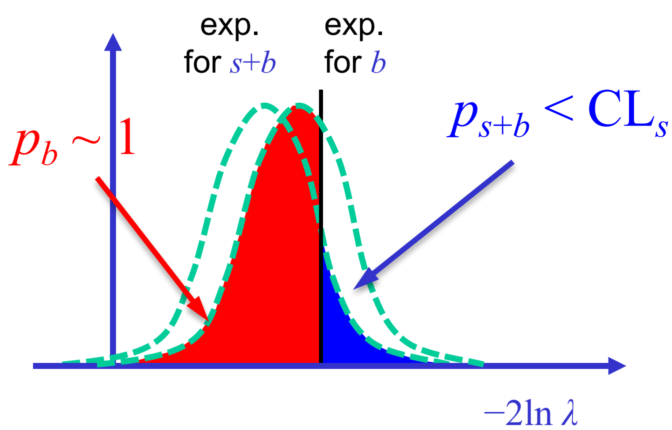
Plot from LEP Higgs combination paper



Main CL_s features

- p_{s+b}: probability to obtain a result which is less compatible with the signal than the observed result, assuming the signal hypothesis
- p_b: probability to obtain a result less compatible with the background-only hypothesis than the observed one
- If the two distributions are very well separated ad H_1 is true, than p_b will be very small ⇒ $1-p_b \sim 1$ and $\mathrm{CL}_s \sim p_{s+b}$, i.e: the ordinary p-value of the $s+b$ hypothesis
- If the two distributions largely overlap, than if p_b will be large ⇒ $1-p_b$ small, preventing CL_s to become very small





- $\mathrm{CL}_s < 1-\alpha$ prevents rejecting cases where the experiment has little sensitivity

$$\mathrm{CL}_s = \frac{p_{s+b}}{1-p_b} = \frac{P(\lambda_{s+b} \le \lambda^{\mathrm{obs}})}{P(\lambda_b \le \lambda^{\mathrm{obs}})}$$

Event counting with CL_s

- Let's consider the previous event counting experiment, using $n = n^{\mathrm{obs}}$ as test statistic
- In this case CL_s can be written as:

$$\mathrm{CL}_s = \frac{P(n \leq n^{\mathrm{obs}} | s + b)}{P(n \leq n^{\mathrm{obs}} | b)}$$

- Explicitating the Poisson distribution, the computation gives the same result as for the Bayesian case with a uniform prior
- In many cases the CL_s upper limits give results that are very close, numerically, to Bayesian computations done assuming a uniform prior
- But the interpretation is very different from Bayesian limits!

$$\alpha = e^{-s^{\mathrm{up}}} \frac{\displaystyle\sum_{m=0}^{n} \frac{(s^{\mathrm{up}} + b)^m}{m!}}{\displaystyle\sum_{m=0}^{n} \frac{b^m}{m!}}$$

Observations on the CL_s method

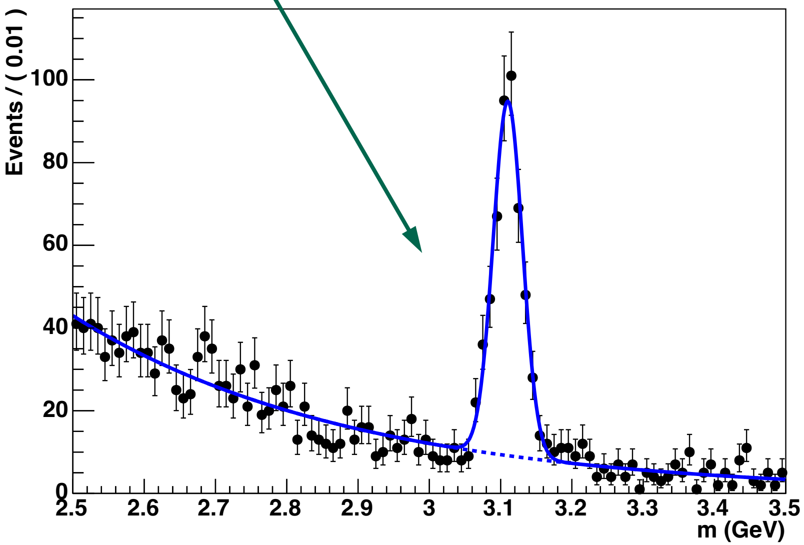
- *"A specific modification of a purely classical statistical analysis is used to avoid excluding or discovering signals which the search is in fact not sensitive to"*
- *"The use of CLs is a conscious decision not to insist on the frequentist concept of full coverage (to guarantee that the confidence interval doesn't include the true value of the parameter in a fixed fraction of experiments)."*
- *"confidence intervals obtained in this manner do not have the same interpretation as traditional frequentist confidence intervals nor as Bayesian credible intervals"*

A. L. Read, Modified frequentist analysis of search results (the CLls method), 1st Workshop on Confidence Limits, CERN, 2000

Nuisance parameters

- Usually, signal extraction procedures (fits, upper limits setting) determine, together with parameters of interest, also nuisance parameters that model effects not strictly related to our final measurement
  - Background yield and shape parameters
  - Detector resolution
  - ...
- Nuisance parameters are also used to model sources of systematic uncertainties
- Often referred to nominal values
  - Examples: cross section × int. lumi
  - $b = \beta\, \sigma_b\, L_{\mathrm{int}}$ with $\beta^{\mathrm{nominal}} = 1$
  - $b = e^{\beta}\, \sigma_b\, L_{\mathrm{int}}$ with $\beta^{\mathrm{nominal}} = 0$ (negative yields not allowed!)

$$L(m; s, b, \mu, \sigma, \lambda) = \frac{e^{-(s+b)}}{n!}\left(s\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(m-\mu)^2}{2\sigma^2}} + b\lambda e^{-\lambda m}\right)$$



Nuisance pars in Bayesian approach

- Notation below: μ = parameter(s) of interest, θ = nuisance parameter(s)
- No special treatment:

$$P(\mu,\theta|x) = \frac{L(x;\mu,\theta)\pi(\mu,\theta)}{\int L(x;\mu',\theta')\pi(\mu',\theta')\mathrm{d}\mu'\mathrm{d}\theta'}$$

- $P(\mu|x)$ obtained as marginal PDF of μ obtained integrating on θ:

$$P(\mu|x) = \int P(\mu,\theta|x)\mathrm{d}\theta = \frac{\int L(x;\mu,\theta)\pi(\mu,\theta)\mathrm{d}\theta}{\int L(x;\mu',\theta)\pi(\mu',\theta)\mathrm{d}\mu'\mathrm{d}\theta}$$

Nuisance pars., frequentist

- Introduce a complementary dataset to constrain the nuisance parameters θ (e.g.: calibration data, background estimates from control sample…)
- Formulate the statistical problem in terms of both the main data sample (x) and the control sample (y)

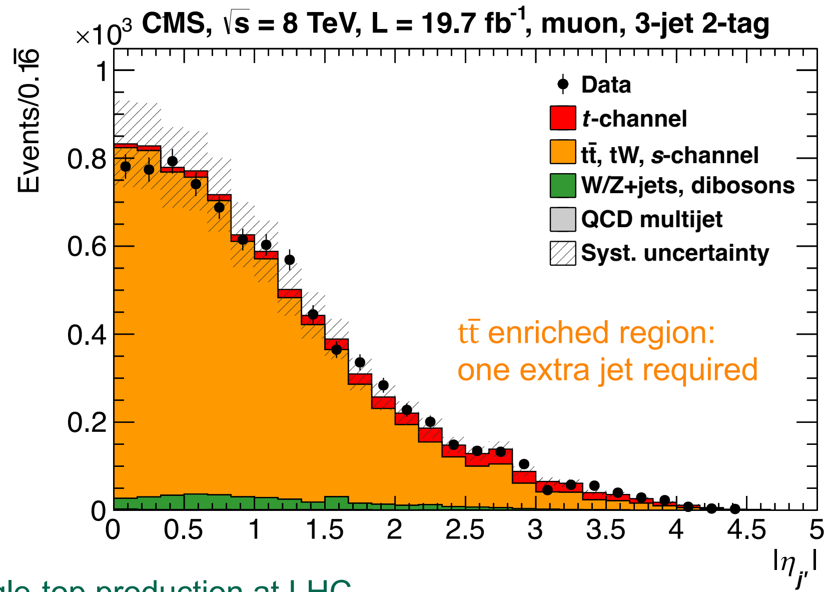
$$L(x, y; \mu, \theta) = L(x; \mu, \theta) L(y; \theta)$$

- Not always the control sample data are available
  - E.g.: calibration from test beam, stored in different formats, control samples analyzed with different software framework…
  - In some cases may be complex and CPU intensive
- Simplest case; assume known PDF for "nominal" value of θ^{nom} (e.g.: estimate with Gaussian uncertainty)

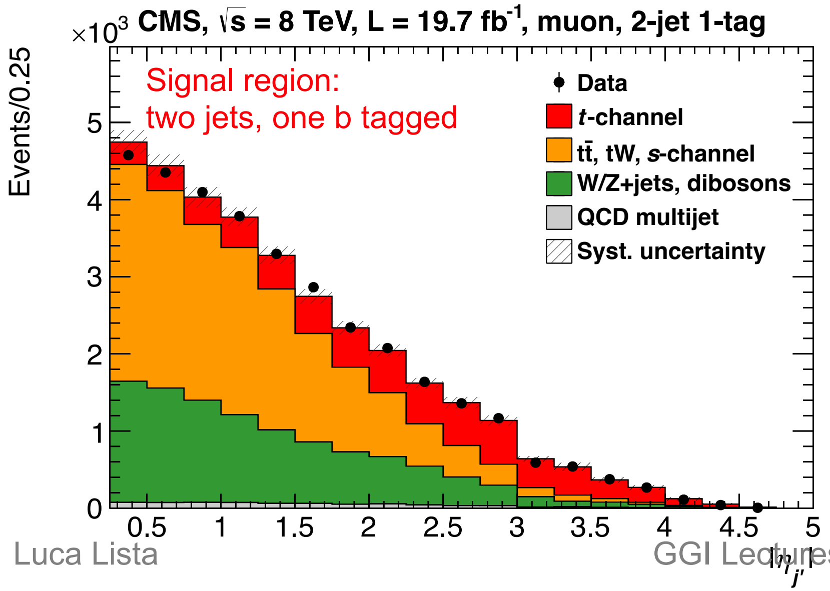
$$L(x, \theta^{\mathrm{nom}}; \mu, \theta) = L(x; \mu, \theta) L(\theta^{\mathrm{nom}}; \theta)$$

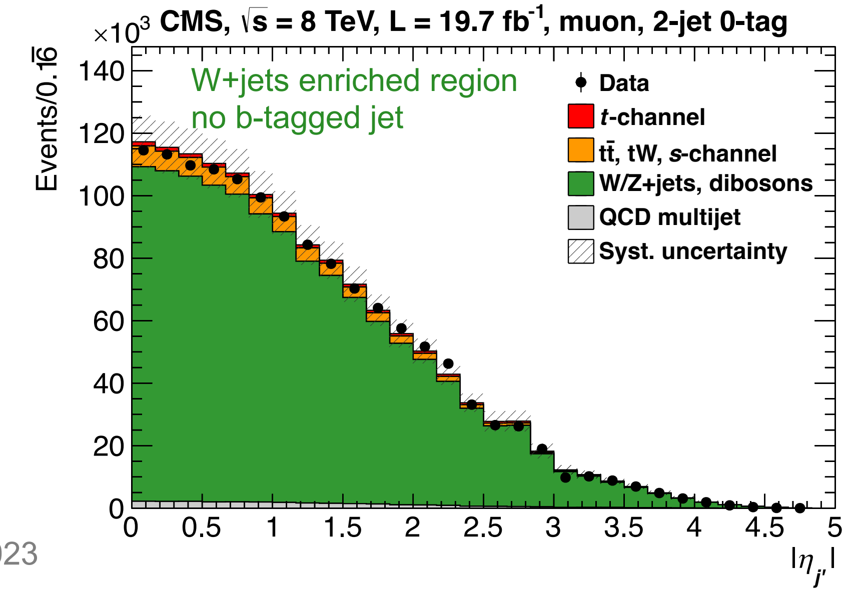
Fitting control regions

- In some cases, background parameters can be constrained from statistically independent control samples
  - Consider possible signal contamination!
- Background yield can be measured in background-enriched regions and extrapolated to signal regions applying scale factors predicted by simulation
- Complete likelihood function = product of likelihood functions in each considered regions, sharing common nuisance parameters
  - Typically: background rates



Measurement of single-top production at LHC





Cousins-Highland hybrid approach

- Method proposed by Cousins and Highland
  - Add posterior from another experiment into the likelihood definition
  - Integrate the likelihood function over the nuisance parameters

$$L^{\text{hybrid}}(x;\mu) = \int L(x;\mu,\theta) L(\theta^{\text{nom}};\theta)\mathrm{d}\theta$$

- Also called “hybrid” approach, because a partial Bayesian approach is implicit in the integration
  - Bayesian integration of PDF, then likelihood used in a frequentist way
- Not guaranteed to provide exact frequentist coverage!
- Numerical studies with pseudo experiments showed that the hybrid CL_s upper limits gives very similar results to Bayesian limit assuming a uniform prior

NIM A320 (1992) 331-335

Profile likelihood

- Define a test statistic based on a likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}$$

Fix μ, fit θ

Fit both μ and θ

- μ is usually the "signal strength" (i.e.: $\sigma/\sigma_{\mathrm{th}}$) in case of a search for a new signal
- Different 'flavors' of test statistics
  - E.g.: deal with unphysical $\mu < 0$, …
- The distribution of $q_\mu = -2 \ln \lambda(\mu)$ may be asymptotically approximated to the distribution of a χ^2 with one degree of freedom (one parameter of interest = μ) due to the Wilks' theorem
  (→ next slide)

Wilks' theorem (1938)

- Consider a likelihood function from N measurements:

$$\prod_{i=1}^{N} L(x_1^i, \cdots, x_n^i; \theta_1, \cdots, \theta_m) = \prod_{i=1}^{N} L(\vec{x}_i; \vec{\theta})$$

- Assume that H_0 and H_1 are two nested hypotheses, i.e.: they can be expressed as:

$$\vec{\theta} \in \Theta_0 \qquad \vec{\theta} \in \Theta_1$$

- Where $\Theta_0 \subseteq \Theta_1$. Then, the following quantity for $N \to \infty$ is distributed as a χ^2 with n.d.o.f. equal to the difference of Θ_0 and Θ_1 dimensionality:

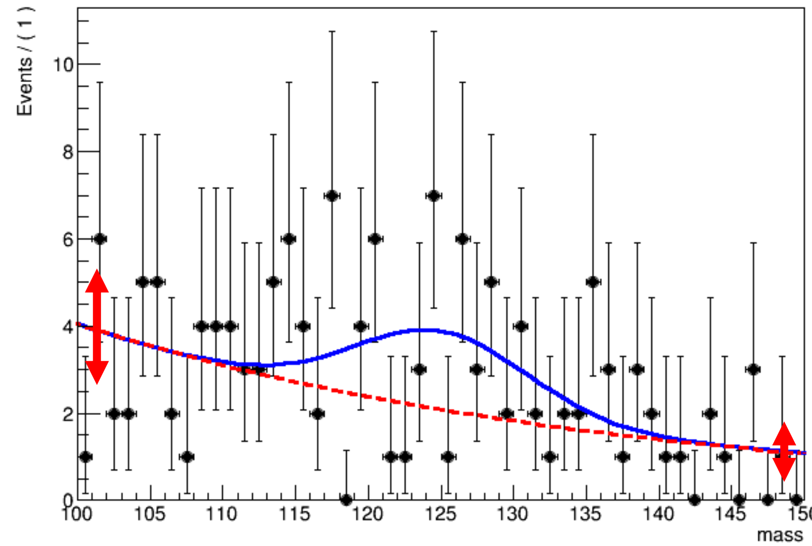
$$\chi_r^2 = -2 \ln \frac{\sup_{\vec{\theta} \in \Theta_0} \prod_{i=1}^{N} L(\vec{x}_i; \vec{\theta})}{\sup_{\vec{\theta} \in \Theta_1} \prod_{i=1}^{N} L(\vec{x}_i; \vec{\theta})}$$

- E.g.: searching for a signal with strength μ, H_0: $\mu = 0$, H_1: $\mu \geq 0$ we have the profile likelihood (supremum = best fit value):

$$\chi_r^2(\mu) = -2 \ln \frac{\sup_{\vec{\theta}} \prod_{i=1}^{N} L(\vec{x}_i; \mu, \vec{\theta})}{\sup_{\mu', \vec{\theta}} \prod_{i=1}^{N} L(\vec{x}_i; \mu', \vec{\theta})}$$

Systematic uncertainties

- Gaussian signal over an exponential background
- Fix all parameters from theory prediction, fit only the signal yield
- Assume a –say– 30% uncertainty on the background yield
- A log normal model may be assumed to avoid unphysical negative yields
  - $b_0 = b\, e^{\beta}$, where our estimate β is known with a Gaussian uncertainty $\sigma_\beta = 0.3$



b_0 = true (unknown) value

b = our estimate

$$L(m; s, \beta) = L_0(m; s, b_0 = be^{\beta}) P(\beta; \sigma_\beta)$$

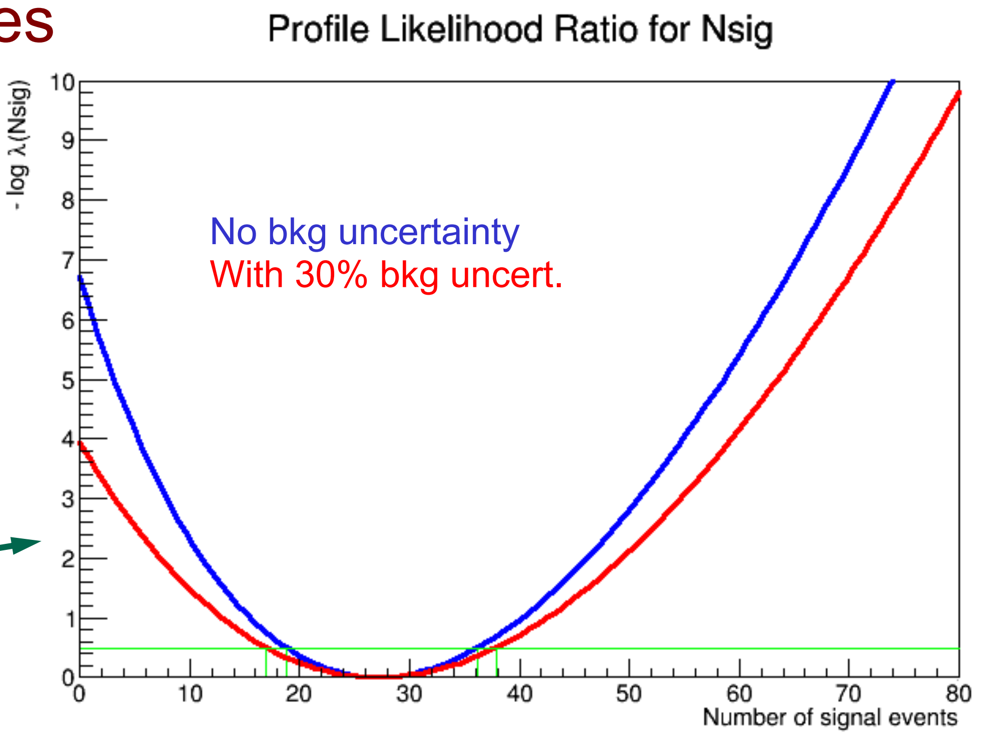
$$L_0(m; s, b_0) = \frac{e^{-(s+b_0)}}{n!}\left(s\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(m-\mu)^2}{2\sigma^2}} + b_0 \lambda e^{-\lambda m} \right)$$

$$P(\beta; \sigma_\beta) = \frac{1}{\sqrt{2\pi}\sigma_\beta} e^{-\frac{\beta^2}{2\sigma_\beta^2}}$$

Systematic uncertainties

- The profile likelihood shape is broadened, with respect to to the usual likelihood function, due to the presence of nuisance parameter β (loss of information) that model systematic uncertainties
- Uncertainty on s increases
- Significance for discovery using s as test statistic decreases

This implementation is based on RooStats, a package, released as optional library with ROOT
http://root.cern.ch



Significance evaluation

- Assume $\mu = 0$, if $q_0 = -2 \ln \lambda(0)$ can be approximated by a χ^2 with one d.o.f., then the significance is approximately equal to:

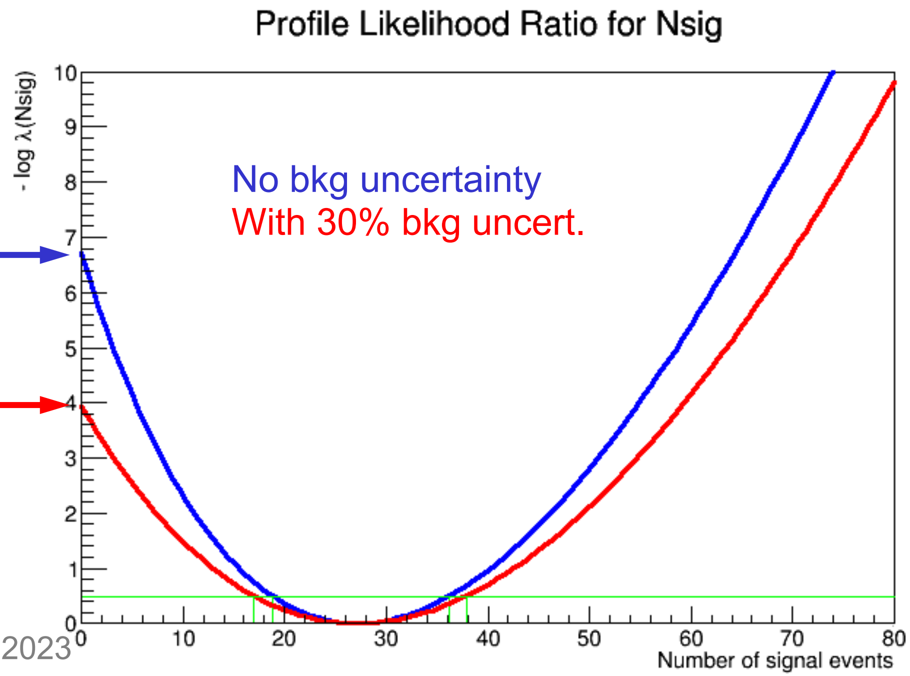
$$Z \cong \sqrt{q_0}$$

- The level of approximation can be verified with a computation done using pseudo experiments:
- Generate a large number of toy samples with zero background and determine the distribution of $q_0 = -2 \ln \lambda(0)$, then count the fraction of cases with values greater than the measured value (p-value), and convert it to Z:

$$Z = \Phi^{-1}(1 - p)$$

$Z \cong \sqrt{2\times 6.66} = 3.66$

$Z \cong \sqrt{2\times 3.93} = 2.81$



- Toy samples may be unpractical for very large Z

Asymptotic approximations

- Asymptotic approximate formulae exist for most of adopted estimators
- If we want to test μ and we suppose data are distributed according to μ', we can write:

$$-2\ln\lambda(\mu) = \frac{(\mu - \hat{\mu})^2}{\sigma_{\hat{\mu}}^2} + \mathcal{O}(1/\sqrt{N})$$

- where $\hat{\mu}$ is distributed according to a Gaussian with average μ' and standard deviation σ (A. Wald, 1943)
- The covariance matrix can be asymptotically approximated by:

$$V_{ij}^{-1} = -\left\langle \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\rangle$$

where μ' is assumed as signal strength value

- Case by case, the estimate of σ (from the inversion of V_{ij}^{-1}) can be determined

A. Wald, Trans. of AMS 54 n.3 (1943) 426-482

G. Cowan et al., EPJ C71 (2011) 1554

Asymptotic approximations

- Under the true hypothesis μ, $\hat{\mu}$ is distributed around μ and the test statistic, neglecting the $\mathcal{O}(1/\sqrt{N})$ term, is distributed according to a χ^2 with one degree of freedom (Wilks' theorem):

$$-2\ln\lambda(\mu) = \frac{(\mu - \hat{\mu})^2}{\sigma_{\hat{\mu}}^2}$$

- If $\hat{\mu}$ is distributed around $\mu' \neq \mu$ the distribution of the test statistic is a non-central χ^2.

Variations on test statistic

G. Cowan et al., EPJ C71 (2011) 1554

- Test statistic for discovery:

$$q_0 = \begin{cases} -2\ln\lambda(0)\,, & \hat{\mu} \geq 0\,, \\ 0\,, & \hat{\mu} < 0\,. \end{cases}$$

  – In case of a negative estimate of μ, set the test statistic to zero: consider only positive μ as evidence against the background-only hypothesis. Approximately: $Z \cong \sqrt{q_0}$.

- Test statistic for upper limits:

$$q_\mu = \begin{cases} -2\ln\lambda(\mu)\,, & \hat{\mu} \leq \mu\,, \\ 0\,, & \hat{\mu} > \mu\,. \end{cases}$$

  – If the estimate is larger than the assumed μ, an upward fluctuation occurred. Don't exclude μ in those cases, hence set the statistic to zero

- Higgs test statistic:

$$\tilde{q}_\mu = \begin{cases} -2\ln\frac{L(\vec{x}|\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(\vec{x}|0,\hat{\hat{\vec{\theta}}}(0))}\,, & \hat{\mu} < 0\,, \\ -2\ln\frac{L(\vec{x}|\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(\vec{x}|\hat{\mu},\hat{\vec{\theta}})}\,, & 0 \leq \hat{\mu} \leq \mu\,, \\ 0\,, & \hat{\mu} > \mu\,. \end{cases}$$

← Protect for unphysical μ<0

← As for upper limits statistic

LEP, Tevatron, LHC Higgs limits

	Test statistic	Profiled?	Test statistic sampling
LEP	$q_\mu = -2\ln\frac{\mathcal{L}(data\|\mu,\tilde{\theta})}{\mathcal{L}(data\|0,\tilde{\theta})}$	no	Bayesian-frequentist hybrid
Tevatron	$q_\mu = -2\ln\frac{\mathcal{L}(data\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data\|0,\hat{\theta}_0)}$	yes	Bayesian-frequentist hybrid
LHC	$\tilde{q}_\mu = -2\ln\frac{\mathcal{L}(data\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data\|\hat{\mu},\hat{\theta})}$	yes $(0 \leq \hat{\mu} \leq \mu)$	frequentist

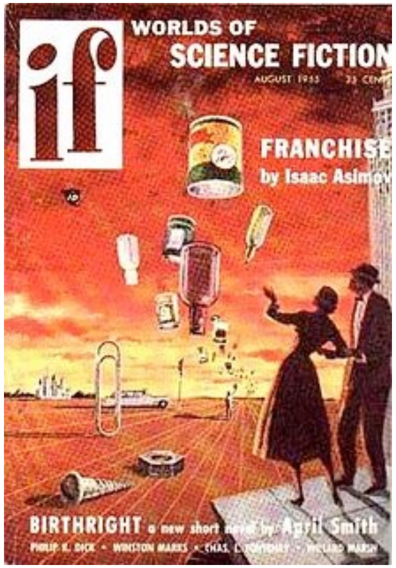
Asimov datasets

- Convenient to compute approximate values:
  *"We define the **Asimov data set** such that when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values"*
- Imagine that our only parameter is μ. We would like to have a dataset where the fit value is the true value μ'. This can be obtained using as number of counts the (non-integer) value $n = \mu' s + b$
- In this case, we have:

$$-2\ln\lambda(\mu) \cong \frac{(\mu-\mu')^2}{\sigma_{\hat{\mu}}^2}$$

- Therefore we can estimate the variance of $\hat{\mu}$ to be used in Wald's approximation of the test statistic:

$$\sigma_{\hat{\mu}}^2 \cong -\frac{(\mu-\mu')^2}{2\ln\lambda(\mu)}$$



G. Cowan et al., EPJ C71 (2011) 1554

Asimov datasets

- In practice: all observables are replaced with their expected value
- Yields expected values are possibly non integer:

$$\lambda_{\mathrm{A}}(\mu) = \frac{L_{\mathrm{A}}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_{\mathrm{A}}(\hat{\mu}, \hat{\boldsymbol{\theta}})} = \frac{L_{\mathrm{A}}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_{A}(\mu', \boldsymbol{\theta})}$$

- The variance of the test statistic, in Wald's approximation, is estimated as:

$$\sigma_{\hat{\mu}}^2 \cong \frac{(\mu - \mu')^2}{-2 \ln \lambda_A}$$

- Median significance for discovery or exclusion (and their $\pm 1\sigma$ bands) can be obtained using the Asimov dataset

$$\mathrm{med}[Z_0|\mu'] = \sqrt{q_{0,\mathrm{A}}}$$ ← For discovery using q_0

$$\mathrm{med}[Z_\mu|0] = \sqrt{q_{\mu,\mathrm{A}}}$$ ← For upper limit using q_μ

$$\mathrm{med}[Z_\mu|0] = \sqrt{\tilde{q}_{\mu,\mathrm{A}}}$$ ← Upper limits using $\tilde{q}_\mu$

In practice: all the interesting formulae are implemented in RooStats package, released as optional library in ROOT

G. Cowan et al., EPJ C71 (2011) 1554

The look-elsewhere effect

- Consider a search for a signal peak over a background distribution that is smoothly distributed over a wide range
- You could either:
  - Know which mass to look at, e.g.: search for a rare decay with a known particle, like $B_s \rightarrow \mu\mu$
  - Search for a peak at an unknown mass value, like for the Higgs boson
- In the former case it's easy to compute the peak significance:
  - Evaluate the test statistics for $\mu = 0$ (background only) at your observed data sample
  - Evaluate the p-value according to the expected distribution of your test statistic q under the background-only hypothesis, convert it to the equivalent area of a Gaussian tail to obtain the significance level:

$$p = \int_{q^{\mathrm{obs}}}^{\infty} f(q|\mu = 0)\mathrm{d}q \qquad Z = \Phi^{-1}(1-p)$$

The look-elsewhere effect

- In case you search for a peak at an unknown mass, the previous p-value has only a local meaning:
  - Probability to find a background fluctuation as large as your signal or more at a fixed mass value m:

$$p(m) = \int_{q^{\mathrm{obs}}(m)}^{\infty} f(q|\mu = 0)\mathrm{d}q$$

  - We need the probability to find a background fluctuation at least as large as your signal at **any** mass value (global)
  - local p-value would be an overestimate of the global p-value
- The chance that an over-fluctuation occurs on at least one mass value increases with the searched range
- Magnitude of the effect:
  - Roughly proportional to the ratio of resolution over the search range, also depending on the significance of the peak
  - Better resolution = less chance to have more events compatible with the same mass value
- Possible approach: let also m fluctuate in the test statistics fit:

$$\hat{q}_0 = -2\ln\frac{L(\mu = 0)}{L(\hat{\mu};\hat{m})}$$

Note: for $\mu=0$
L doesn't depend on m
Wilks' theorem doesn't apply

$$p^{\mathrm{glob}} = \int_{\hat{q}_0^{\mathrm{obs}}}^{\infty} f(\hat{q}_0|\mu = 0)\mathrm{d}\hat{q}_0$$

Estimate LEE

- The effect can be evaluated with brute-force Toy Monte Carlo:
  - Run N experiments with background-only
  - Find the maximum $\hat{q}$ of the test statistic q in the entire search range
  - Determine its distribution, hence compute the observed global p-value
  - Requires very large toy Monte Carlo samples (5σ: $p = 2.87 \times 10^{-7}$)

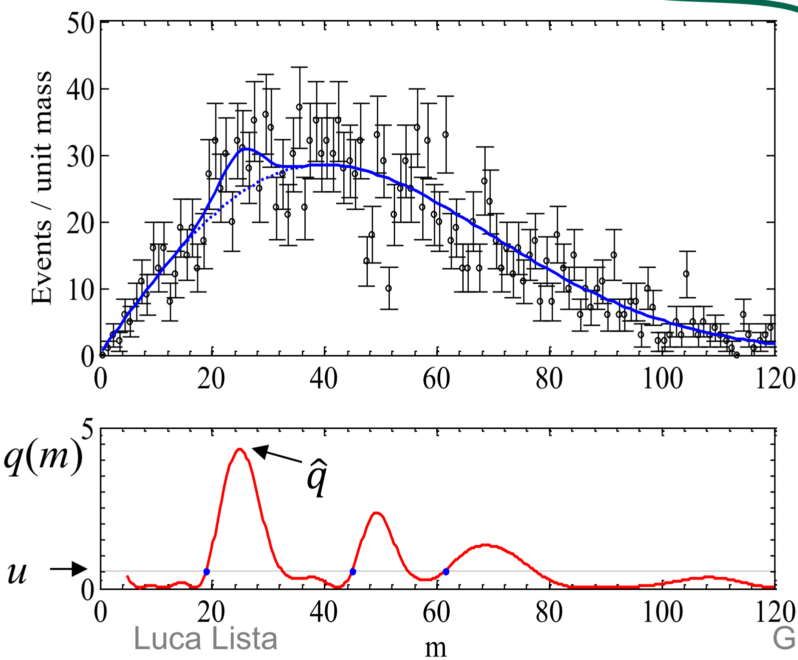
$$\hat{q} = \max_m q(m)$$

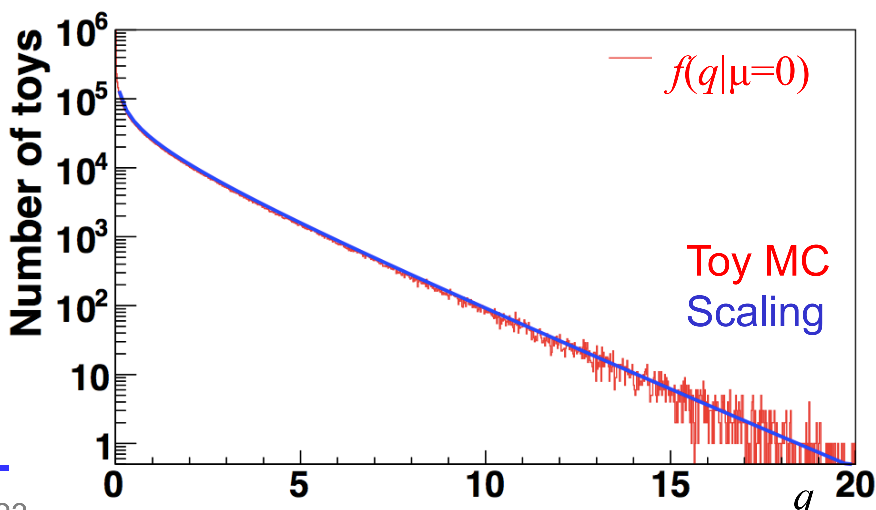
- Approximate evaluation based on local p-value, times correction factors ("trial factors", Gross and Vitells, EPJC 70:525-530,2010)

$$p^{\mathrm{glob}} = P(\hat{q} > u) \leq \langle N_u \rangle + \frac{1}{2} P(\chi^2 > u)$$

$\langle N_u \rangle$ is the average number of up-crossings of the test statistic, can be evaluated at some lower reference level (toy MC) and scaled by:

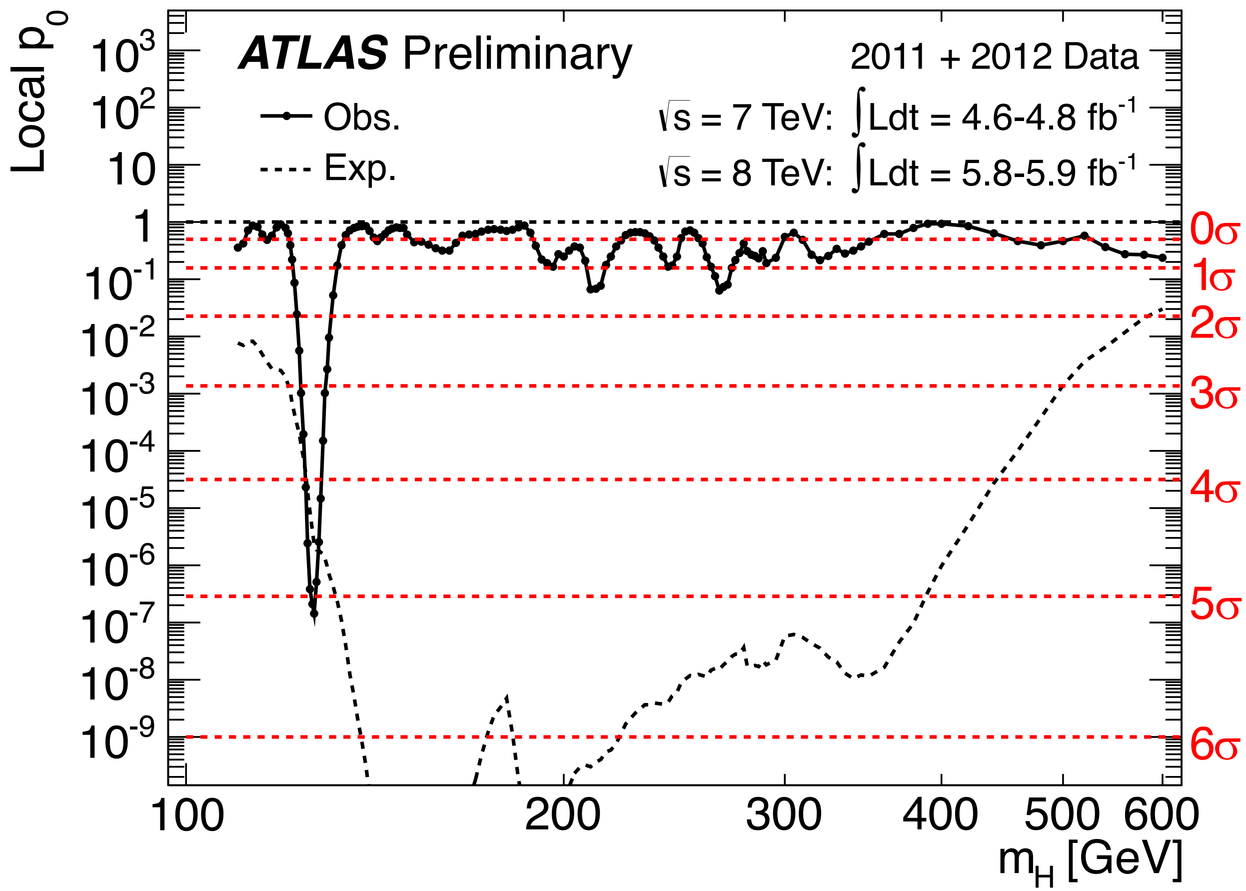
$$\langle N_u \rangle = \langle N_{u_0} \rangle \, e^{-\frac{u-u_0}{2}}$$





Practical application

- Higgs search at ATLAS

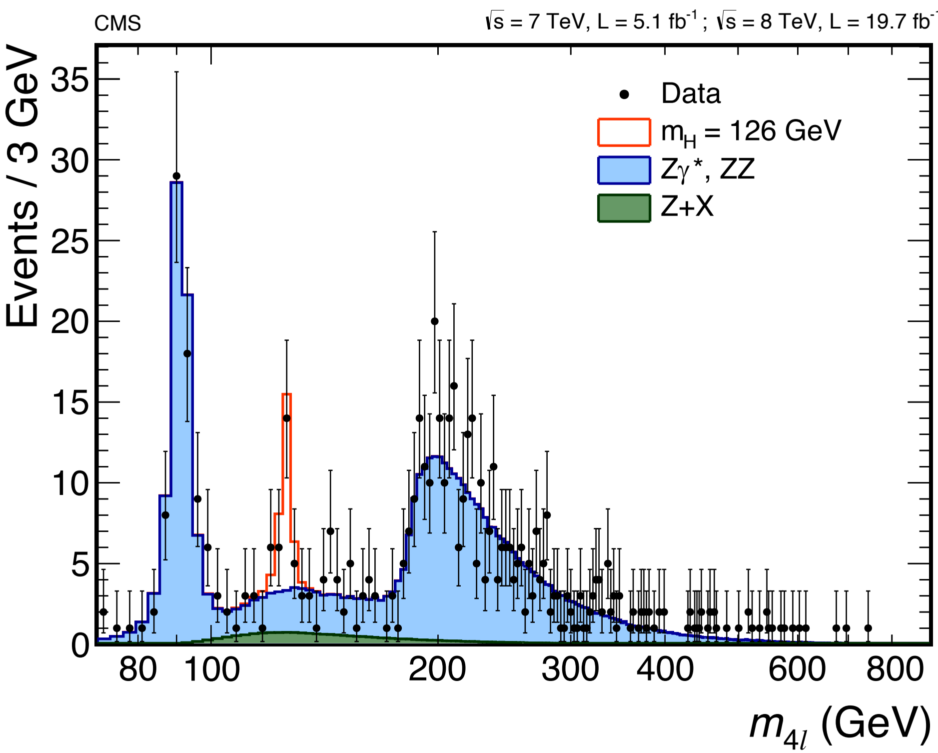


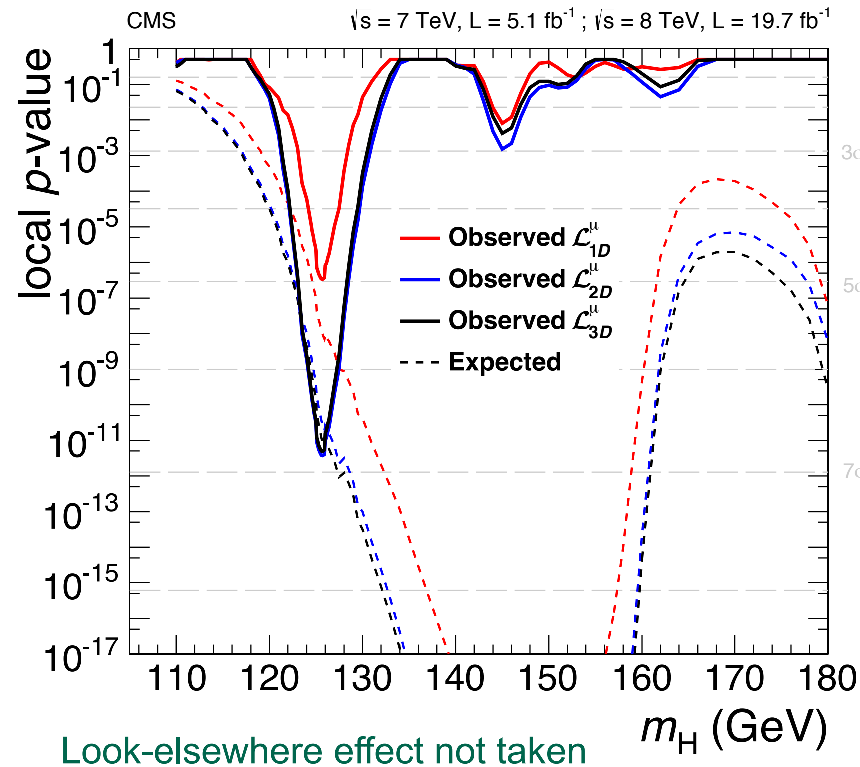
Practical application

- Use the number of σs Z as test statistic: $u = Z^2$ behaves as a chi-square
- Use the 0σ level ($p = 0.5$) as level u^0, then extrapolate to the minimum p value, where $Z \cong 5$, i.e.: $u = Z^2 \cong 5^2 = 25$
- The number of upcrossing can be counted from the plot, and is equal to $N_0 = 9$, which allows us to estimate: $\langle N_0 \rangle = 9 \pm 3$
- Estimate the global p-value as:
- $p^{glob} \cong \langle N_u \rangle + \frac{1}{2} P(\chi^2 > u) \cong \langle N_u \rangle + 3\times 10^{-7}$
- $\langle N_u \rangle \cong \langle N_0 \rangle e^{-(5^2 - 0^2)/2}$
- $\langle N_u \rangle \cong (9 \pm 3) e^{-25/2} \cong (3 \pm 1)\times 10^{-5}$
- $p^{glob} \cong 3\times 10^{-5} + 3\times 10^{-7} \cong 3\times 10^{-5} \implies Z \cong 4\sigma$ instrad of 5σ

Putting all together

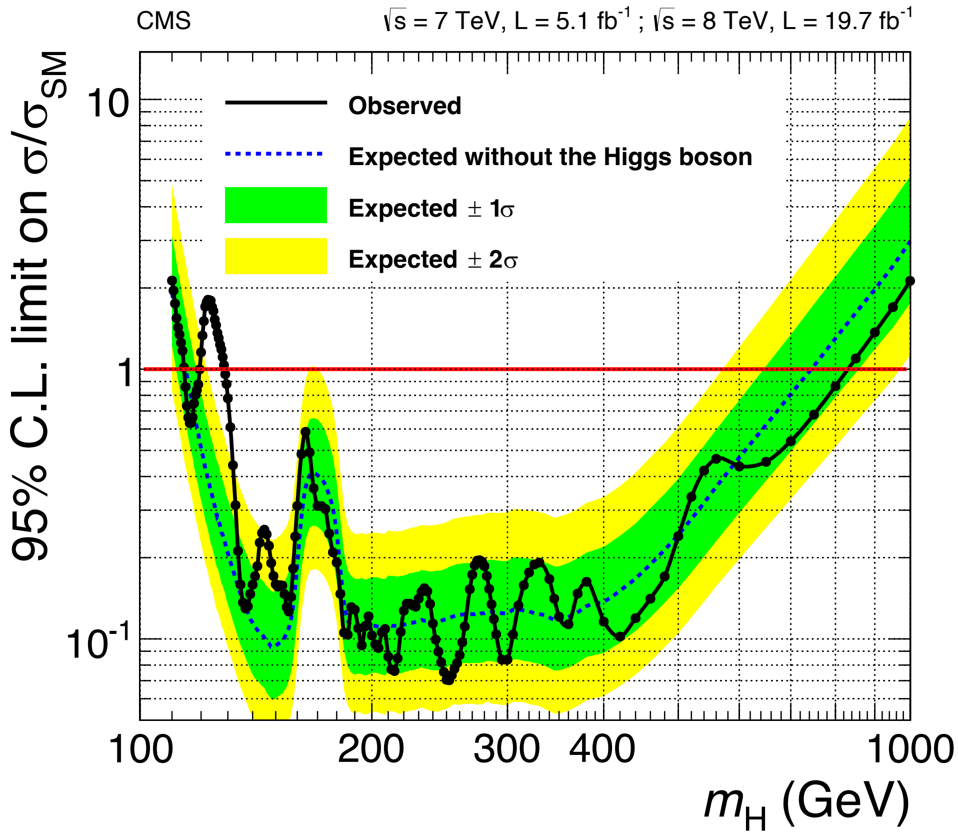
- Search for Higgs boson in H→4l at LHC
- 1D, 2D, 3D: different test statistics using 4l invariant mass plus other discriminating variables based on the event kinematics

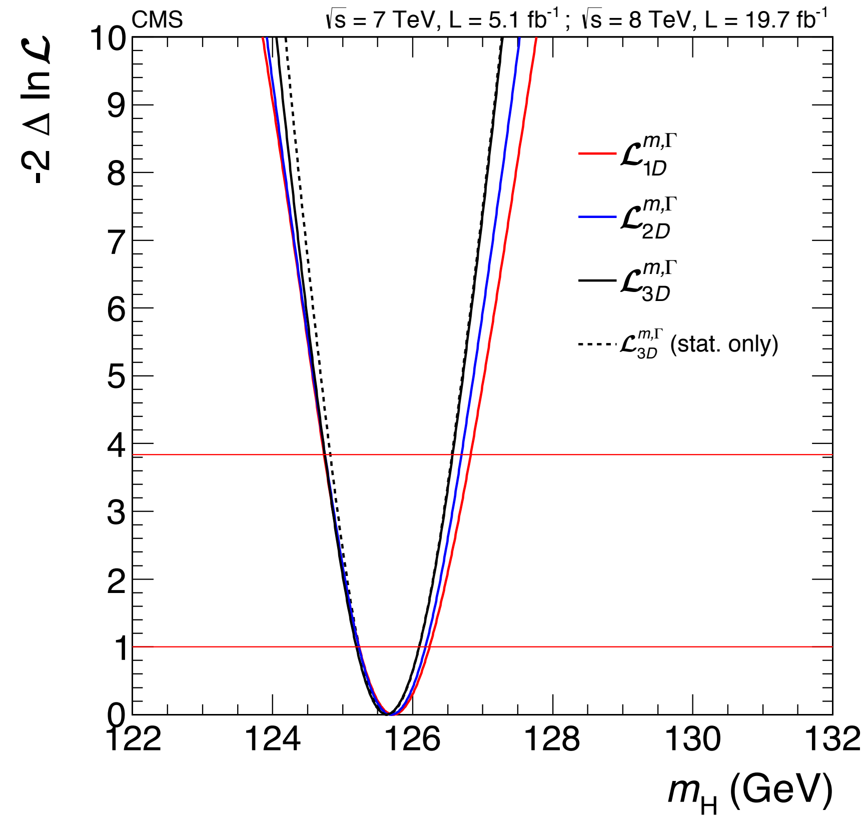




Look-elsewhere effect not taken into account here

Higgs exclusion





"The modified frequentist construction CLs is adopted as the primary method for reporting limits. As a complementary method to the frequentist construction, a Bayesian approach yields consistent results."

Agreed statistical procedure described in:
ATLAS and CMS Collaborations,
LHC Higgs Combination Group
ATL-PHYS-PUB 2011-11/CMS NOTE 2011/005, 2011.