Illuminating the dark side of statistics: Bayesian inference in particle physics

Lorenz Gärtner[1]

[1] *LMU Munich*

March 10, 2025

About me

- BSc @ University of Manchester
  Physics with theoretical physics
- MSc @ LMU Munich
  More theory . . .

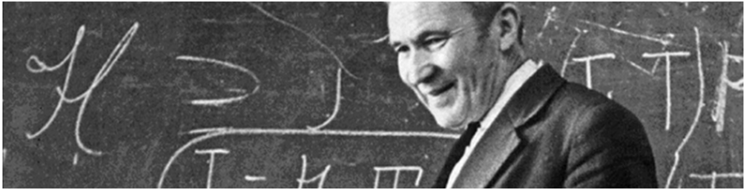
— — — almost no stats — — —

- Currenly PhD @ LMU Munich
  A lot of stats

➜ Never too late to start

What is a probability?

Kolmogorov probability axioms



1. $p(\Omega) = 1$, where Ω is the sample space.
2. $p(x) \geq 0$ for any event $x \subseteq \Omega$.
3. For any sequence of disjoint events $x_1, x_2, \ldots$,

$$p\left(\bigcup_{i=1} x_i\right) = \sum_{i=1} p(x_i)$$

Probability interpretations

Axioms:

old probabilities ➜ new probabilities

To assign probabilities we need probability interpretations.

The interpretations **share the same mathematical framework**, but the meaning of $p(x)$ is different.

Frequentist interpretation

Assign a probability as relative frequency

$$p(x) = \lim_{N \to \infty} \frac{N_x}{N}$$

- "The data is random."
- For repeatable experiments only.
- No probabilities for single events.

Bayesian interpretation

Assign a probability $p(x)$ as *degree of belief*.

- "Parameters are random."
- Inference results are subjective.

Bayes' theorem

The **posterior** is

$$p(\text{theory}|\text{data}) = \frac{p(\text{data}|\text{theory})p(\text{theory})}{p(\text{data})}$$

- **Likelihood** $p(\text{data}|\text{theory})$
- **Prior** $p(\text{theory})$
- **Marginal likelihood**
  $p(\text{data}) = \int p(\text{data}|\text{theory})p(\text{theory})$

Can you derive it?

[...] *nearly all physicists tend to misinterpret frequentist results as statements about the theory given with the data.*

Presentation of search results: the CL_s technique,
A. L. Read

Bayesian beliefs

$$p(SM|data)$$

$$p(Higgs|data)$$

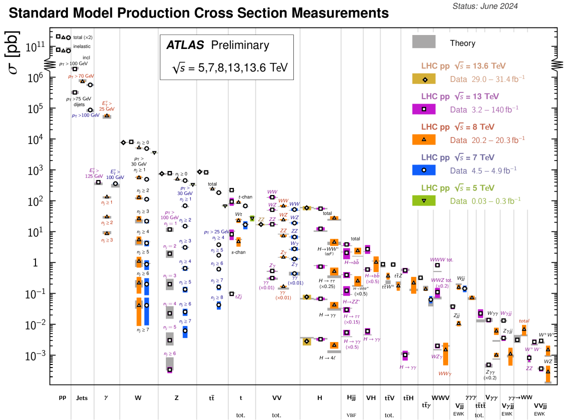
$$p(SUSY|data)$$

Bayesian beliefs

$p(SM|data)$

$p(Higgs|data)$

$p(SUSY|data)$

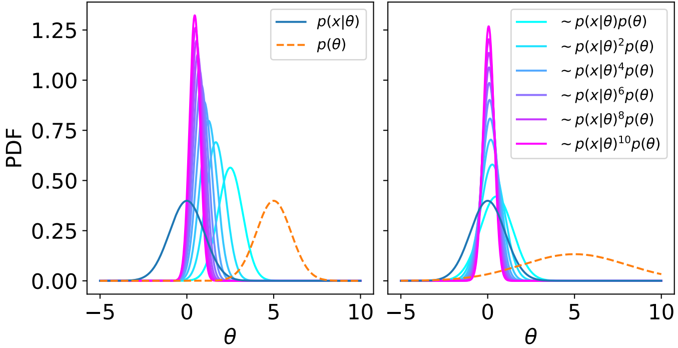


ATLAS

Bayesian updating

Can generally use measurements to *update* our posterior

$$p(\theta|x_1, x_2) = \frac{p(x_2|\theta)}{p(x_2)} p(\theta|x_1) = \frac{p(x_2|\theta)}{p(x_2)} \frac{p(x_1|\theta)}{p(x_1)} p(\theta).$$

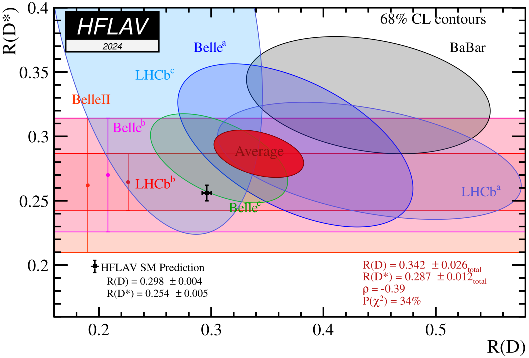


Bayesian updating in real life

PDFs for parameters are manifestly Bayesian.
Combinations of $\mu \pm \sigma$ assume an underlying PDF for μ.

$$\mathcal{R}(D) = \frac{\mathcal{B}(\bar{B} \to D\tau^- \bar{\nu}_\tau)}{\mathcal{B}(\bar{B} \to D l^- \bar{\nu}_l)}$$

$$\mathcal{R}(D^*) = \frac{\mathcal{B}(\bar{B} \to D^*\tau^- \bar{\nu}_\tau)}{\mathcal{B}(\bar{B} \to D^* l^- \bar{\nu}_l)}$$



Actual combinations a bit more involved ➜ HFLAV 2024.

The common ground

Frequentist inference is based on

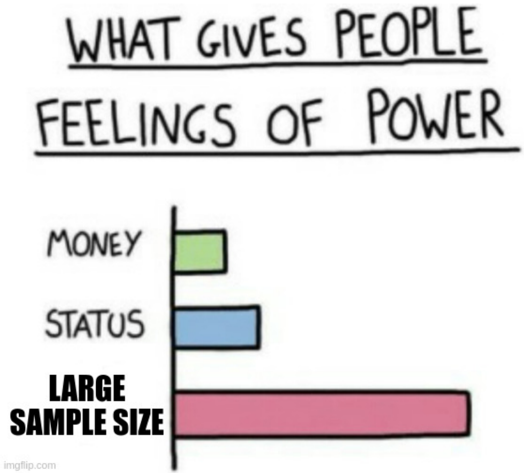
$$\boxed{p(x|\theta)}$$

Bayesian inference is based on

$$p(\theta|x) = \frac{\boxed{p(x|\theta)}\, p(\theta)}{p(x)}$$

You want the best possible $\boxed{p(x|\theta)}$!

The best possible model...



Physics does not care...

... about our interpretation

For many inference problems,

$$\text{results} \mid \text{frequentist} \approx \text{results} \mid \text{Bayesian},$$

even though they answer different questions.

Parameter inference

Parameter inference

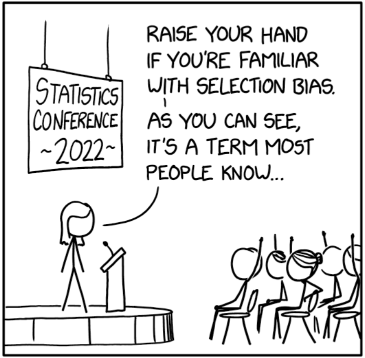
Point estimates
Identify the most probable parameter point.

Interval estimation
Identify extended regions in parameter space based on compatibility with the data.

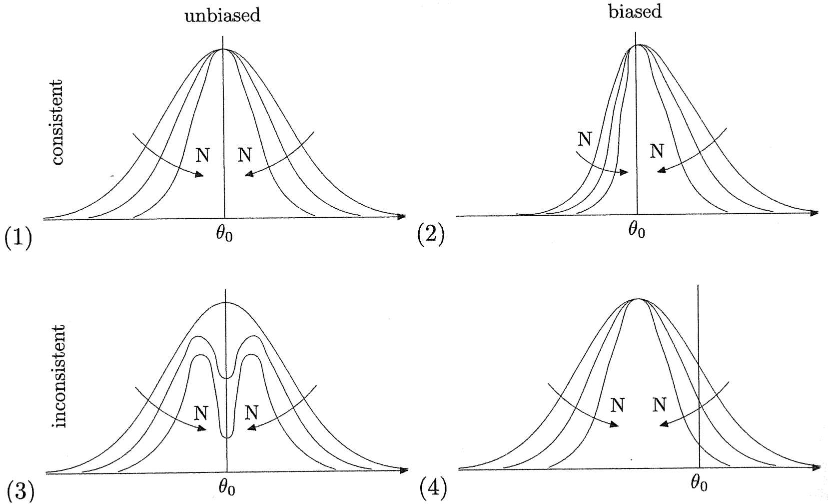
Frequentist point estimates: estimators

Estimator is a *statistic* $\hat{\theta}(x)$, with desired properties

- **consistency**
  $\lim_{N_x \to \infty} E_x[\hat{\theta}] = \theta_{true}$
- **unbiasedness**
  $b = E_x[\hat{\theta}] - \theta_{true}$
- **efficiency**
  $$V(\hat{\theta}) = I(\theta)^{-1} = E_x\left[\left(\frac{\partial \ln p(x|\theta)}{\partial \theta}\right)^2\right]^{-1}$$
- ...



A good data set is the key to success ...



Method of maximum likelihood

We find maximum likelihood estimators $\hat{\theta}$ by solving

$$\boxed{\hat{\theta} = \text{argmax}_\theta \, p(x|\theta)}$$

It has the property

$$\lim_{N\to\infty} p\left(\sqrt{N}(\hat{\theta} - \theta_{true})\right) = \mathcal{N}\left(0, I^{-1}(\theta)\right),$$

implying consistency, asymptotic unbiasedness and efficiency.

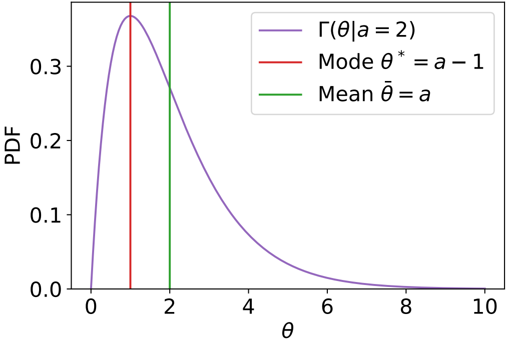
Bayesian point estimates

Mode
Value of θ with maximum a-posteriori probability

$$\theta^* = \mathrm{argmax}_\theta \, p(\theta|x)$$

Mean
Expected value of θ under the posterior

$$\bar{\theta} = E_{p(\theta|x)}[\theta]$$



Bayesian point estimates

Mode
Value of θ with maximum a-posteriori probability

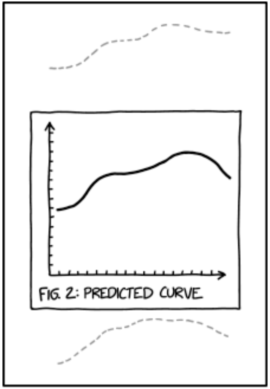
$$\theta^* = \text{argmax}_\theta\, p(\theta|x)$$

Note

$$0 = \left.\frac{\partial p(\theta|x)}{\partial\theta}\right|_{\theta=\theta^*} \propto \left(\frac{\partial p(x|\theta)}{\partial\theta} p(\theta) + p(x|\theta)\frac{\partial p(\theta)}{\partial\theta}\right)\Bigg|_{\theta=\theta^*}$$

$$\implies \theta^* = \hat{\theta} \quad \text{if} \quad \left.\frac{\partial p(\theta)}{\partial\theta}\right|_{\theta=\theta^*} = 0$$

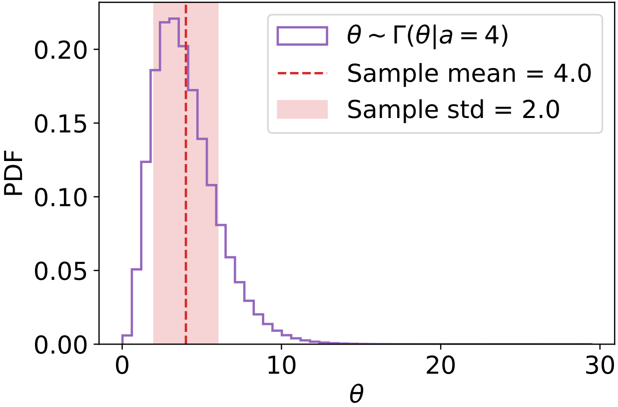
Intervals and limits



SCIENCE TIP: IF YOUR MODEL IS BAD ENOUGH, THE CONFIDENCE INTERVALS WILL FALL OUTSIDE THE PRINTABLE AREA.

Intervals and limits

If an estimator PDF is not Normal, $\hat{\theta} \pm \sigma_\theta$ is meaningless.



Frequentist intervals

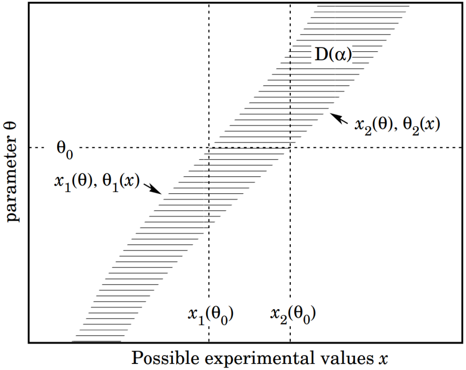
Neyman confidence belt

$$\int_{x_1}^{x_2} dx\, p(x|\theta) = 1 - \alpha$$

Not unique ➜ *central interval*

$$\int_{-\infty}^{x_1} dx\, p(x|\theta) = \int_{x_2}^{\infty} dx\, p(x|\theta) = \alpha/2$$

or *upper/lower interval*



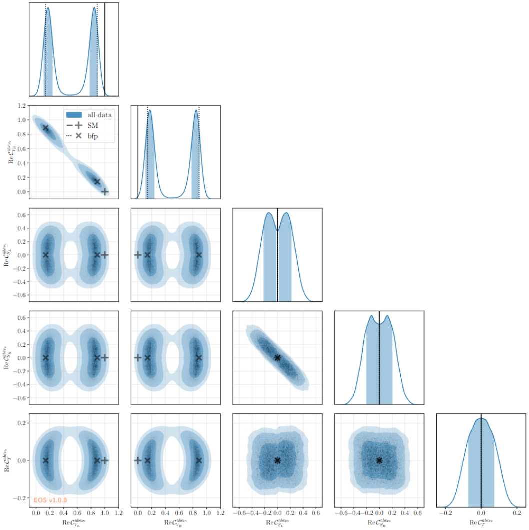
Bayesian intervals

Credible intervals $[\theta_1, \theta_2]$ cover $1 - \alpha$ of the posterior

$$\int_{\theta_1}^{\theta_2} d\theta \, p(\theta|x) = 1 - \alpha$$

- For upper/lower limits: set θ_1 or θ_2 to boundary
- *Smallest possible interval*
  ➜ Highest (posterior) density intervals (HDI)

$b \to ul^-\bar{\nu}$ in the Weak Effective Theory



- Corner plots are great for visualization.
- Posterior for Wilson coefficients
- Identify modes, credible intervals

arXiv:2302.05268v2 [hep-ph]

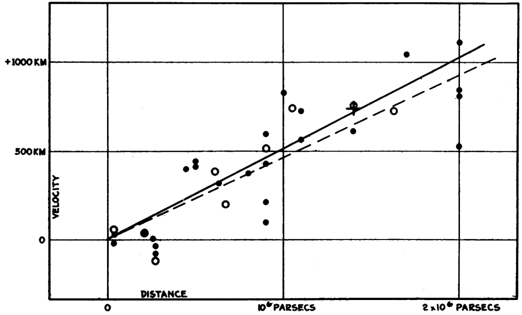
Nuisance parameters and priors

Nuisance parameters

Models are not perfect
➜ **systematic bias**

Solution:
Nuisance parameters ν,

$$p(x|\psi, \nu)$$



Hubble 1929

Usually, we want to constrain nuisance parameters, but in the frequentist language **everything is data**.

Frequentist "priors"

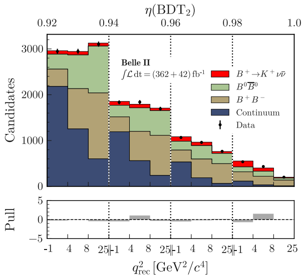
Constrain nuisance parameters using *auxiliary data a*,

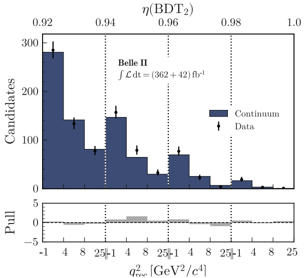
$$p(x|\psi,\nu)p(a|\nu)$$

$p(a|\nu)$ represents *degree of belief* in ν.

Often *auxiliary data* is *created* to match our desired constraint term.

Belle II 2024





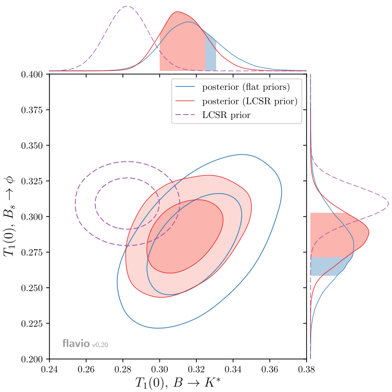
Bayesian nuisance parameters

Priors can be defined using auxiliary data

$$p(\nu|a) \propto p(a|\nu)p_0(\nu)$$

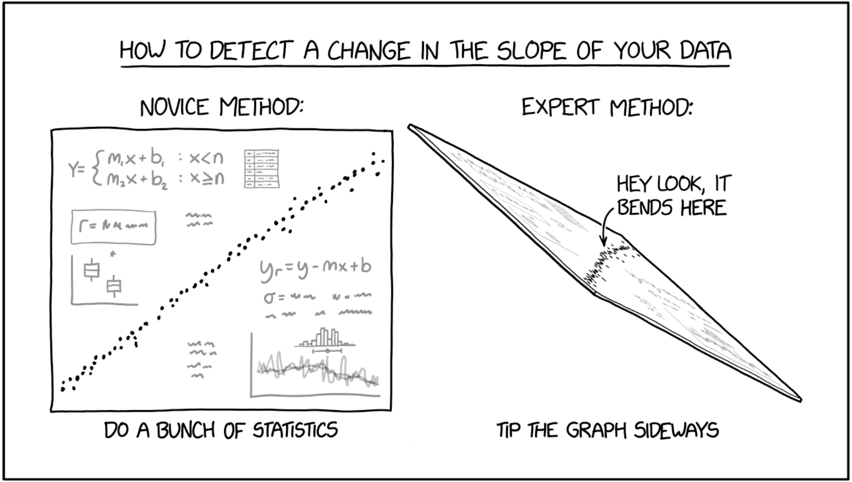
Only in the Bayesian case, other prior choices are also allowed

$$p(\nu) = \mathcal{N}(\nu|\nu_0, \sigma_\nu^2)$$



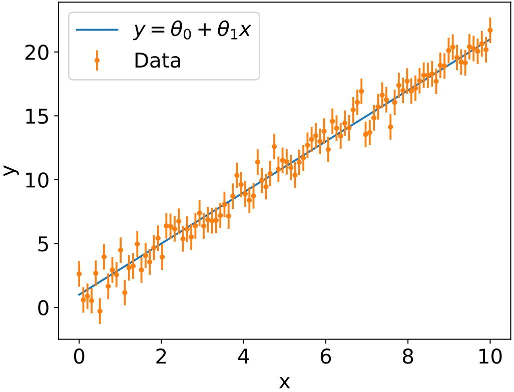
Paul 2017

A simple linear model



A simple linear model

- Our independent data : $\boldsymbol{X} = (x_i, y_i, \sigma_i)$



A simple linear model

- Our independent data: $\boldsymbol{X} = (x_i, y_i, \sigma_i)$
- Our model:

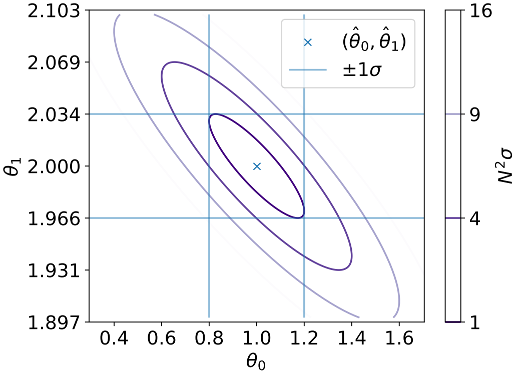
$$p(\boldsymbol{X}|\theta_0, \theta_1) = \prod_{x_i, y_i, \sigma_i \in \boldsymbol{X}} \mathcal{N}(y_i | \mu(x_i|\theta_0, \theta_1), \sigma_i^2)$$

$$\mu(x_i|\theta_0, \theta_1) = \theta_0 + \theta_1 x_i$$

➜ We want to know about θ_0, do not care about θ_1.

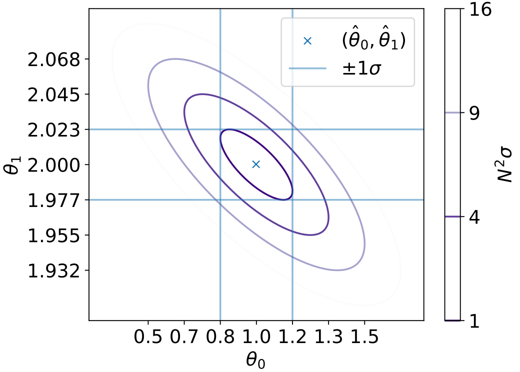
Unconstrained likelihood

$$-2\log p(\boldsymbol{X}|\theta_0,\theta_1) = \sum_{x_i,y_i,\sigma_i\in\boldsymbol{X}} \frac{(y_i-\mu(x_i|\theta_0,\theta_1))^2}{\sigma_i^2}$$



Including a measurement of θ_1: t_1, σ_{t_1}

$$-2\log p(\boldsymbol{X}|\theta_0,\theta_1) = \sum_{x_i,y_i,\sigma_i\in\boldsymbol{X}} \frac{(y_i-\mu(x_i|\theta_0,\theta_1))^2}{\sigma_i^2} + \frac{(\theta_1-t_1)^2}{\sigma_{t_1}^2}$$

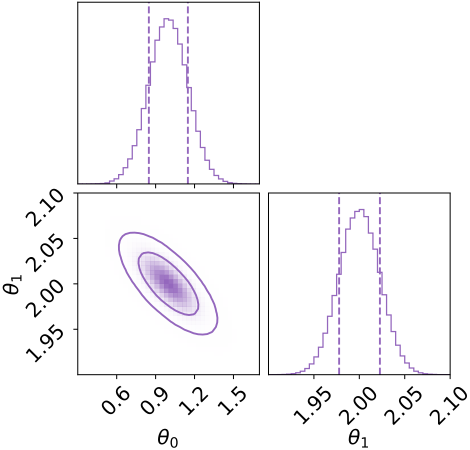


Posterior

$$p(\theta_0, \theta_1 | \boldsymbol{X}) \propto p(\boldsymbol{X} | \theta_0, \theta_1) p(\theta_0) p(\theta_1)$$

$$p(\theta_0) = \text{Uniform}(0, 2)$$

$$p(\theta_1) = \mathcal{N}(\theta_1 | t_1, \sigma_{t_1}^2)$$

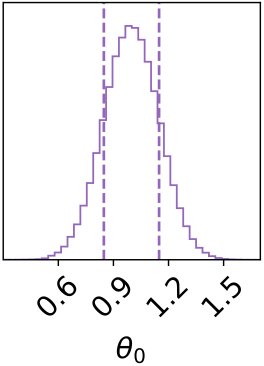


Marginal posterior

$$p(\theta_0|\boldsymbol{X}) = \int d\theta_1 \, p(\theta_0, \theta_1|\boldsymbol{X}) = \mathcal{N}(\theta_0|\theta_0^*, \sigma_{\theta_0})$$

In this example, we get

- $\theta_0^* = \hat{\theta}_0$
- 68% HDI = $\hat{\theta}_0 \pm \sigma_{\theta_0}$



MCMC

The hard part ...

- We only want the posterior for ψ alone.
- Remove nuisance parameters by integrating over ν.

➜ The *marginal posterior* is

$$p(\psi|x) = \int d\nu \, p(\psi, \nu|x)$$

- Commonly a high dimensional integral
  ➜ compute with Monte Carlo methods.

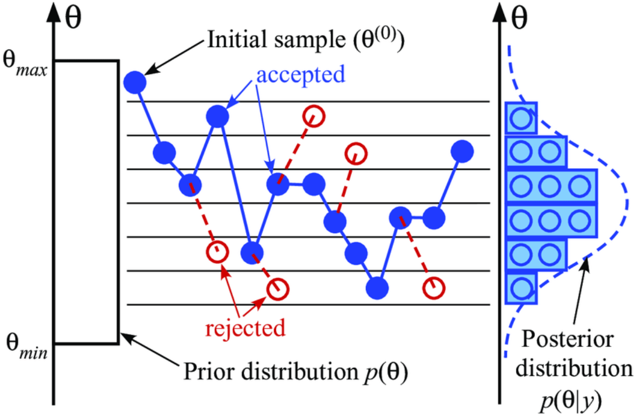
Markov chain

A sequence of events, where probability of the next state depends solely on the current state

$$\ldots \to \theta_i \sim g(\theta_i|\theta_{i-1}) \to \theta_{i+1} \sim g(\theta_{i+1}|\theta_i) \to \ldots$$

for some *proposal distribution* g.

Markov Chain Monte Carlo (MCMC)



Metropolis-Hastings

We loop

1. Generate $\theta \sim g(\theta|\theta_i)$
2. Update

$$\theta_{i+1} = \begin{cases} \theta & u \leq \min\left(1, \frac{p(\theta)g(\theta|\theta_i)}{p(\theta_i)g(\theta_i|\theta)}\right) \\ \theta_i & \text{otherwise} \end{cases}$$

where $u \sim \text{Uniform}(0, 1)$

Note: Need to define a proposal distribution $g(\theta|\theta_0)$.

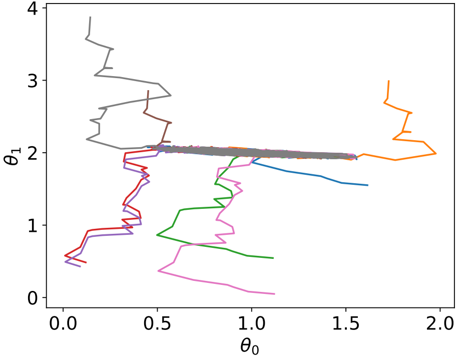
Chains

In MCMC we generate a sequence

$$\theta_0 \to \theta_1 \to \theta_2 \to \dots$$

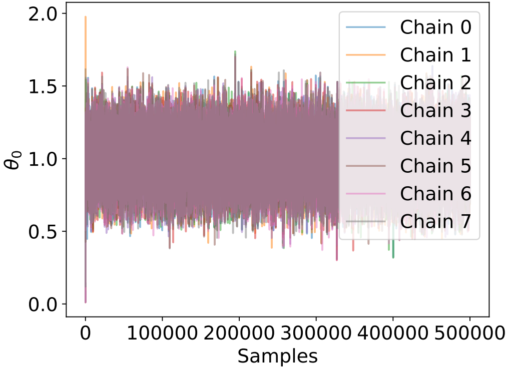
Only one start can land you in local minima.

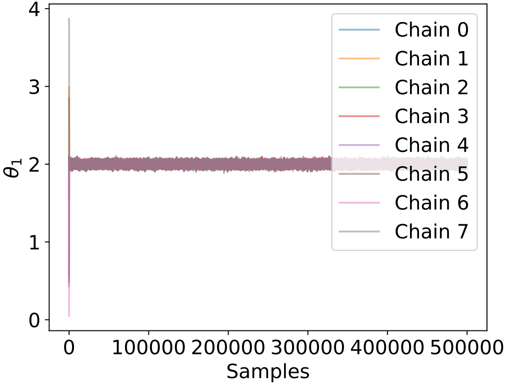
$$\theta_0^0 \to \theta_1^0 \to \theta_2^0 \to \dots$$
$$\theta_0^1 \to \theta_1^1 \to \theta_2^1 \to \dots$$
$$\theta_0^2 \to \theta_1^2 \to \theta_2^2 \to \dots$$
$$\dots$$



Convergence

Trace plots are a useful convergence diagnostic



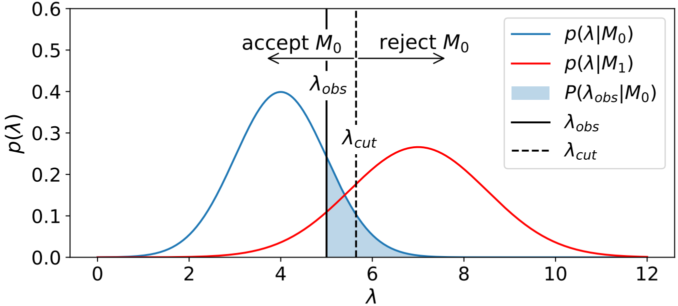


... but one can become more fancy.

Model comparison

Point-wise: P-values

$$P(\lambda_{obs}|M_0) = \int_{\lambda_{obs}}^{\infty} d\lambda \, p(\lambda|M_0), \quad \lambda = -2 \ln \frac{p(x|\hat{\theta}, M_0)}{p(x|\hat{\theta}, M_1)} \; ^{\dagger}$$



† Likelihood ratio = optimal test statistic ➜ Newman-Pearson lemma

Averaged: Bayes factor

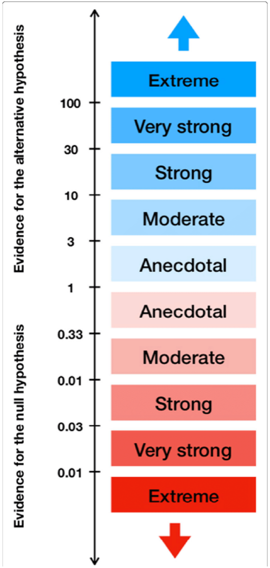
Compare the probabilities of the observed data being produced by a given model.

$$p(\theta|x,M) = \frac{p(x|\theta,M)\,p(\theta|M)}{\color{red}{p(x|M)}}$$

$$\color{red}{p(x|M)} = \int d^n\theta\, p(x|\theta,M)\,p(\theta|M)$$

$$B = \frac{p(x|M_1)}{p(x|M_0)}$$

Do you see a potential hazard?



$b \to u l^- \bar{\nu}$ in the Weak Effective Theory

fit model M	χ^2	d.o.f.	p value [%]	$\ln Z(M)$
	goodness of fit			
SM	44.18	48	63.03	372.5 ± 0.4
CKM	43.75	47	60.78	372.4 ± 0.4
WET	36.13	43	76.17	376.5 ± 0.4

Table 1. Goodness-of-fit values for the three main fits conducted as part of this analysis. We provide $\chi^2 = -2 \ln P(\text{data} \,|\, \vec{x}^*)$ at the best-fit point $\vec{x}^*$ next to the p value and the natural logarithm of the evidence $\ln Z$. We find that the p values associated with each individual likelihood are larger than 42%.

$$B = \exp(\ln Z(WET) - \ln Z(SM)) = 54.6$$

arXiv:2302.05268v2 [hep-ph]

Tools to try

BAT
BAYESIAN ANALYSIS TOOLKIT

PyMC

ArviZ