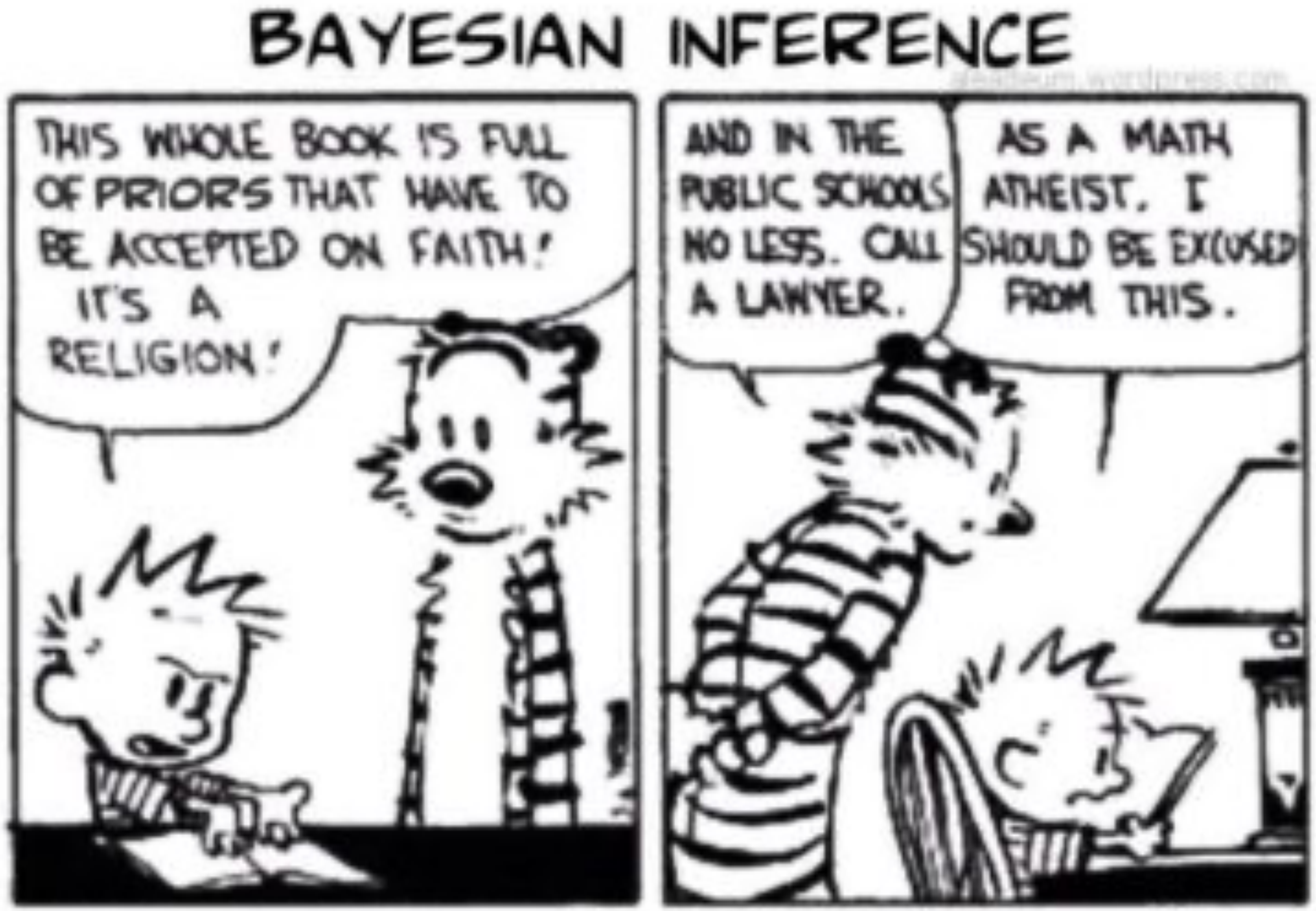
Data Analysis and Bayesian Methods
Lecture 4

Maurizio Pierini

mPP

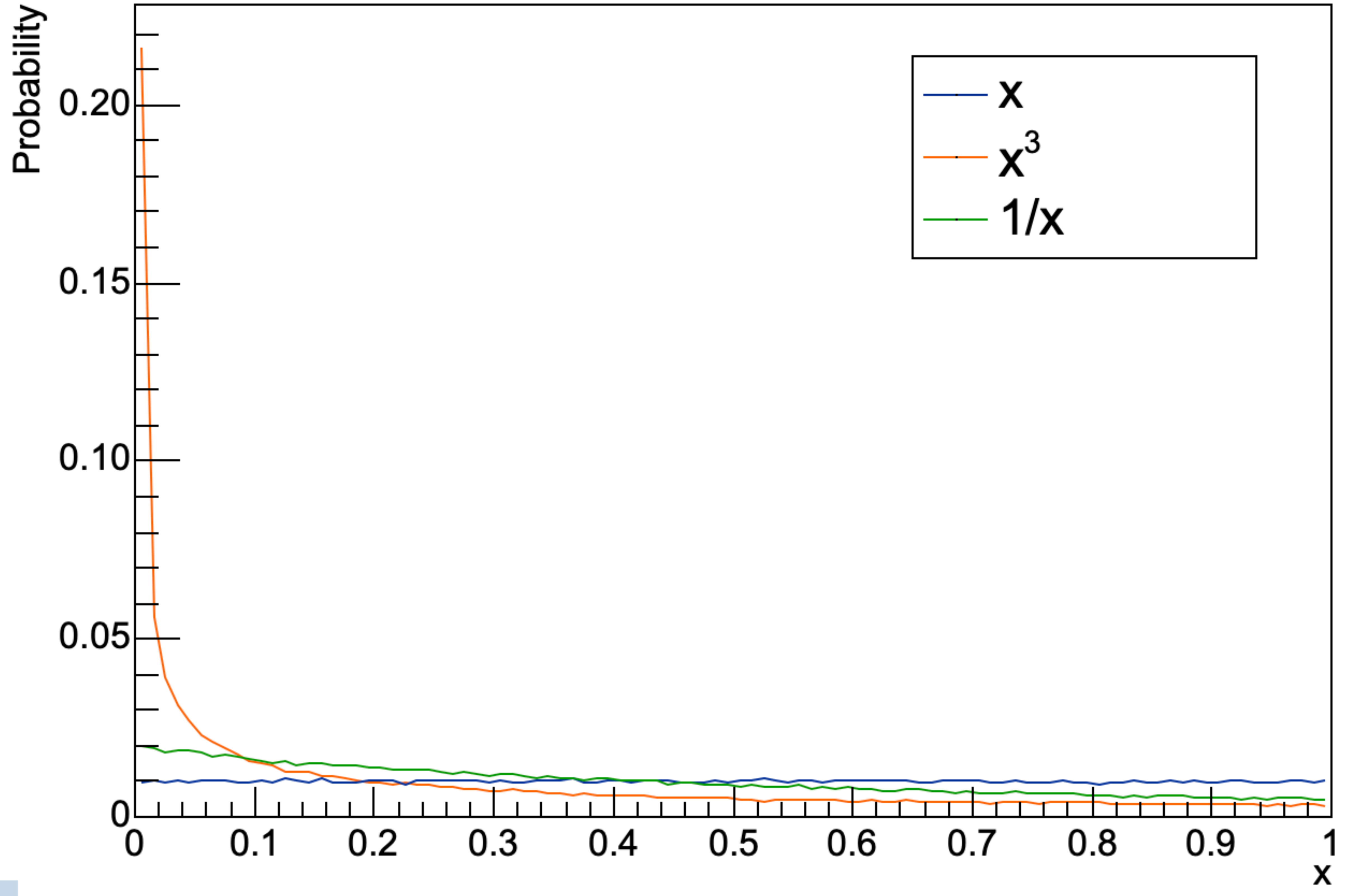
CERN


BAYESIAN INFERENCE
THIS WHOLE BOOK IS FULL OF PRIORS THAT HAVE TO BE ACCEPTED ON FAITH! IT'S A RELIGION!
AND IN THE PUBLIC SCHOOLS NO LESS. CALL A LAWYER.
AS A MATH ATHEIST, I SHOULD BE EXCUSED FROM THIS.

Which Prior?

Which prior?

- *The choice of a prior is a crucial aspect of a Bayesian analysis. There are two classes of issues*
  - *Modeling knowledge: when something is known about a quantity, one needs to find a prior that models that knowledge*
  - *Modeling ignorance: this is where troubles start. Even an innocent assumption on a quantity x (e.g., a flat prior) can become a strong statement on some function of x (mind the Jacobian)*



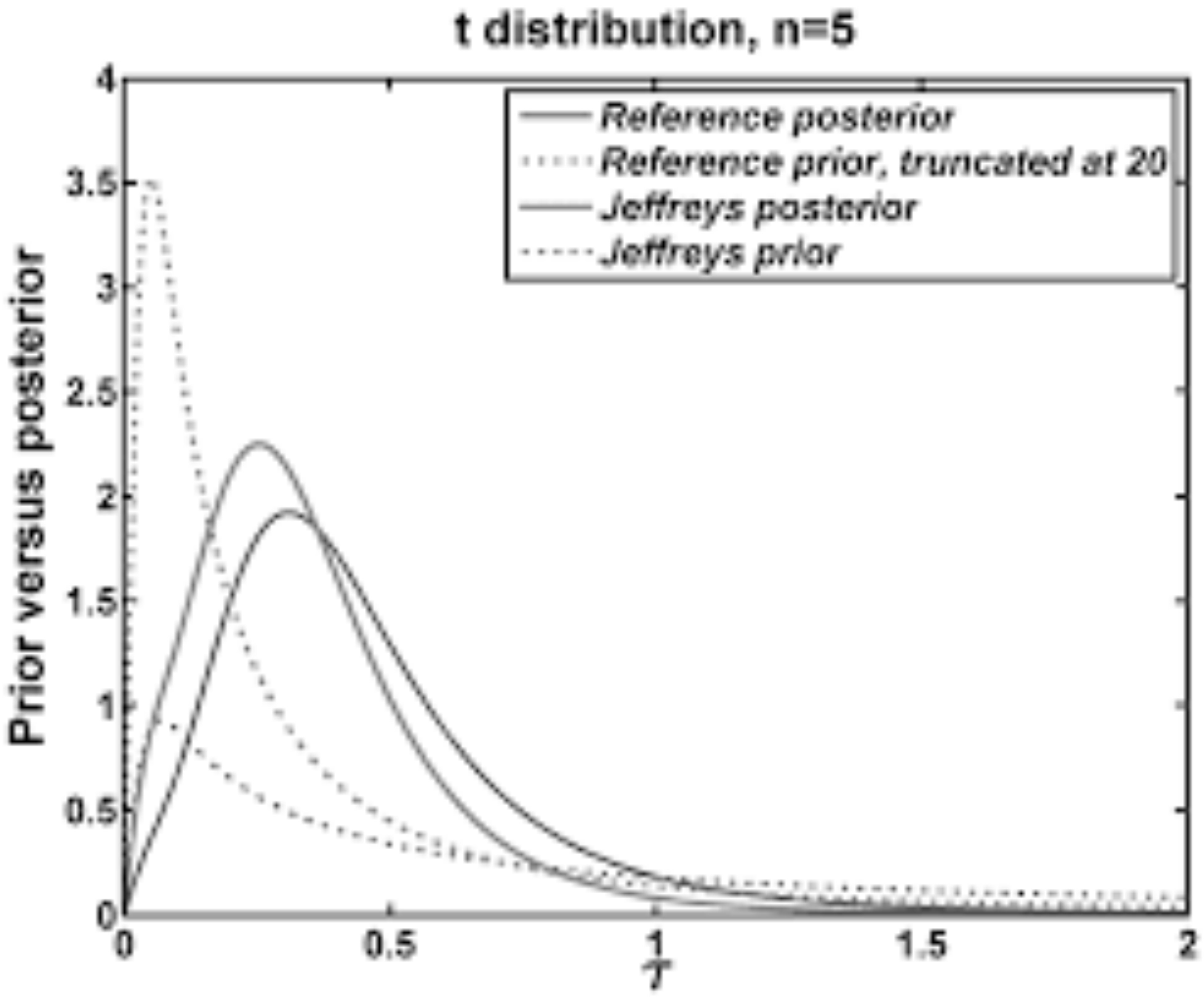
Objective priors

- *We said that a prior is the modeling on a-priori knowledge (subjective prior)*

- *When modeling ignorance, it was suggested to instead fix the prior by a formal rule (objective prior)*

  - *We are trading the instability of a "democratic" flat function with a desired property, e.g., invariance under specific reparameterization*

- *Often called non-informative priors. Misleading name, since they carry a lot of information (they prefer certain values over others)*



Principle of Indifference

- *The principle of indifference states that in the absence of any relevant evidence, agents should distribute their credence (or 'degrees of belief') equally among all the possible outcomes under consideration.*
  - *Starting point for the first writers on probability (Laplace, Bernoulli, etc.)*
  - *The simplest example of non-informative priors*
  - *Obvious application in case of discrete outcomes (dice, etc.)*
  - *Doesn't work for continuous variables: flat on what?*

https://en.wikipedia.org/wiki/Principle_of_indifference

- Suppose there is a cube hidden in a box. A label on the box says the cube has a side length between 3 and 5 cm.
- We don't know the actual side length, but we might assume that all values are equally likely and simply pick the mid-value of 4 cm.
- The information on the label allows us to calculate that the volume of the cube is between 27 and 125 cm^3. We don't know the actual volume, but we might assume that all values are equally likely and simply pick the mid-value of 76 cm^3.
- However, **we have now reached the impossible conclusion that the cube has a side length of 4 cm and a volume of 76 cm^3**

Principle of Transformation Group

- *Generalization of principle of indifference, by E. T. Jaynes*
  - *One should have an indifferent choice between equivalent problems (i.e., equivalent quantities of interest) rather than between different outcomes*
  - *In practice, given two quantities x and y, one looks for a prior f such that solving for f(x) or f(y) gives the same result*
  - *Reduces to principle of indifference for a discrete problem (which has to be permutation invariant)*
- *The prior to choose depends on what one is ignorant about*

<u>Reparameterization invariance:</u> consider two possible parameterisations θ and ϕ of a given model. Assume that one is a smooth function of the other.

A reparameterization invariant prior is a prior that transforms under the usual rule of the change-of-variables theorem

$$p_\phi(\phi) = p_\theta(\theta) \left| \frac{d\theta}{d\phi} \right|$$

Jeffrey's prior

- *The definition of Jeffrey's prior starts from Fisher information* $I(\vec{\theta})$
  - $I(\vec{\theta})$ *measures the information that a random value carries about a parameter*
  - *The Jeffrey's prior is (up to a normalization factor) the sqrt of the determinant of the Fisher information matrix*

$$p(\vec{\theta}) \propto \sqrt{\det I(\vec{\theta})}$$

<u>Score:</u> the partial derivative of the log Likelihood wrt the parameters

$$\frac{\partial}{\partial\theta} log \mathscr{L}(x|\theta)$$

The score has 0 expectation value

$$E\left[\frac{\partial}{\partial\theta}\log\mathscr{L}(x|\theta)\right] = \int dx \mathscr{L}(x|\theta)\frac{\partial}{\partial\theta}\log\mathscr{L}(x|\theta) = \int dx \mathscr{L}(x|\theta)\frac{\frac{\partial}{\partial\theta}\mathscr{L}(x|\theta)}{\mathscr{L}(x|\theta)} = \frac{\partial}{\partial\theta}\int dx \mathscr{L}(x|\theta) = \frac{\partial}{\partial\theta}1$$

One can then write the score's variance as

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log\mathscr{L}(x|\theta)\right)^2\right] = \int dx \mathscr{L}(x|\theta)\left(\frac{\partial}{\partial\theta}\log\mathscr{L}(x|\theta)\right)^2$$

Jeffrey's prior

- (you can convince yourself that) $I(\vec{\theta})$ transformation rules guarantee the desired invariance under reparameterization
- Jeffrey's prior coincides with the priors satisfying the principle of transformation group
- But its definition is formally tight to information theory and it offers the opportunity to further generalisations

Mean of a Gaussian: consider a Gaussian likelihood with fixed σ

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

$$I(\mu) = E\left[\left(\frac{d}{d\mu}\log G(x\,|\,\mu,\sigma)\right)^2\right] = E\left[\left(\frac{d}{d\mu}\frac{(x-\mu)^2}{2\sigma^2}\right)^2\right] =$$

$$E\left[\frac{(x-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^4}\int dx G(x\,|\,\mu,\sigma)(x-\mu)^2$$

So that $p(\mu) \propto \sqrt{I(\mu)}$ = constant

$$E[(x-\mu)^{2p}]\int dx G(x\,|\,\mu,\sigma)(x-\mu)^{2p} = (2p-1)!\sigma^{2p}$$

Jeffrey's prior

- *(you can convince yourself that) $I(\vec{\theta})$ transformation rules guarantee the desired invariance under reparameterization*
- *Jeffrey's prior coincides with the priors satisfying the principle of transformation group*
- *But its definition is formally tight to information theory and it offers the opportunity to further generalisations*

<u>RMS of a Gaussian:</u> consider a Gaussian likelihood with fixed μ

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

$$I(\sigma) = E\left[\left(\frac{d}{d\sigma}\log G(x\,|\,\mu,\sigma)\right)^2\right] = E\left[\left(\frac{d}{d\sigma}\frac{(x-\mu)^2}{2\sigma^2}\right)^2\right] =$$

$$E\left[\frac{(x-\mu)^4}{4}\frac{d}{d\sigma}\sigma^{=2}\right] = \frac{1}{4}\left(\frac{d}{d\sigma}\sigma^{-2}\right)^2 E\left[(x-\mu)^4\right] = \frac{1}{\sigma^6}\sigma^4 3!$$

So that $p(\sigma) \propto \sqrt{I(\sigma)} = \dfrac{1}{\sigma}$

$$E[(x-\mu)^{2p}]\int dx G(x\,|\,\mu,\sigma)(x-\mu)^{2p} = (2p-1)!\sigma^{2p}$$

Information and Entropy

- *In Information theory, the Kullback-Leibler divergence measures the relative entropy between two functions*

$$D_{KL}(p(x), q(x)) = E\left[\log\frac{p(x)}{q(x)}\right]_p = \int dx\; p(x)\log\frac{p(x)}{q(x)}$$

- *It's not a distance between two functions*

$$D_{KL}(p(x), q(x)) \neq D_{KL}(q(x), p(x))$$

- *But it can be generalized to the smooth and symmetric Jensen–Shannon divergence (sometimes used in ML HEP literature)*

$$D_{SJ}[p(x), q(x)] = \frac{1}{2}\left[D_{KL}\left(p(x), \frac{p(x)+q(x)}{2}\right) + D_{KL}\left(q(x), \frac{p(x)+q(x)}{2}\right)\right]$$

Information and Entropy

- *Consider a family of probability density functions, characterized by different values of some parameter θ*

- *Compute the KL divergence between two of these pdfs*

$$D_{KL}(\theta_1, \theta_2) = D_{KL}[f(x|\theta_1), f(x|\theta_2)] = \int dx\, f(x|\theta_1) \log \frac{f(x|\theta_1)}{f(x|\theta_2)}$$

- *For small variation, i.e., for small differences between θ_1 and θ_2, one can expand θ_1 around θ_2*

- *It can be shown that the second therm of the expansion is the Fisher information*

$$\left(\frac{\partial^2}{\partial \theta'_i\, \partial \theta'_j} D(\theta, \theta') \right)_{\theta'=\theta} = -\int f(x;\theta) \left(\frac{\partial^2}{\partial \theta'_i\, \partial \theta'_j} \log(f(x;\theta')) \right)_{\theta'=\theta} dx = [\mathcal{I}(\theta)]_{i,j}$$

Jayne's Maximum Entropy Principle

- *The principle of maximum entropy states that the probability distribution which best represents the current state of knowledge about a system is the one with largest entropy*

$$H(\Pi) = -\int \Pi(\theta) \log \Pi(\theta) d\theta$$

- *Based on this principle, one can build a prescription to choose a prior in a Bayesian application*
- *Notice that maximizing $H(\Pi)$ corresponds to minimising the KL diverge between Π and a flat distribution*
- *In this respect, the MEP is a generalization of the indifferent principle, based on information theory*

Reference priors

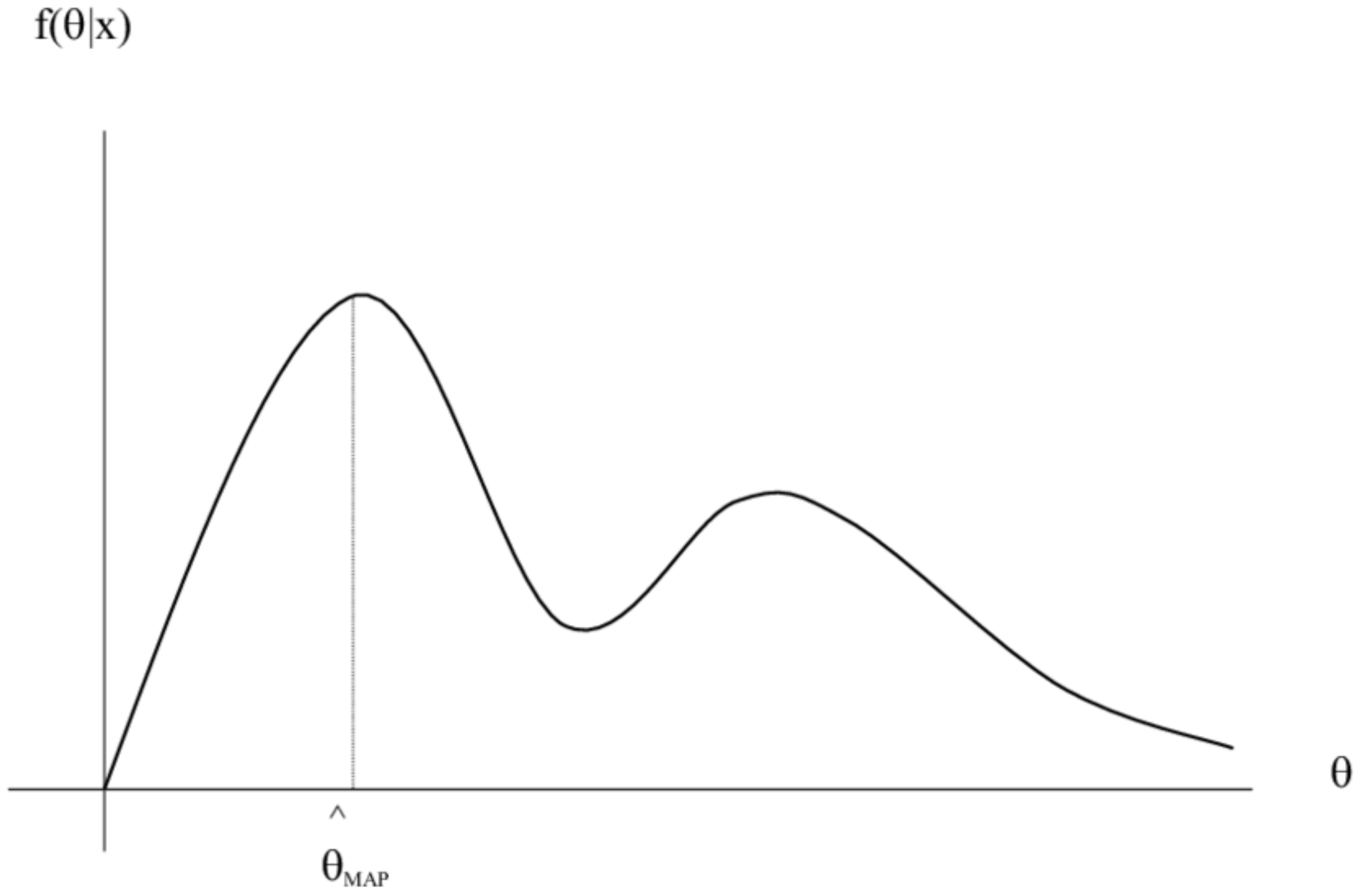
- The KL divergence can be used as a metric to define a set of priors (reference priors) which generalise Jeffrey's priors
- Given a likelihood $\mathscr{L}$, one looks for the prior Π that maximises the expected distance between the prior and the posterior (as a way to minimise the role of the prior in the inference)
  - The KL divergence is used as a metric for the distance
  - The maximisation has to be done across all possible experiment outcomes, since the prior has to be chosen regardless of the experiment result

$$I(\Theta, T) = \int p(t) \int p(\theta|t) \log \frac{p(\theta|t)}{p(\theta)} d\theta dt \quad = \int \int p(\theta, t) \log \frac{p(\theta, t)}{p(\theta)p(t)} d\theta dt$$

What did we measure?

Bayesian Estimator

- Once the posterior is derived, all the information on the parameter of interest x is there
- We can then use it to make estimates on it
  - Maximum A Posterior (MAP) estimation: estimate the value of x using the max of the posterior
- Similar to max-likelihood estimator in frequentists statistics
  - In the limit of large statistics, likelihood becomes narrow, prior less important, and the two estimators converge



Maximum Likelihood Estimate (MLE)

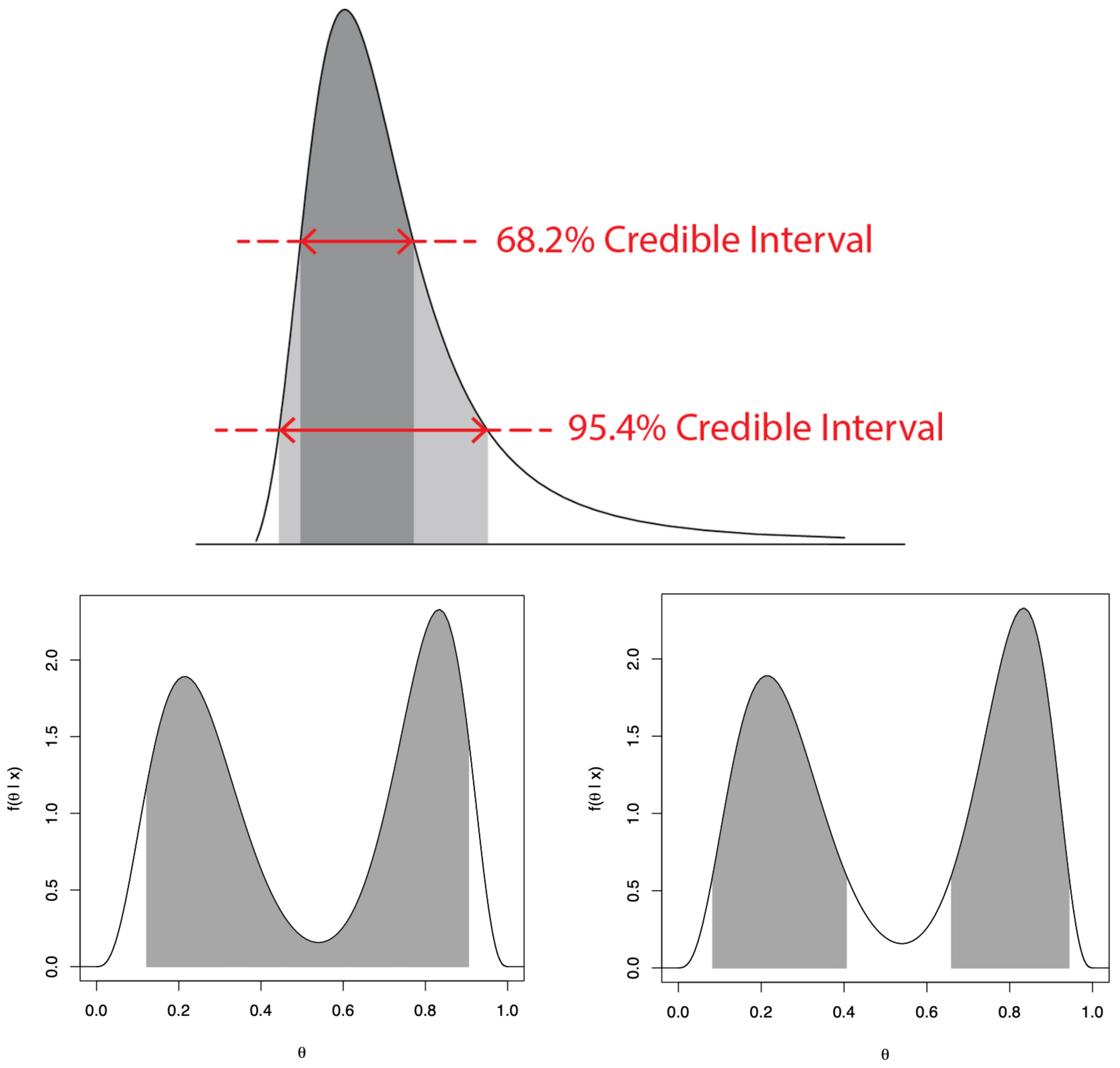
$$\boldsymbol{\theta}^{\mathrm{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)} | \boldsymbol{\theta})$$

Maximum *a posteriori* (MAP) estimate

$$\boldsymbol{\theta}^{\mathrm{MAP}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)} | \boldsymbol{\theta}) \underbrace{p(\boldsymbol{\theta})}_{\text{Prior}}$$

Credible Intervals

- By integrating the posterior one can define a credible interval
- Not uniquely defined
  - Can integrate around the median, the MAP estimator, etc
  - Different strategies adapt to different distribution
    - e.g., don't use the median for bur-modal distributions



MAP Hypothesis Testing

- As for frequentist statistics, one can use Bayesian statistics to decide between two hypotheses
  - Two hypotheses (H_0 and H_1) and their probability models $p(D|H_0)$ and $p(D|H_1)$
- Using Bayesian probability

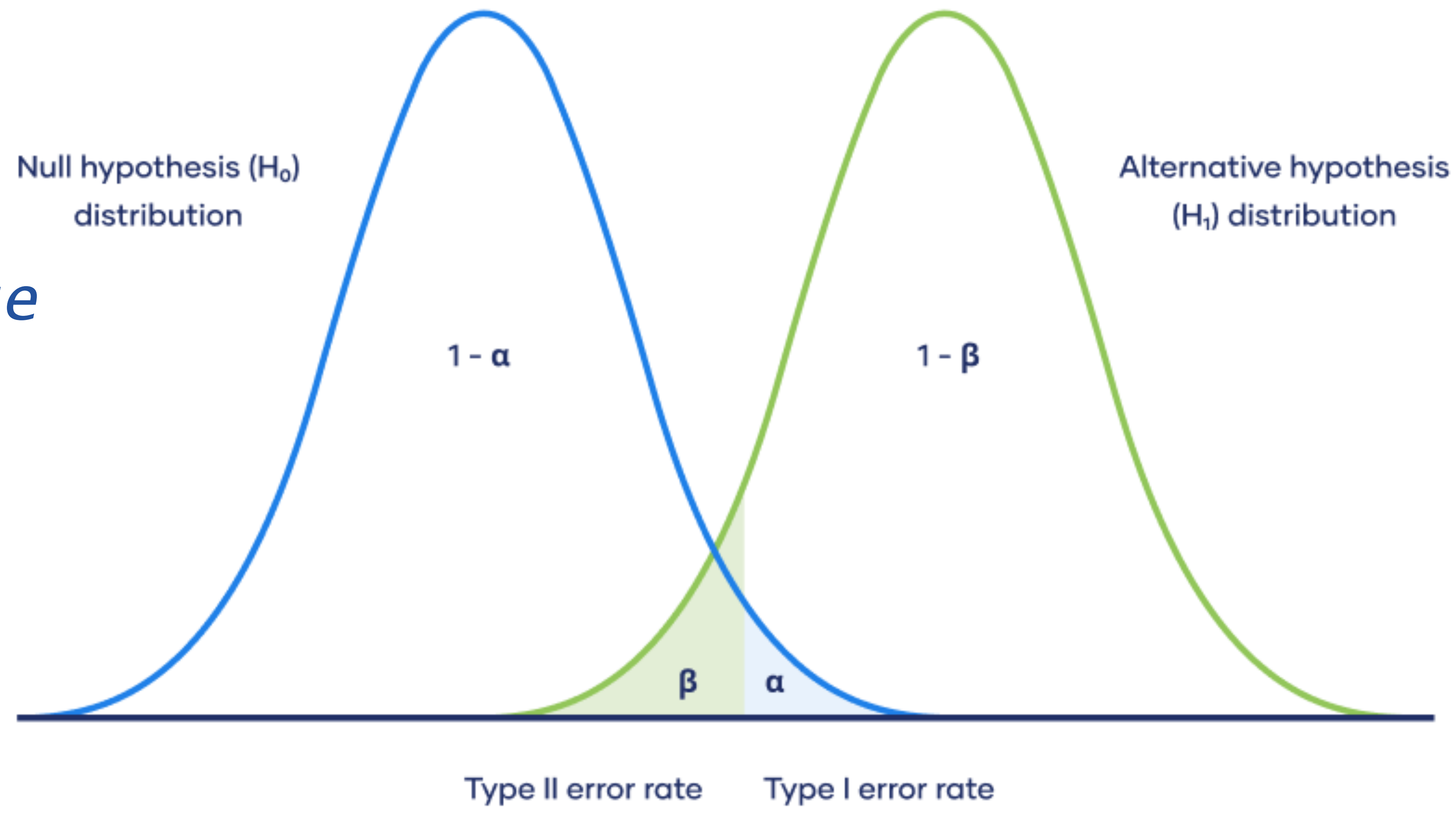
$$p(H_0|D) = \frac{P(D|H_0)p(H_0)}{P(D)} \qquad p(H_1|D) = \frac{P(D|H_1)p(H_1)}{P(D)}$$

- Maximum A Posteriori (MAP) test:
  - Choose the hypothesis with largest $p(H|D)$
  - This choice minimises the probability of a mistake

Type I and Type II errors

- *In a hypothesis test*
  - *Type I error (false positive) consists in rejecting the null hypothesis H_0 (e.g., there is no new physics) when H_0 is true*
  - *Type II error (false negative) consists in rejecting the alternative hypothesis H_1 (e.g., there is new physics) when H_1 is true*
- *The two errors have different implications*
  - *A Type II error corresponds to missing a discovery. With more data and a stronger evidence, the discovery is just postponed*
  - *A Type I error corresponds to a false claim, that would ruin your scientific reputation*
- *HEP is (rightfully) a very conservative field: the community is willing to expose itself to Type II errors in order to minimise the chance of a Type I error*

Probability of making Type I and Type II errors



Minimum Cost Hypothesis Test

- *A MAP hypothesis testing gives same weight to the two errors*
- *One might have different costs for the two euros*
  - C_{10}: *cost of choosing H1 when H0 is true*
  - C_{01}: *cost of choosing H0 when H1 is true*
- *In that case, one would choose between H0 and H1 weighting for the costs. In that case one would choose the maximum between*

$$p(H_0|D)C_{10} \qquad p(H_1|D)C_{01}$$

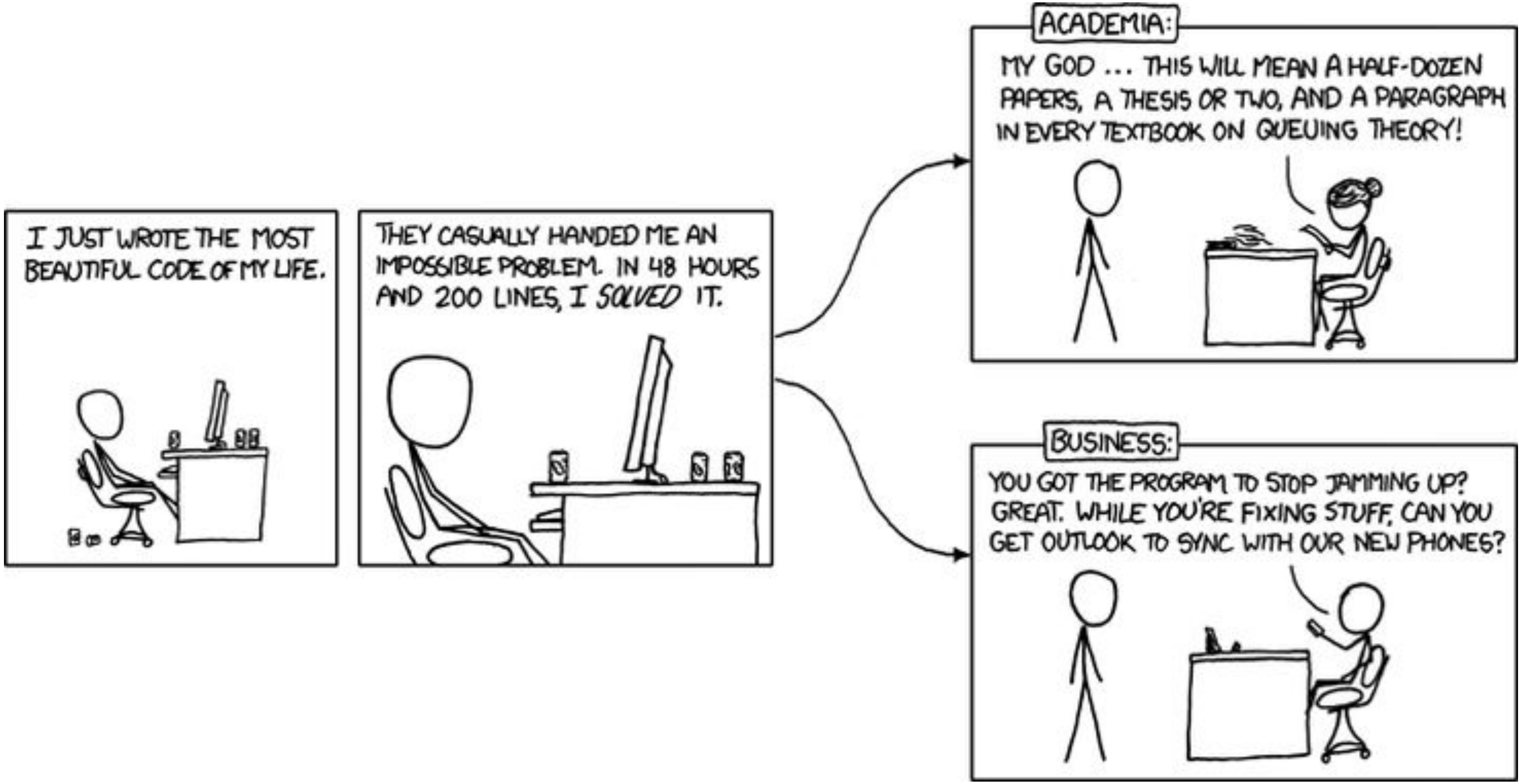
Bayes Factor

- *Bayesian evidence: integral of a posterior model over all the parameters of the hypothesis (i.e., the denominator in posterior formula)*

$$p_M(D) = \int d\alpha P_M(D|\alpha)P(\alpha)$$

- *Bayes factor is the ratio of evidences and it's used to select among different hypotheses*

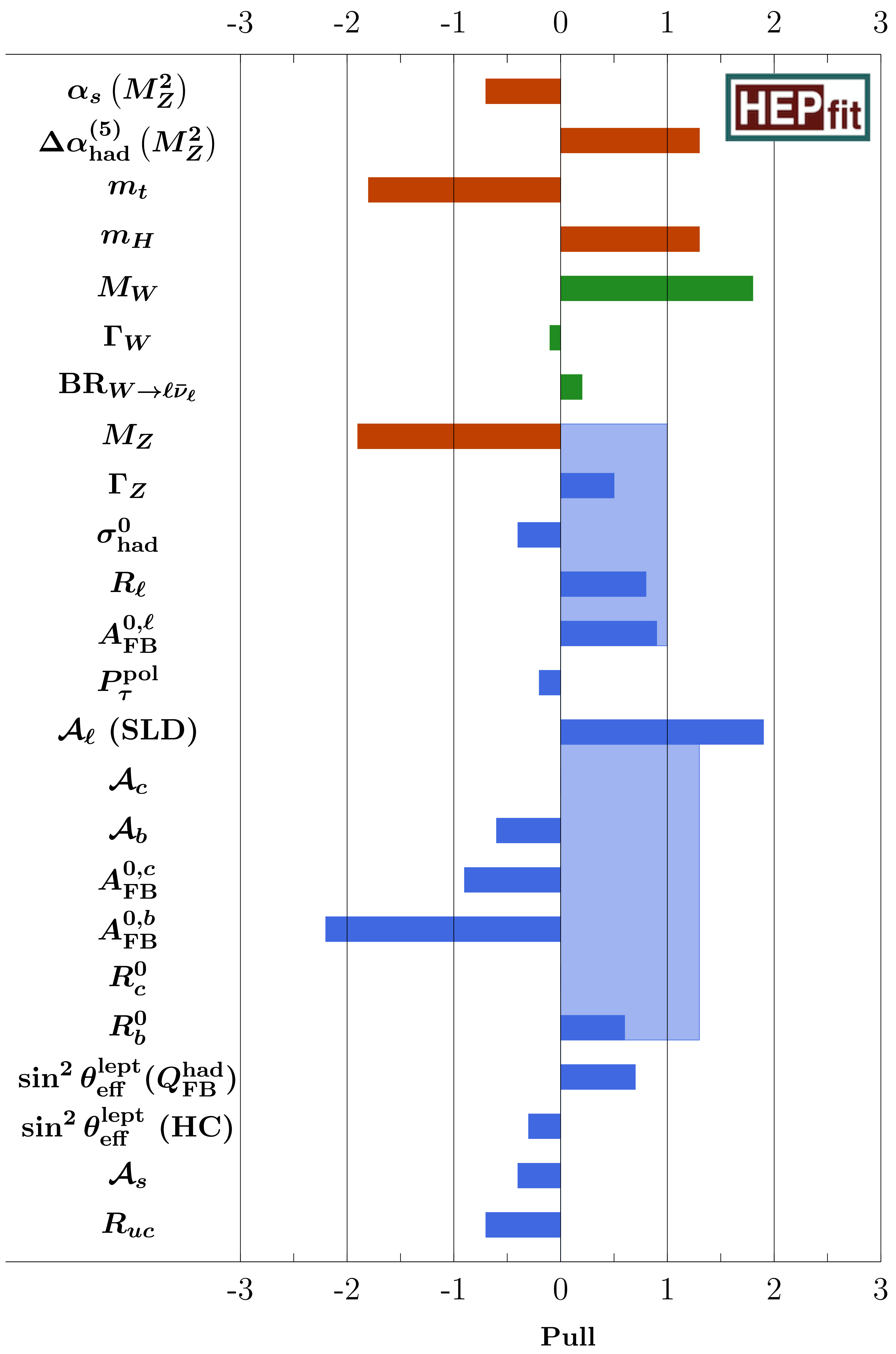
$$K = \frac{P_{M_1}(D)}{P_{M_2}(D)} = \frac{\int d\alpha P_{M_1}(D|\alpha)P(\alpha)}{\int d\beta P_{M_2}(D|\beta)P(\beta)}$$



Efficient strategy for a Bayesian analysis

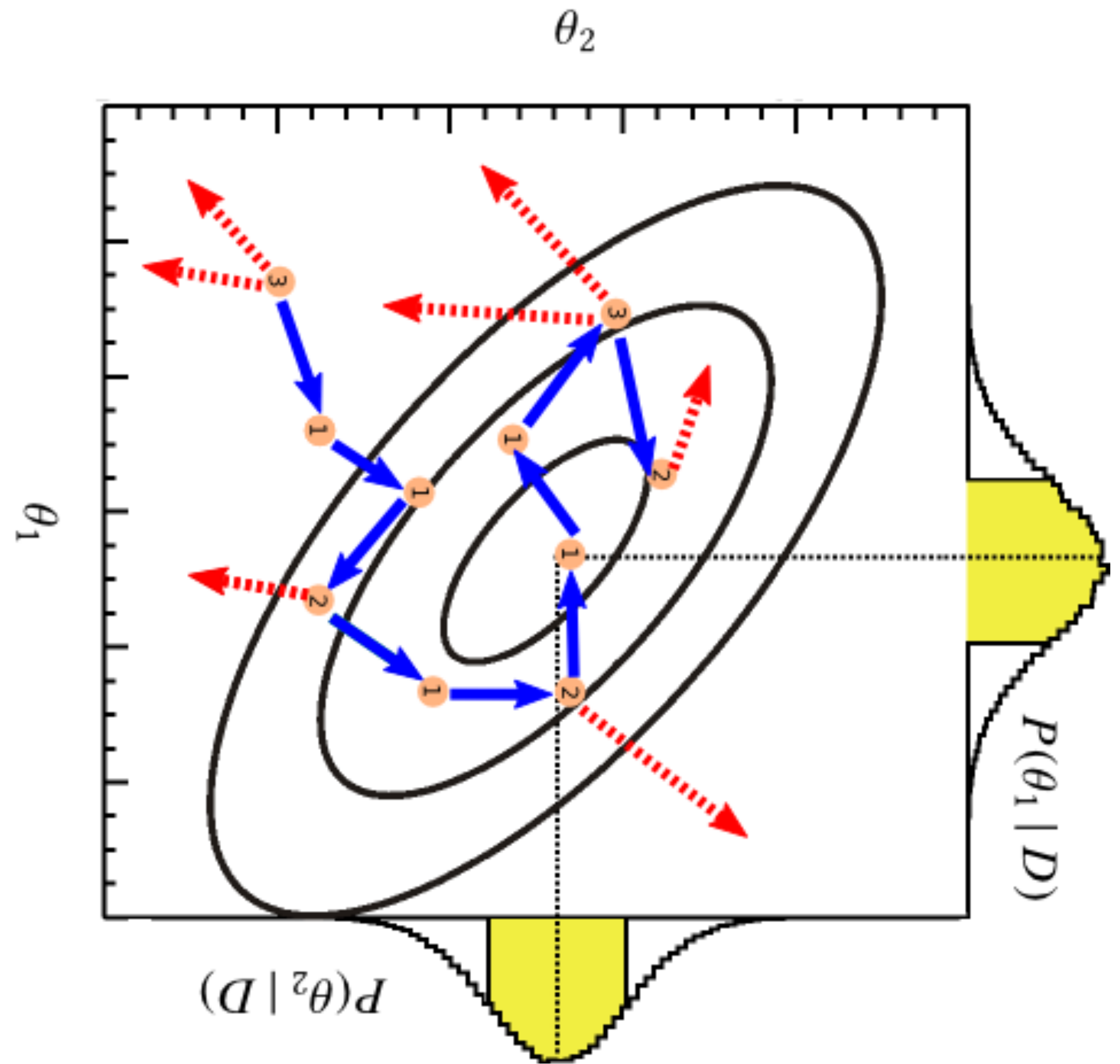
Example: The EW fit

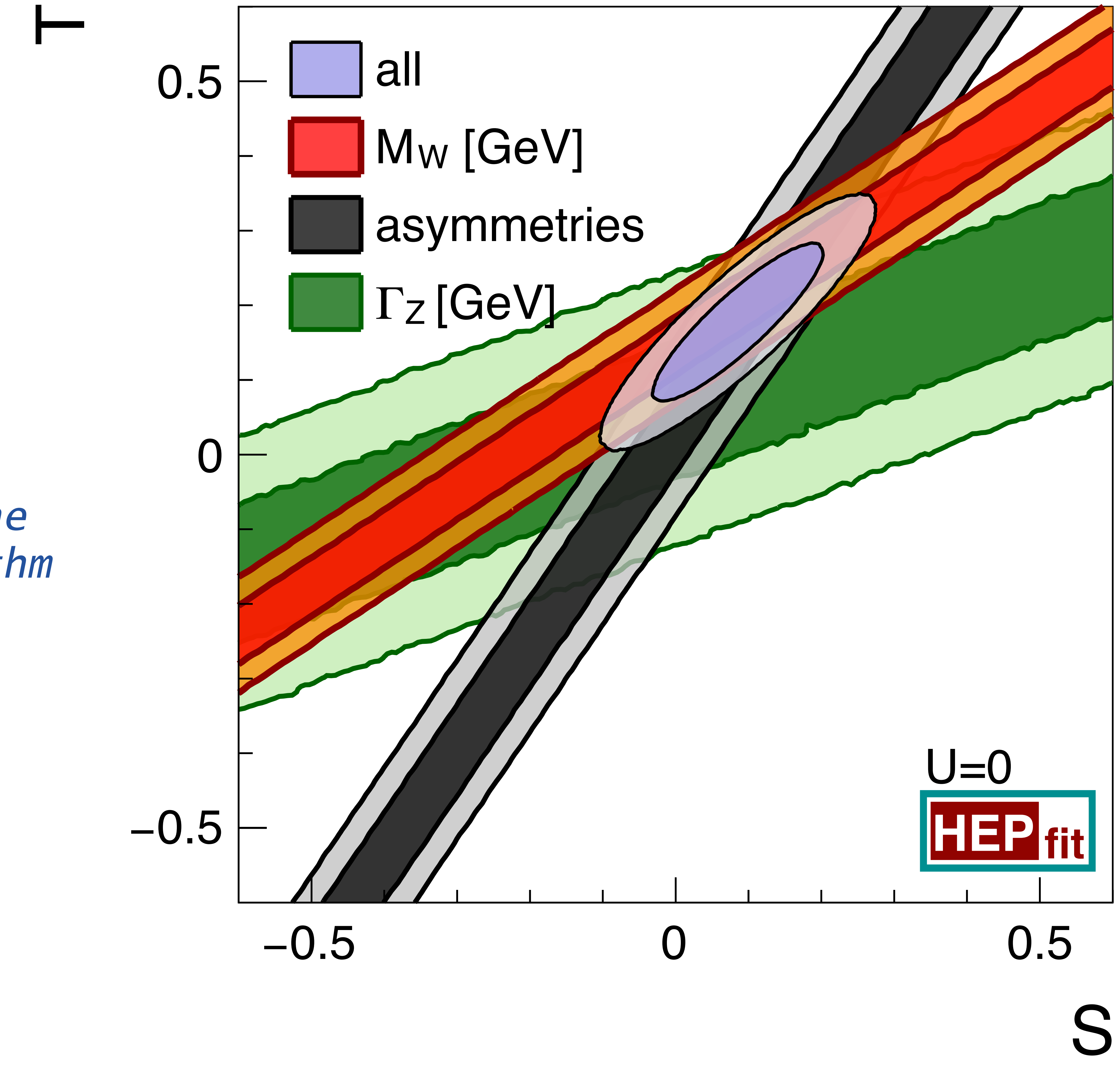
- Exploit the over-constrained EW sector (dictated by rigid symmetry structure) to perform consistency tests of the SM with EW precision observables
- Set of input parameters (α scheme): GF, α, mZ, mH, mt, αS(mZ), Δαhad(5)
- Compute EW precision observables as functions of these quantities
- Z-pole observables
- W observables
- Compare computations to experimental data to learn the values of the
- input quantities
- Extend relations to include BSM effects and determine bounds on New Physics
- Oblique parameters: S, T, U, …
- Effective interactions: SMEFT
- …



Example: The EW fit

- Input parameters $\underbrace{\alpha, G_F}_{\text{fixed}}, \alpha_s(M_Z), M_Z, M_H, m_t, \Delta\alpha_{had}^{(5)}$
- Observables LEP, Tevatron, and LHC
  - EW precision observables
  - top mass
  - W mass
  - Higgs couplings
- HEPfit uses BAT (Bayesian Analysis Toolkit) as the underlying engine for MCMC with Metropolis algorithm



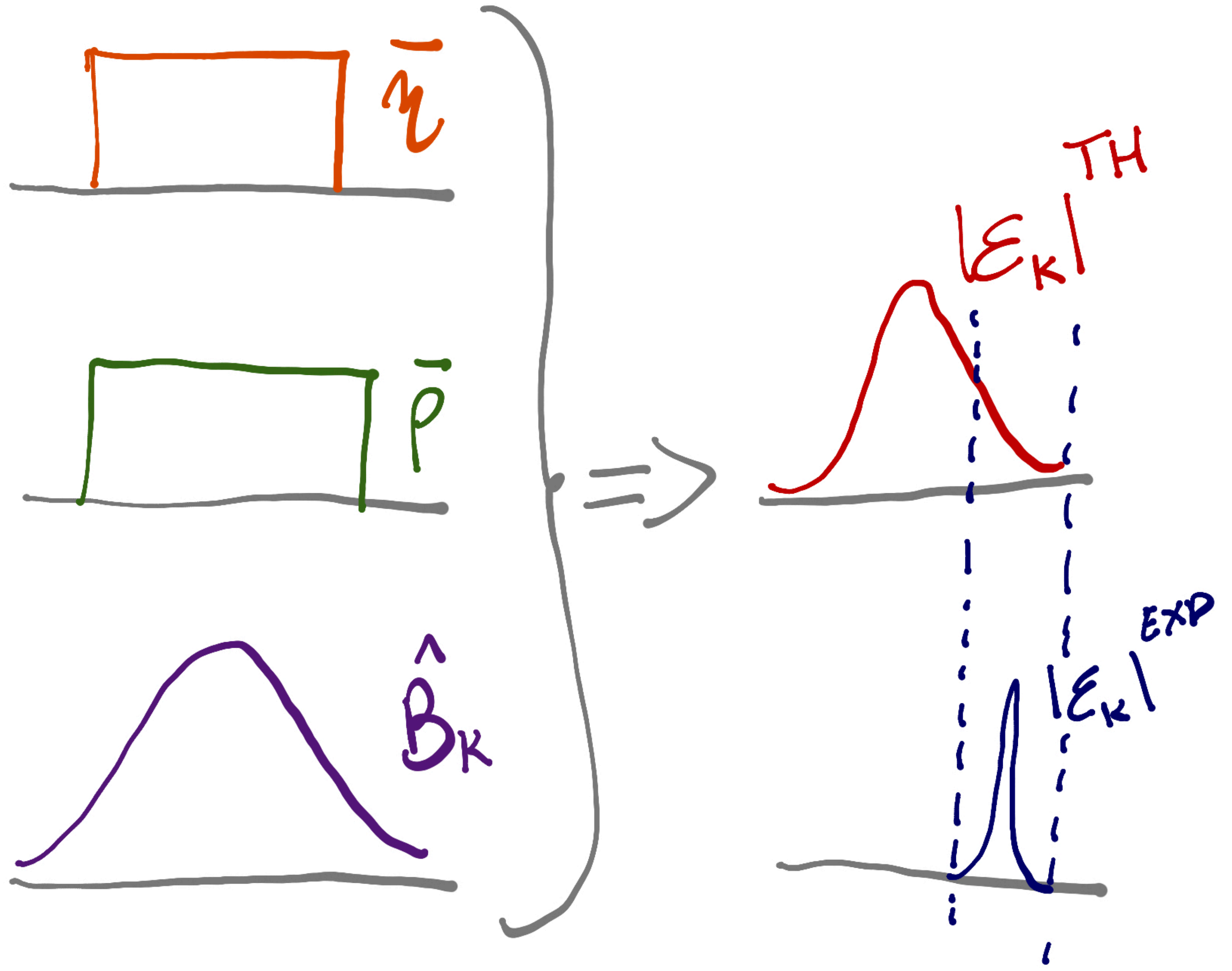


Computational complexity

- A posterior is computed with three elements
  - A prior
  - A likelihood
  - The evidence (the denominator)

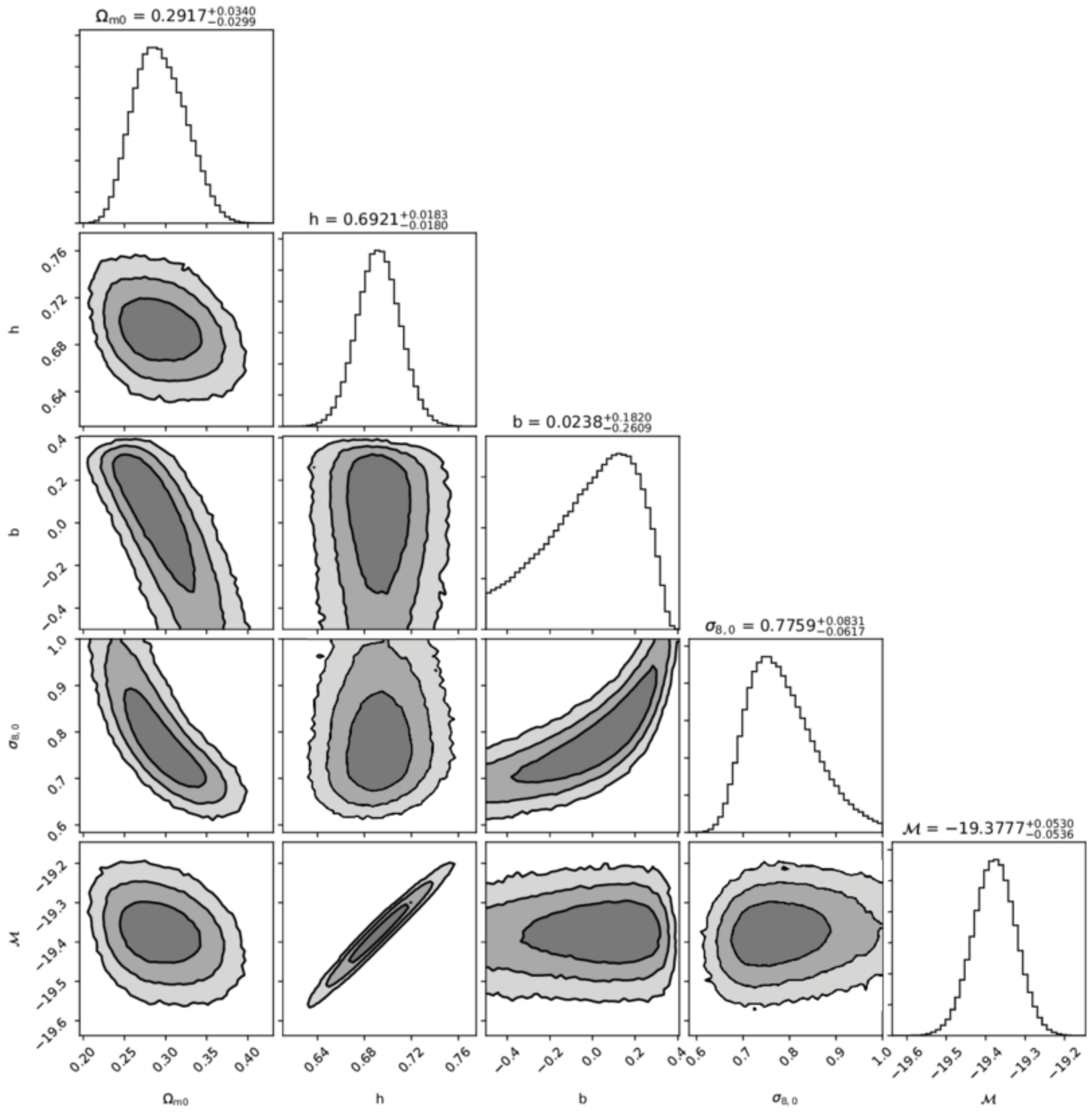
$$p(\vec{\theta}) = p(\vec{\theta}\,|\,D) = \frac{p(D\,|\,\vec{\theta})\pi(\vec{\theta})}{\int_{\theta} p(D\,|\,\vec{\theta})\pi(\vec{\theta})}$$

- The evidence also enters many applications, e.g., Bayesian hypothesis testing
- Computing the evidence implies computational expensive N-dim integration

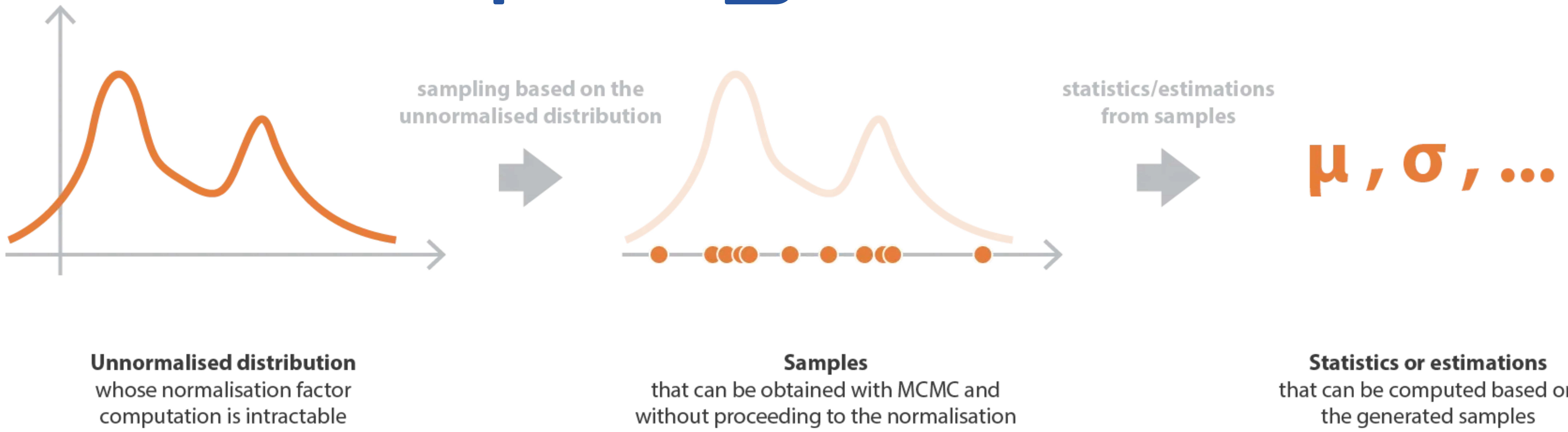


Sampling Methods

- At low dimension, the integration might be affordable
- At high dimension, the integral might be intractable
  - In this case, sampling approaches are typically adopted to speed up the computation
  - With the UT analysis, we already saw a MC-based integration based on random sampling from a low dimension (~10) distribution
  - With higher dimensionality, smarter sampling techniques might be used (e.g., Markov Chain Monte Carlo)



Sampling Methods



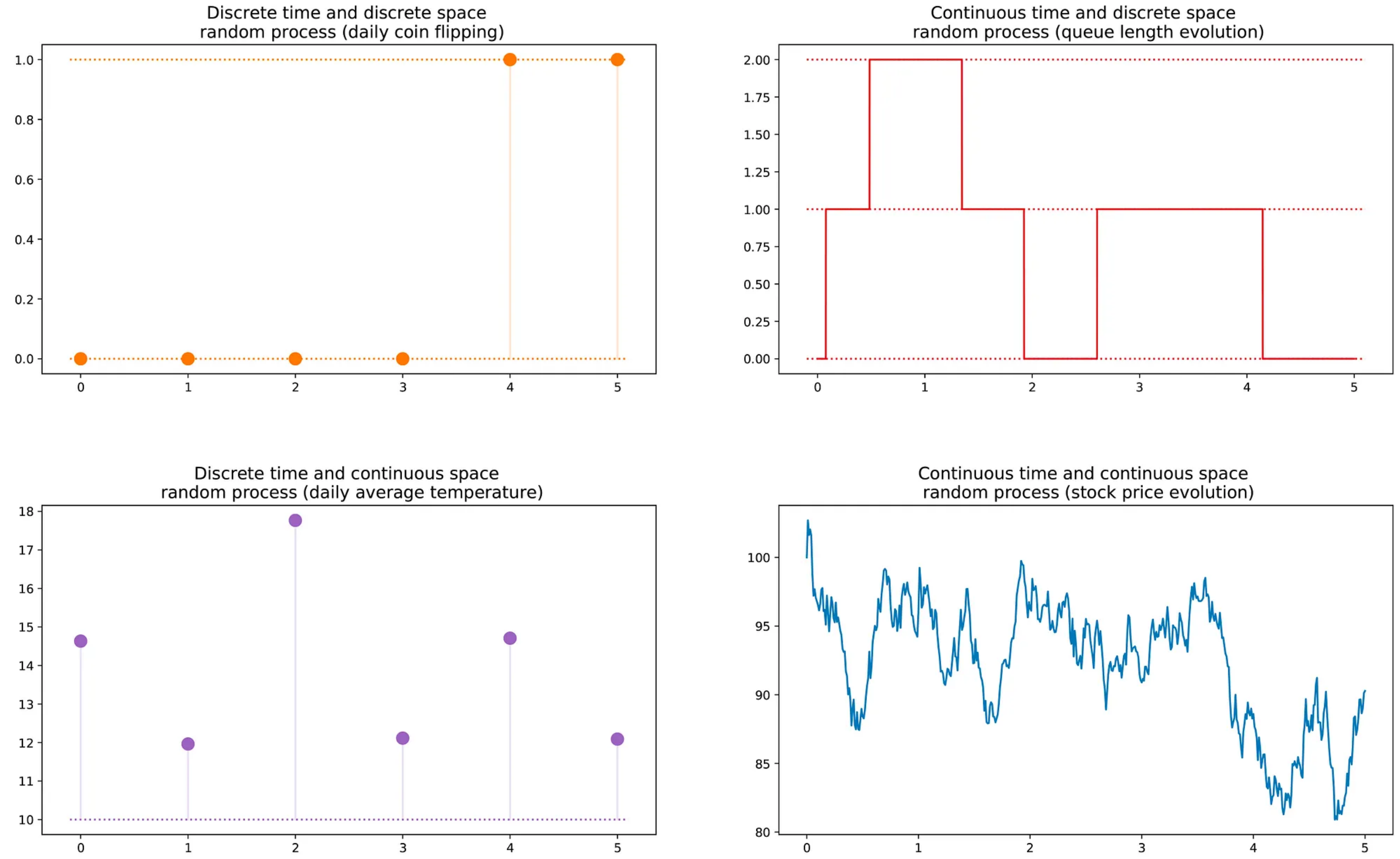
- *With a sampling method, we give up the idea of deriving a normalized posterior*
  - *Instead, the target is to sample from the unnormalised distribution and use the sampled dataset to derive parameters estimates*
    - *mean values, credibility intervals, etc.*
    - *typically using histograms*
- *MCMCs are computational effective ways to do so*

Markov Chains

- To explain Markov Chain techniques, we need a few concepts
  - ***A random process***: an ordered sequence of random variables, the ordered being given by some index T (typically some discrete time index)
  - ***A Markov process***: a random process in which the knowledge of the value taken by the process at some T_0 doesn't provide information about the evolution of the process at $T>T_0$
  - ***A homogenous discrete time Markov chain*** is a Markov process in which the space state at discrete time is also discrete
- In practice, one can discretize a continuous problem by using a Monte Carlo technique

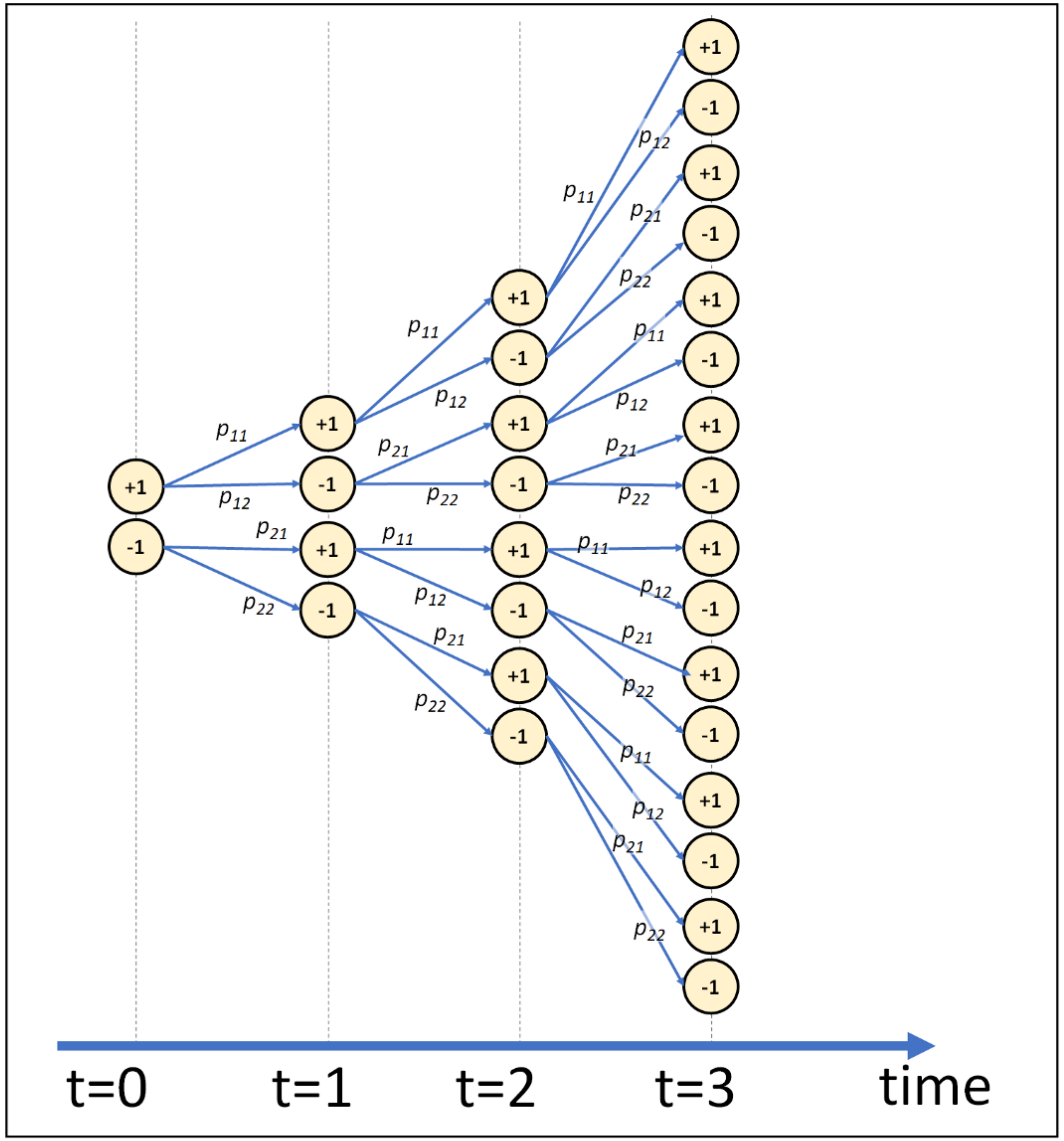
$$P(\text{ future} \mid \text{present, past }) = P(\text{ future} \mid \text{present, } \cancel{\text{past}})$$

Markov property



Some notation

- $E = \{e_i, e_2, \ldots, e_N\}$ is the state space
  - When I write e_i, the i index runs across the discrete states of the state space
- $X = \{x_1, x_2, \ldots, x_N\}$ is the array of values assumed by x at different discrete time instances
  - When I write x_j, the index j labels the value taken by X at time T_j

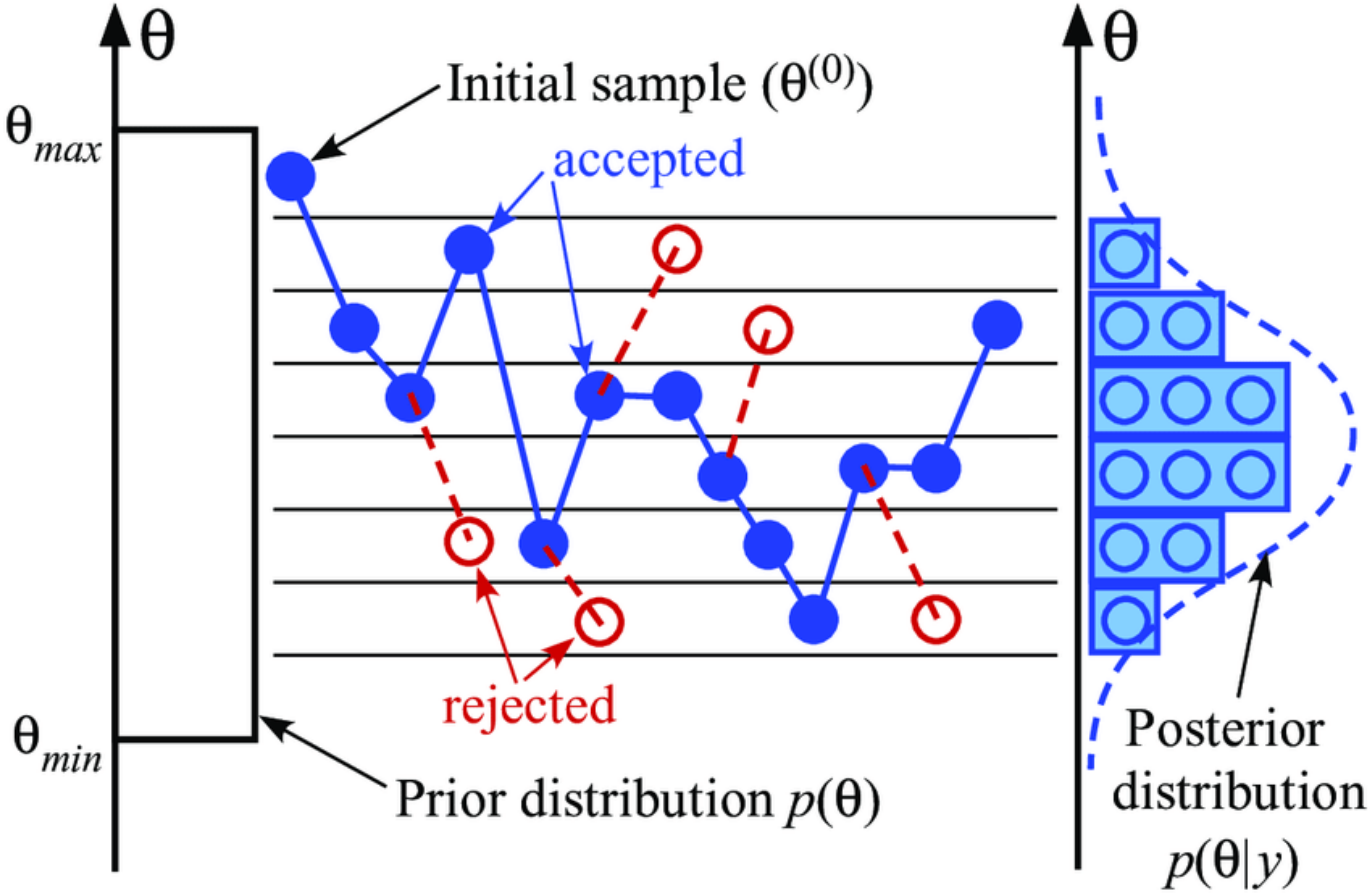


Markov Chain evolution

- *A Markov chain is fully specified by two ingredients*
  - *The* ***initial probability distribution*** *of being in a certain state at T=0*

$$p(x_0 = s) = q_0(s) \qquad \forall s \in E$$

  - *The* ***transition probability kernel*** *= probability model to transition from state x to state x'*

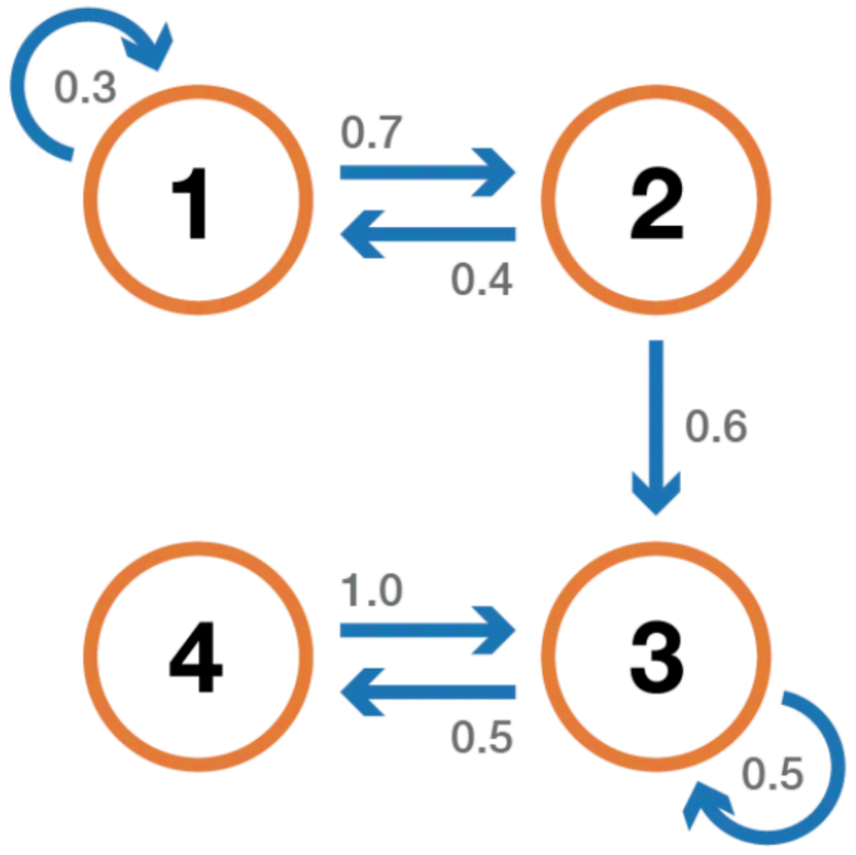


$$P(x_{n+1} = s_{n+1} \,|\, x_n = s_n) = p(s_n, s_{n+1}) \qquad \forall s_n, s_{n+1} \in E \times E$$

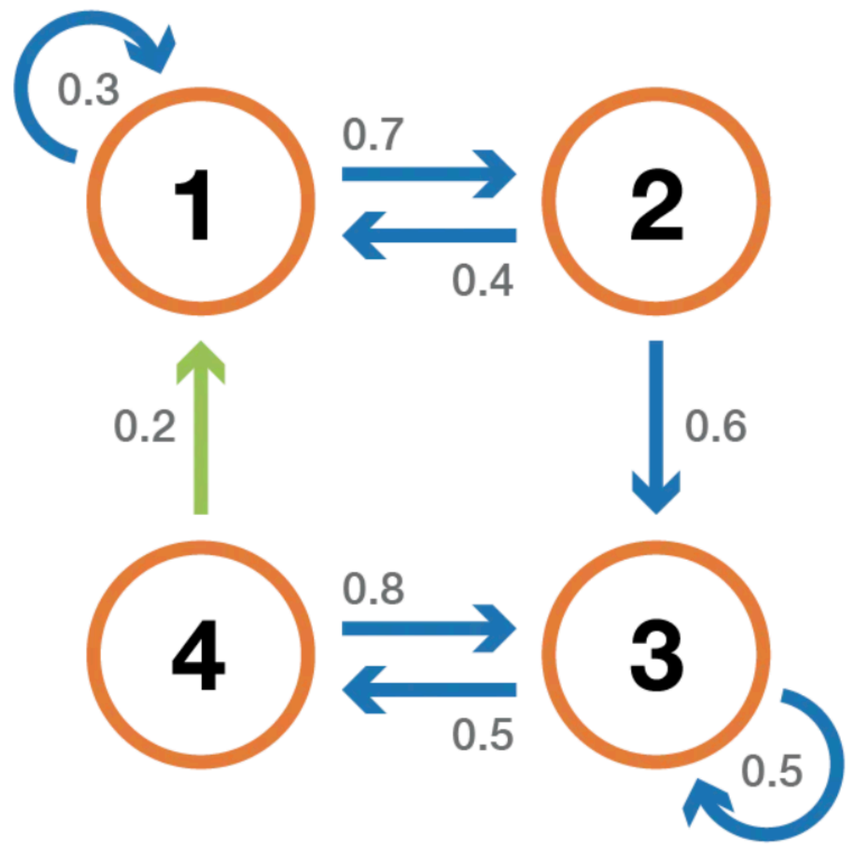
Chain properties

Reducibility

Reducible chain: some state cannot be reached from other states

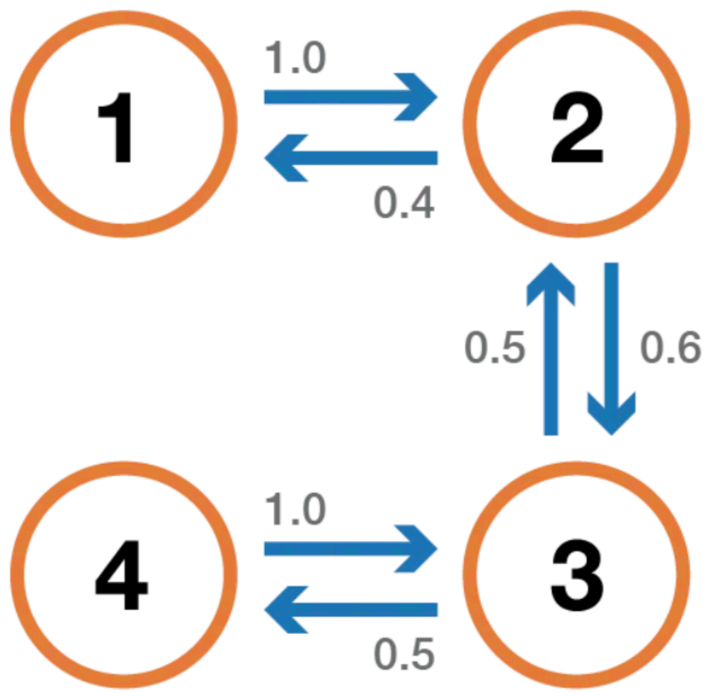


Irreducible chain: all states can be reached from any other states

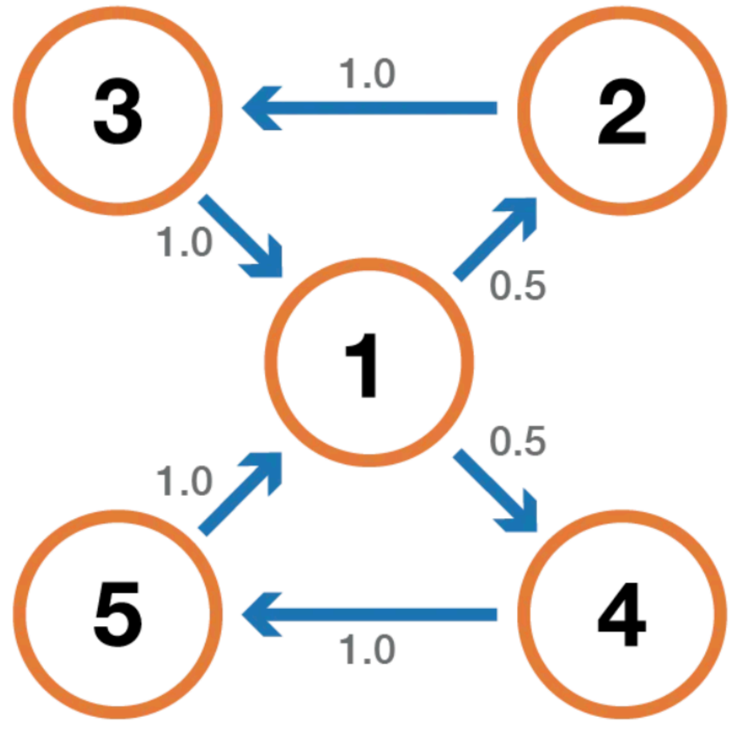


k-periodicity

2-periodicity: it takes 2n steps to go back to a given state

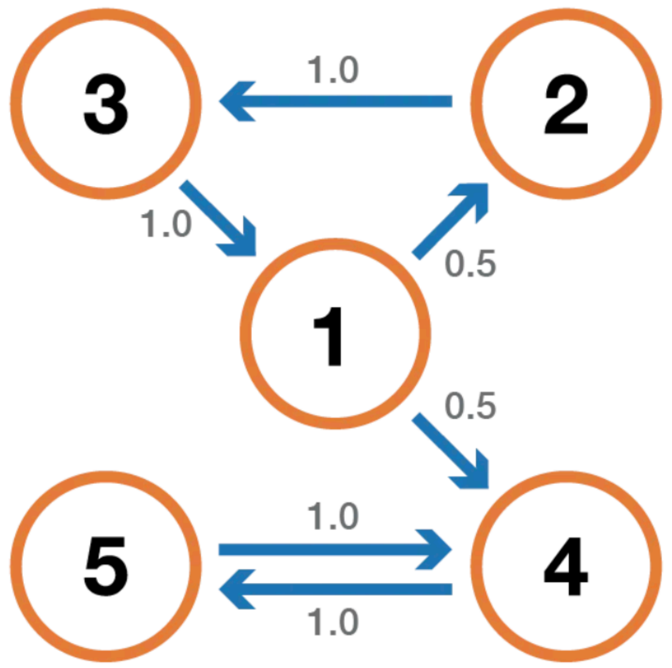


3-periodicity: it takes 3n steps to go back to a given state

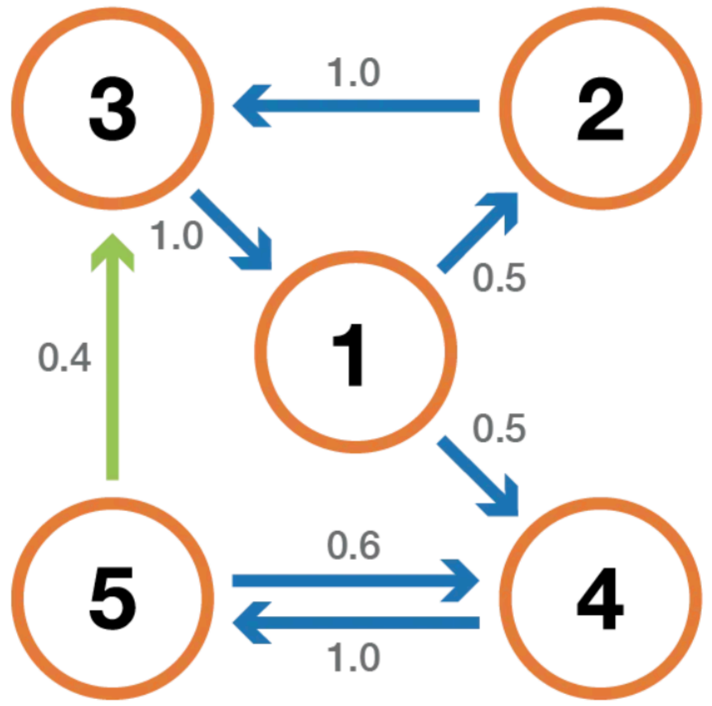


recurrence/transience

transient: a state (e.g., 1,2,3 here below) is transient if there is a >0 probability not to go back to it



recurrent: a state is recurrent (e.g., 4 and 5 above) if we know that we will go back to it



Stationary distribution

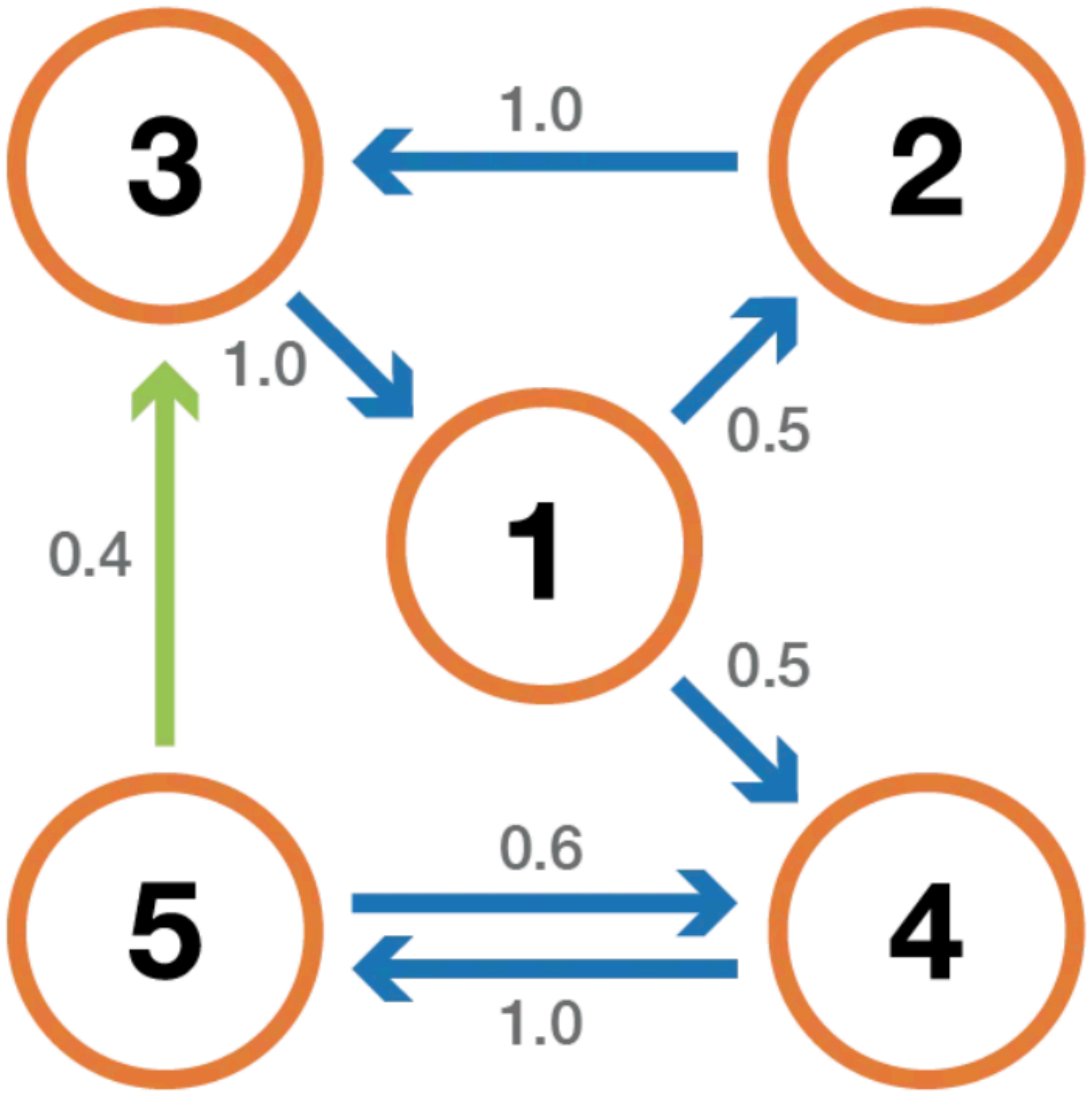
- *A pdf $\pi(x_i)$ over the space of states X is a stationary distribution if it relates to the probability transition kernel as*

$$\pi(x') = \sum_{x \in X} \pi(x) p(x, x')$$

  - $\pi(x')$ *= probability of being at x' a the current step*
  - $\sum_{x \in X} \pi(x) p(x, x')$ *= probability of being at x' a the next iteration*

- *By definition, a stationary distribution does not evolve in time*

- *An irreducible Markov chain has a stationary probability distribution if and only if all of its states are positive recurrent.*



positive recurrent: recurrent within a finite time (as opposed to zero recurrent)

MCMC and Posterior

- We have a target function $f(x)$ that we want to “learn” and then sample from (our posterior)
- We want to build a Markov Chain whose stationary solution is $f(x)$
- Once we have such a chain, we can sample a random sequence of states from the chain, long enough to reach the steady state solution
- A fraction of those generated sates are then kept

Build a Markov Chain whose stationary distribution is the distribution we want to sample from

Generate a sequence from that Markov Chain long enough to reach the steady state

Keep some well chosen states from that sequence as samples to be returned

Reversibility

- *We need an algorithm to build the Markov Chain that has our target pdf as stationary distribution*
- *To do so, we can exploit reversibility*
- *A MC is reversible if there exist a function* $\gamma(x)$ *such that*

Oner can show that $\gamma(x)$ **<u>is a stationary distribution</u>**

$$p(x', x)\gamma(x') = p(x, x')\gamma(x)$$

$$\int_E p(x', x)\gamma(x')dx = \int_E p(x, x')\gamma(x)dx$$

$$\gamma(x')\int_E p(x', x)dx = \int_E p(x, x')\gamma(x)dx$$

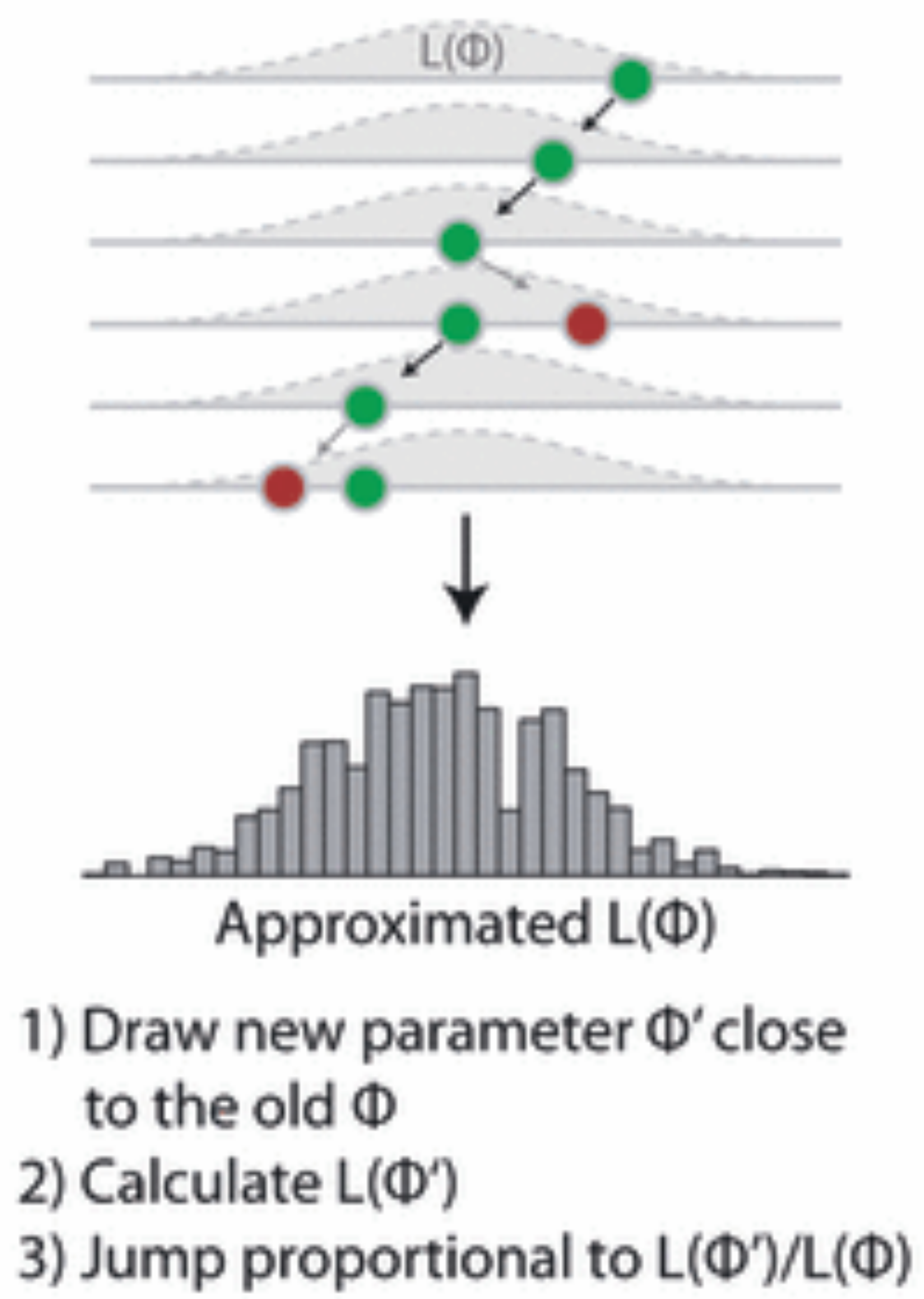
$$\gamma(x') = \int_E p(x, x')\gamma(x)dx$$

which is the definition of stationary distribution for a continuous x.
If the chain is irreducible, then this distribution is unique

$$p(x' = s', x = s)\gamma(x' = s') = p(x = s, x' = s')\gamma(x = s) \qquad \forall s, s' \in E$$

Metropolis-Hasting

- *We want to sample values from a function $g()$, from which we cannot normally sample from (but we can evaluate the function)*
- *The initial ingredients are*
  - *the target function $g()$*
  - *A suggested transition to a new state x, sampled from a transition probability h: $x \sim h(X_n)$ (very often, a Gaussian kernel is used)*
- *For a given suggested transition,* $r = \min\left(1, \frac{g(x)h(x, X_n)}{g(X_N)h(X_n, x)}\right)$ *is the probability to accept the suggested transition,*



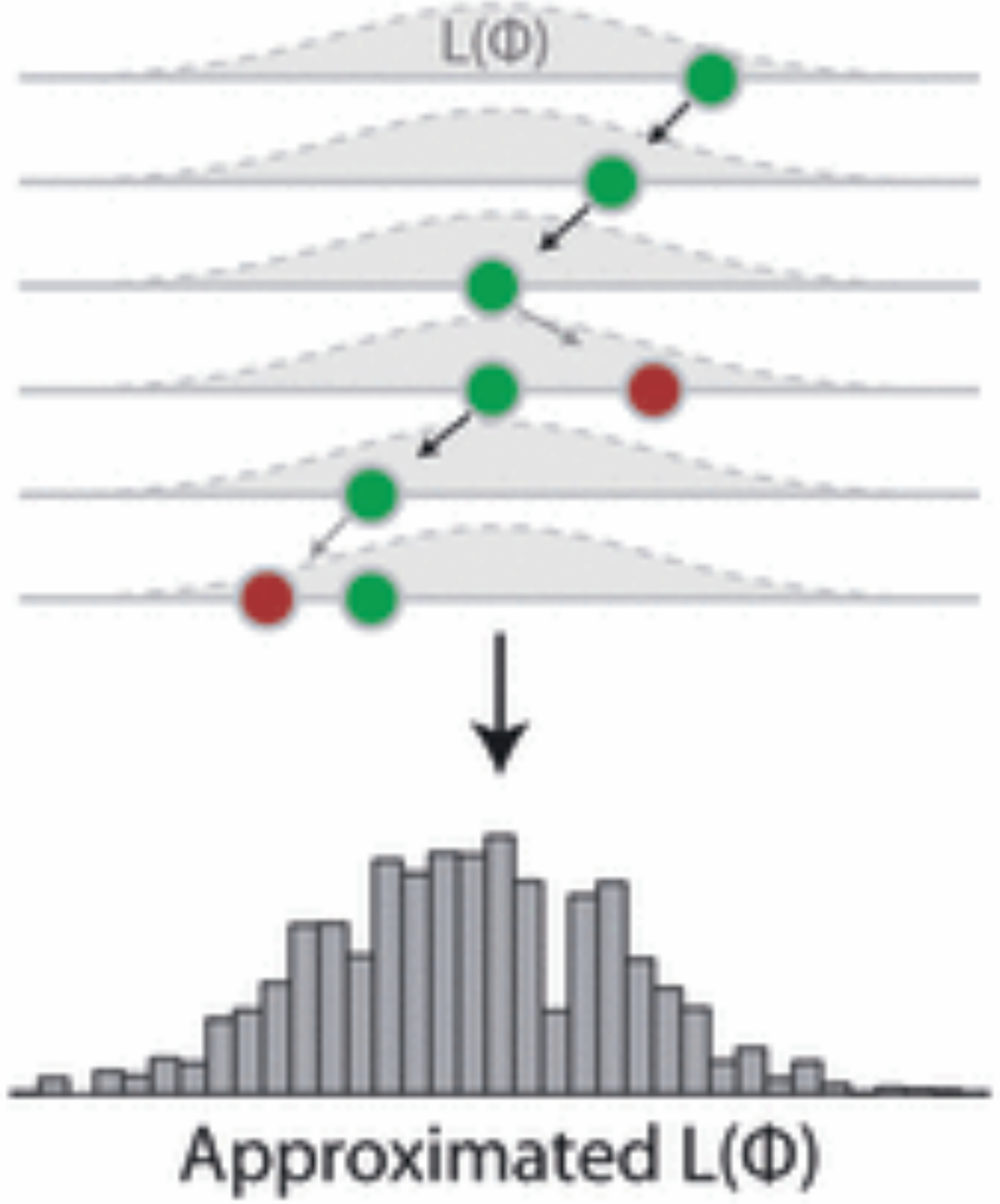
Metropolis-Hasting

- *Based on this, the transition will be*
  - $X_n \to X_{n+1} = x$ *with probability* r
  - $X_n \to X_{n+1} = X_n$ *with probability* $1 - r$
- *The transition probability is then*

$$k(\alpha,\beta) = r \cdot h(\alpha,\beta) = h(\alpha,\beta)\min\left(1, \frac{g(\beta)h(\beta,\alpha)}{g(\alpha)h(\alpha,\beta)}\right)$$

- *The reversibility condition is verified*

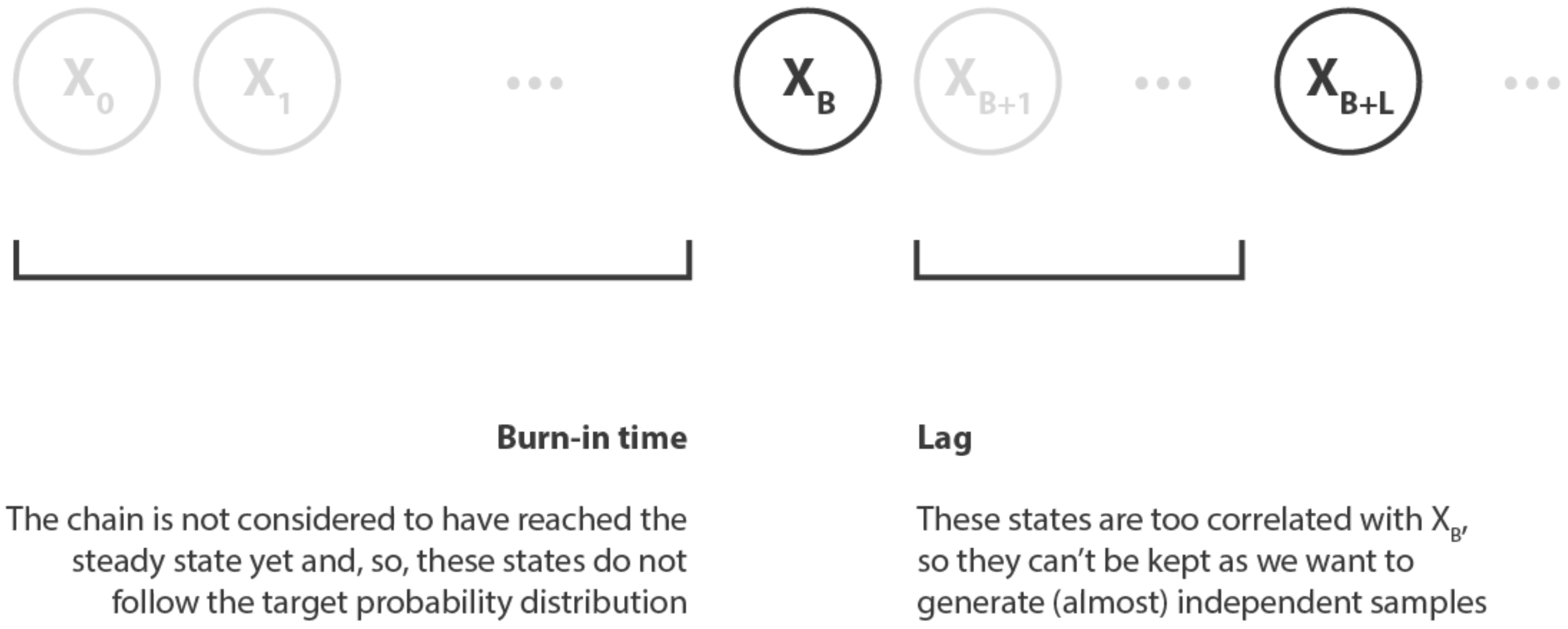
$$\boxed{\begin{aligned} g(\alpha)k(\alpha,\beta) &= g(\alpha)h(\alpha,\beta)\min\left(1, \frac{g(\beta)h(\beta,\alpha)}{g(\alpha)h(\alpha,\beta)}\right) = \min\left(g(\alpha)h(\alpha,\beta), g(\beta)h(\beta,\alpha)\right) \\ &= g(\beta)h(\beta,\alpha)\min\left(1, \frac{g(\alpha)h(\alpha,\beta)}{g(\beta)h(\beta,\alpha)}\right) = g(\beta)k(\beta,\alpha) \end{aligned}}$$



1) Draw new parameter Φ′ close to the old Φ
2) Calculate L(Φ′)
3) Jump proportional to L(Φ′)/L(Φ)

Sampling Process

- Once the chain is defined, we can simulate a random sequence of state across it
- One has to guarantee that steady state conditions are reached. This is done removing the first N samples (during time)
- One has to avoid taking two consecutive states, since they are correlated. Instead, one should wait a few steps (lag) before selecting the extra element of the chain



A simple example

◉ *Example: sampling from an exponential (the target)*

```
target = function(x){
  return(ifelse(x<0,0,exp(-x)))
}
```

Initial condition

proposed transition

acceptance probability

```
x = rep(0,10000)
x[1] = 3     #initialize; I've set arbitrarily set this to 3
for(i in 2:10000){
  current_x = x[i-1]
  proposed_x = current_x + rnorm(1,mean=0,sd=1)
  A = target(proposed_x)/target(current_x)
  if(runif(1)<A){
    x[i] = proposed_x       # accept move with probabily min(1,A)
  } else {
    x[i] = current_x        # otherwise "reject" move, and stay where we are
  }
}
```

Notice that the choice of a symmetric kernel for the exchange of the two arguments allows to simplify the definition of the acceptance rate. The transition is accepted with a probability = ratio of the probabilities of the old and newly proposed states

Summary

- *We reviewed various generalisation of the principle of indifference to choose an objective prior, based on information theory*
- *We discussed how to extract information from a posterior*
  - *estimator*
  - *credibility interval*
  - *hypothesis testing*
- *We discussed how to use Markov Chain Monte Carlo to sample from an intractable posterior*

Backup

Gibbs Sampling

- *The Gibbs Sampling method is based on the assumption that, even if the joint probability is intractable, the conditional distribution of a single dimension given the others can be computed.*

- *First we randomly choose an integer d among the K dimensions of $\vec{\theta}$. Then we sample a new value for that dimension according to the corresponding conditional probability given that all the other dimensions are kept fixed*

The initial condition is some arbitrary set of values $\vec{\theta}^{(0)}$

One also needs the i-th full conditional posterior distribution

$$\pi(\theta_i \,|\, \theta_{-i}, y) = \pi(\theta_i \,|\, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_K, y) \quad \forall i \in [1,K]$$

For a generic iteration

$$
\begin{array}{ll}
\text{Step 1.} & \text{draw } \theta_1^{(s+1)} \sim \pi(\theta_1 | \theta_2^{(s)}, \theta_3^{(s)}, \cdots, \theta_K^{(s)}, y) \\
\text{Step 2.} & \text{draw } \theta_2^{(s+1)} \sim \pi(\theta_2 | \theta_1^{(s+1)}, \theta_3^{(s)}, \cdots, \theta_K^{(s)}, y) \\
\vdots & \\
\text{Step i.} & \text{draw } \theta_i^{(s+1)} \sim \pi(\theta_i | \theta_1^{(s+1)}, \theta_2^{(s+1)}, \cdots, \theta_{i-1}^{(s+1)}, \theta_{i+1}^{(s)}, \cdots, \theta_K^{(s)}, y) \\
\text{Step i+1.} & \text{draw } \theta_{i+1}^{(s+1)} \sim \pi(\theta_{i+1} | \theta_1^{(s+1)}, \theta_2^{(s+1)}, \cdots, \theta_i^{(s+1)}, \theta_{i+2}^{(s)}, \cdots, \theta_K^{(s)}, y) \\
\vdots & \\
\text{Step K.} & \text{draw } \theta_K^{(s+1)} \sim \pi(\theta_K | \theta_1^{(s+1)}, \theta_2^{(s+1)}, \cdots, \theta_{K-1}^{(s+1)}, y)
\end{array}
$$

Gibbs Sampling