Introduction to Bioinformatics for Computer Scientists

Lecture 12

Outline

- Bayesian statistics
- Monte-Carlo simulations
- Markov-Chain Monte-Carlo (MCMC) methods
- Metropolis-coupled MCMC-methods
- Some phylogenetic proposals
- Reversible jump MCMC

Bayesian and Maximum Likelihood Inference

- In phylogenetics Bayesian and ML (Maximum Likelihood) methods have **a lot** in common
- Computationally, both approaches re-evaluate the phylogenetic likelihood *over and over and over* again for different tree topologies, branch lengths, and model parameters
- Bayesian and ML codes spend approx. 80-95% of their total run time in likelihood calculations on trees
- Bayesian methods sample the posterior probability **distribution**
- ML methods strive to find a **point estimate** that maximizes the likelihood

Bayesian Phylogenetic Methods

- The methods used perform stochastic searches, that is, they do not strive to maximize the likelihood, but rather integrate over it
- Thus, no numerical optimization methods for model parameters and branch lengths are needed, parameters are proposed at random
- It is substantially easier to infer trees under complex models using Bayesian statistics than using Maximum Likelihood

A Review of Probabilities

Hair color

Eye color	brown	blonde	Σ
light	5/40	15/40	20/40
dark	15/40	5/40	20/40
Σ	20/40	20/40	**40/40**

A Review of Probabilities

Hair color

Eye color	brown	blonde	Σ
light	5/40	15/40	20/40
dark	15/40	5/40	20/40
Σ	20/40	20/40	**40/40**

Joint probability: probability of observing both A and B: *Pr(A,B)*
For instance, *Pr(brown, light) = 5/40 = 0.125*

A Review of Probabilities

Hair color

Eye color

	brown	blonde	Σ
light	5/40	15/40	20/40
dark	15/40	5/40	20/40
Σ	20/40	20/40	**40/40**

Marginalize over hair color

Marginal Probability: *unconditional* probability of an observation *Pr(A)*
For instance, *Pr(dark) = Pr(dark,brown) + Pr(dark,blonde) = 15/40 + 5/40 = 20/40 = 0.5*

A Review of Probabilities

Hair color

Eye color	brown	blonde	Σ
light	5/40	15/40	20/40
dark	15/40	5/40	20/40
Σ	20/40	20/40	**40/40**

Conditional Probability: The probability of observing A given that B has occurred:
Pr(A|B) is the fraction of cases *Pr(B)* in which *B* occurs where *A* also occurs with *Pr(AB)*
Pr(A|B) = Pr(AB) / Pr(B)

For instance, *Pr(blonde|light) = Pr(blonde,light) / Pr(light) = (15/40) / (20/40) = 0.75*

A Review of Probabilities

Hair color

Eye color	brown	blonde	Σ
light	5/40	15/40	20/40
dark	15/40	5/40	20/40
Σ	20/40	20/40	**40/40**

Statistical Independence: Two events A and B are independent
If their joint probability *Pr(A,B)* equals the product of their marginal probability *Pr(A) Pr(B)*

For instance, *Pr(light,brown) ≠ Pr(light) Pr(brown)*, that is, the events are not independent!

A Review of Probabilities

Conditional Probability:

$Pr(A|B) = Pr(A,B) / Pr(B)$

Joint Probability:

$Pr(A,B) = Pr(A|B)\ Pr(B)$

and

$Pr(A,B) = Pr(B|A)\ Pr(A)$

Problem:

If I can compute $Pr(A|B)$ how can I get $Pr(B|A)$?

A Review of Probabilities

Conditional Probability:

$Pr(A|B) = Pr(A,B) / Pr(B)$

Joint Probability:

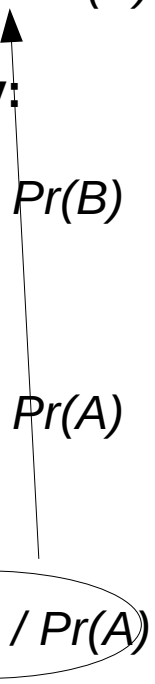
$Pr(A,B) = Pr(A|B)\ Pr(B)$

and

$Pr(A,B) = Pr(B|A)\ Pr(A)$

Bayes Theorem:

$Pr(B|A) = Pr(A,B) / Pr(A)$



A Review of Probabilities

Conditional Probability:

$Pr(A|B) = Pr(A,B) / Pr(B)$

Joint Probability:

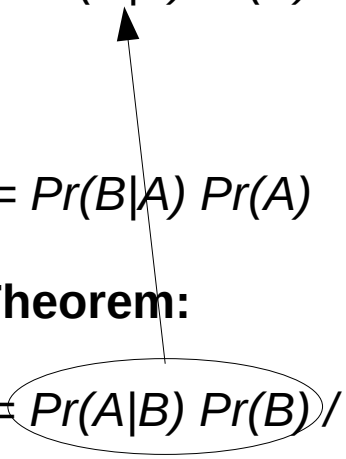
$Pr(A,B) = Pr(A|B)\ Pr(B)$

and

$Pr(A,B) = Pr(B|A)\ Pr(A)$

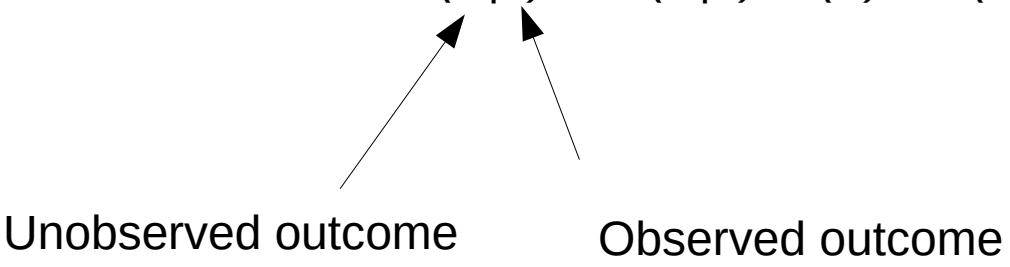
Bayes Theorem:

$Pr(B|A) = Pr(A|B)\ Pr(B) / Pr(A)$



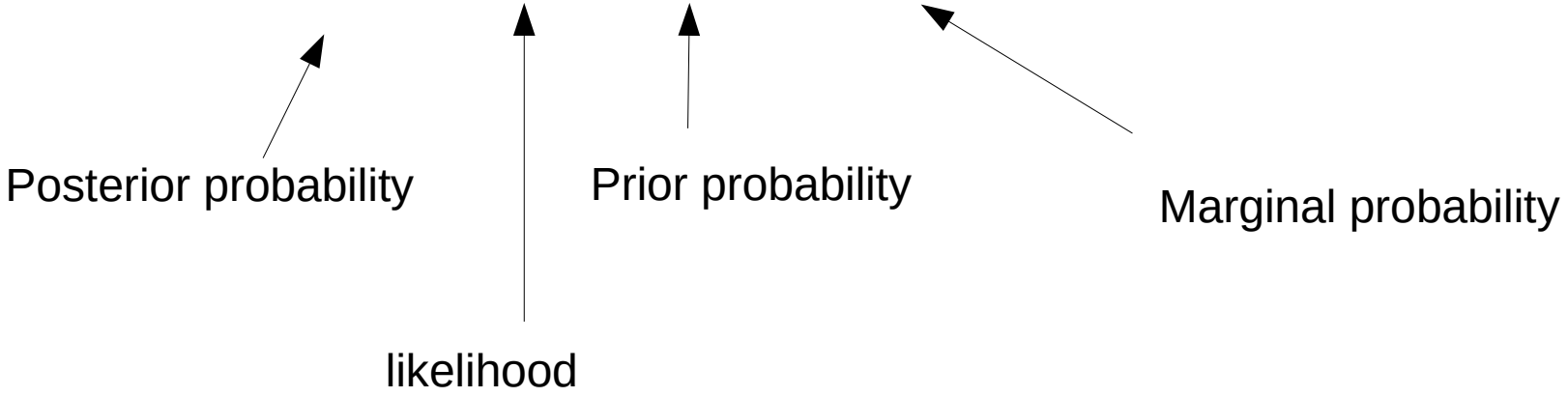
Bayes Theorem

$$Pr(B|A) = Pr(A|B)\ Pr(B) / Pr(A)$$

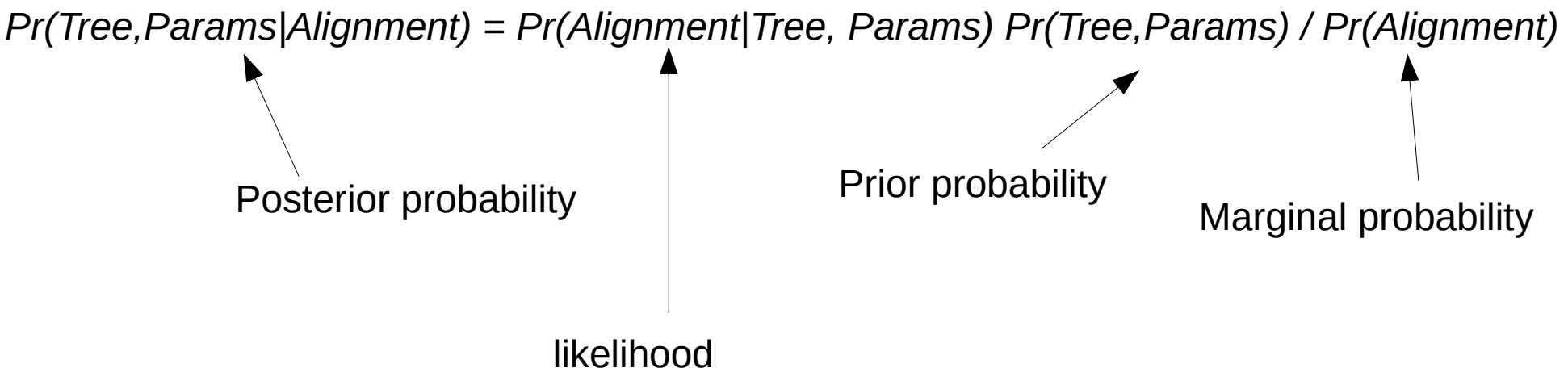


Bayes Theorem

$$Pr(B|A) = Pr(A|B)\ Pr(B) / Pr(A)$$



Bayes Theorem: Phylogenetics

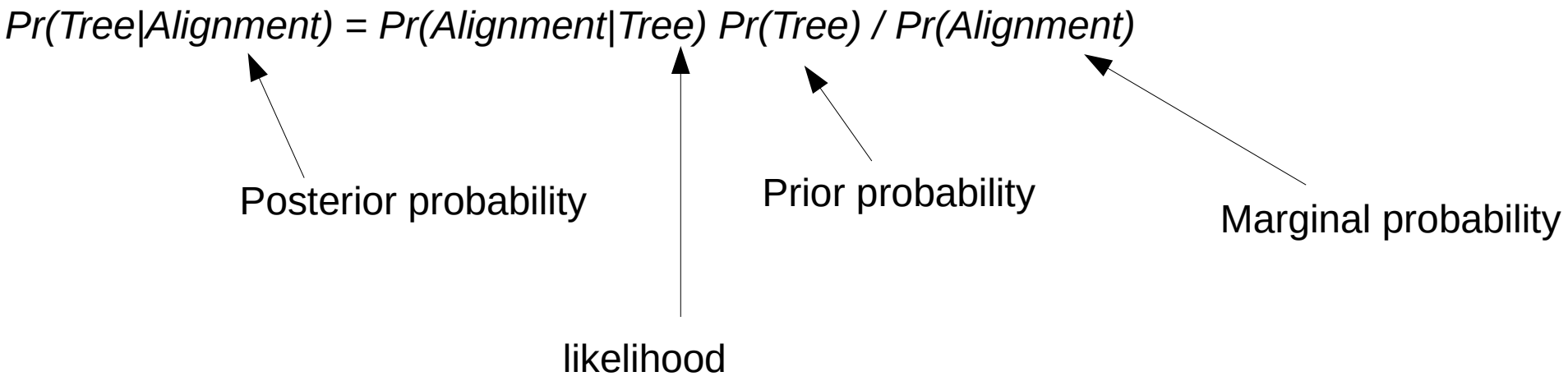


Posterior probability: distribution over all possible trees and all model parameter values
Likelihood: does the alignment fit the tree and model parameters?
Prior probability: introduces prior knowledge/assumptions about the probability distribution of trees and model parameters (e.g., GTR rates, α shape parameter).

For instance, we typically assume that all possible tree topologies are equally probable → uniform prior

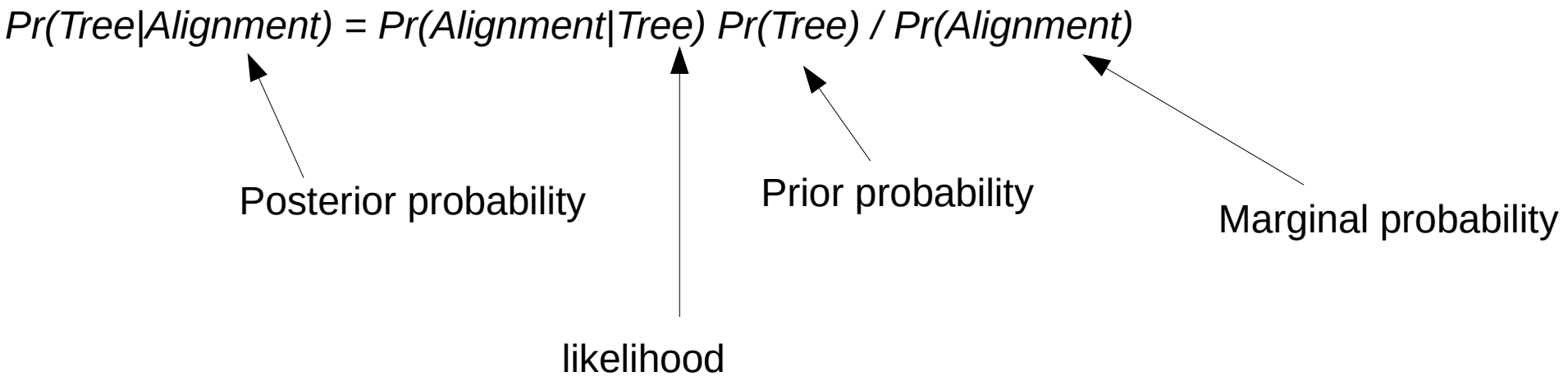
Marginal probability: how do we obtain this?

Bayes Theorem: Phylogenetics



Marginal probability: Assume that our only model parameter is the tree and marginalizing means summing over all unconditional probabilities, thus
Pr(Alignment)
can be written as
$Pr(Alignment) = Pr(Alignment, t_0) + Pr(Alignment, t_1) + \ldots + Pr(Alignemnt, t_n)$
where *n+1* is the number of possible trees!

Bayes Theorem: Phylogenetics



Marginal probability: Assume that our only model parameter is the tree and marginalizing means summing over all unconditional probabilities, thus

Pr(Alignment)

can be written as

$Pr(Alignment) = Pr(Alignment, t_0) + Pr(Alignment,t_1) + \ldots + Pr(Alignemnt, t_n)$

where *n+1* is the number of possible trees!

This can be re-written as

$Pr(Alignment) = Pr(Alignment|t_0)\ Pr(t_0) + Pr(Alignment|t_1)\ Pr(t_1) + \ldots + Pr(Alignment|t_n)\ Pr(t_n)$

Bayes Theorem: Phylogenetics

$$Pr(Tree|Alignment) = Pr(Alignment|Tree)\ Pr(Tree) / Pr(Alignment)$$

Posterior probability

likelihood

Prior probability

Marginal probability

Marginal probability:

$$Pr(Alignment) = Pr(Alignment|t_0)\ Pr(t_0) + Pr(Alignment|t_1)\ Pr(t_1) + \ldots + Pr(Alignment|t_n)\ Pr(t_n)$$

likelihood

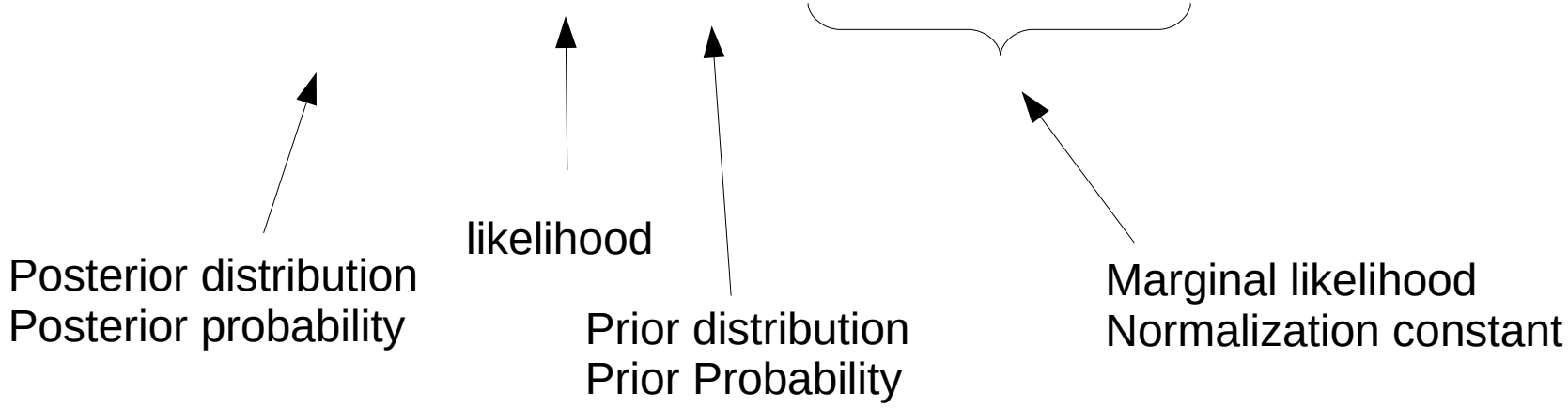
Prior := *1 / (n+1)*
→ this is a uniform prior!

Now, we have all the ingredients for computing *Pr(Tree|Alignment)*, however computing *Pr(Alignment)* is prohibitive due to the large number of trees!

With continuous parameters the above equation for obtaining the marginal probability becomes an integral. Usually, all parameters we integrate over (tree topology, model parameters, etc.) are lumped into a parameter vector denoted by θ

Bayes Theorem General Form

$$f(\theta|A) = f(A|\theta)\, f(\theta) \,/\, \int f(\theta) f(A|\theta) d\theta$$



We know how to compute $f(A|\theta)$ → the likelihood of the tree

Problems:

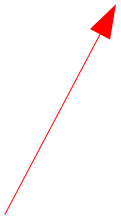
Problem 1: $f(\theta)$ is given a priori, but how do we chose an appropriate distribution?
→ biggest strength and weakness of Bayesian approaches

Problem 2: How can we calculate/approximate $\int f(\theta) f(A|\theta) d\theta$?
→ to explain this we need to introduce additional machinery

However, let us first look at an example for $f(\theta|A)$ *in* phylogenetics

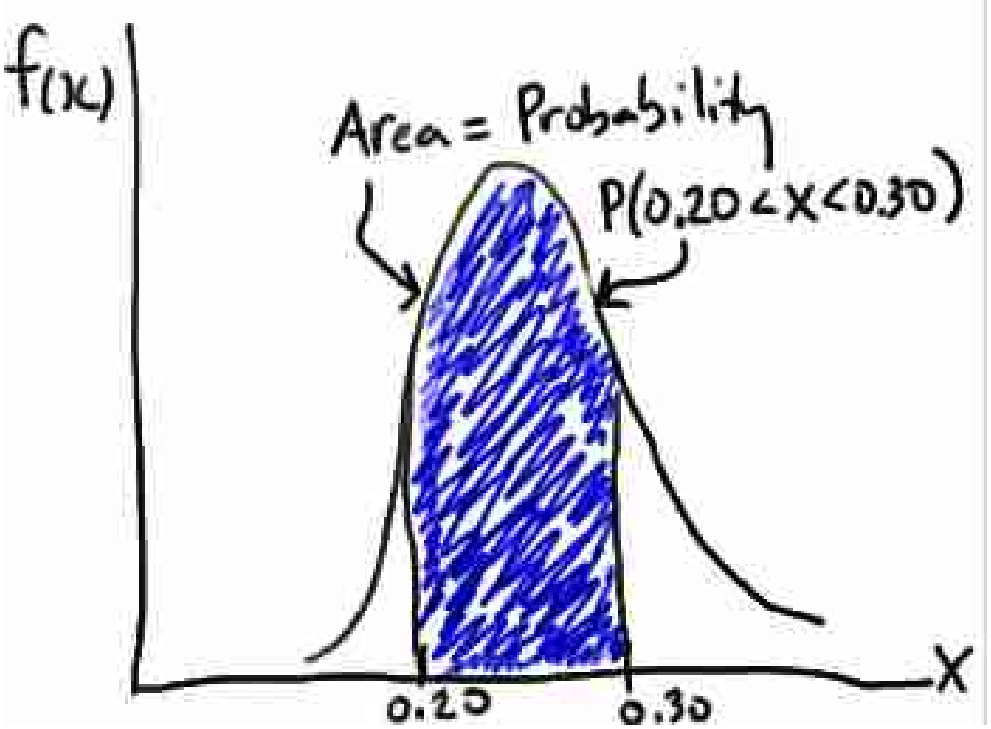
Bayes Theorem General Form

$$f(\theta|A) = f(A|\theta)\, f(\theta) \,/\, \int f(\theta) f(A|\theta) d\theta$$



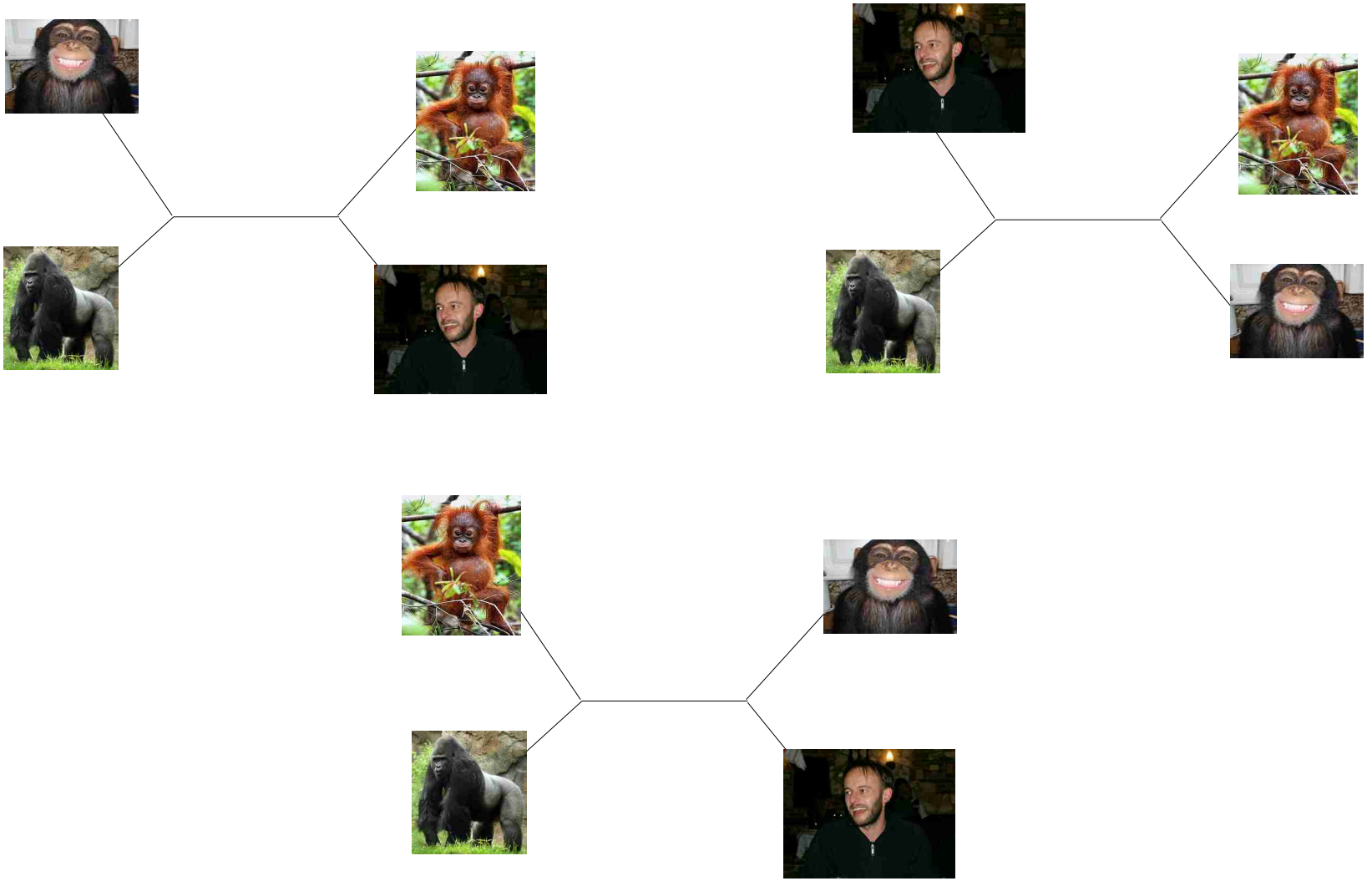
Note that, in the continuous case *f()* is called probability density function

Probability Density Function

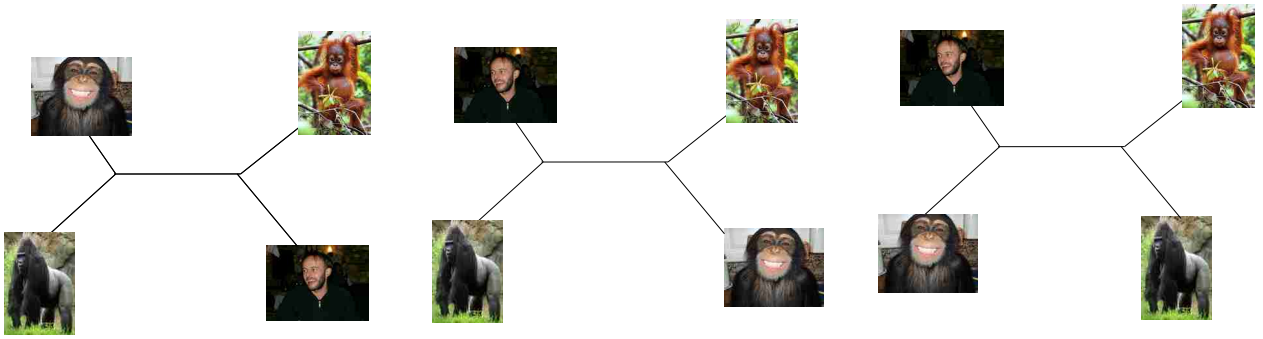


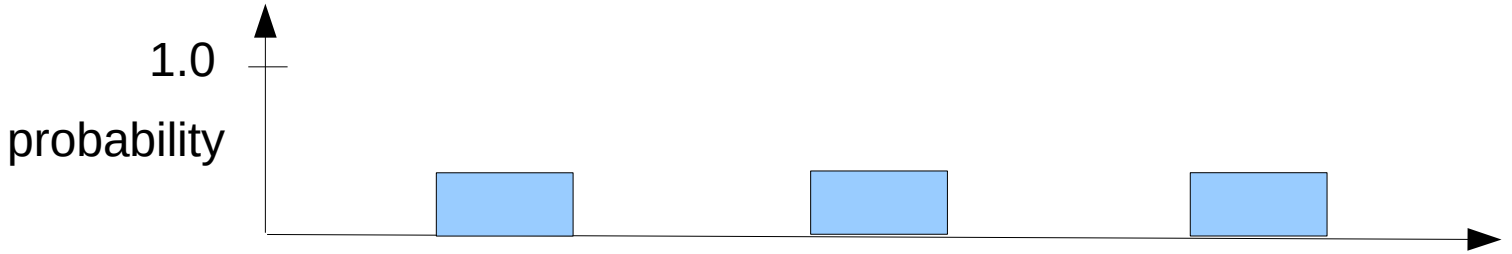
Properties:
1. *f(x) > 0* for all allowed values *x*
2.The area under *f(x)* is *1.0*
3.The probability that *x* falls into an interval (e.g. *0.2 – 0.3*) is given by the integral of *f(x)* over this interval

An Example



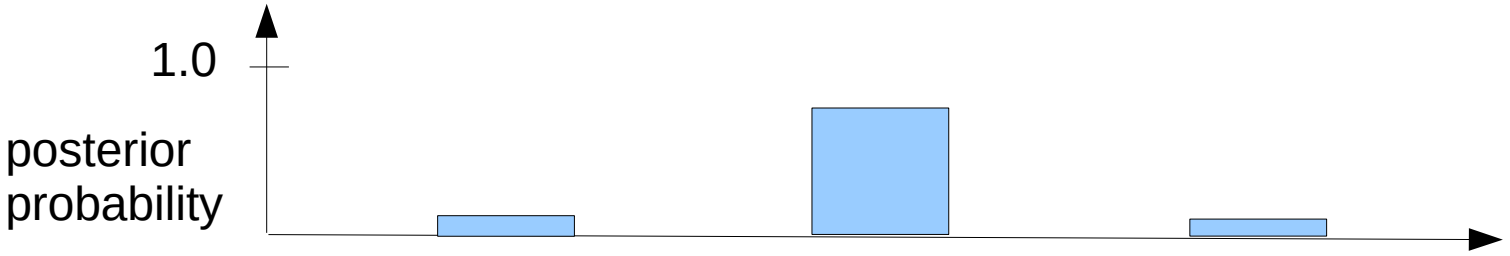
An Example





Prior distribution

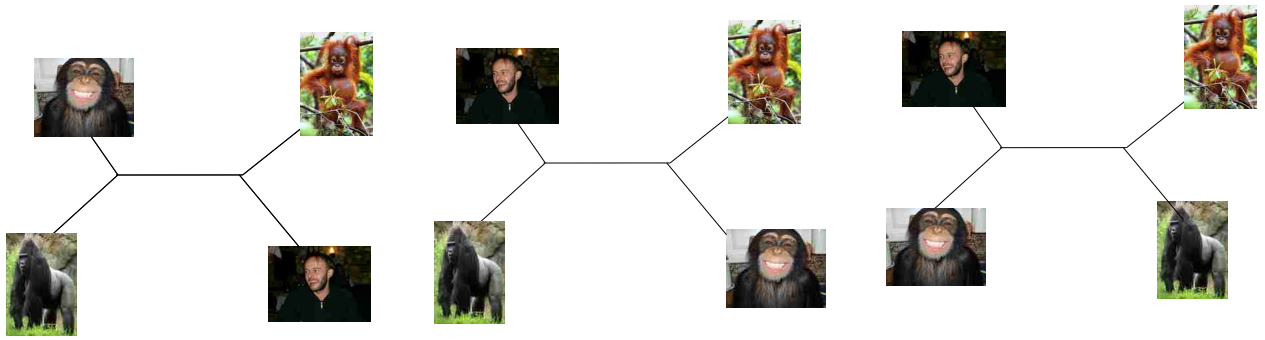
Data (observations → sequences)

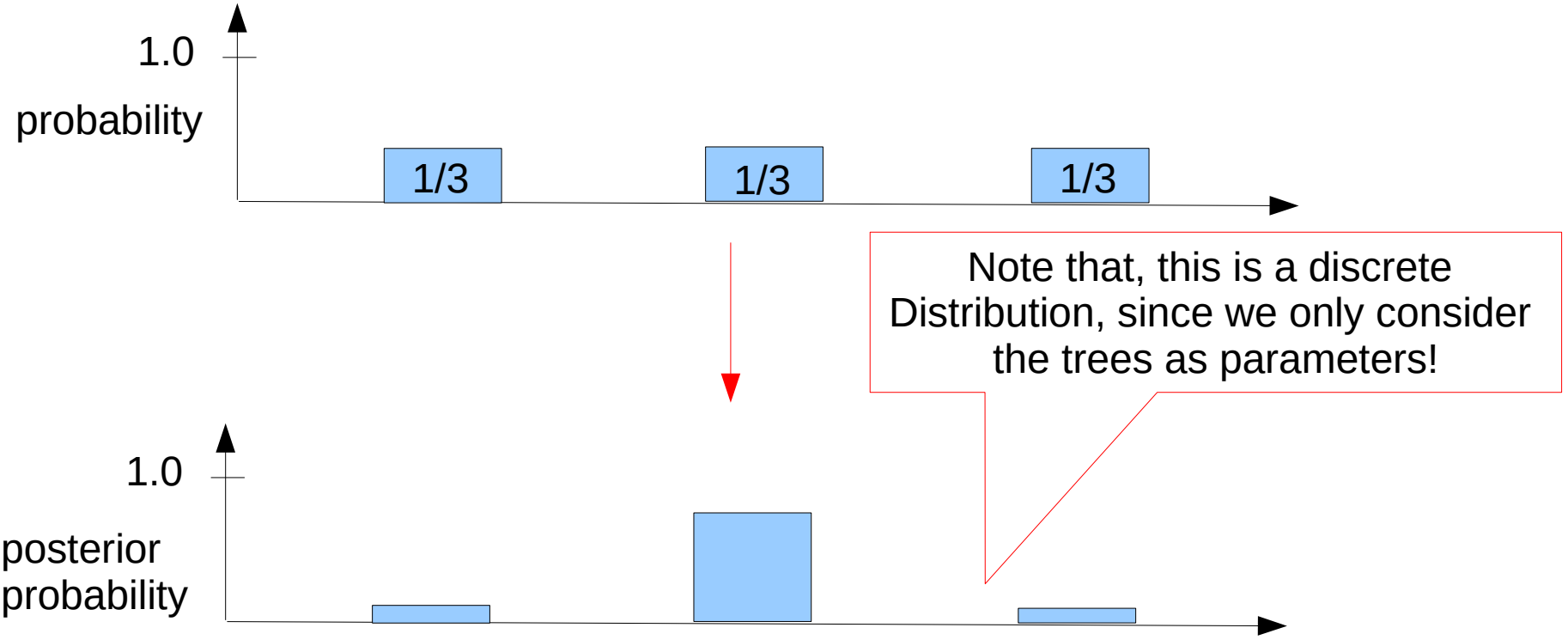


Posterior distribution

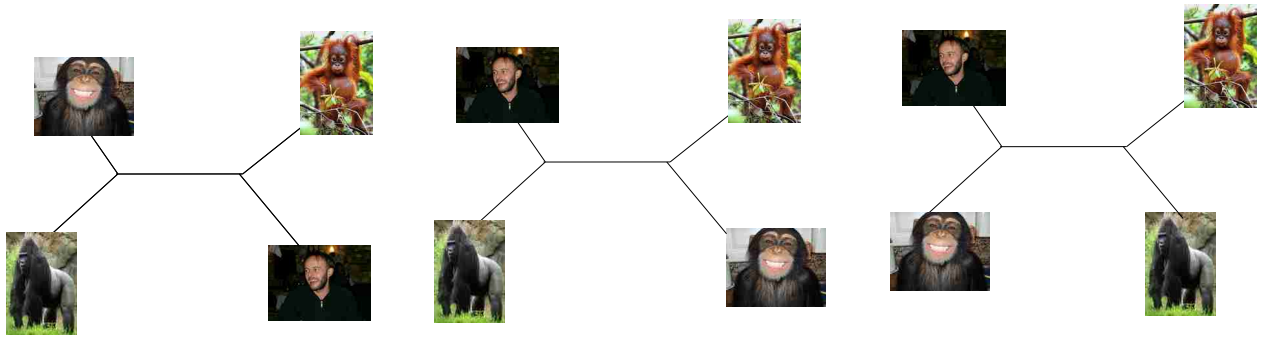
Parameter space → 3 distinct tree topologies

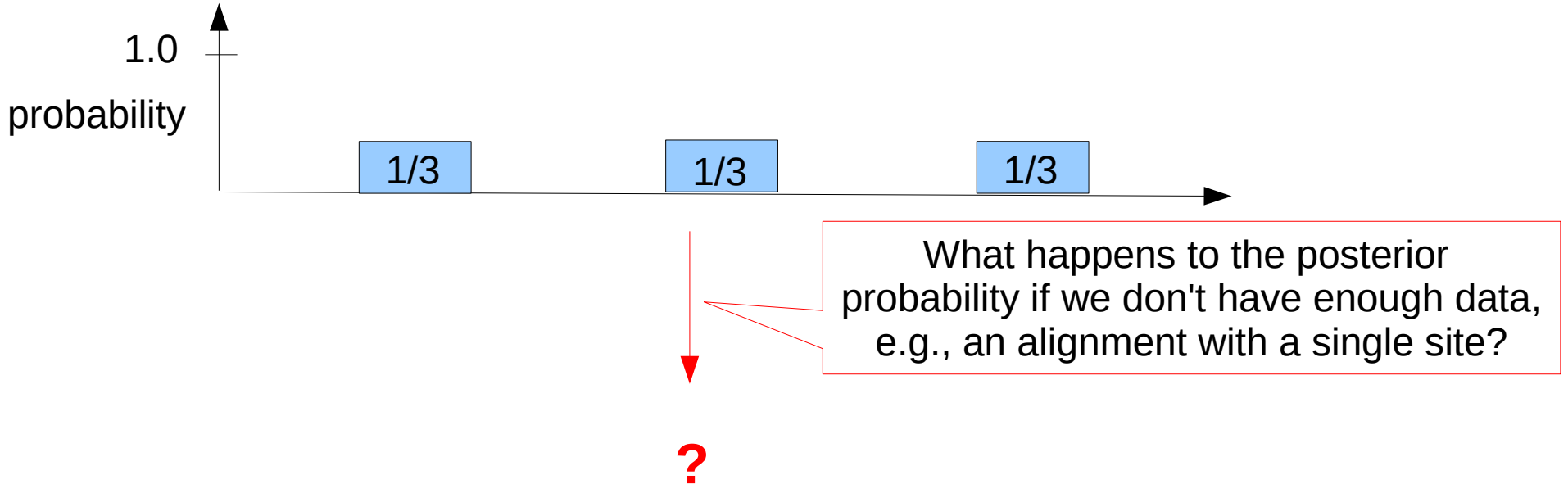
An Example




1.0
probability
1/3
1/3
1/3
Note that, this is a discrete Distribution, since we only consider the trees as parameters!
1.0
posterior probability
Parameter space → 3 distinct tree topologies

An Example



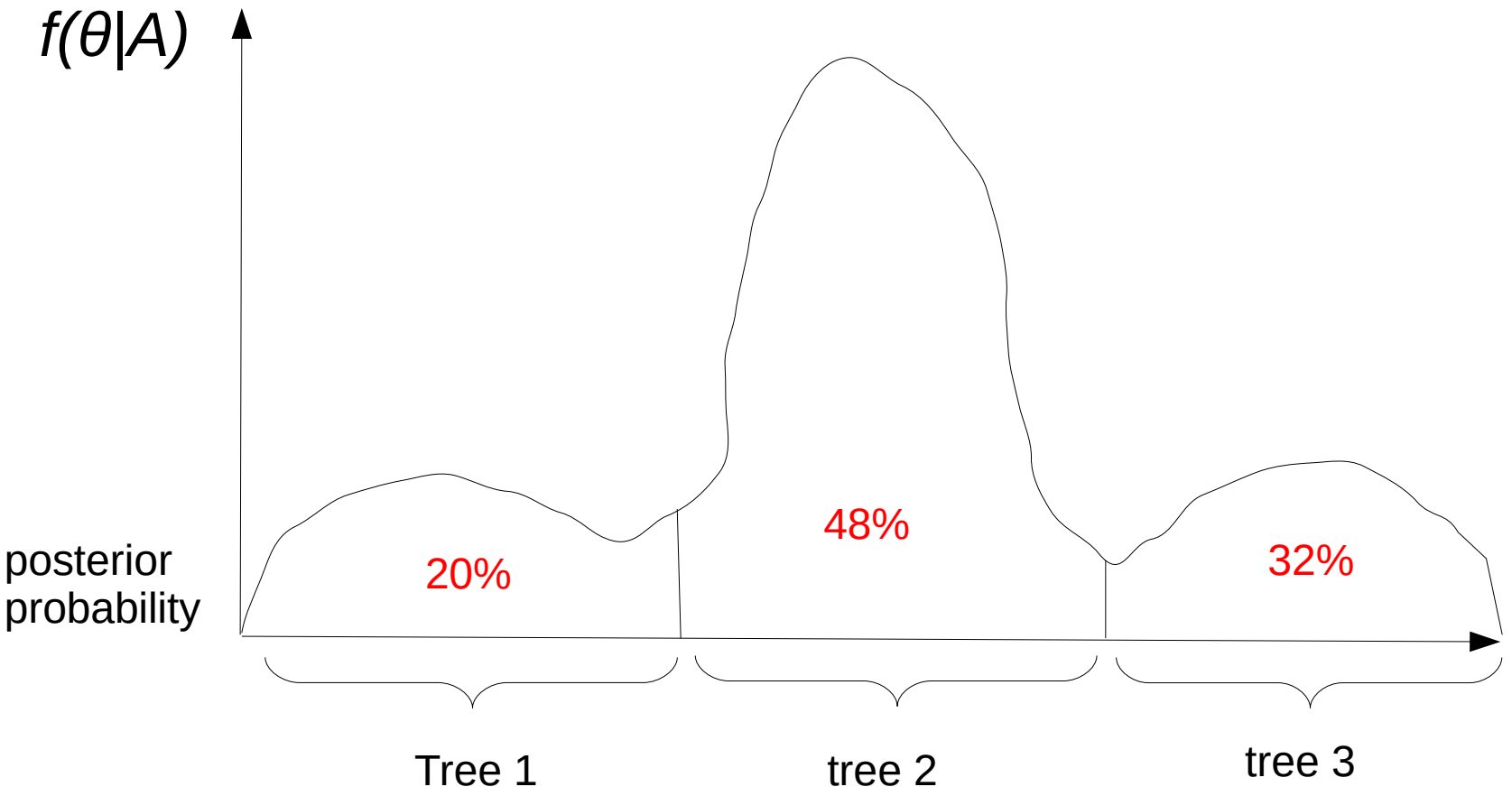

1.0
probability
1/3
1/3
1/3
What happens to the posterior probability if we don't have enough data, e.g., an alignment with a single site?
?
posterior probability

An Example

Include additional model parameters such as branch lengths,GTR rates, and the α-shape paremeter of the *Γ* distribution into the model:
θ = (tree, α, branch-lengths, GTR-rates)

An Example

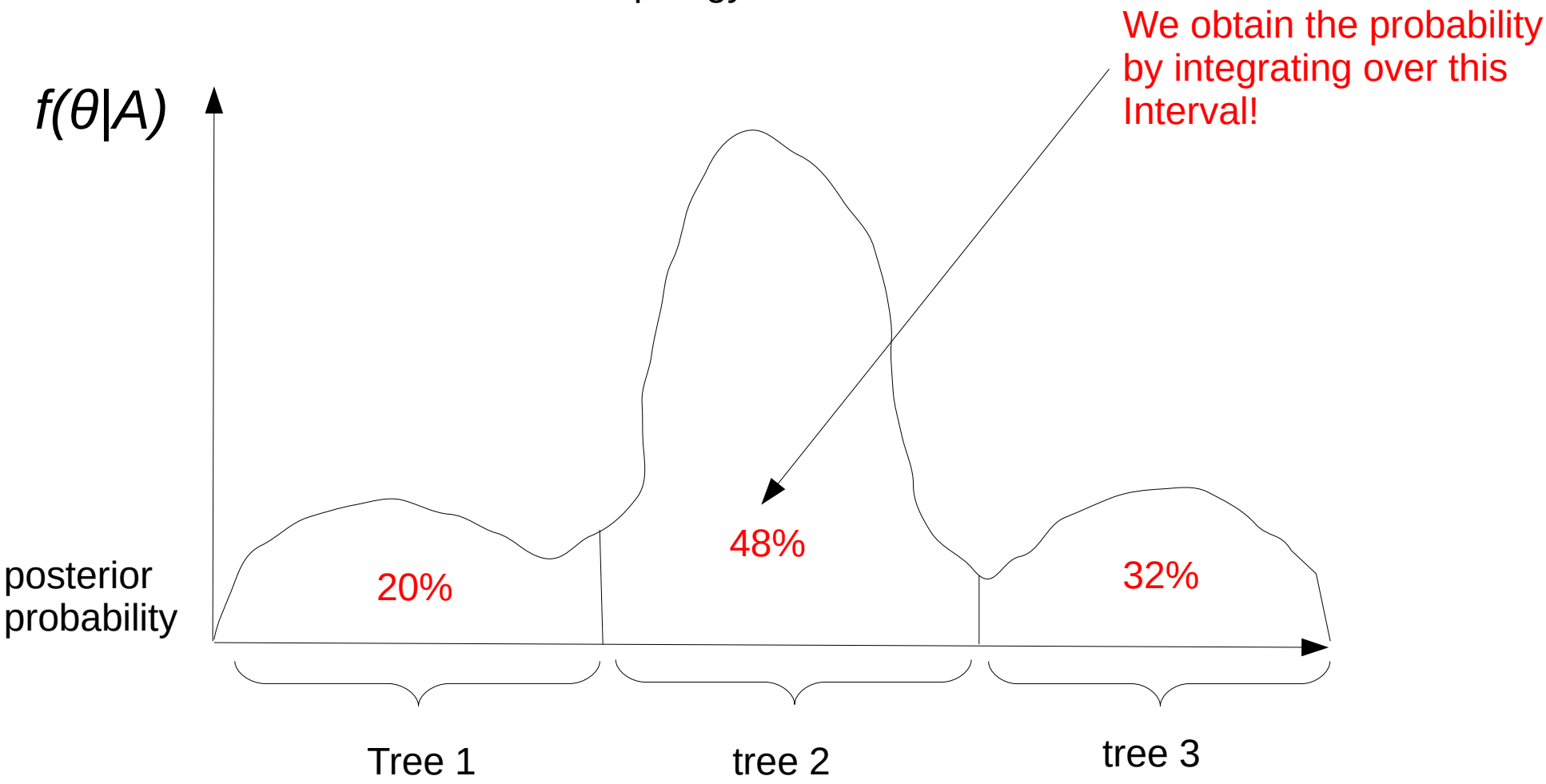
We can look at this distribution for any parameter of interest by marginalizing (integrating out) all other parameters.
Here we focus on the tree topology.



Marginal probability distribution of trees

An Example

We can look at this distribution for any parameter of interest by marginalizing (integrating out) all other parameters.
Here we focus on the tree topology.



Marginal probability distribution of trees

Marginalization

Marginal probabilities of α values

trees

	t_1	t_2	t_3	
$\alpha_1 = 0.5$	0.10	0.07	0.12	0.29
$\alpha_2 = 1.0$	0.05	0.22	0.06	0.33
$\alpha_3 = 5.0$	0.05	0.19	0.14	0.38
	0.20	0.48	0.32	**1.0**

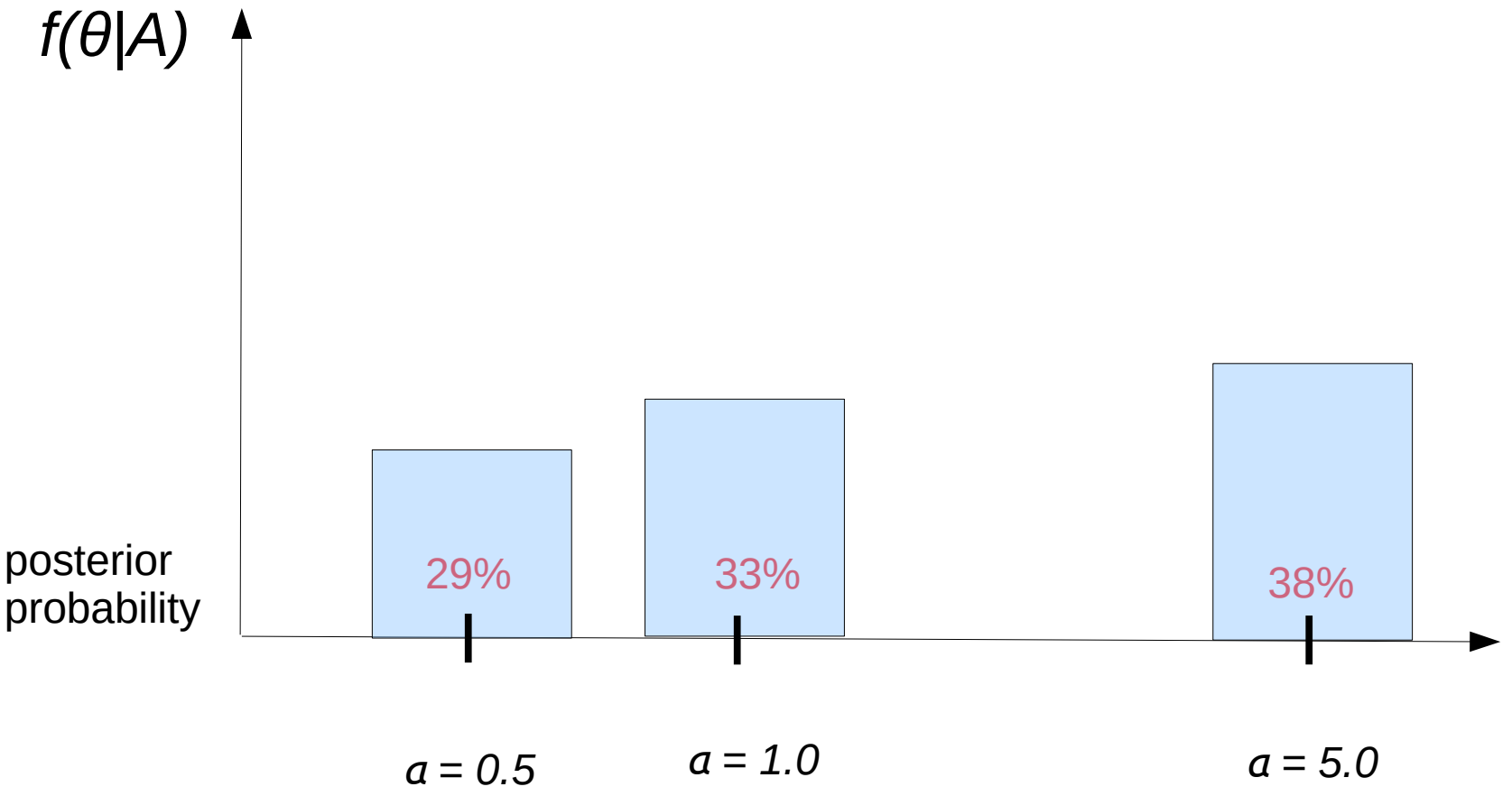
Three discrete Values of the α-shape parameter

Joint probabilities

Marginal probabilities of trees

An Example

We can look at this distribution for any parameter of interest by marginalizing (integrating out) all other parameters.
Here we focus on the three discrete α values.



Marginal probability distibution of *α*

Bayes versus Likelihood

likelihood

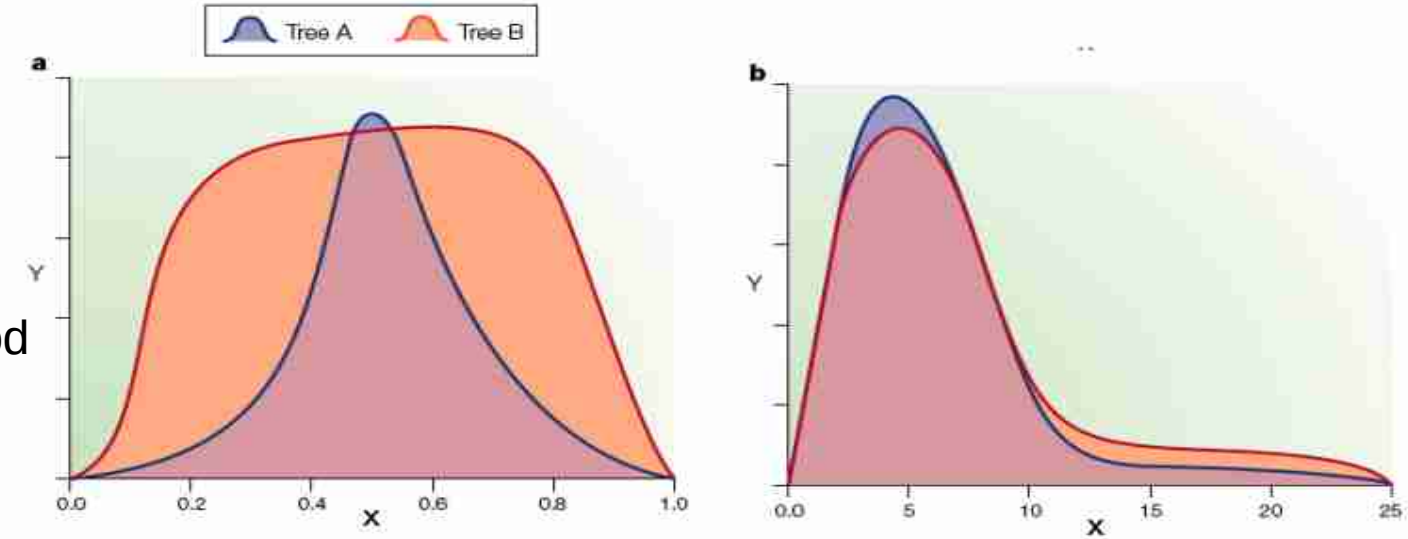


Figure 1 | **Contrast between marginal and joint estimation.** Panels **a** and **b** depict the likelihood profile for two trees versus a hypothetical parameter *x*. The *x* axis represents some nuisance parameter (for example, the ratio of the rate of transitions to the rate of transversions). The *y* axis represents the likelihood in the case of ML, or the posterior-probability density in a Bayesian approach. The area under the likelihood curve for tree A is shown in light blue, the area for tree B is shown in orange. Mauve regions are under the curve for both trees. In both cases, jointly estimating *x* and the tree favours tree A (that is, the highest peak is blue in both cases), but marginalizing over *x* favours tree B (that is, the orange area is greater than the blue area).

ML: Joint estimation
Bayesian: Marginal estimation

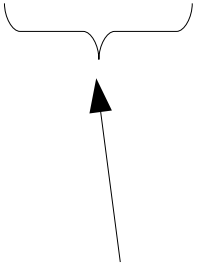
See: Holder & Lewis
“Phylogeny Estimation: traditional & Bayesian Approaches” Link to paper

Outline

- Bayesian statistics
- Monte-Carlo simulation & integration
- Markov-Chain Monte-Carlo methods
- Metropolis-coupled MCMC-methods
- Some phylogenetic proposals
- Reversible jump MCMC

Bayes Theorem General Form

$$f(\theta|A) = (likelihood * prior) / ouch$$



Marginal likelihood
Normalization constant
→ difficult to calculate

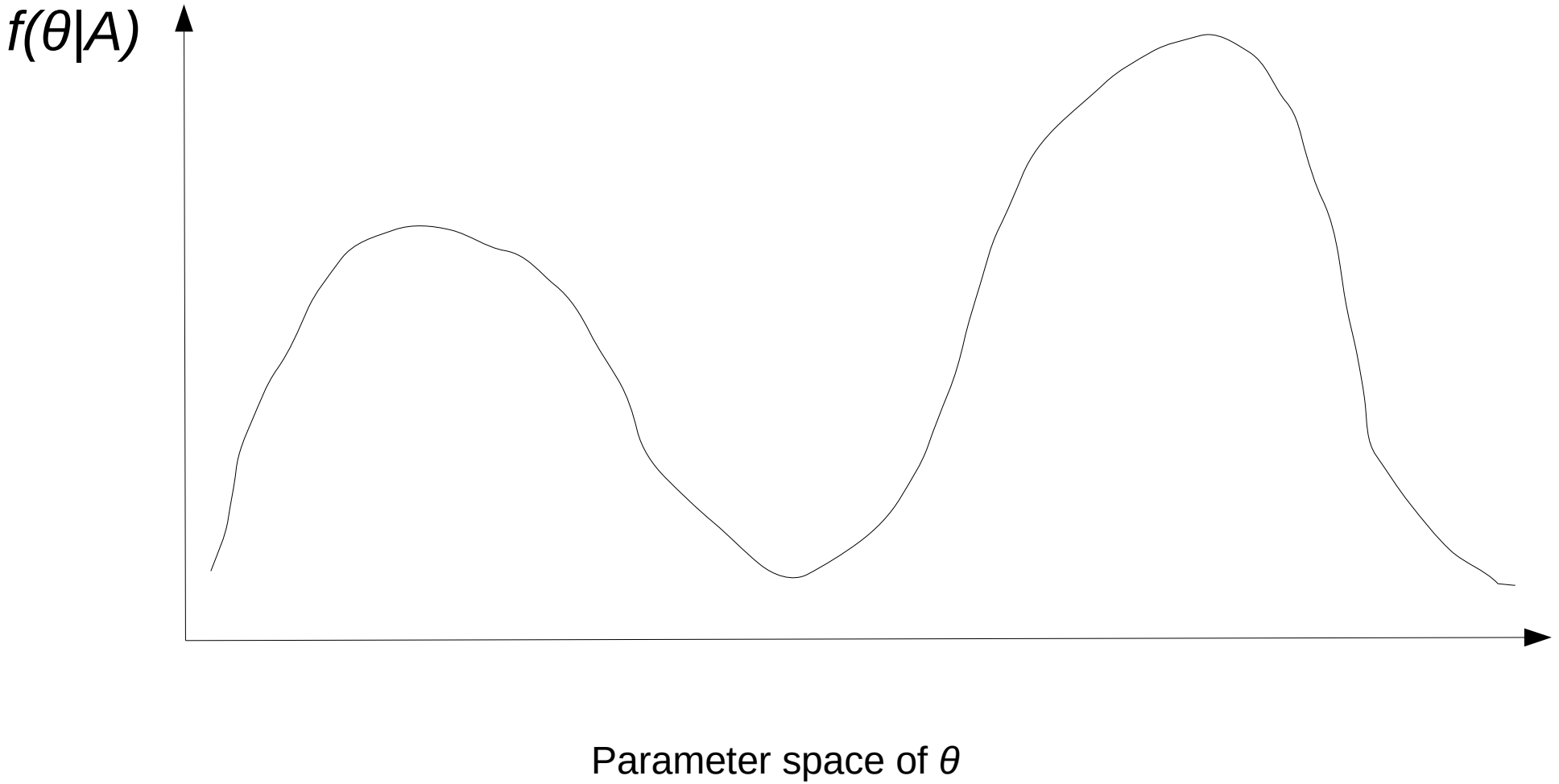
We know how to compute $f(A|\theta)$ → the likelihood of the tree

Problems:

Problem 1: $f(\theta)$ is given a priori, but how do we chose an appropriate distribution?
→ biggest strength and weakness of Bayesian approaches

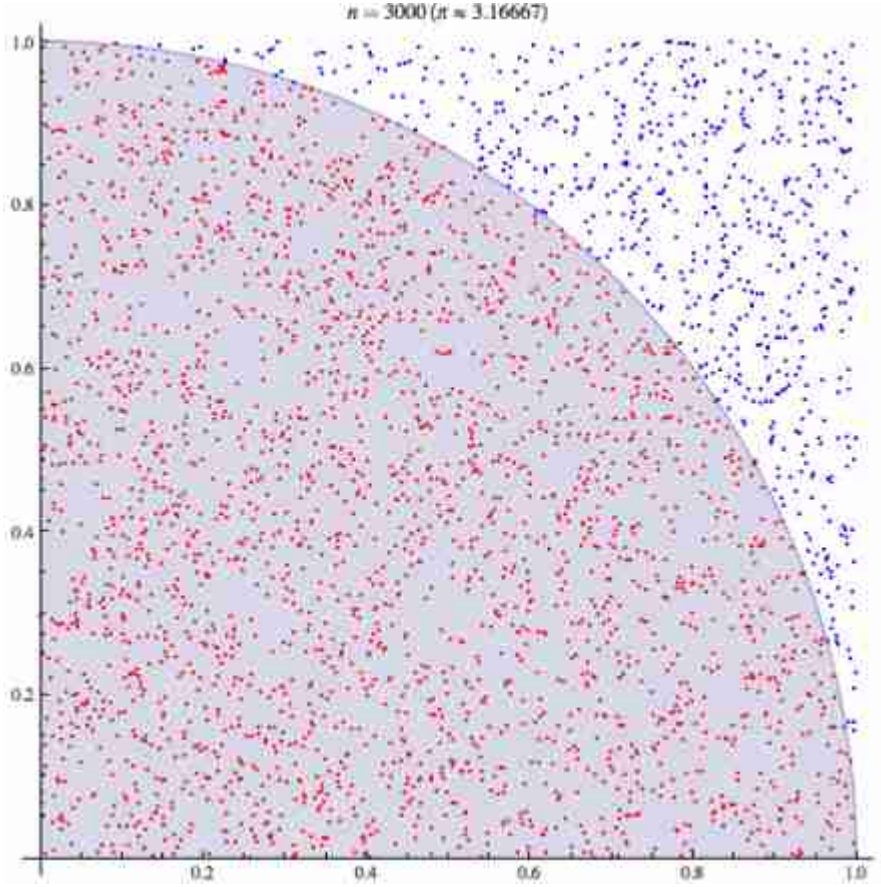
Problem 2: How can we calculate/approximate $\int f(\theta)f(A|\theta)d\theta$
→ to explain this we need to introduce additional machinery to design methods for numerical integration

How can we compute this integral?



The Classic Example

- Calculating π (the geometric constant!) with Monte-Carlo



Procedure:
1. Randomly throw points onto the rectangle n times
2. Count how many points fall into the circle n_i
3. determine π as the ratio n / n_i

→ this yields an approximation of the ratio of the areas (the square and the circle)

Monte Carlo Integration

- Method for numerical integration of *m*-dimensional integrals over *R*:

  $\int f(\theta)d\theta \approx 1/N \sum f(\theta_i)$

  where θ *is from domain* R^m

- More precisely, if the integral $\int$ *is* defined over a domain/volume *V* the equation becomes: $V * 1/N * \sum f(\theta_i)$

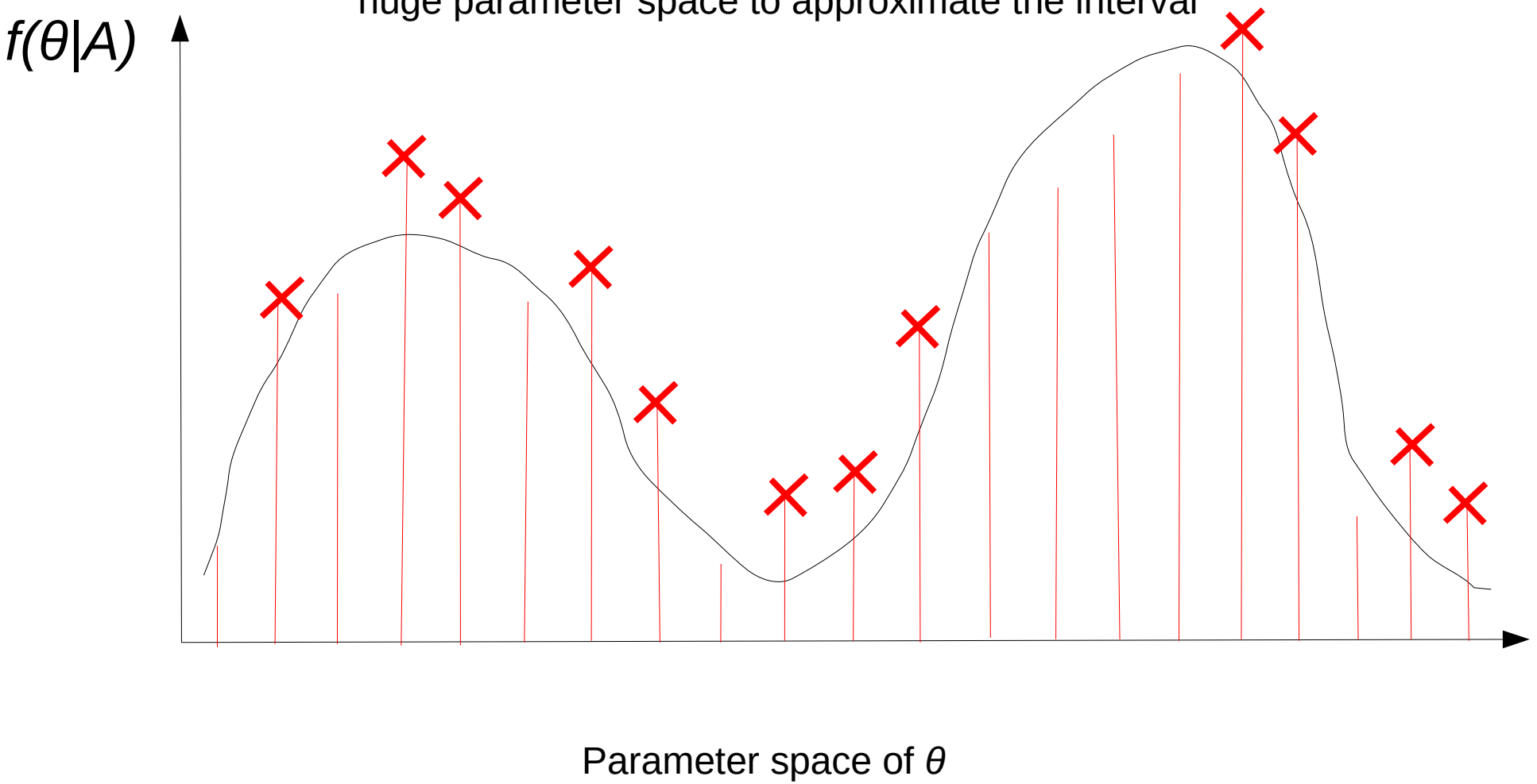
- Key issues:
  - Monte Carlo simulations draw samples θ_i of function *f()* completely at random → random grid
  - How many points do we need to sample for a 'good' approximation?
  - Domain R^m might be too large for random sampling!

Outline

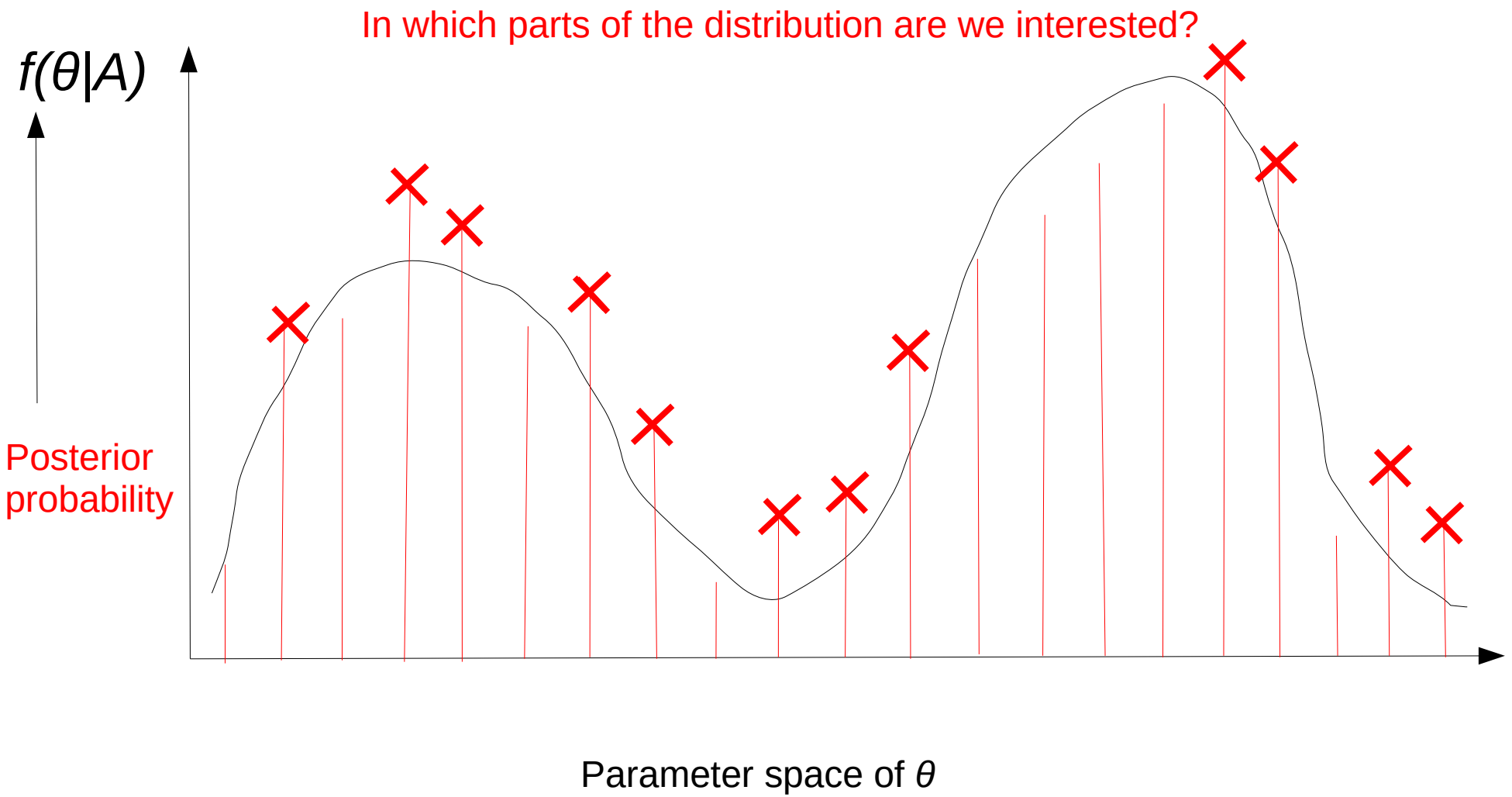
- Bayesian statistics
- Monte-Carlo simulation & integration
- Markov-Chain Monte-Carlo methods
- Metropolis-coupled MCMC-methods
- Some phylogenetic proposals
- Reversible jump MCMC

How can we compute this integral?

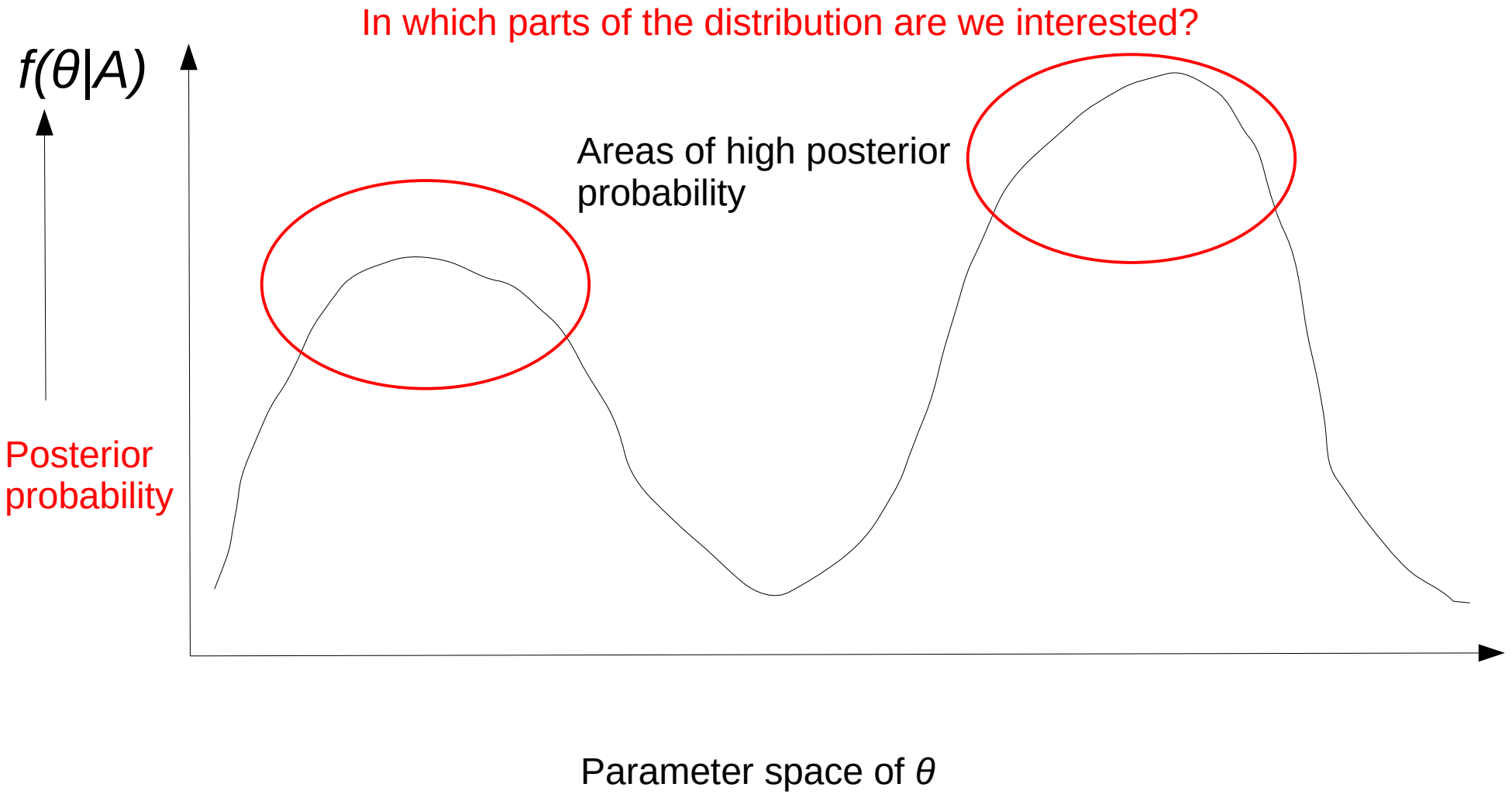
Monte-Carlo Methods: randomly sample data-points in this huge parameter space to approximate the interval



How can we compute this integral?

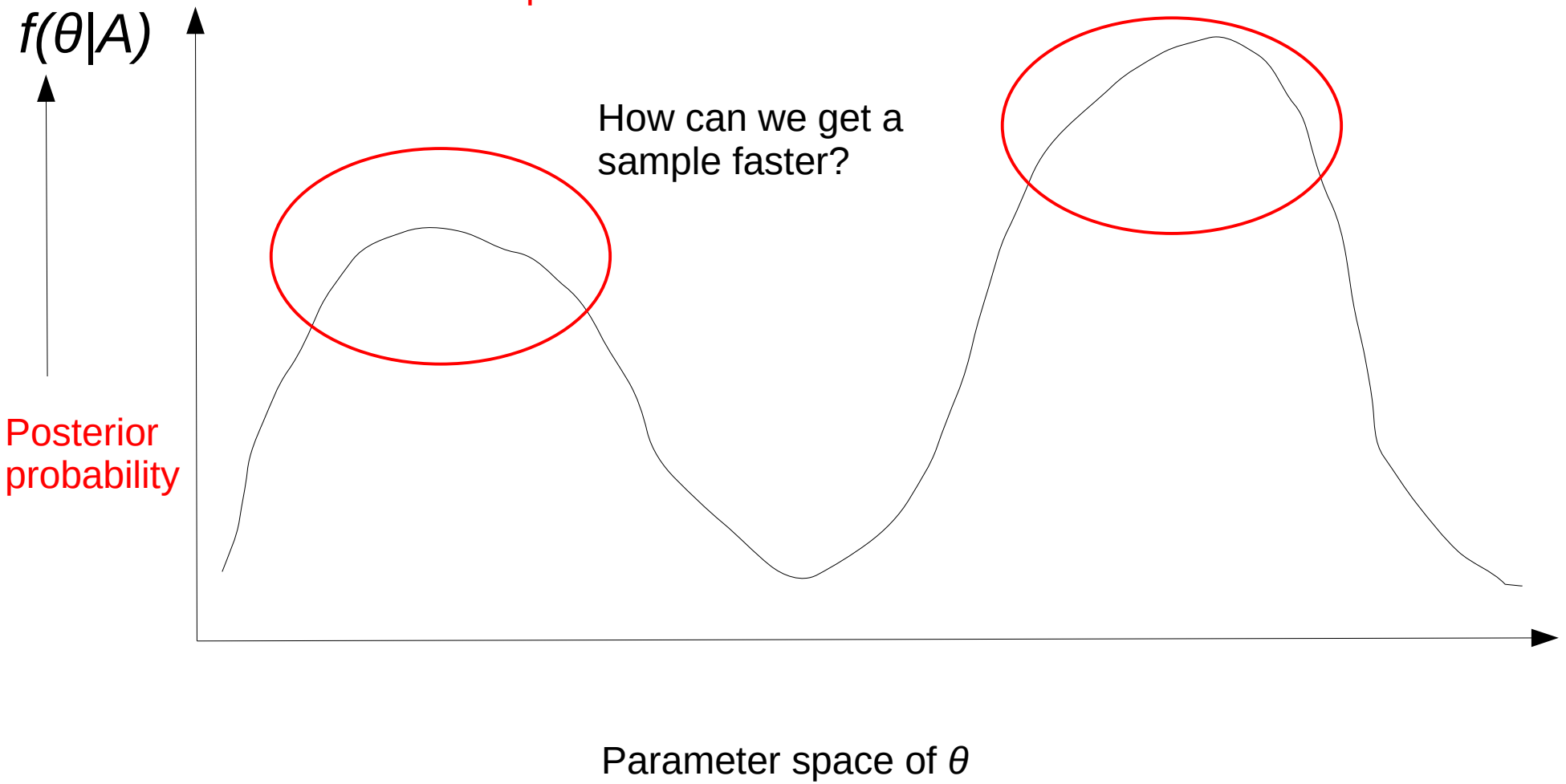


Distribution Landscape



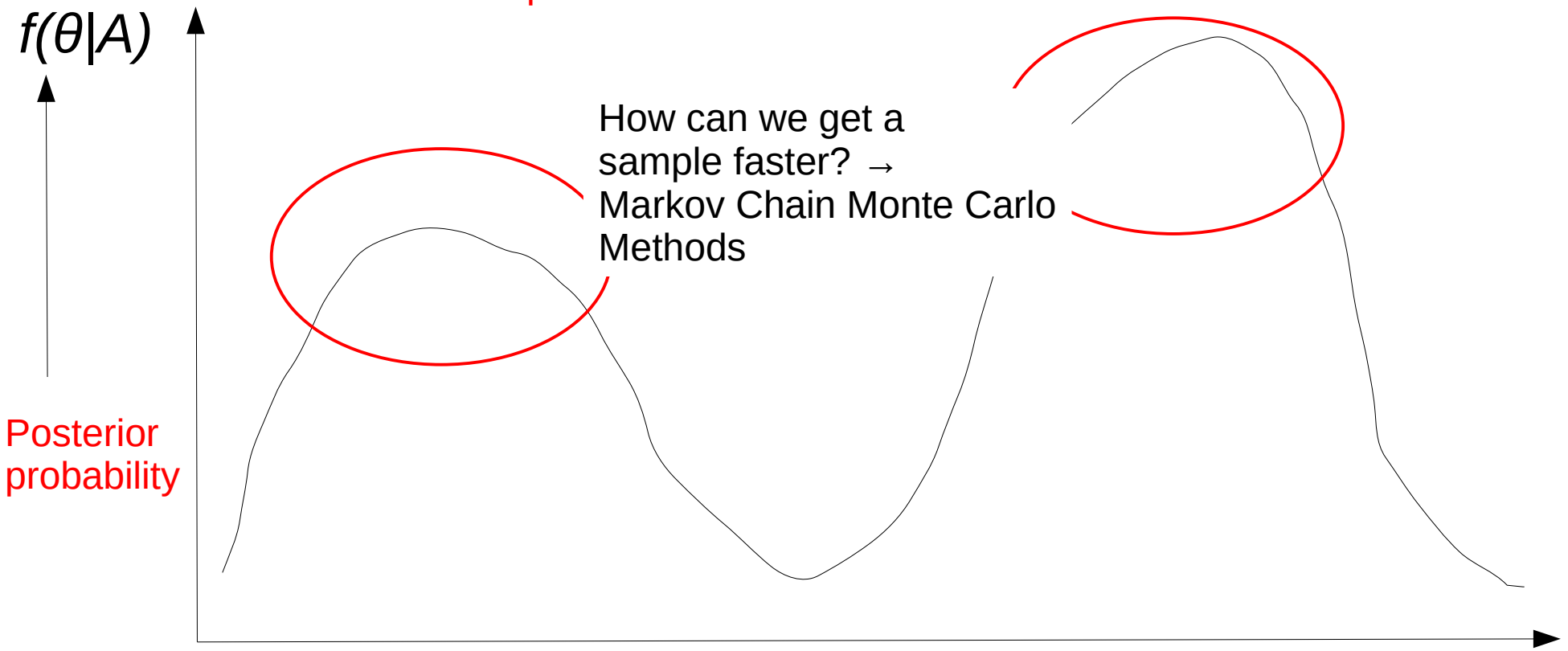
Distribution Landscape

In which parts of the distribution are we interested?



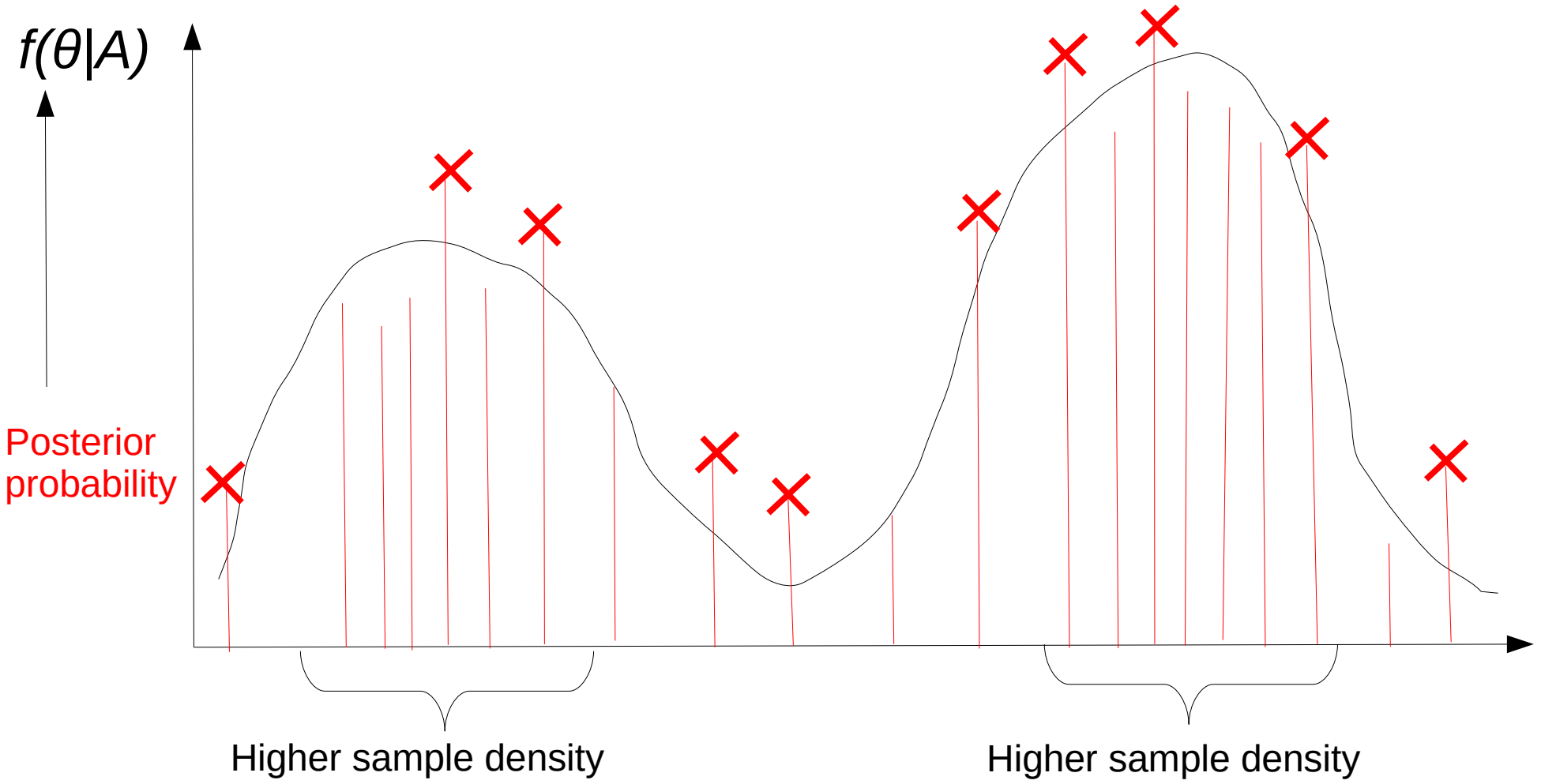
Distribution Landscape

In which parts of the distribution are we interested?



Distribution Landcsape

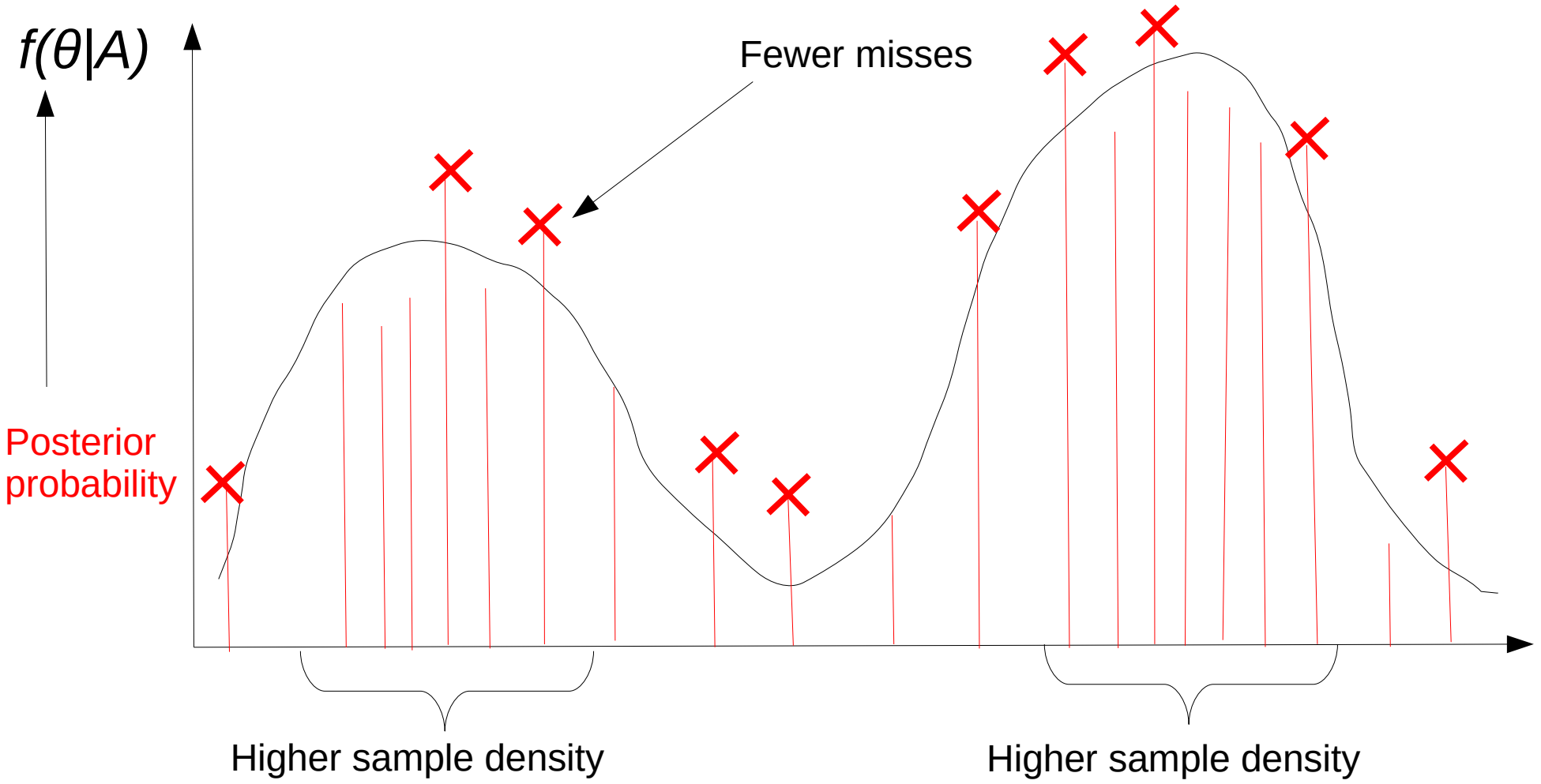
In which parts of the distribution are we interested?



Parameter space of θ

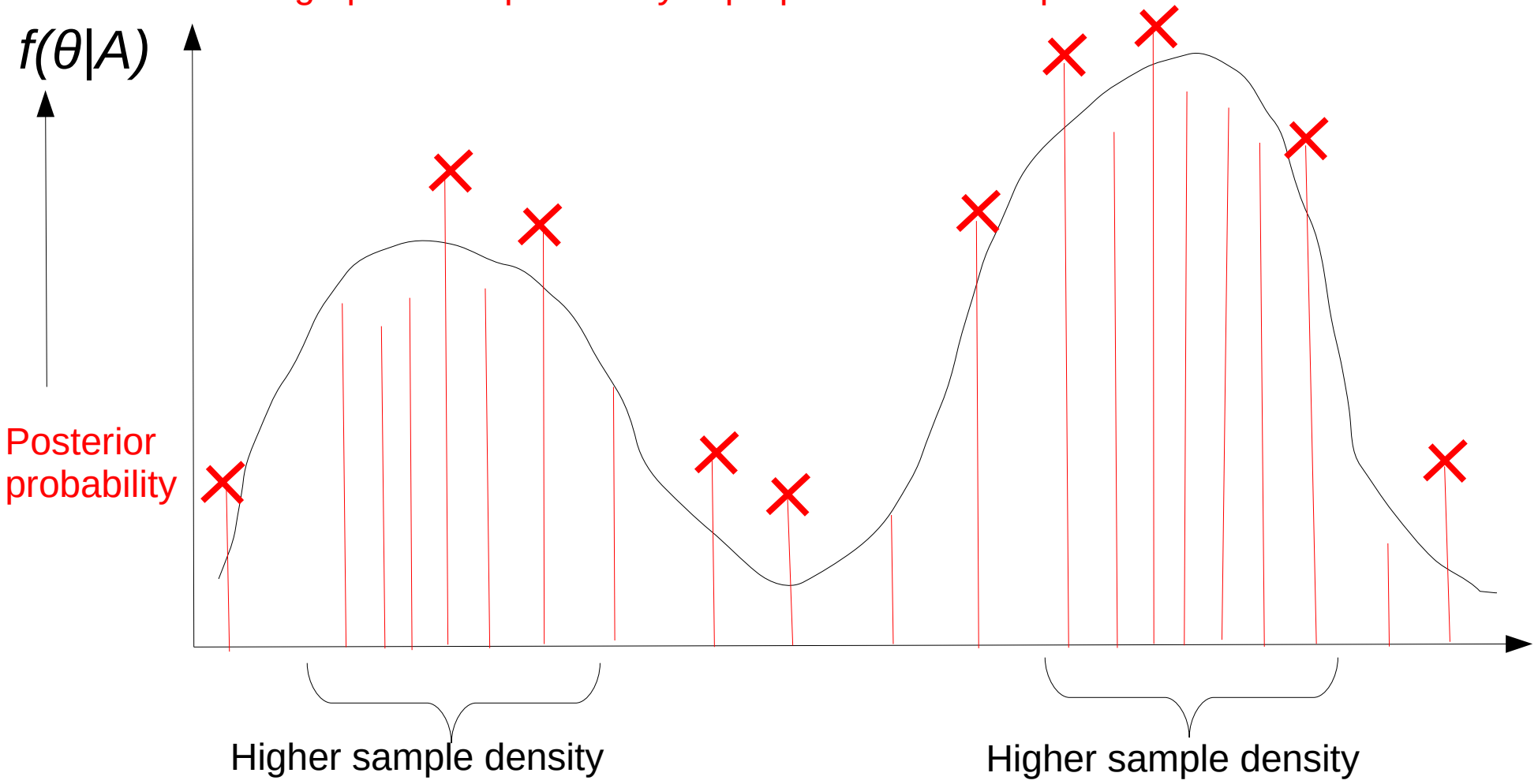
Distribution Landcsape

In which parts of the distribution are we interested?



Markov-Chain Monte-Carlo

MCMC → biased random walks: the probability to evaluate/find a sample in an area with high posterior probability is proportional to the posterior distribution



Markov-Chain Monte-Carlo

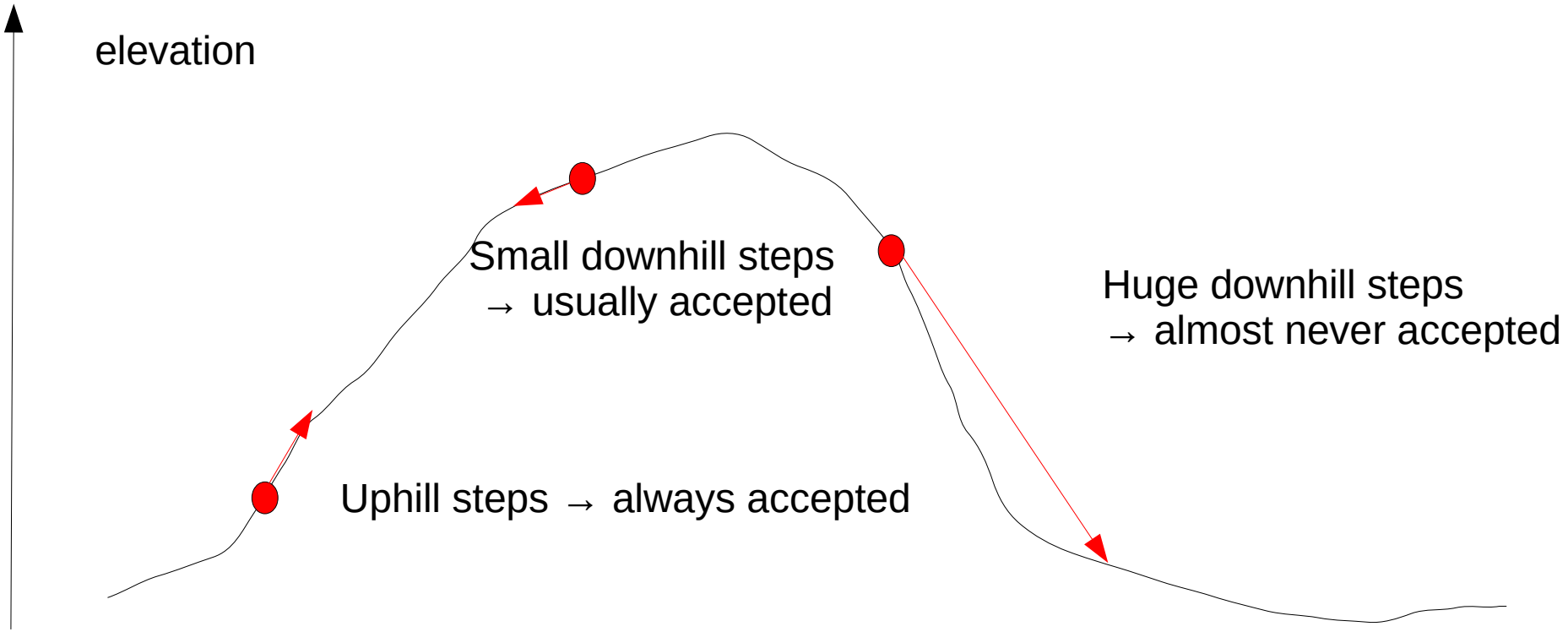
- Idea: Move the grid/samples into regions of high probability
- Construct a Markov Chain that generates samples such that more time is spent (more samples are evaluated) in the most interesting regions of the state space
- MCMC can also be used for hard CS optimization problems, for instance, the knapsack problem
- Note that, MCMC is similar to Simulated Annealing → there's no time to go into the details though here!

The Robot Metaphor

Crete, December 2019

The Robot Metaphor

- Drop a robot onto an unknown planet to explore its landscape
- Teaching idea and slides adapted from Paul O. Lewis



How to accept/reject proposals

- Decision to accept/reject a proposal to go from *Point 1* → *Point 2* is based on the ratio *R* of posterior densities of the two points/samples

  $R = Pr(Point2|data) / Pr(point1|data) =$

  $(Pr(Point2)Pr(data|point2) / Pr(data)) / (Pr(Point1)Pr(data|point1) / Pr(data))$

  $= Pr(point2)Pr(data|point2) / Pr(point1)Pr(data|point1)$

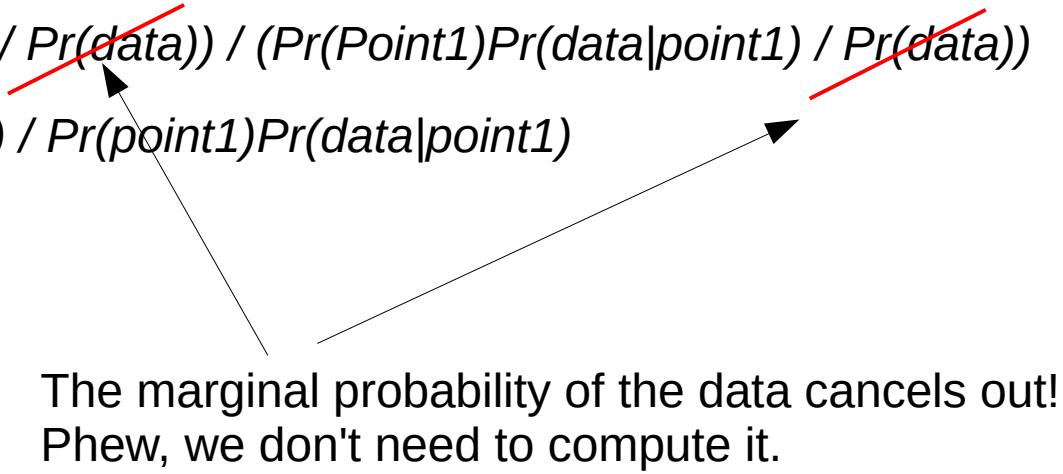
How to accept/reject proposals

- Decision to accept/reject a proposal to go from *Point 1* → *Point 2* is based on the ratio *R* of posterior densities of the two points/samples

R = Pr(Point2|data) / Pr(point1|data) =

(Pr(Point2)Pr(data|point2) / Pr(data)) / (Pr(Point1)Pr(data|point1) / Pr(data))

= Pr(point2)Pr(data|point2) / Pr(point1)Pr(data|point1)



The marginal probability of the data cancels out!
Phew, we don't need to compute it.

How to accept/reject proposals

- Decision to accept/reject a proposal to go from *Point 1* → *Point 2* is based on the ratio *R* of posterior densities of the two points/samples

$$R = Pr(Point2|data) / Pr(point1|data) =$$

$$(Pr(Point2)Pr(data|point2) / Pr(data)) / (Pr(Point1)Pr(data|point1) / Pr(data)) =$$

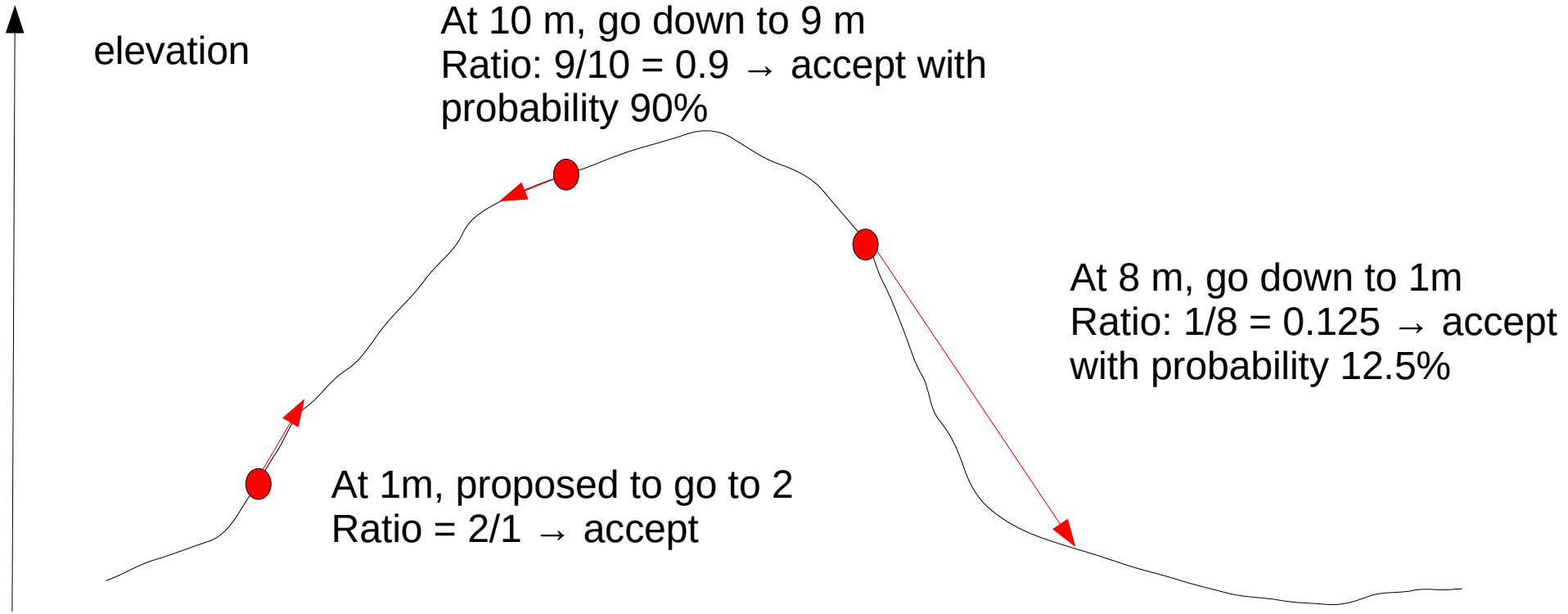
$$(Pr(point2)/Pr(point1)) * (Pr(data|point2) / Pr(data|point1))$$

Prior ratio: for uniform priors this is 1 !

Likelihood ratio

The Robot Metaphor

- Drop a robot onto an unknown planet to explore its landscape

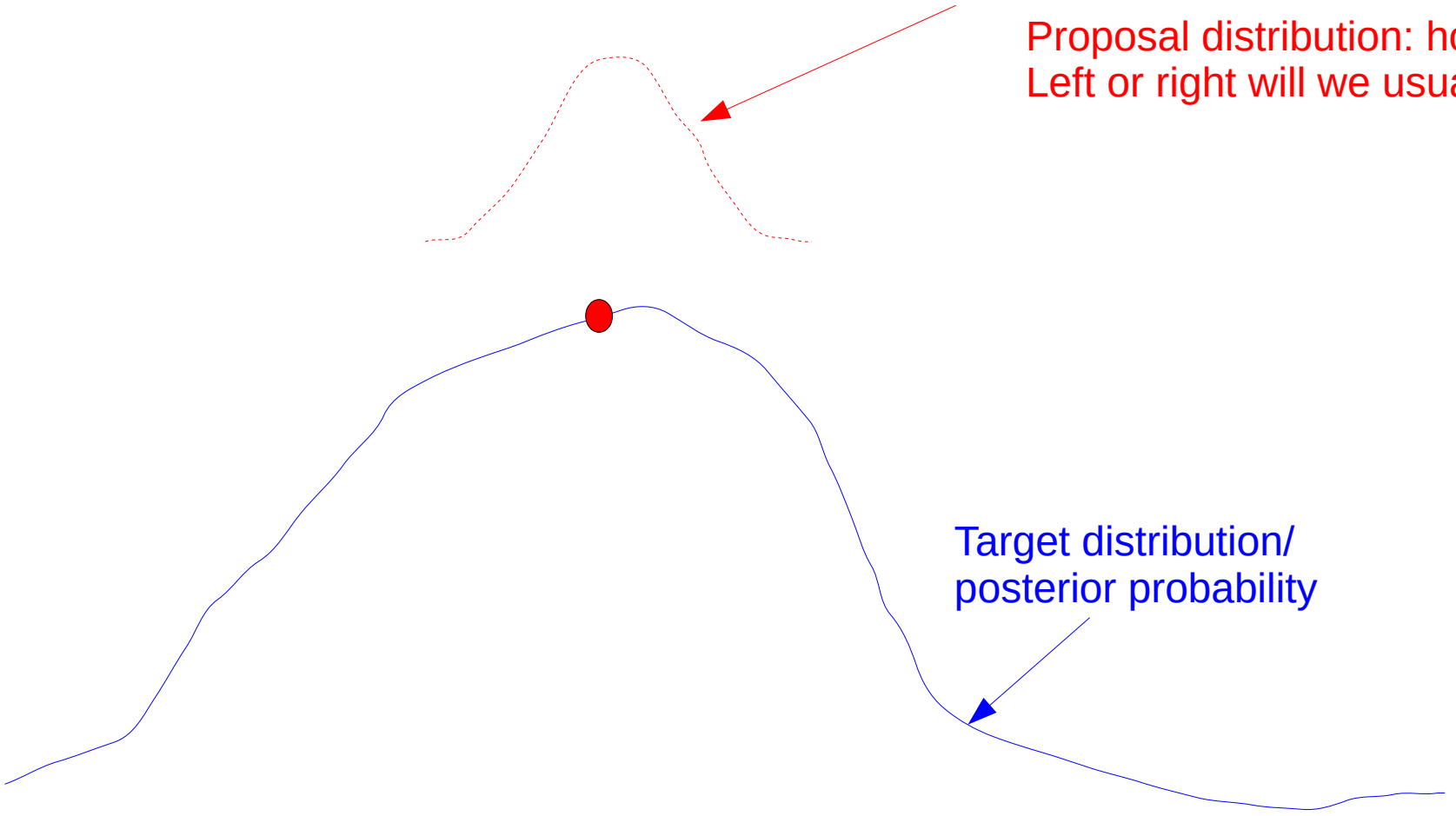


Distributions

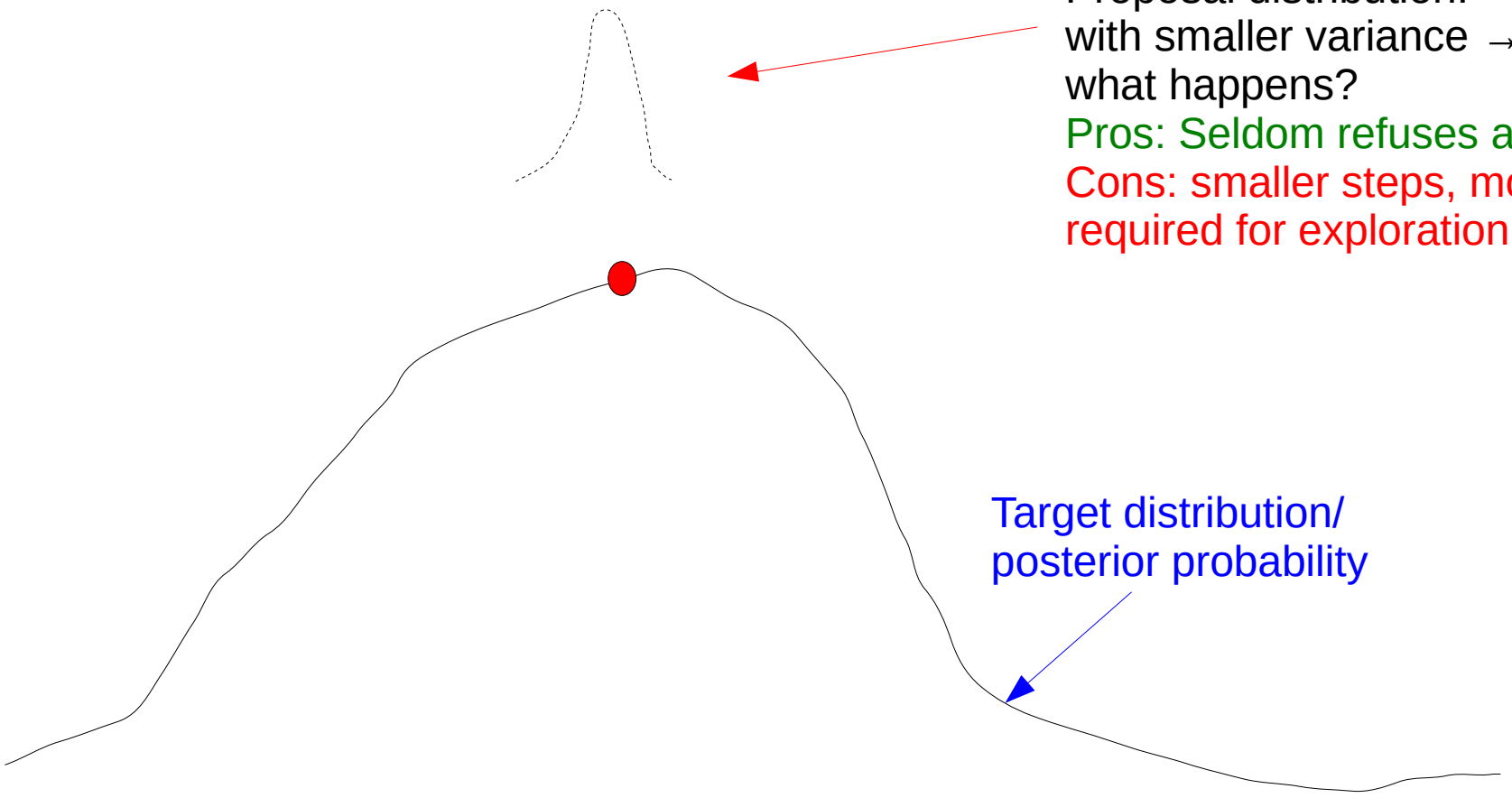
- The target distribution is the posterior distribution we are trying to sample (integrate over)!
- The proposal distribution decides which point (how far/close) in the landscape to randomly go to/try next:

  → The choice has an effect on the efficiency of the MCMC algorithm, that is, how fast it will get to these interesting areas we want to sample

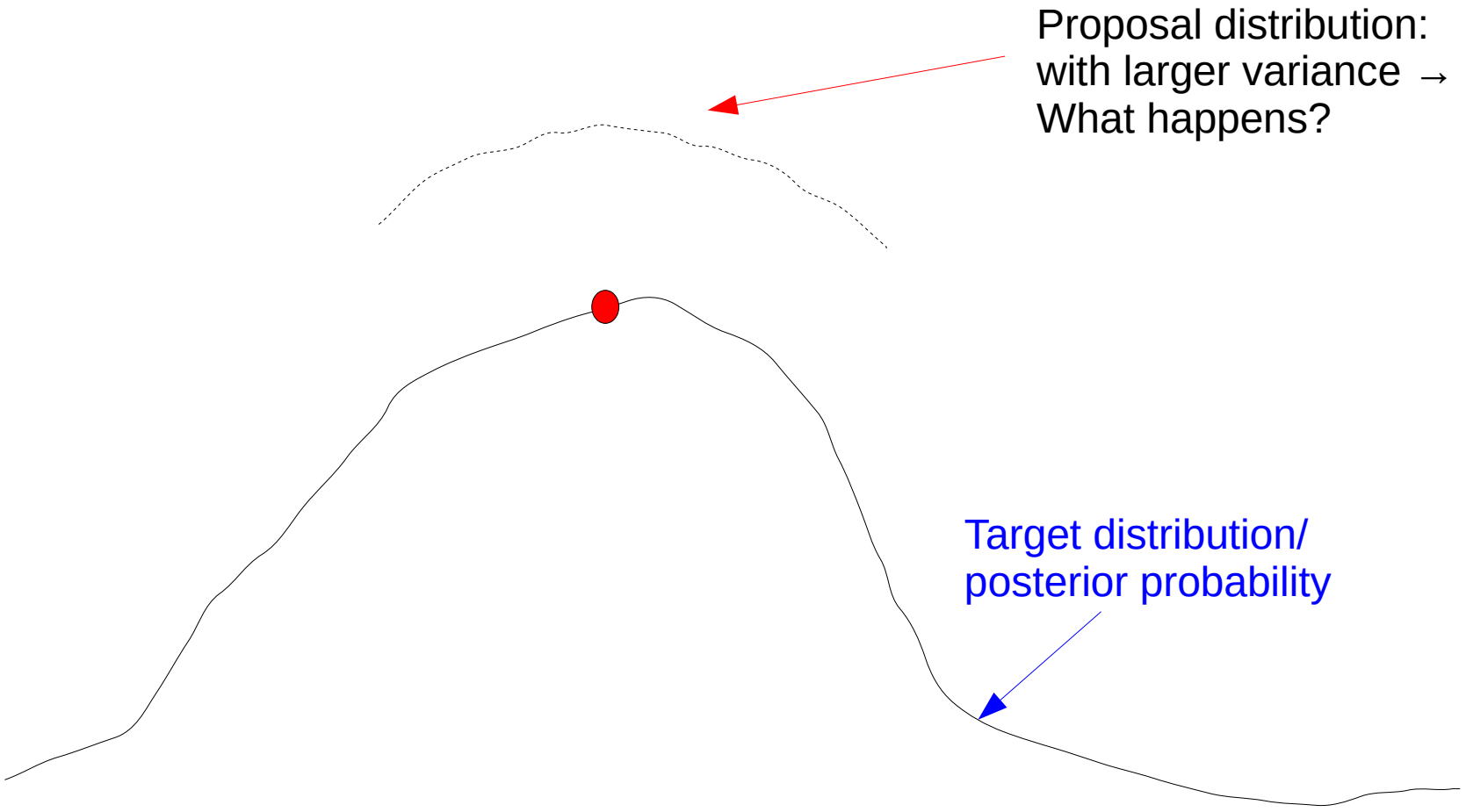
The Robot Metaphor



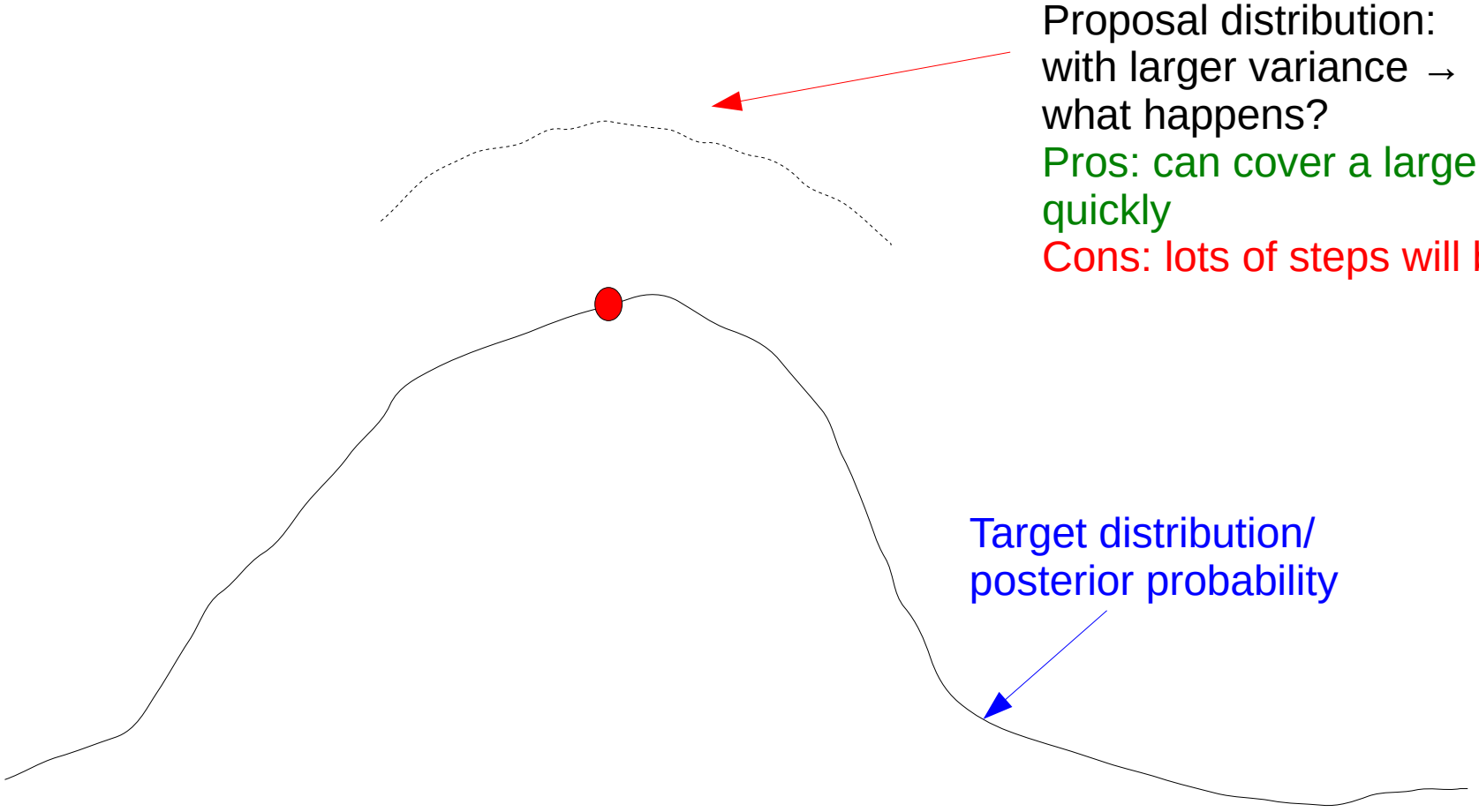
The Robot Metaphor



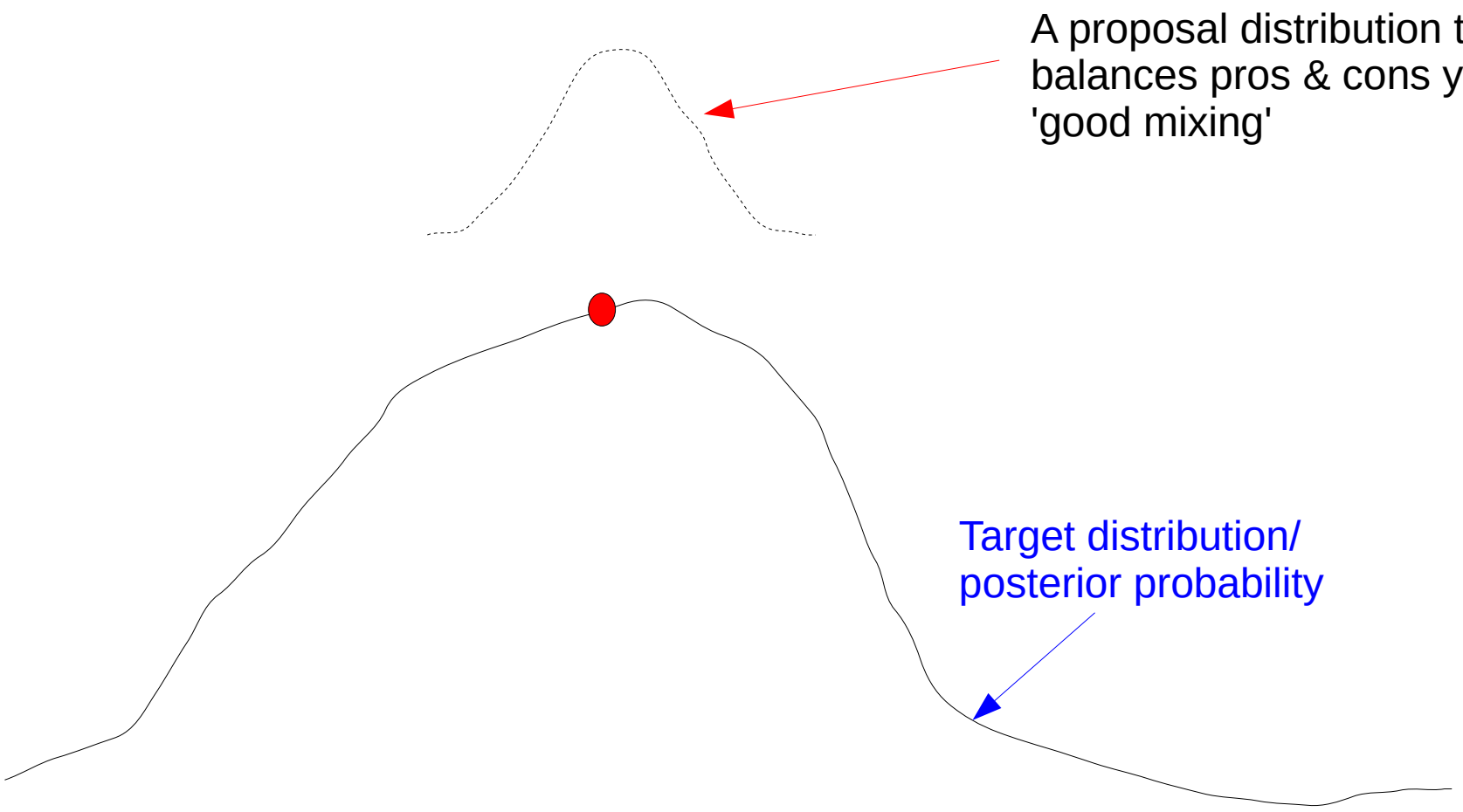
The Robot Metaphor



The Robot Metaphor



The Robot Metaphor

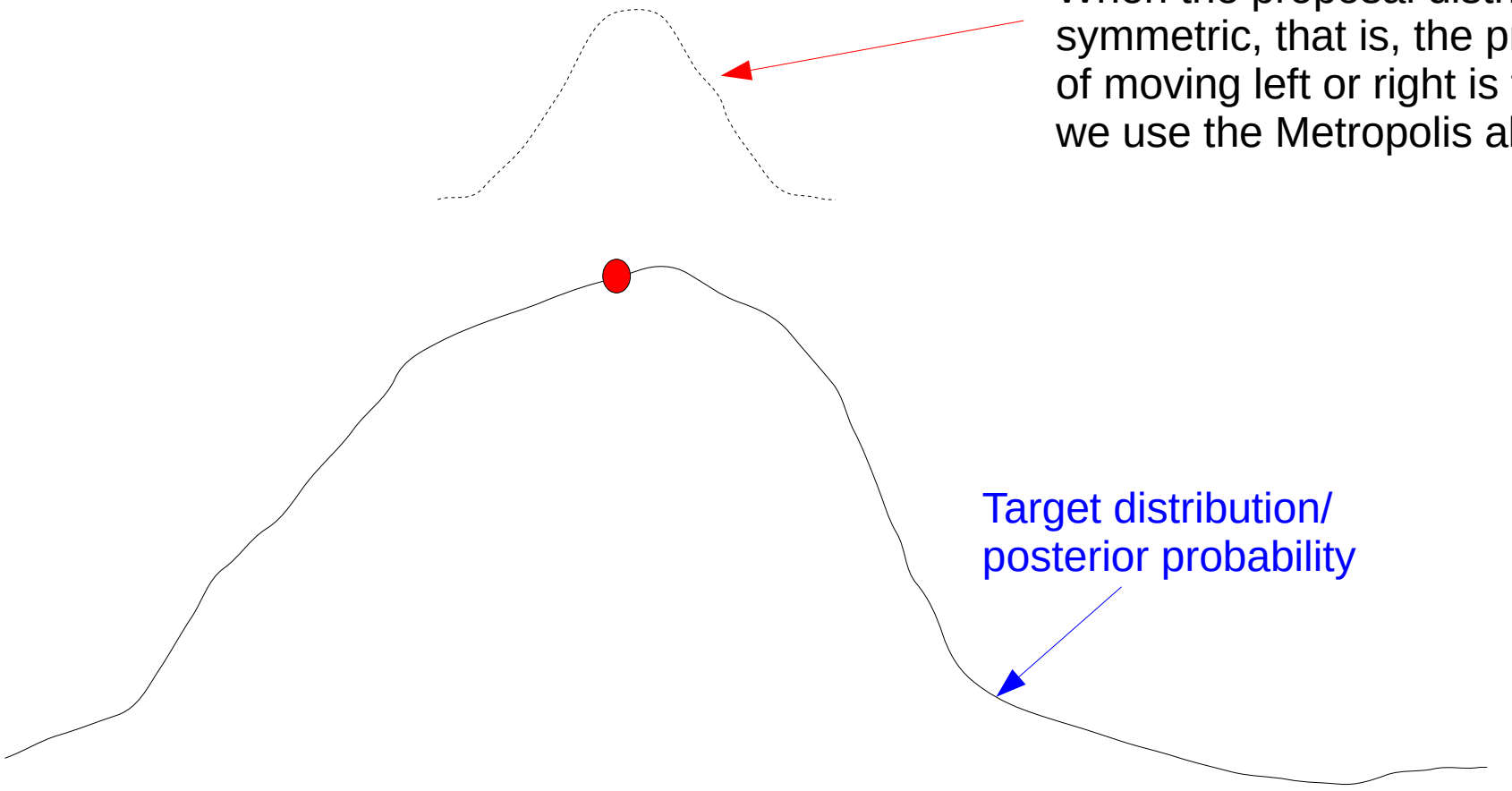


Mixing

- A well-designed chain will require a few steps until reaching convergence, that is, approximating the underlying probability density function 'well-enough' from a random starting point
- It is a somewhat fuzzy term, refers to the proportion of accepted proposals (acceptance ratio) generated by a proposal mechanism → should be neither too low, nor too high
- The real art in designing MCMC methods consists
  - building & tuning good proposal mechanisms
  - selecting appropriate proposal distributions
  - such that they quickly approximate the distribution we want to sample from

The Robot Metaphor

When the proposal distribution is symmetric, that is, the probability of moving left or right is the same, we use the Metropolis algorithm



The Metropolis Algorithm

- Metropolis *et al.* 1953 http://www.aliquote.org/pub/metropolis-et-al-1953.pdf
- Initialization: Choose an arbitrary point θ_0 as first sample
- Choose an arbitrary probability density $Q(\theta_{i+1}|\theta_i)$ which suggests a candidate for the next sample θ_{i+1} given the previous sample θ_i.
- For the Metropolis algorithm, $Q()$ must be symmetric:

  it must satisfy $Q(\theta_{i+1}|\theta_i) = Q(\theta_i|\theta_{i+1})$
- For each iteration i:
  - Generate a candidate θ^* for the next sample by picking from the distribution $Q(\theta^*|\theta_i)$
  - Calculate the acceptance ratio $R = Pr(\theta^*)Pr(data|\theta^*) / Pr(\theta_i)Pr(data/\theta_i)$
    - If R ≥ 1, then θ^* is more likely than θ_i → automatically accept the candidate by setting $\theta_{i+1} := \theta^*$
    - Otherwise, accept the candidate θ^* with probability R → if the candidate is rejected: $\theta_{i+1} := \theta_i$

The Metropolis Algorithm

- Metropolis *et al.* 1953 http://www.aliquote.org/pub/metropolis-et-al-1953.pdf
- Initialization: Choose an arbitrary point θ_0 as first sample
- Choose an arbitrary probability density $Q(\theta_{i+1}|\theta_i)$ which suggests a candidate for the next sample θ_{i+1} given the previous sample θ_i.
- For the Metropolis algorithm, $Q()$ must be symmetric:

  it must satisfy $Q(\theta_{i+1}|\theta_i) = Q(\theta_i|\theta_{i+1})$
- For each iteration i:
  - Generate a candidate θ^* for the next sample by picking from the distribution $Q(\theta^*|\theta_i)$
  - Calculate the acceptance ratio $R = Pr(\theta^*)Pr(data|\theta^*) / Pr(\theta_i)Pr(data/\theta_i)$
    - If R ≥ 1, then θ^* is more likely than θ_i → automatically accept the candidate by setting $\theta_{i+1} := \theta^*$
    - Otherwise, accept the candidate θ^* with probability R → if the candidate is rejected: $\theta_{i+1} := \theta_i$

Conceptually this is the same Q we saw for substitution models and in the Markov Chain lecture!

Phylogenetic Metropolis Algorithm

- Initialization: Choose a random tree with random branch lengths as first sample
- For each iteration *i*:
  - Propose either
    - a new tree topology
    - a new branch length

      and re-calculate the likelihood
  - Calculate the acceptance ratio of the proposal
  - Accept the new tree/branch length or reject it
  - Print current tree with branch lengths to file only every *k* (e.g. 1000) iterations

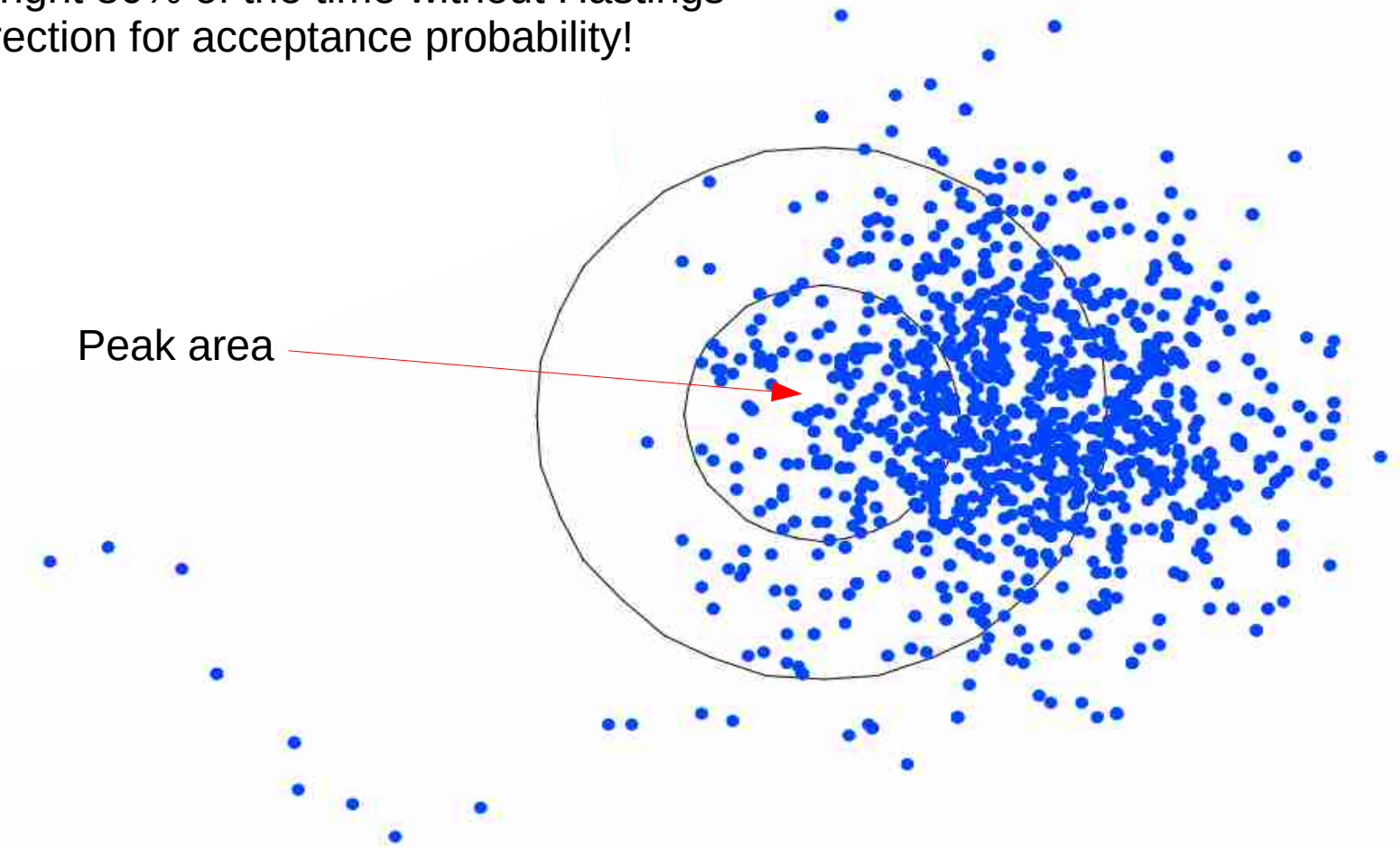
    → to generate a sample from the chain

    → to avoid writing TBs of files

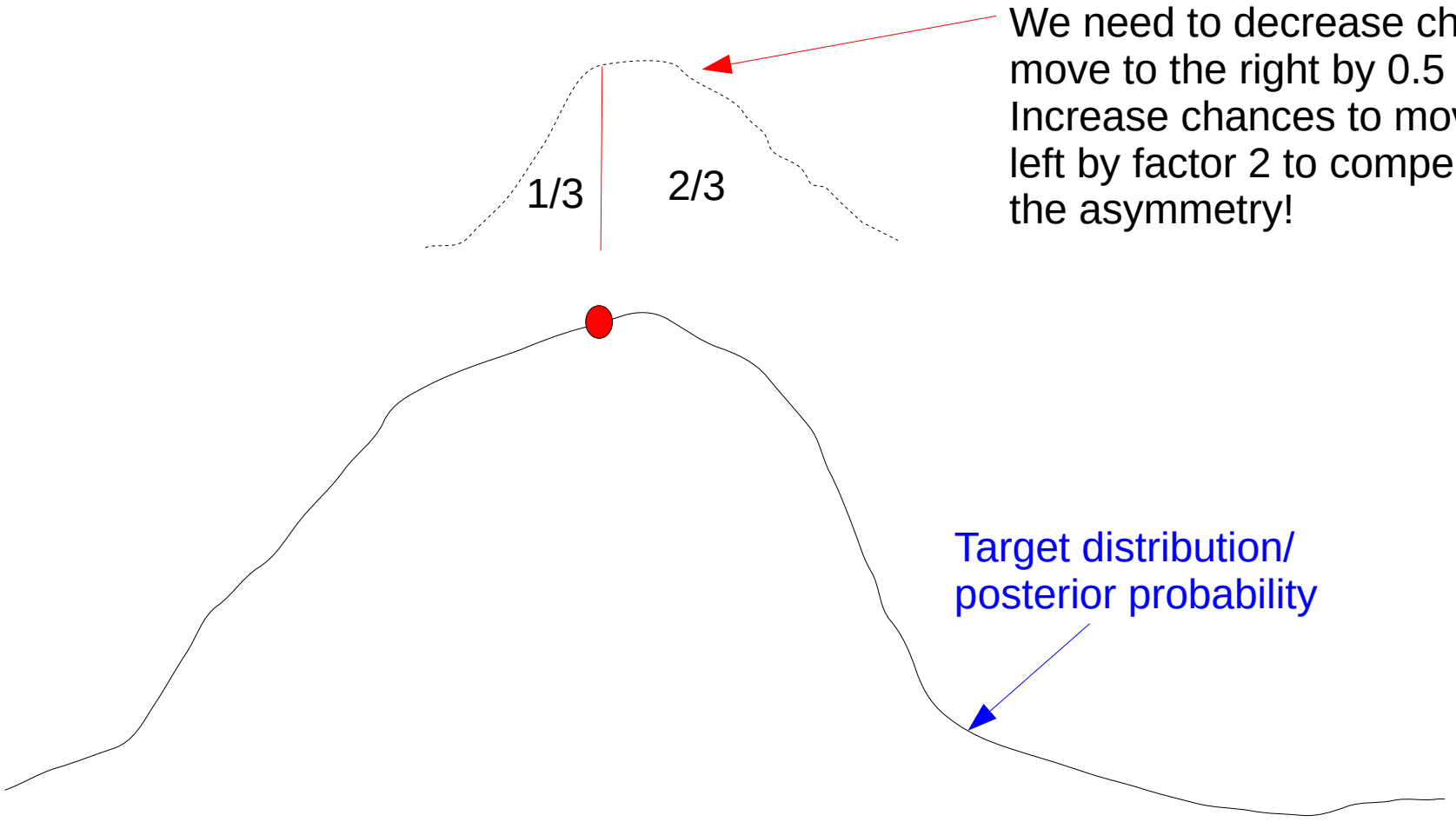
    → also known as thinning
- Summarize the sample using means, histograms, credible intervals, consensus trees, etc.

Uncorrected Proposal Distribution A Robot in 3D

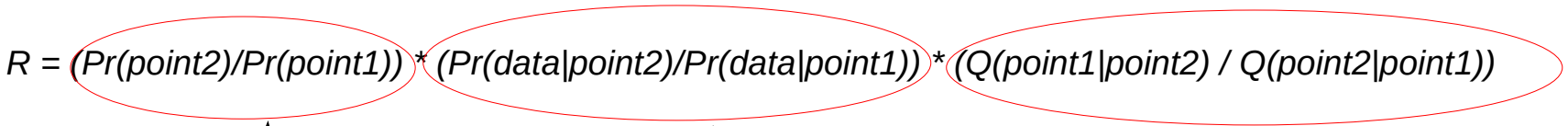
Example: MCMC proposed moves to the right 80% of the time without Hastings correction for acceptance probability!



Hastings Correction



Hastings Correction

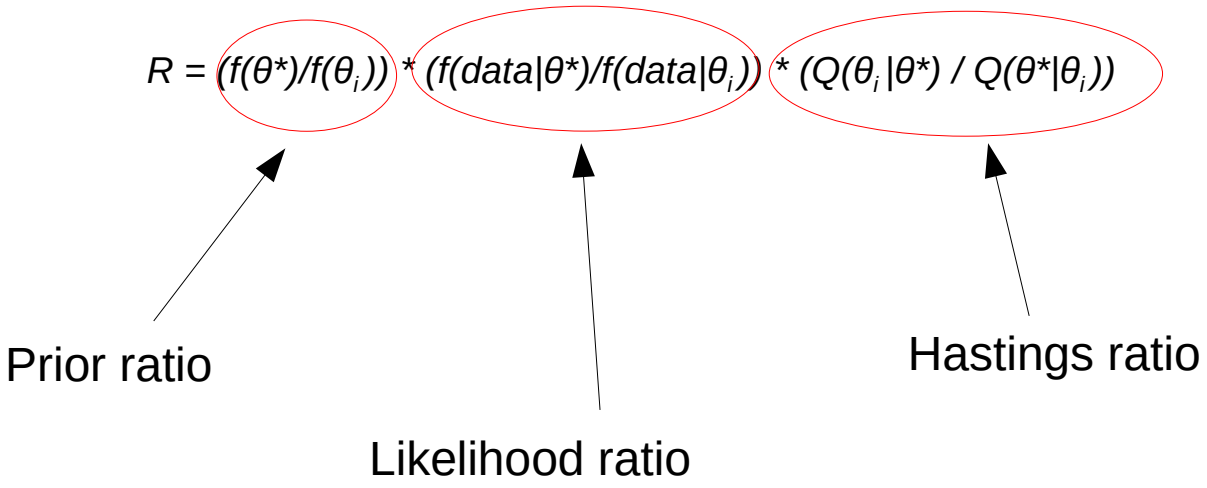


Prior ratio: for uniform priors this is 1 !

Likelihood ratio

Hastings ratio: if *Q* is symmetric *Q(point1|point2) = Q(point2|point)* and the hastings ratio is 1 → we obtain the normal Metropolis algorithm

Hastings Correction more formally



Hastings Correction is not trivial

- Problem with the equation for the hastings correction
- M. Holder, P. Lewis, D. Swofford, B. Larget. 2005. **Hastings Ratio of the LOCAL Proposal Used in Bayesian Phylogenetics.** *Systematic Biology*. 54:961-965. http://sysbio.oxfordjournals.org/content/54/6/961.full

  "As part of another study, we estimated the marginal likelihoods of trees using different proposal algorithms and discovered repeatable discrepancies that implied that the published Hastings ratio for a proposal mechanism used in many Bayesian phylogenetic analyses is incorrect."
- Incorrect Hastings ratio used from 1999-2005