Monte Carlo Markov Chain

Bayesian methods

Simone Paradiso

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Most of the material in this presentation is taken from

“**Information Theory, Inference, and Learning Algorithms**”

by **David J. C. McKay**.

Free pdf at https://www.inference.org.uk/itprnn/book.pdf

Suggested reading: “Handbook of Markov chain Monte Carlo” by S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Recap of the Bayes' theorem

$$P(x\,|\,y) = \frac{p(y\,|\,x)p(x)}{p(y)}$$

Recap of the Bayes' theorem

$$\boxed{P(x|y)} = \frac{p(y|x)p(x)}{p(y)}$$

The conditional probability of x given that y is true. Also known as **Posterior probability**

Recap of the Bayes' theorem

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

The conditional probability of x given that y is true. Also known as **Posterior probability**

This is also a conditional probability. This is the probability of y given that x is true; it can also be interpreted as the **Likelihood** of x given y

WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

Recap of the Bayes' theorem

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

The conditional probability of x given that y is true. Also known as **Posterior probability**

This is also a conditional probability. This is the probability of y given that x is true; it can also be interpreted as the **Likelihood** of x given y

This is the probability of observing x without further condition. It is also known as the **prior probability**.

WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

Recap of the Bayes' theorem

$$\boxed{P(x|y)} = \frac{\boxed{p(y|x)}\boxed{p(x)}}{\boxed{p(y)}}$$

The conditional probability of x given that y is true. Also known as **Posterior probability**

This is also a conditional probability. This is the probability of y given that x is true; it can also be interpreted as the **Likelihood** of x given y

This is the probability of observing x without further condition. It is also known as the **prior probability**.

This is the probability of y without any given conditions; this is known as the **evidence**. It can be written as: $P(y) = \int p(y|x)p(x)dx$

Monte Carlo methods

The goal

Monte Carlo methods are computational techniques that make use of random numbers. The aims of MC methods is to solve one or both of the following problems:

Monte Carlo methods

The goal

Monte Carlo methods are computational techniques that make use of random numbers. The aims of MC methods is to solve one or both of the following problems:

- Generate samples $\{\mathbf{x}_\mathrm{i}\}_{\mathrm{i}=1}^{\mathrm{N}}$ from a probability distribution $P(\mathbf{x})$

Monte Carlo methods

The goal

Monte Carlo methods are computational techniques that make use of random numbers. The aims of MC methods is to solve one or both of the following problems:

- Generate samples $\{\mathbf{x}_\mathrm{i}\}_{\mathrm{i}=1}^{\mathrm{N}}$ from a probability distribution $P(\mathbf{x})$
- Generate expectation of functions under this distribution. For instance:

$$\Phi = \langle \phi(\mathbf{x}) \rangle = \int d^N P(\mathbf{x}) \phi(\mathbf{x})$$

Monte Carlo methods

The goal

Monte Carlo methods

The goal

We call $P(\mathbf{x})$ the *target density* is what we usually want to characterize in a physical problem: the posterior distribution of some parameters given the data.

Monte Carlo methods

The goal

We call $P(\mathbf{x})$ the *target density* is what we usually want to characterize in a physical problem: the posterior distribution of some parameters given the data.

Some examples of $\phi(\mathbf{x})$ can be the first order moments of the target density, i.e. mean and variance.

Monte Carlo methods

The goal

We call $P(\mathbf{x})$ the *target density* is what we usually want to characterize in a physical problem: the posterior distribution of some parameters given the data.

Some examples of $\phi(\mathbf{x})$ can be the first order moments of the target density, i.e. mean and variance.

So let's assume that $P(\mathbf{x})$ is too complicated and its moments can't be evaluated in some exact way. **We want to use Monte Carlo methods**.

Monte Carlo methods

The sampling problem

- Let's focus on the **sampling problem**, because if we solve this, then generating the expectation function is straightforward using the random samples $\{\mathbf{x}^{(\mathrm{i})}\}_{\mathrm{i}=1}^{\mathrm{N}}$ to build the estimator:

$$\hat{\Phi} = \frac{1}{N} \sum \phi_i(\mathbf{x})$$

Monte Carlo methods

The sampling problem

- Let's focus on the **sampling problem**, because if we solve this, then generating the expectation function is straightforward using the random samples $\{\mathbf{x}^{(\mathrm{i})}\}_{\mathrm{i}=1}^{\mathrm{N}}$ to build the estimator:

$$\hat{\Phi} = \frac{1}{N}\sum \phi_i(\mathbf{x})$$

- The estimator is unbiased as long as $\{\mathbf{x}^{(\mathrm{i})}\}_{\mathrm{i}=1}^{\mathrm{N}} \sim P(\mathbf{x})$.

Monte Carlo methods

The sampling problem

$$\hat{\Phi} = \frac{1}{N} \sum \phi_i(\mathbf{x})$$

Monte Carlo methods

The sampling problem

$$\hat{\Phi} = \frac{1}{N} \sum \phi_i(\mathbf{x})$$

- If the number of samples increases, the variance of $\hat{\Phi}$ goes down as σ^2/N.

Monte Carlo methods

The sampling problem

$$\hat{\Phi} = \frac{1}{N} \sum \phi_i(\mathbf{x})$$

- If the number of samples increases, the variance of $\hat{\Phi}$ goes down as σ^2/N.

Property of MC methods: The accuracy of an MC estimate depends only on the variance of ϕ, not on the parameter space dimension. Precisely, the variance of $\hat{\Phi}$ decreases with the number of samples.

Monte Carlo methods

The sampling problem

$$\hat{\Phi} = \frac{1}{N} \sum \phi_i(\mathbf{x})$$

- If the number of samples increases, the variance of $\hat{\Phi}$ goes down as σ^2/N.

Property of MC methods: The accuracy of an MC estimate depends only on the variance of ϕ, not on the parameter space dimension. Precisely, the variance of $\hat{\Phi}$ decreases with the number of samples.

- However, sampling from high-dimensionality can be a problem in MC methods; drawing samples from $P(\mathbf{x})$ is not easy in general.

Monte Carlo methods

The sampling problem

- Let’s assume we can evaluate $P(\mathbf{x})$ up to a normalization constant, i.e. we can evaluate $P^*(\mathbf{x}) = P(\mathbf{x})/Z$.

- We want to draw samples from $P(\mathbf{x})$, but:

Monte Carlo methods

The sampling problem

- Let's assume we can evaluate $P(\mathbf{x})$ up to a normalization constant, i.e. we can evaluate $P^*(\mathbf{x}) = P(\mathbf{x})/Z$.

- We want to draw samples from $P(\mathbf{x})$, but:

   A. We do not know the normalization constant $Z = \int P^*(\mathbf{x}) d^N\mathbf{x}$, and

Monte Carlo methods

The sampling problem

- Let's assume we can evaluate $P(\mathbf{x})$ up to a normalization constant, i.e. we can evaluate $P^*(\mathbf{x}) = P(\mathbf{x})/Z$.

- We want to draw samples from $P(\mathbf{x})$, but:

  A. We do not know the normalization constant $Z = \int P^*(\mathbf{x}) d^N \mathbf{x}$, and

  B. Even if we knew Z, it would still be challenging to sample in a high-dimensional space without exploring most of all the possible states.

Monte Carlo methods

The sampling problem

- To draw samples from $P(\mathbf{x})$ we need to explore the space where $P(\mathbf{x})$ is large, and we can only guess where it happens by evaluating the density *everywhere*.

WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

Monte Carlo methods

The sampling problem

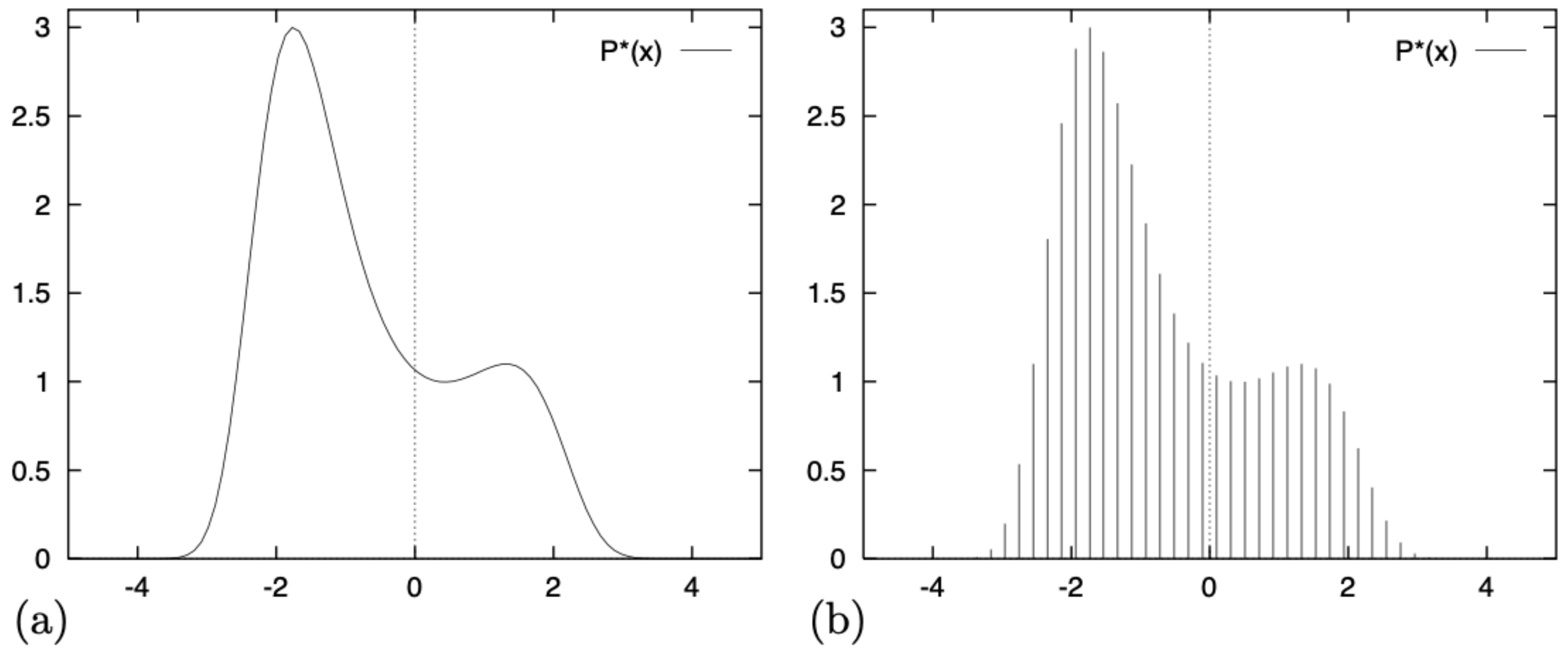
- To draw samples from $P(\mathbf{x})$ we need to explore the space where $P(\mathbf{x})$ is large, and we can only guess where it happens by evaluating the density *everywhere*.
- We know a convenient way to draw samples from very few distributions; the Normal distribution is of course one of these.

Monte Carlo methods

A simple 1D example

- We want to draw samples from $P^*(x) = P(x)/Z$, with:

$$P^*(x) = e^{0.4(x-0.4)^2 - 0.08x^4}$$



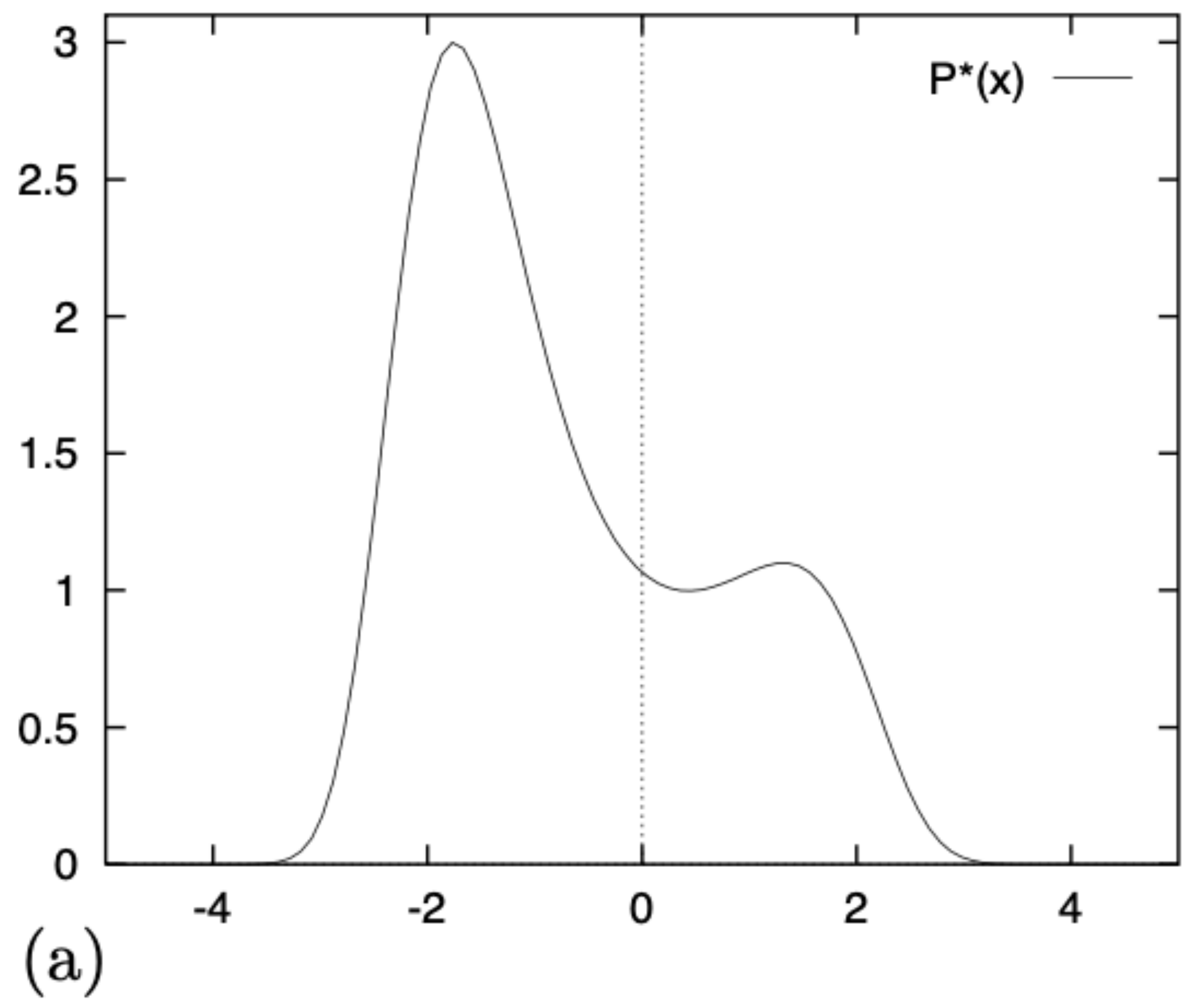
Monte Carlo methods

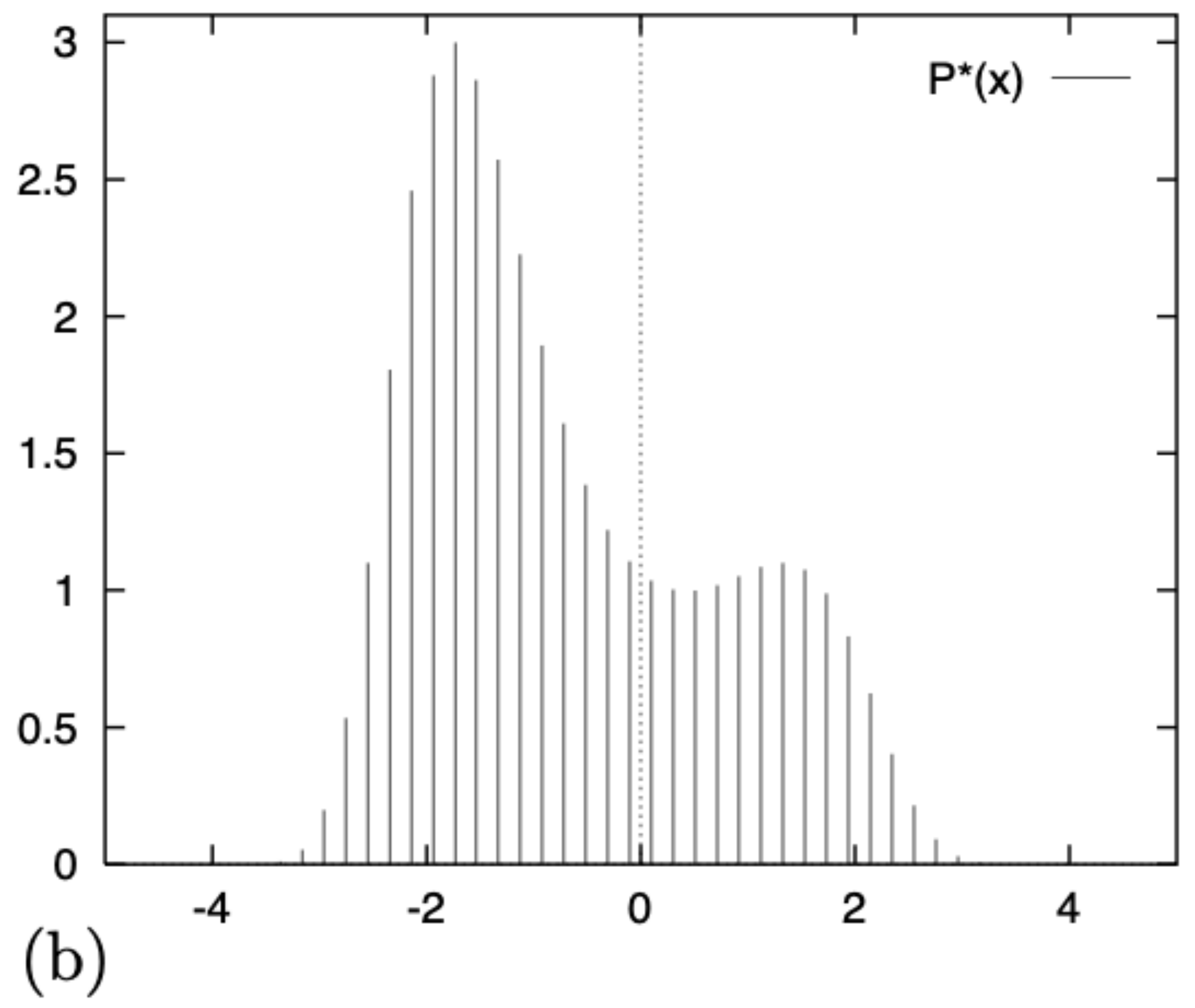
A simple 1D example

- We want to draw samples from $P^*(x) = P(x)/Z$, with:

$$P^*(x) = e^{0.4(x-0.4)^2 - 0.08x^4}$$

- We can plot it, but don't know how to draw samples from it.





Monte Carlo methods

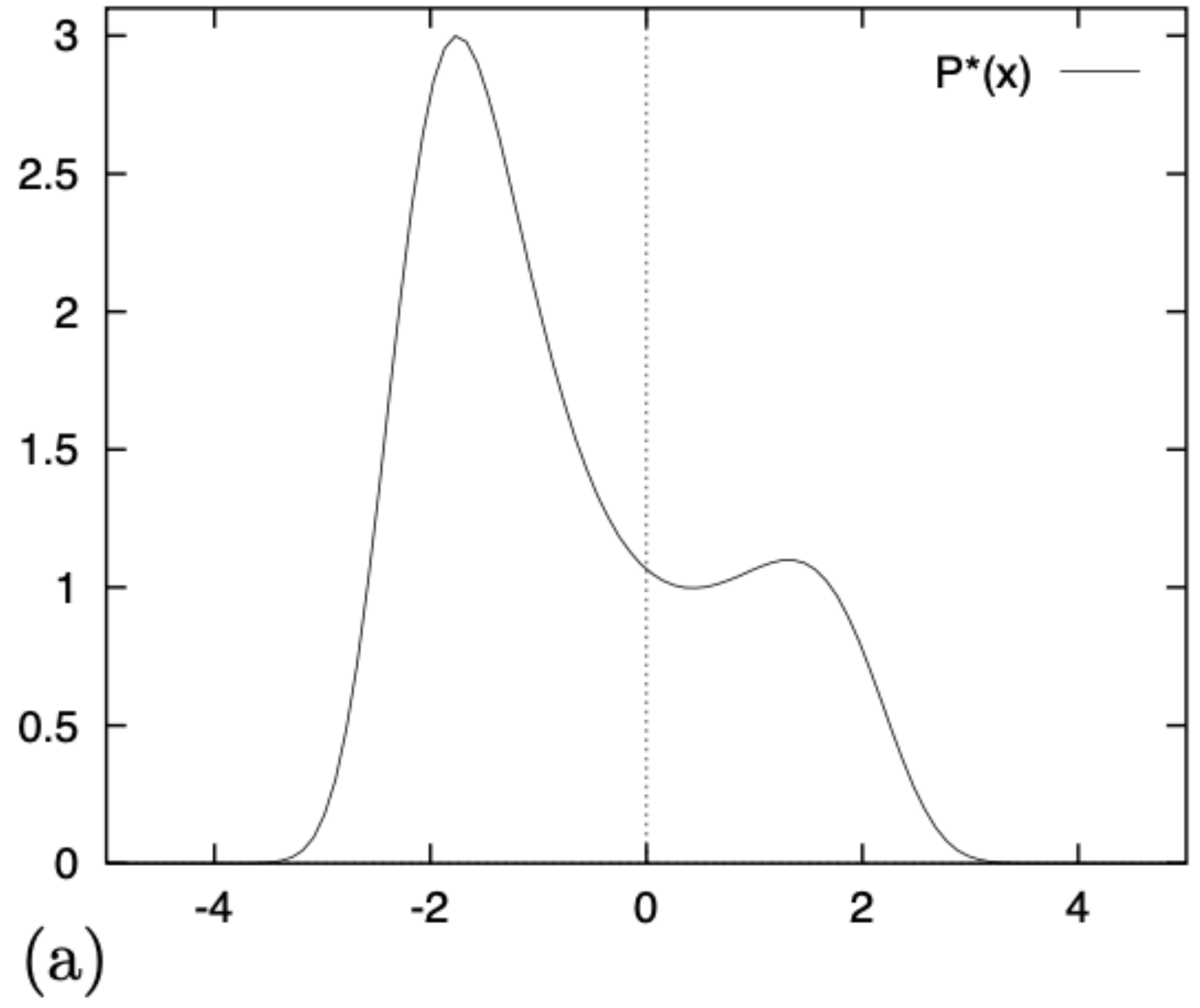
A simple 1D example

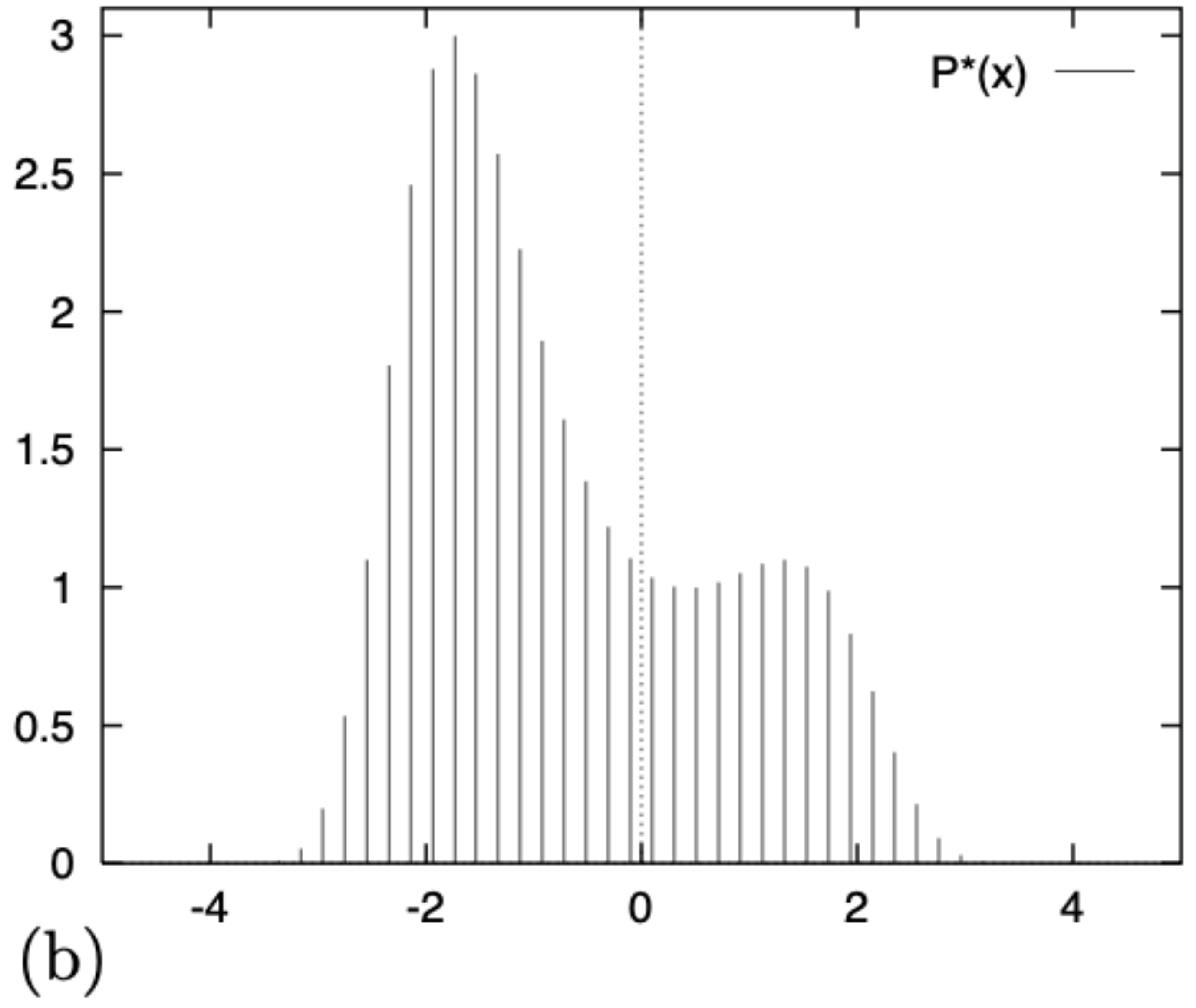
- We want to draw samples from $P^*(x) = P(x)/Z$, with:

$$P^*(x) = e^{0.4(x-0.4)^2 - 0.08x^4}$$

- We can plot it, but don't know how to draw samples from it.

- We first need to know Z.





Monte Carlo methods

A simple 1D example

- Let's try with a brute-force approach: we can evaluate discretize $P^*(x)$ by evaluating it on equally spaced values $\{x^{(i)}\}$, and then compute:

$$Z = \sum_i P^*(x^{(i)}) \rightarrow P(x^{(i)}) = P^*(x^{(i)})/Z$$

- We can now sample from $P(x^{(i)})$. A basic sampling algorithm would just be to draw many times a number in $U(0,1)$ and compare to $P(x^{(i)})$ for each $x_i \in \{x^{(i)}\}$.

Monte Carlo methods

A simple 1D example: computational cost

- To compute Z we have to visit every point in the space. Depending on the x spacing (n), and the number of dimensions (d), this operation requires n^d computations.

- A typical cosmological problem involves at least 6 parameters, and the typical spacing is ~1000 points. This means we need to perform $1000^6 \approx 10^{18}$ operations. A theory code for cosmological purposes (CAMB, CLASS) take ~1s (best scenario); this means a computer would take $10^{18}s \approx 30Gyrs$.

Monte Carlo methods

Uniform sampling

- Acknowledged that we can't exhaustively visit every location x in the parameter space, let's try to solve the second problem: **estimating the expectation of a function** $\phi(x)$.

Monte Carlo methods

Uniform sampling

- Acknowledged that we can't exhaustively visit every location x in the parameter space, let's try to solve the second problem: **estimating the expectation of a function** $\phi(x)$.
- We could try to draw random samples of $\{x^{(i)}\}_{i \in N}$ uniformly and evaluate $P^*(x^{(i)})$ at those points.

Monte Carlo methods

Uniform sampling

- Acknowledged that we can't exhaustively visit every location x in the parameter space, let's try to solve the second problem: **estimating the expectation of a function** $\phi(x)$.
- We could try to draw random samples of $\{x^{(i)}\}_{i \in N}$ uniformly and evaluate $P^*(x^{(i)})$ at those points.
- We can then normalize by $Z_N = \sum_{i=1}^{N} P^*(x^{(i)})$ and estimate $\Phi(x) = \int \phi(x) P(x) dx$ via:

$$\hat{\Phi} = \sum_{i=1}^{N} \phi(x^{(i)}) \frac{P^*(x^{(i)})}{Z_N}$$

WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

Monte Carlo methods

Uniform sampling

- Efficiency here depends on $\phi(x)$ and $P^*(x)$.

Monte Carlo methods

Uniform sampling

- Efficiency here depends on $\phi(x)$ and $P^*(x)$.
- A high dimensional (d) distribution is often concentrated in a small region of the parameters space whose volume is given by $|T| \approx 2^{H(x)}$, where $H(x)$ is the entropy of $P(x)$.

Monte Carlo methods

Uniform sampling

- Efficiency here depends on $\phi(x)$ and $P^*(x)$.
- A high dimensional (d) distribution is often concentrated in a small region of the parameters space whose volume is given by $|T| \approx 2^{H(x)}$, where $H(x)$ is the entropy of $P(x)$.
- We need a sufficiently high number of samples N to ensure we hit at least a couple of times the typical volume T. How many?

Monte Carlo methods

Uniform sampling

- Efficiency here depends on $\phi(x)$ and $P^*(x)$.
- A high dimensional (d) distribution is often concentrated in a small region of the parameters space whose volume is given by $|T| \approx 2^{H(x)}$, where $H(x)$ is the entropy of $P(x)$.
- We need a sufficiently high number of samples N to ensure we hit at least a couple of times the typical volume T. How many?
- Each sample has a chance $2^H/2^d$ to fall in the typical set. We therefore need at least $N \approx 2^{d-H}$ samples.

Monte Carlo methods

Uniform sampling

- Efficiency here depends on $\phi(x)$ and $P^*(x)$.
- A high dimensional (d) distribution is often concentrated in a small region of the parameters space whose volume is given by $|T| \approx 2^{H(x)}$, where $H(x)$ is the entropy of $P(x)$.
- We need a sufficiently high number of samples N to ensure we hit at least a couple of times the typical volume T. How many?
- Each sample has a chance $2^H/2^d$ to fall in the typical set. We therefore need at least $N \approx 2^{d-H}$ samples.
- **This means that unless $H \sim d$, or $P(x)$ is uniform, corresponding to quite boring cases, uniform sampling is unlikely to be useful, or efficient.**

Monte Carlo methods

overview

- Drawing samples from $P(x) = P^*(x)/Z$ is complicated, even if we knew Z and P^* is easy to evaluate.

Monte Carlo methods

overview

- Drawing samples from $P(x) = P^*(x)/Z$ is complicated, even if we knew Z and P^* is easy to evaluate.

We need some workaround.

Monte Carlo methods

Importance Sampling

Let’s try to make the uniform sampling a little bit more general.

- Let’s assume a target distribution $P(x)$ we are able to evaluate at each x:

$$P(x) = P^*(x)/Z$$

Monte Carlo methods

Importance Sampling

Let's try to make the uniform sampling a little bit more general.

- Let's assume a target distribution $P(x)$ we are able to evaluate at each x:

$$P(x) = P^*(x)/Z$$

- But $P(x)$ is too complicated to sample directly from it…

Monte Carlo methods

Importance Sampling

Let's try to make the uniform sampling a little bit more general.

- Let's assume a target distribution $P(x)$ we are able to evaluate at each x:

$$P(x) = P^*(x)/Z$$

- But $P(x)$ is too complicated to sample directly from it…

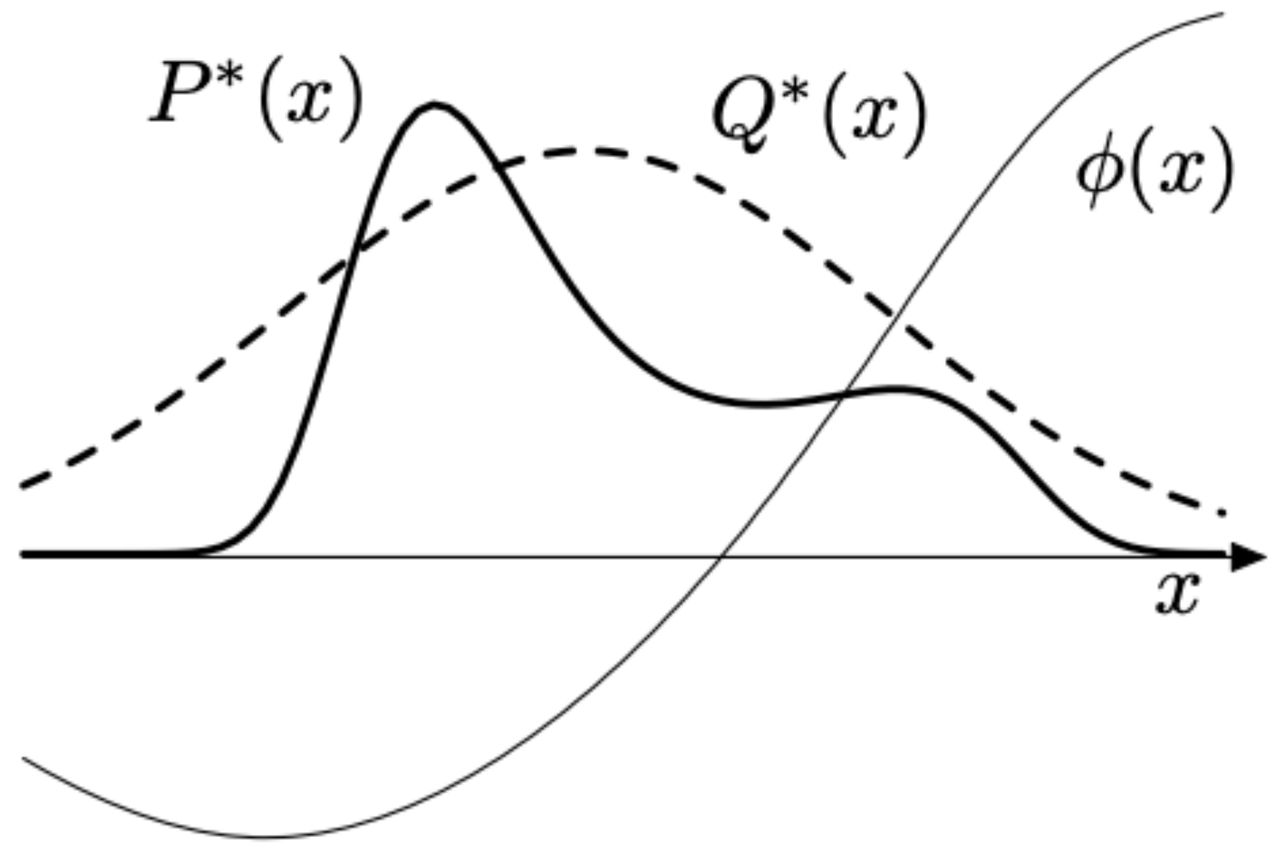
- But let's assume we have a simpler density $Q(x)$, from which we can generate samples and which we can evaluate up to a normalization constant:

$$Q(x) = Q^*(x)/Z$$

Monte Carlo methods

Importance Sampling

- We can generate N samples from $Q(x)$.

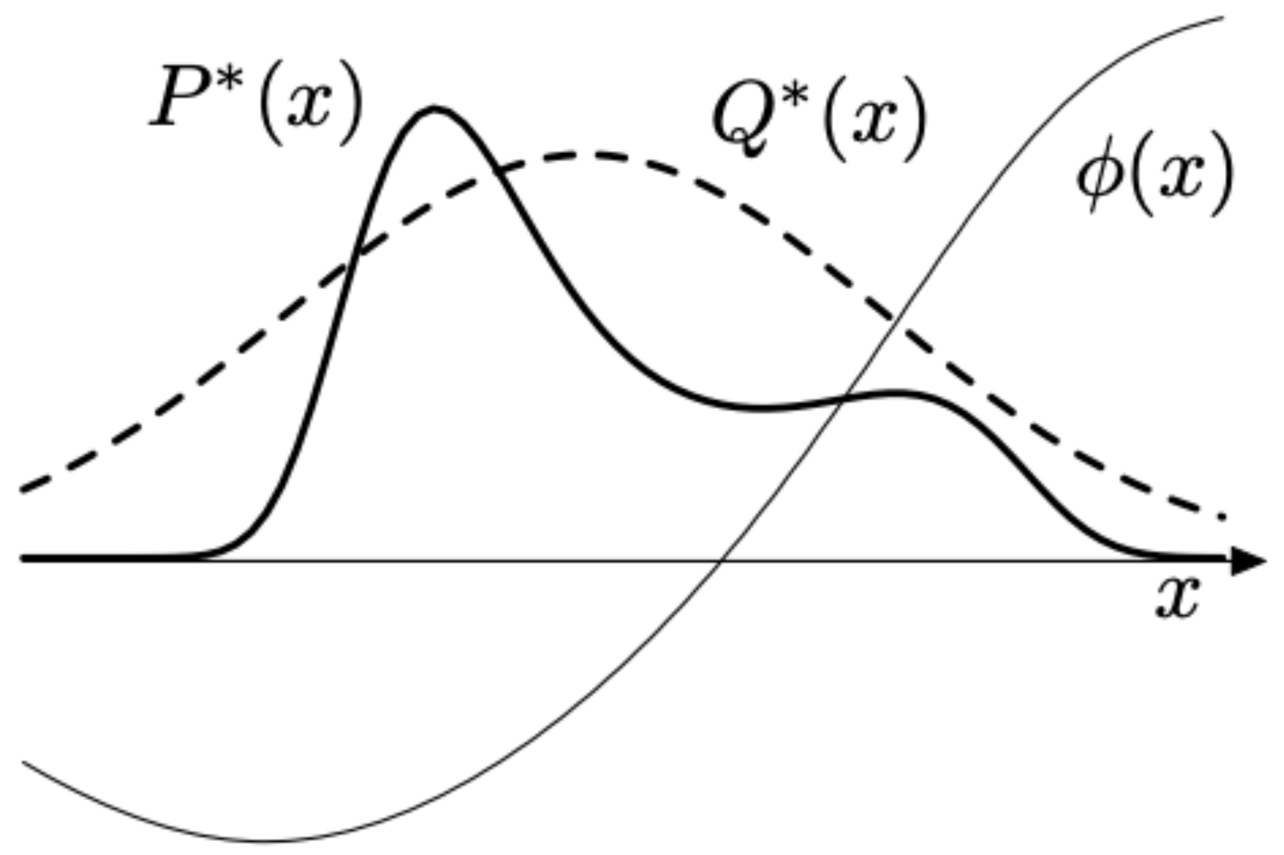


Monte Carlo methods

Importance Sampling

- We can generate N samples from $Q(x)$.
- To take into account for the fact that our samples have been sampled from the wrong distribution we introduce weights:

$$w^{(i)} = \frac{P^*(x^{(i)})}{Q^*(x^{(i)})}$$



Monte Carlo methods

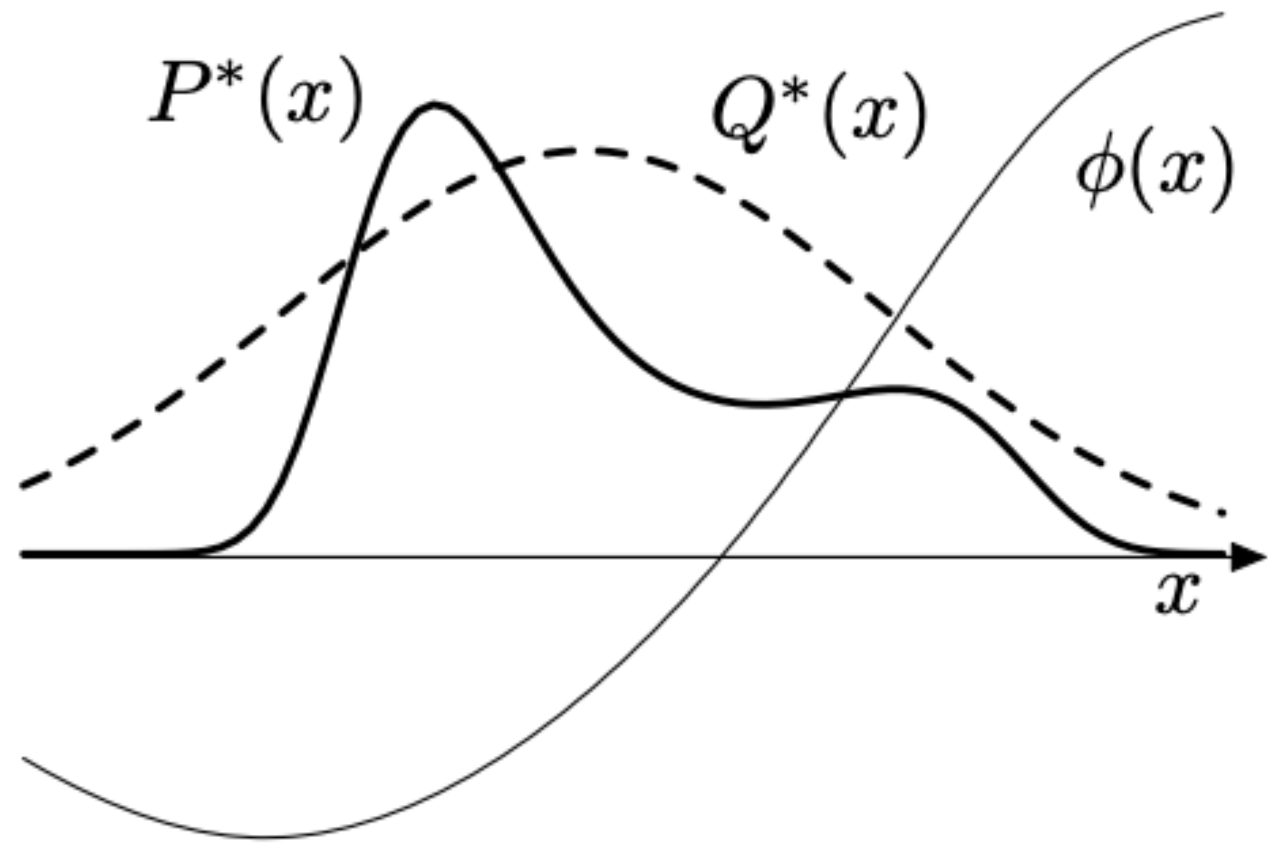
Importance Sampling

- We can generate N samples from $Q(x)$.
- To take into account for the fact that our samples have been sampled from the wrong distribution we introduce weights:

$$w^{(i)} = \frac{P^*(x^{(i)})}{Q^*(x^{(i)})}$$

- Which we can use to adjust the importance of each point in our estimator such that:

$$\hat{\Phi} = \frac{\sum_i w^{(i)} \phi(x^{(i)})}{\sum_i w^{(i)}}$$



Monte Carlo methods

Importance Sampling

WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

Monte Carlo methods

Importance Sampling

Drawbacks of importance sampling:

Monte Carlo methods

Importance Sampling

Drawbacks of importance sampling:

- It is hard to estimate how reliable $\hat{\Phi}$ is.

Monte Carlo methods

Importance Sampling

Drawbacks of importance sampling:

- It is hard to estimate how reliable $\hat{\Phi}$ is.
- The variance of the estimator is hard to estimate because it depends on an integral over x and $P^*(x)$.

Monte Carlo methods

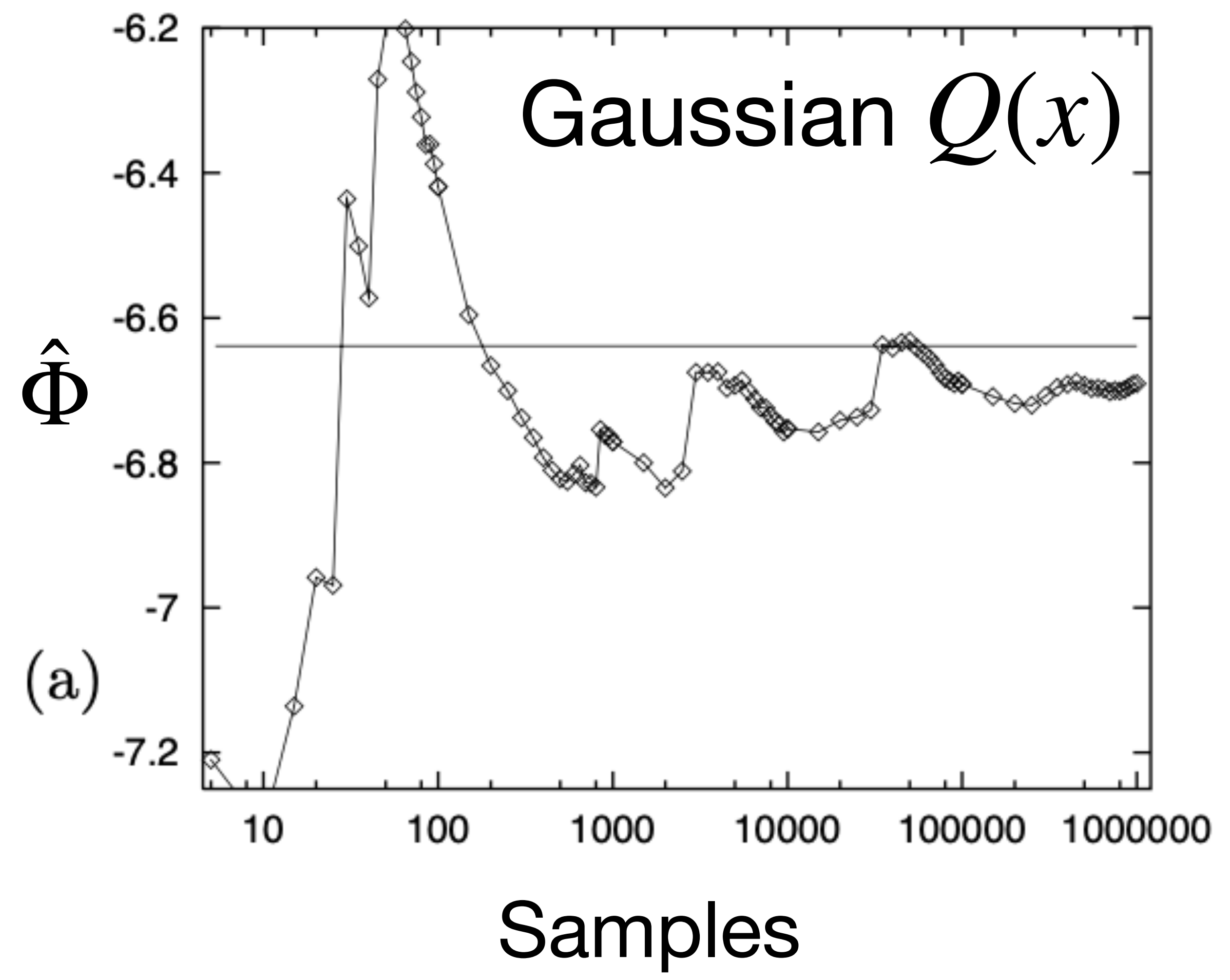
Importance Sampling

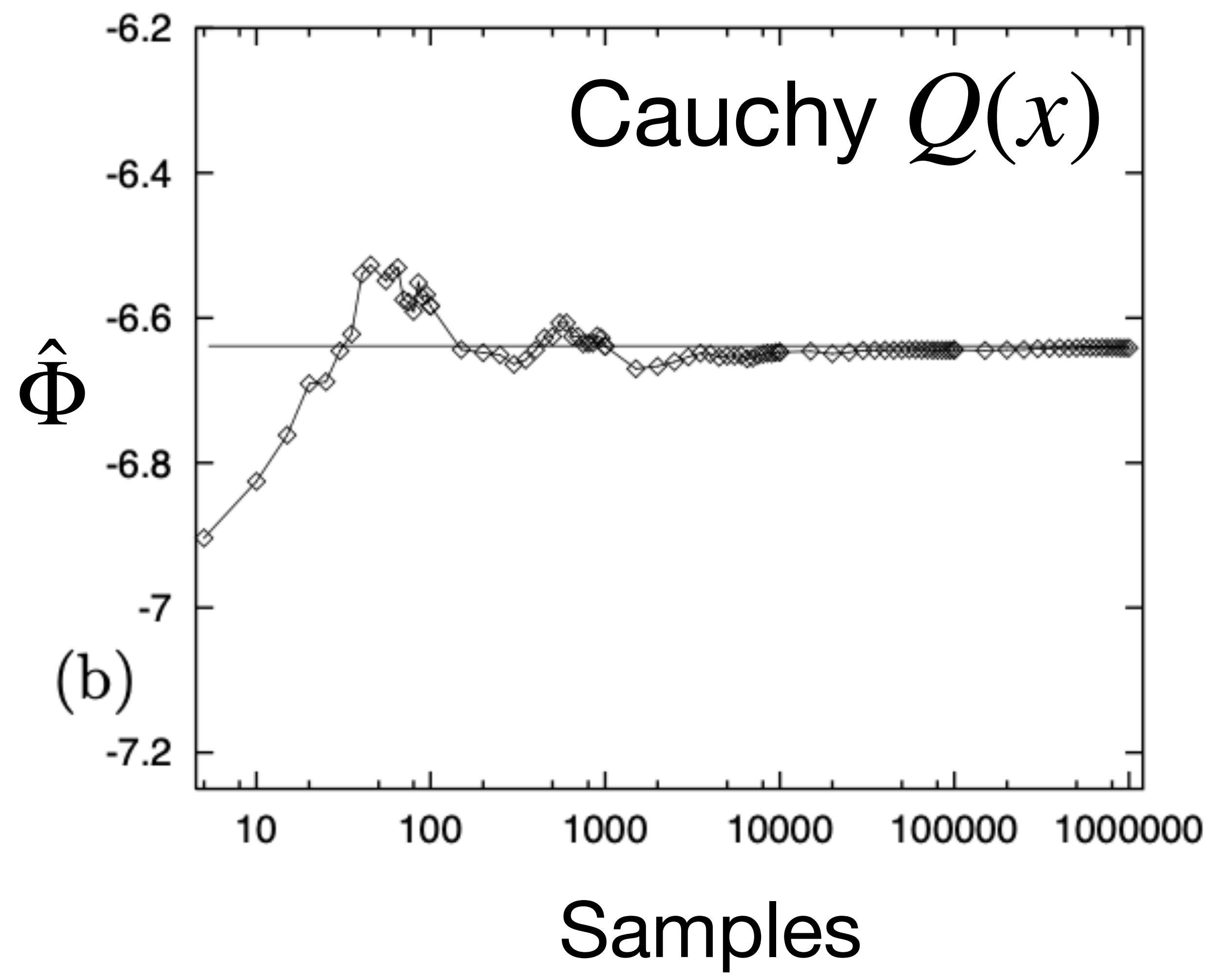
Drawbacks of importance sampling:

- It is hard to estimate how reliable $\hat{\Phi}$ is.
- The variance of the estimator is hard to estimate because it depends on an integral over x and $P^*(x)$.
- The variance of $\hat{\Phi}$ is not trivial because only relies on the empirical variances of $w^{(i)}$ and $w^{(i)}\phi(x^{(i)})$.

Monte Carlo methods

Importance Sampling

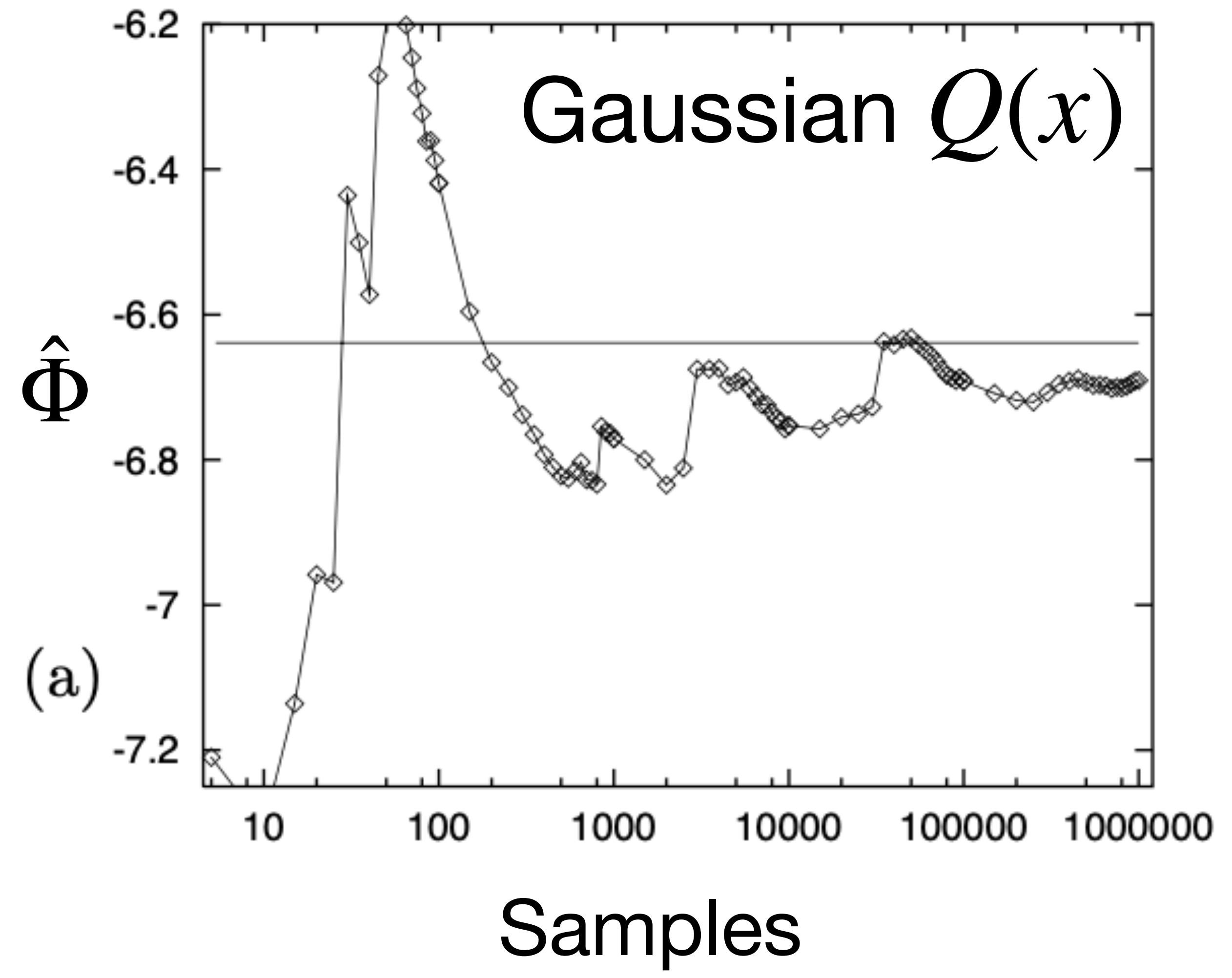


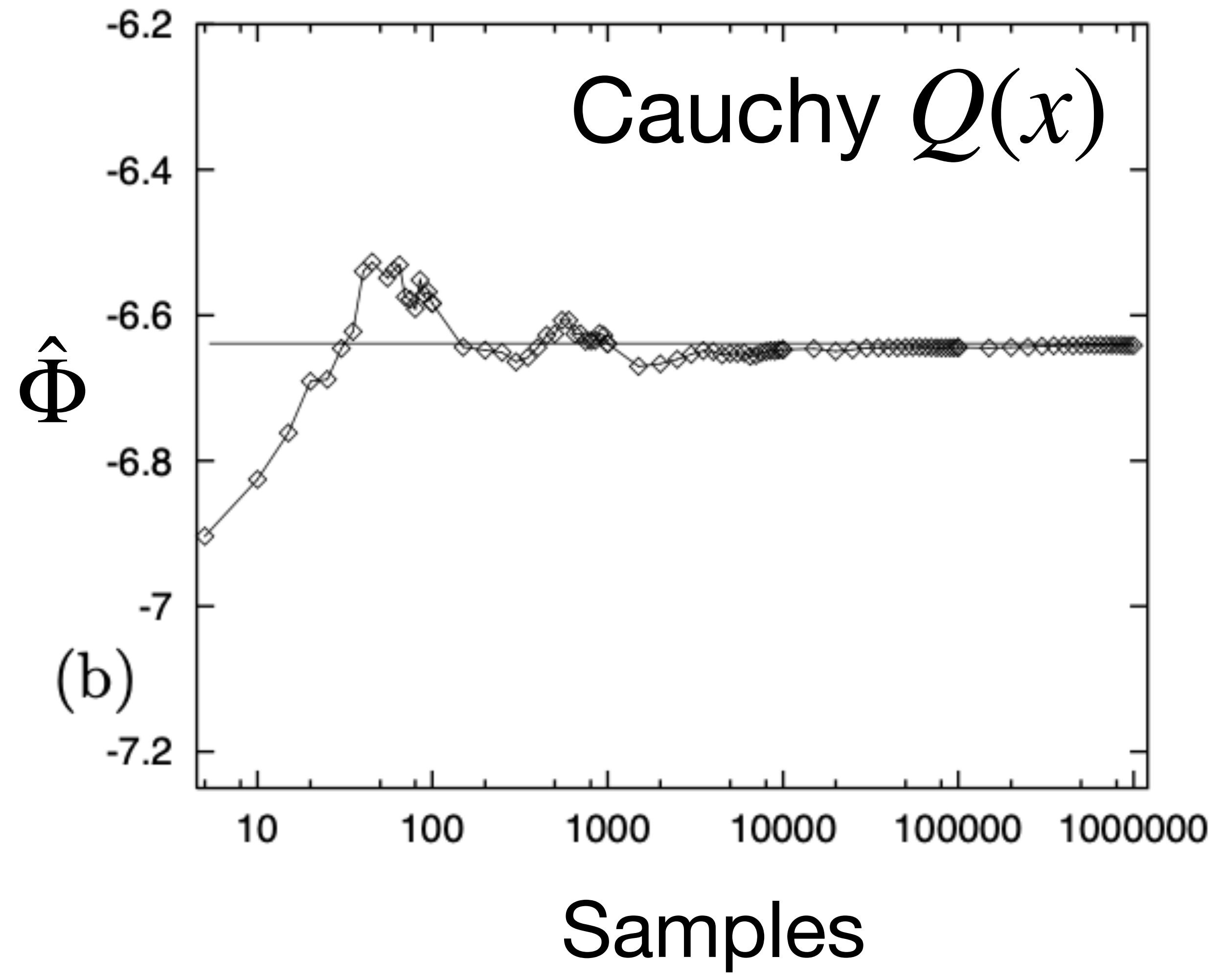


Monte Carlo methods

Importance Sampling

- The variance of $\hat{\Phi}$ is not trivial because only relies on the empirical variances of w_i and $w_i\phi(x_i)$.

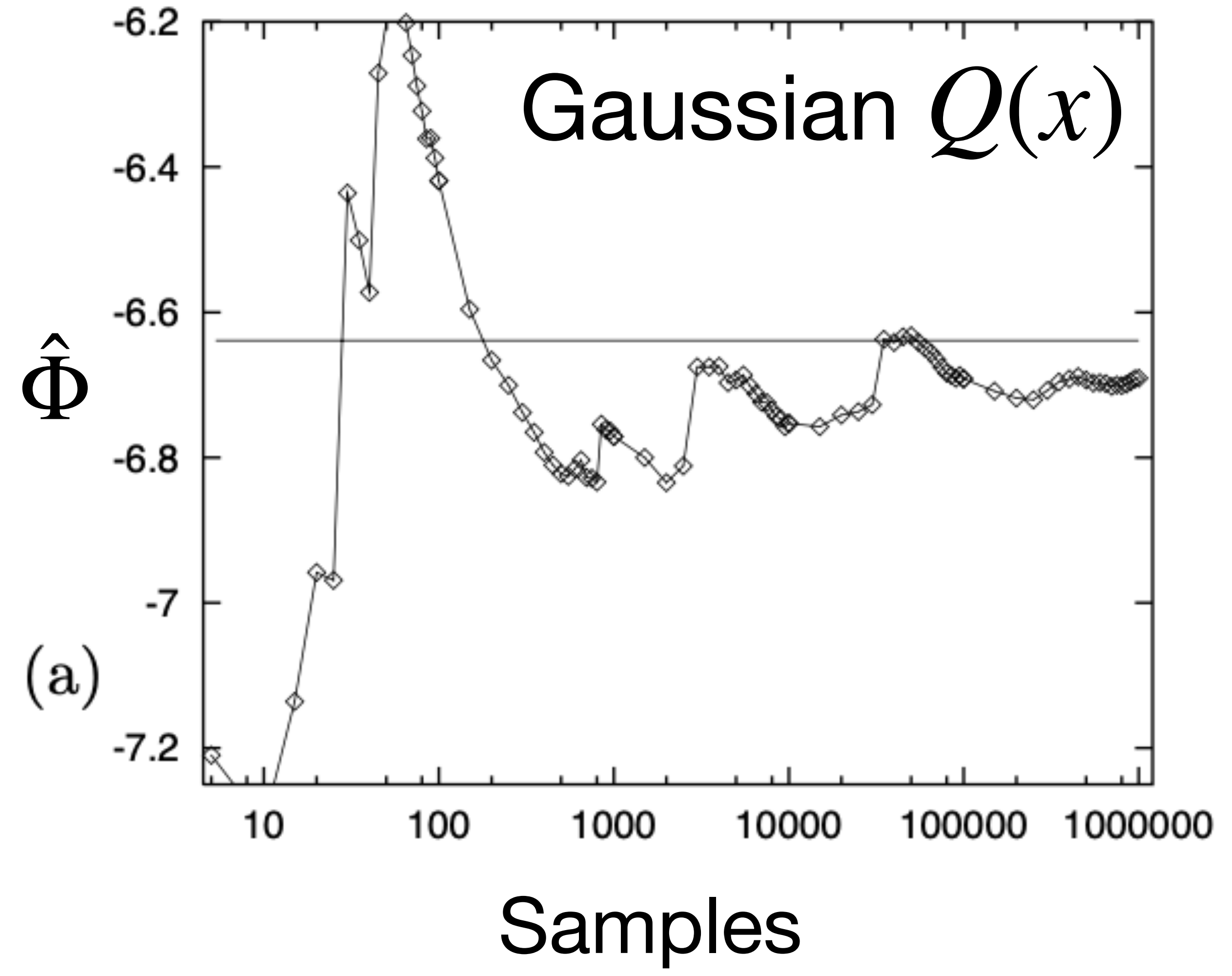


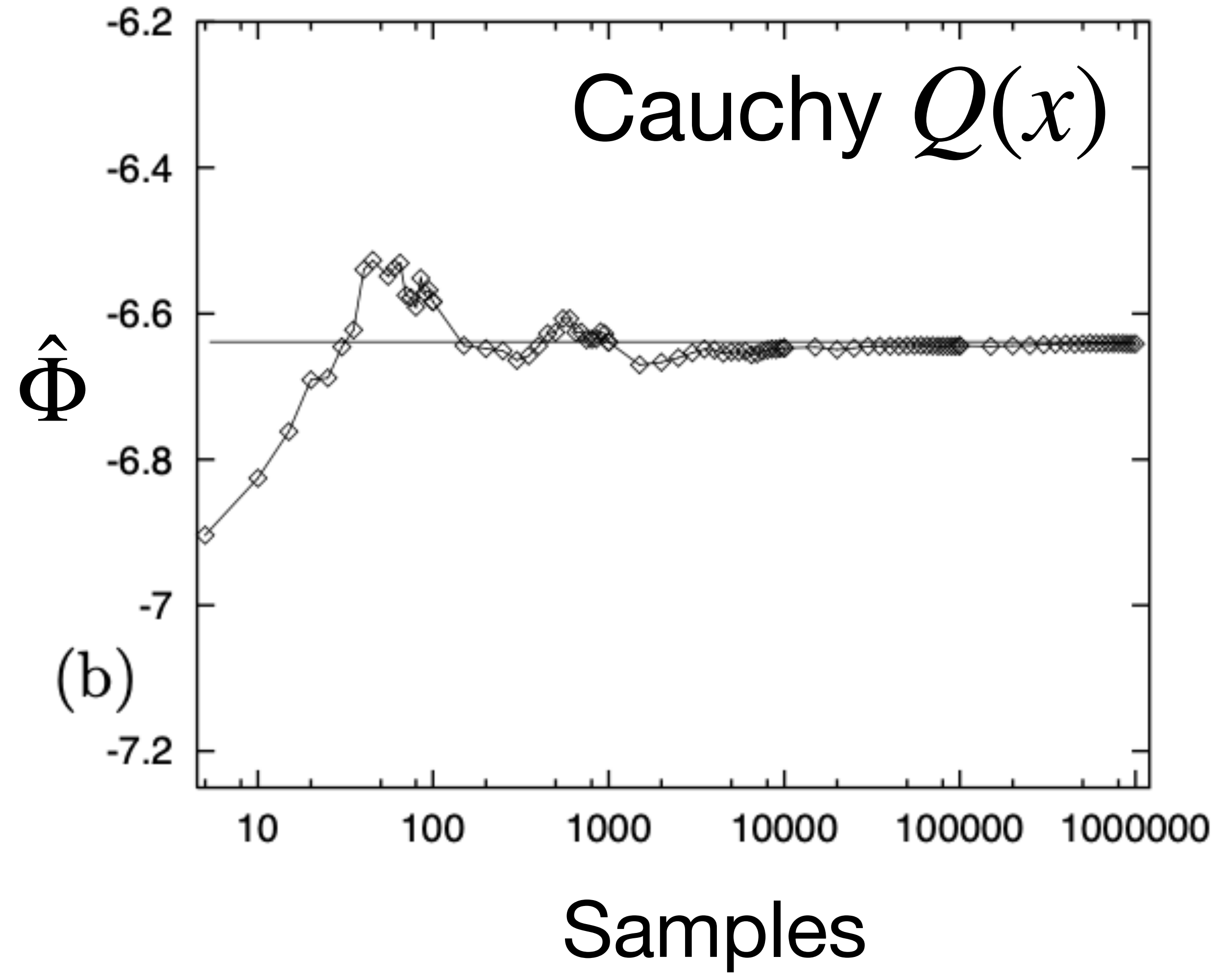


Monte Carlo methods

Importance Sampling

- The variance of $\hat{\Phi}$ is not trivial because only relies on the empirical variances of w_i and $w_i\phi(x_i)$.
    - If the $Q(x)$ is small where $|\phi(x)P^*(x)|$ is large, it may happen that, even for a large number of samples, none of them will fall in that region.

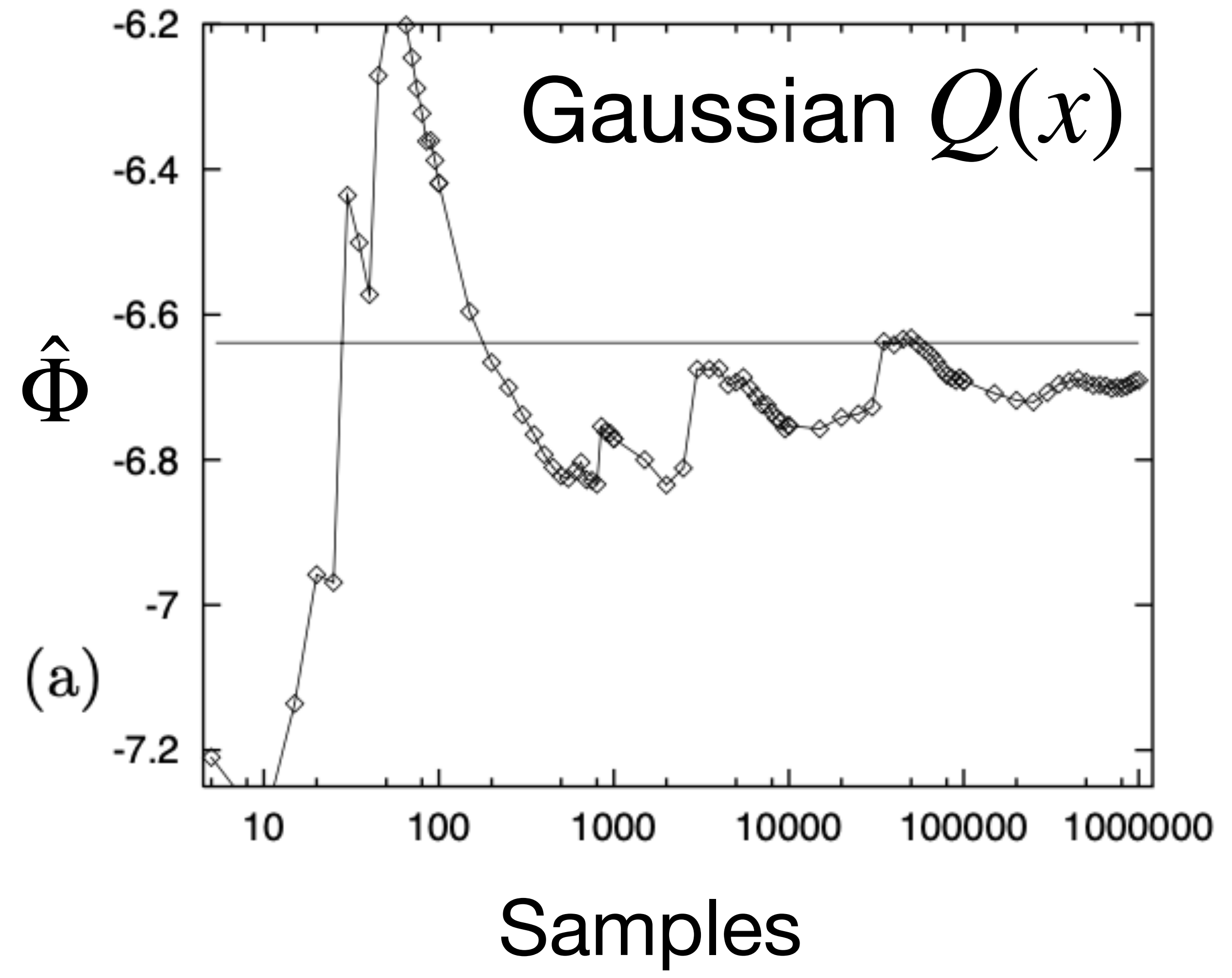


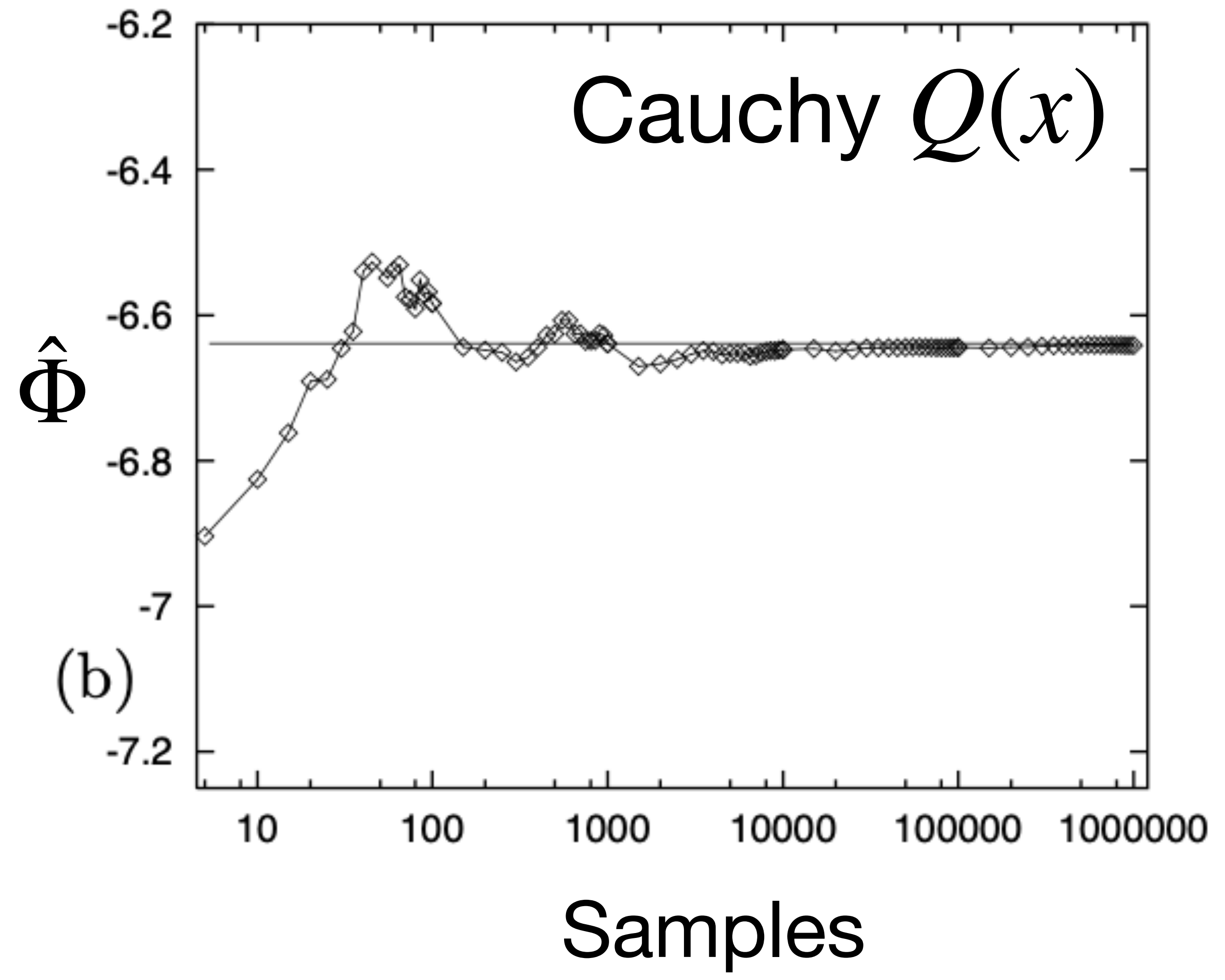


Monte Carlo methods

Importance Sampling

- The variance of $\hat{\Phi}$ is not trivial because only relies on the empirical variances of w_i and $w_i\phi(x_i)$.
  - If the $Q(x)$ is small where $|\phi(x)P^*(x)|$ is large, it may happen that, even for a large number of samples, none of them will fall in that region.
  - In this case $\hat{\Phi}$ will be wrong, with no indication in the empirical variance variance that the true variance of $\hat{\Phi}$ is large.





Monte Carlo methods

Rejection sampling

- Let's once more assume a target distribution $P(x)$ that is too complicated for us to be able to sample from it directly.

Monte Carlo methods

Rejection sampling

- Let's once more assume a target distribution $P(x)$ that is too complicated for us to be able to sample from it directly.

- Let's assume we have a proposal density $Q(x)$ we can evaluate, and from which we can generate samples.

Monte Carlo methods

Rejection sampling

- Let's once more assume a target distribution $P(x)$ that is too complicated for us to be able to sample from it directly.

- Let's assume we have a proposal density $Q(x)$ we can evaluate, and from which we can generate samples.

- We also assume that we know the value of a constant c such that:

$$cQ^*(x) > P^*(x)$$

Monte Carlo methods

Rejection sampling

- We generate two random numbers: the first is $x \sim Q(x)$.

Monte Carlo methods

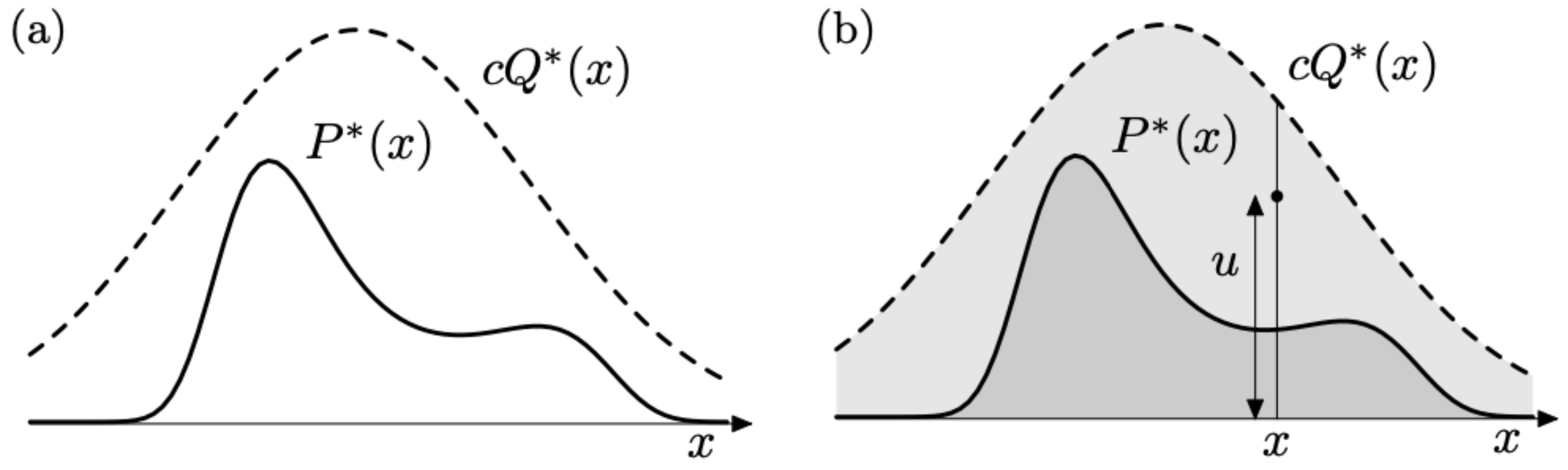
Rejection sampling

- We generate two random numbers: the first is $x \sim Q(x)$.
- We evaluate $cQ^*(x)$ and generate $u \sim U(0, cQ^*(x))$.

Monte Carlo methods

Rejection sampling

- We generate two random numbers: the first is $x \sim Q(x)$.
- We evaluate $cQ^*(x)$ and generate $u \sim U(0, cQ^*(x))$.
- We evaluate $P^*(x)$ and accept or reject x by comparing the value of u with $P^*(x)$. If $u > P^*(x)$ then x is rejected, otherwise is accepted and added to our collection of samples $\{x^{(i)}\}_N$.



Monte Carlo methods

Rejection sampling

- This method works best if Q is a good approximation to P.

Monte Carlo methods

Rejection sampling

- This method works best if Q is a good approximation to P.

- If Q and P are very different, we need a very large value of c to ensure $cQ > P$ everywhere, and therefore the rejection frequency will be large —> need a lot of samples.

Markov chains

Definitions

- A sequence $X_1, X_2, \ldots, X_n$ of random elements of some set is a **Markov chain** if the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ depends on X_n only.
- A Markov chain has **stationary transition probabilities** if $P(X_{n+1} \,|\, X_n)$ does not depend on n.

Markov chains

Definitions

The joint distribution of a Markov chain is determined by:

Markov chains

Definitions

The joint distribution of a Markov chain is determined by:

- The marginal distribution of X_1, called the **initial distribution**.

Markov chains

Definitions

The joint distribution of a Markov chain is determined by:

- The marginal distribution of X_1, called the **initial distribution**.
- $P(X_{n+1} \,|\, X_n)$, called the **transition probability distribution**.

Markov chains

Definitions

The joint distribution of a Markov chain is determined by:

- The marginal distribution of X_1, called the **initial distribution**.
- $P(X_{n+1} \mid X_n)$, called the **transition probability distribution**.

Moreover:

Markov chains

Definitions

The joint distribution of a Markov chain is determined by:

- The marginal distribution of X_1, called the **initial distribution**.
- $P(X_{n+1} \,|\, X_n)$, called the **transition probability distribution**.

Moreover:

- A transition probability distribution is **reversible** with respect to an initial distribution if, for the Markov chain $X_1, X_2, \ldots$, the distribution of pairs m (X_i, X_{i+1}) is exchangeable.

Markov chains

Reversibility

A Markov chain is reversible if its transition probability is reversible with respect to its initial distribution

Markov chains

Reversibility

A Markov chain is reversible if its transition probability is reversible with respect to its initial distribution

- Reversibility —> Stationarity (not vice versa).

Markov chains

Reversibility

A Markov chain is reversible if its transition probability is reversible with respect to its initial distribution

- Reversibility —> Stationarity (not vice versa).

Reversibility plays two roles in Markov chain theory:

Markov chains
Reversibility

A Markov chain is reversible if its transition probability is reversible with respect to its initial distribution

- Reversibility —> Stationarity (not vice versa).

Reversibility plays two roles in Markov chain theory:

- The Markov chain Central Limit theorem is much sharper and conditions much simpler when reversibility applies.

Markov chains

Reversibility

A Markov chain is reversible if its transition probability is reversible with respect to its initial distribution

- Reversibility —> Stationarity (not vice versa).

Reversibility plays two roles in Markov chain theory:

- The Markov chain Central Limit theorem is much sharper and conditions much simpler when reversibility applies.

- It allows for constructing efficient probability mechanisms for MCMC (we'll see shortly…).

MCMC algoritms

Metropolis-Hastings

In the Metropolis-Hastings sampling our proposal density $Q(x)$ depends on the current state $x^{(i)}$.

MCMC algoritms

Metropolis-Hastings

In the Metropolis-Hastings sampling our proposal density $Q(x)$ depends on the current state $x^{(i)}$.

- This is an MCMC process, and $Q(x'; x^{(i)})$ does not need to be close to the true $P(x)$ to be useful!

MCMC algoritms

Metropolis-Hastings

In the Metropolis-Hastings sampling our proposal density $Q(x)$ depends on the current state $x^{(i)}$.

- This is an MCMC process, and $Q(x'; x^{(i)})$ does not need to be close to the true $P(x)$ to be useful!

Let's assume we can evaluate $P*(x)$ at any given x. How do we draw a new sample x' from the proposal density $Q(x'; x^{(i)})$?

MCMC algoritms

Metropolis-Hastings

Let's assume we can evaluate $P^*(x)$ at any given x. How do we draw a new sample x' from the proposal density $Q(x'; x_i)$?

MCMC algoritms

Metropolis-Hastings

Let's assume we can evaluate $P^*(x)$ at any given x. How do we draw a new sample x' from the proposal density $Q(x'; x_i)$?

- We generate a new sample x' from $Q(x'; x_i)$;

MCMC algoritms

Metropolis-Hastings

Let’s assume we can evaluate $P^*(x)$ at any given x. How do we draw a new sample x' from the proposal density $Q(x'; x_i)$?

- We generate a new sample x' from $Q(x'; x_i)$;
- Compute the **acceptance ratio:**

$$a = \frac{P^*(x')Q(x_i; x')}{P^*(x_i)Q(x'; x_i)}$$

MCMC algoritms

Metropolis-Hastings

Let's assume we can evaluate $P^*(x)$ at any given x. How do we draw a new sample x' from the proposal density $Q(x'; x_i)$?

- We generate a new sample x' from $Q(x'; x_i)$;
- Compute the **acceptance ratio:**

$$a = \frac{P^*(x')Q(x_i; x')}{P^*(x_i)Q(x'; x_i)}$$

- If $a \geq 1$ the new state is accepted; otherwise it is accepted with probability a;

MCMC algoritms

Metropolis-Hastings

Let's assume we can evaluate $P^*(x)$ at any given x. How do we draw a new sample x' from the proposal density $Q(x'; x_i)$?

- We generate a new sample x' from $Q(x'; x_i)$;
- Compute the **acceptance ratio:**

$$a = \frac{P^*(x')Q(x_i; x')}{P^*(x_i)Q(x'; x_i)}$$

- If $a \geq 1$ the new state is accepted; otherwise it is accepted with probability a;
- If the new state is accepted $x_{i+1} = x'$, otherwise $x_{i+1} = x_i$.

MCMC algoritms

Metropolis-Hastings

Let’s assume we can evaluate $P^*(x)$ at any given x. How do we draw a new sample x' from the propos

Note the difference from rejection sampling: here a rejection cause the current state to be written again into the samples list.

- We generate a
- Compute the **acceptance ratio:**

$$a = \frac{P^*(x')Q(x_i; x')}{P^*(x_i)Q(x'; x_i)}$$

- If $a \geq 1$ the new state is accepted; otherwise it is accepted with probability a;
- If the new state is accepted $x_{i+1} = x'$, otherwise $x_{i+1} = x_i$.

WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

MCMC algoritms

Gibbs sampling

Gibbs sampling can be viewed as a MH method in which a sequence of proposal distributions Q are defined in terms of the conditional distributions of the joint distribution $P(\mathbf{x})$.

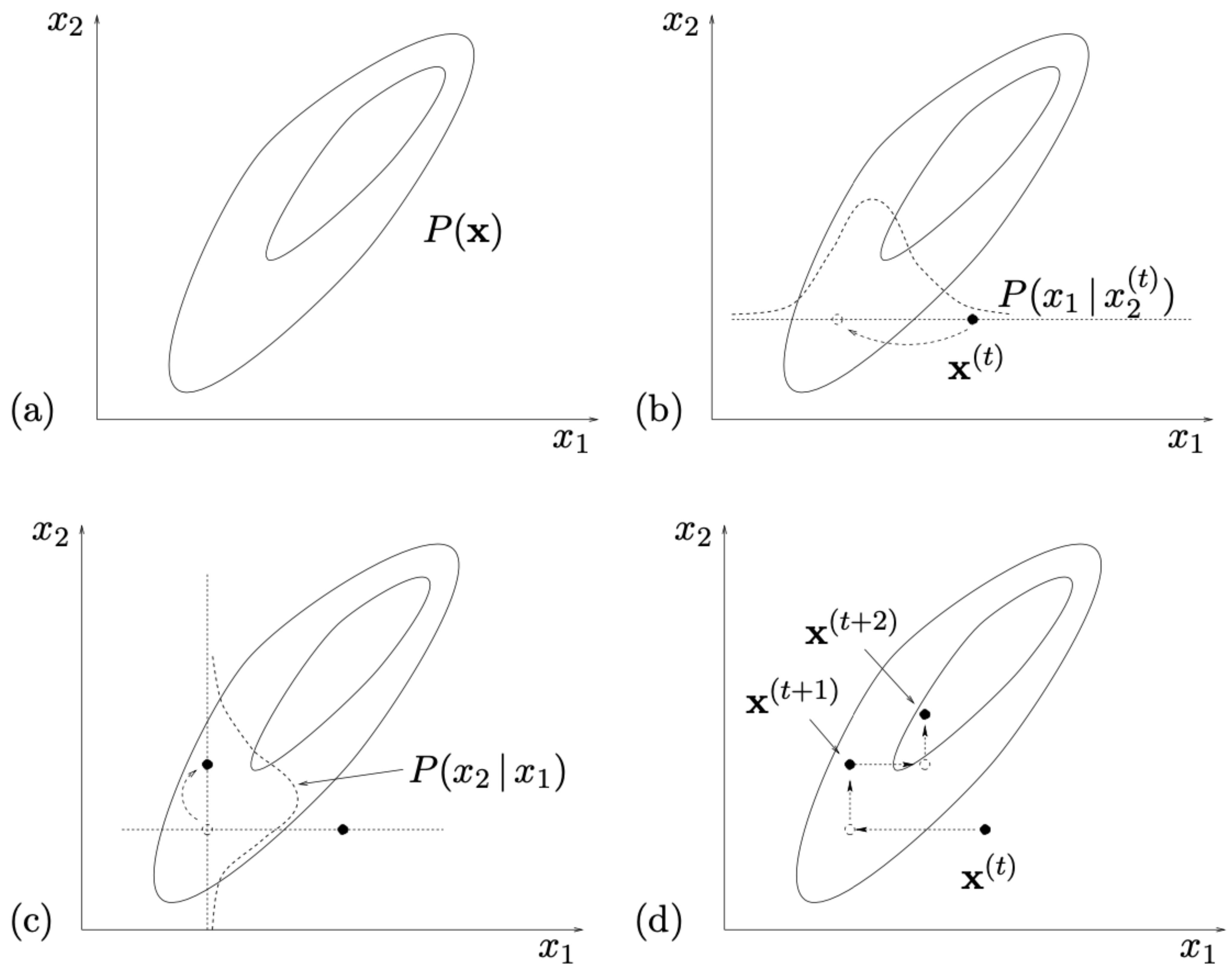
It is assumed that $P(\mathbf{x})$ is too complex to draw samples from, but its conditional distributions $P(x_i \,|\, x_{j\neq i})$ are tractable to work with.

PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

A simple 2D example:



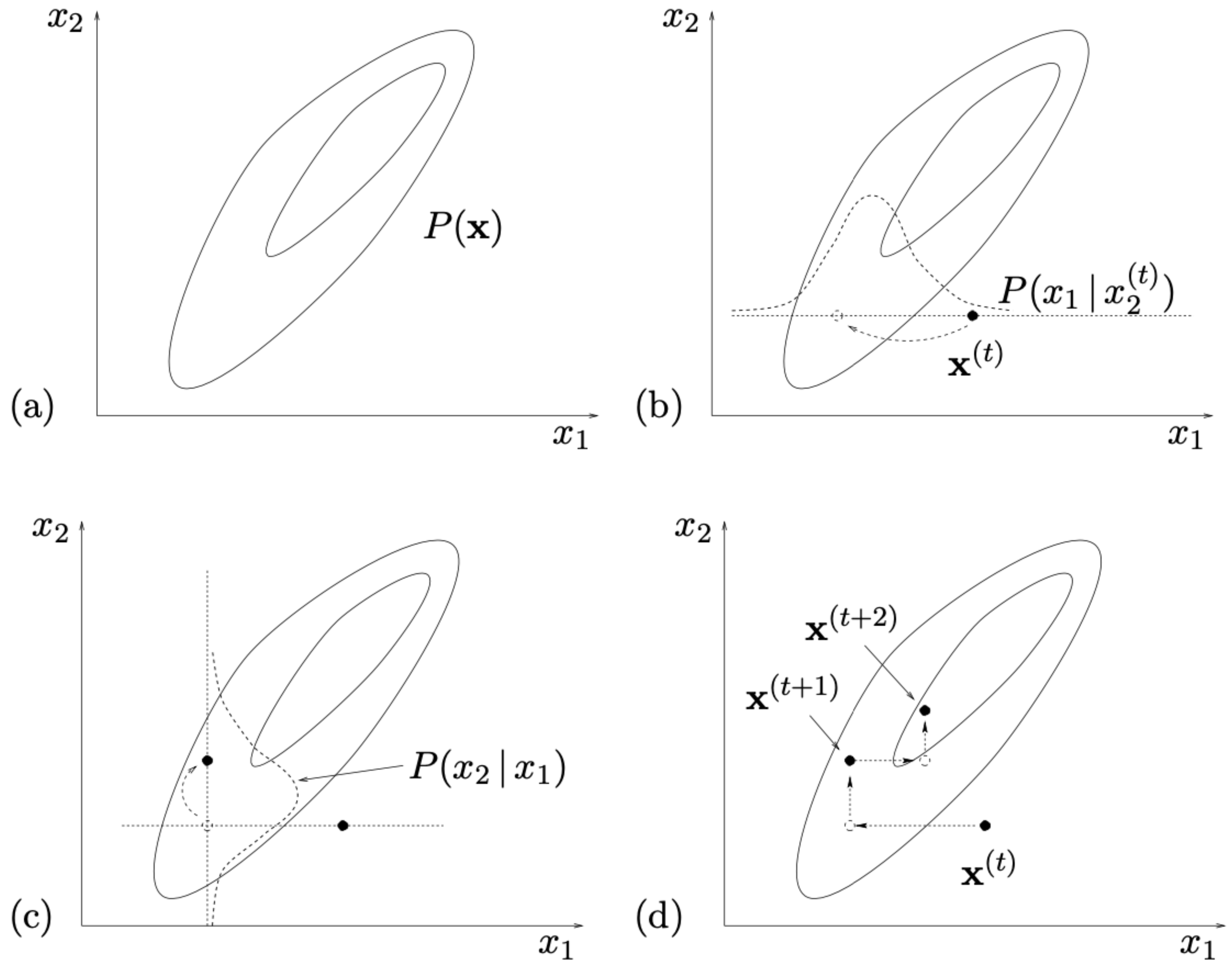
PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

A simple 2D example:

- Two variables $(x_1, x_2) = \mathbf{x}$



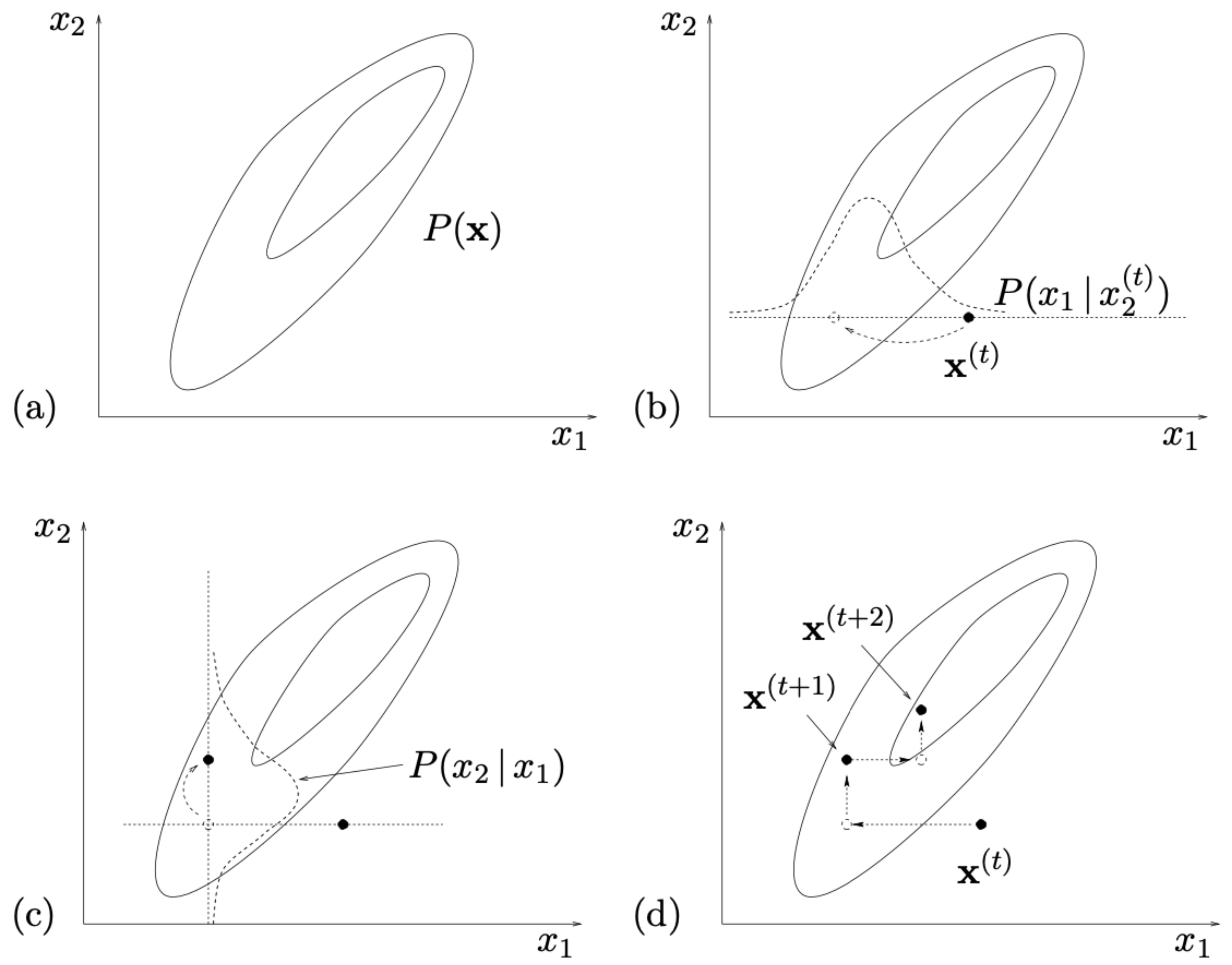
PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

A simple 2D example:

- Two variables $(x_1, x_2) = \mathbf{x}$
- On each iteration, we start from the current state $\mathbf{x}^{(i)}$, and sample:



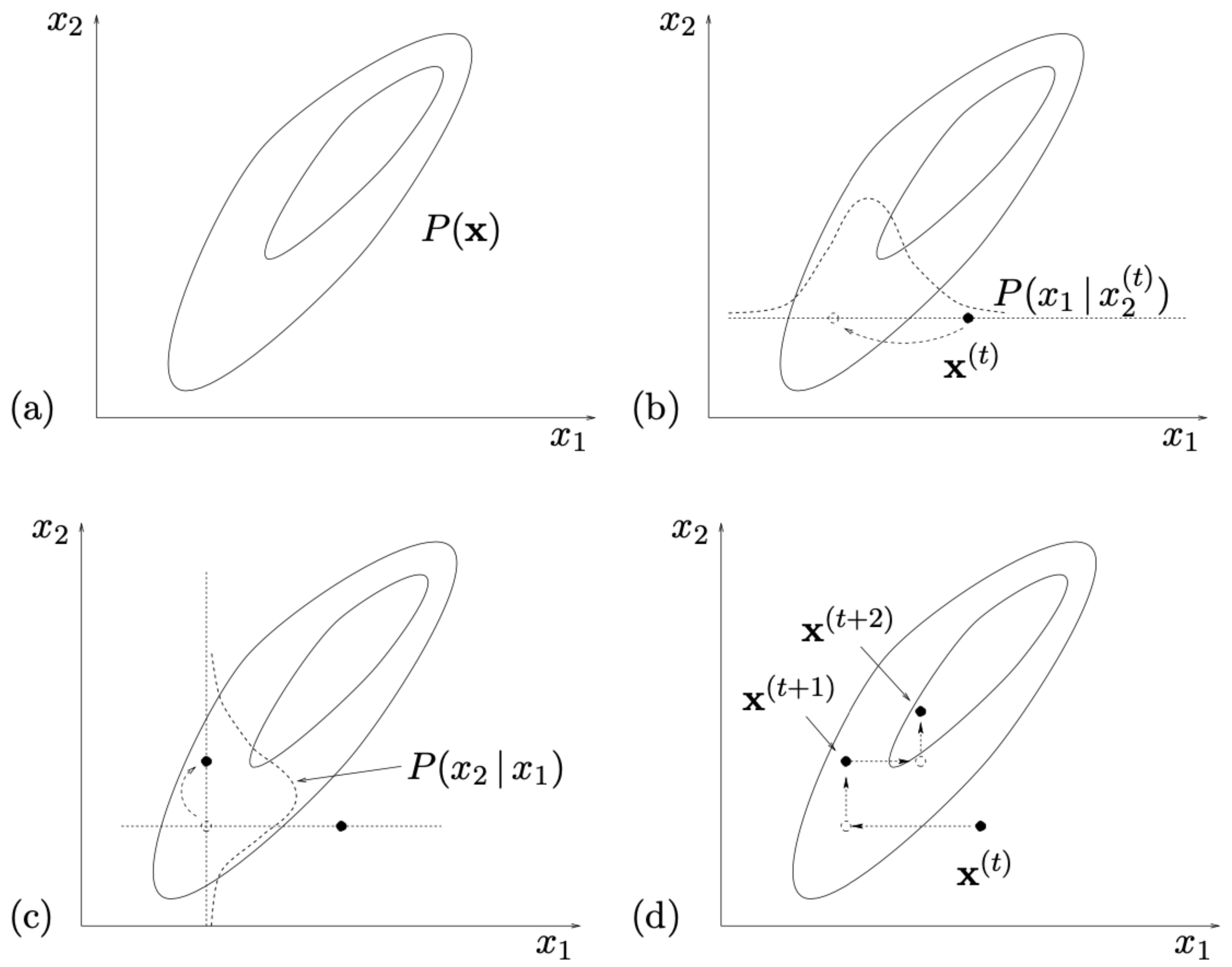
PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

A simple 2D example:

- Two variables $(x_1, x_2) = \mathbf{x}$
- On each iteration, we start from the current state $\mathbf{x}^{(i)}$, and sample:
  - $x_1^{(i+1)} \sim P(x_1 \,|\, x_2^{(i)})$



WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

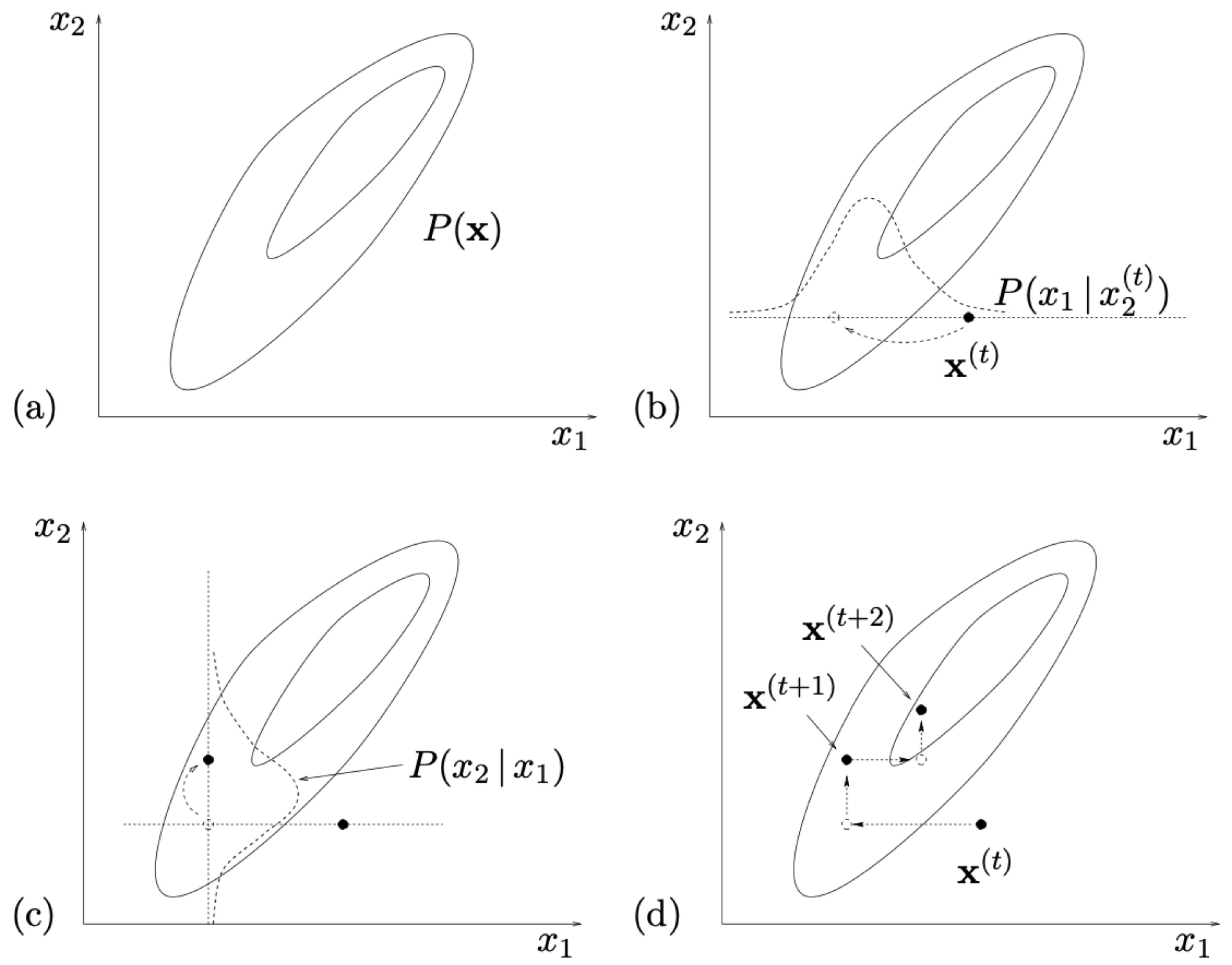
PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

A simple 2D example:

- Two variables $(x_1, x_2) = \mathbf{x}$
- On each iteration, we start from the current state $\mathbf{x}^{(i)}$, and sample:
    - $x_1^{(i+1)} \sim P(x_1 \,|\, x_2^{(i)})$
    - $x_2^{(i+1)} \sim P(x_2 \,|\, x_1^{(i+1)})$



WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

In general, for a system of K variables, a single iteration involves sampling one parameter at a time:

$$\begin{aligned}
x_1^{(t+1)} &\sim P(x_1 \,|\, x_2^{(t)}, x_3^{(t)}, \dots, x_K^{(t)}) \\
x_2^{(t+1)} &\sim P(x_2 \,|\, x_1^{(t+1)}, x_3^{(t)}, \dots, x_K^{(t)}) \\
x_3^{(t+1)} &\sim P(x_3 \,|\, x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_K^{(t)}), \text{ etc.}
\end{aligned}$$

PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

In general, for a system of K variables, a single iteration involves sampling one parameter at a time:

$$\begin{aligned} x_1^{(t+1)} &\sim P(x_1 \,|\, x_2^{(t)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\ x_2^{(t+1)} &\sim P(x_2 \,|\, x_1^{(t+1)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\ x_3^{(t+1)} &\sim P(x_3 \,|\, x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_K^{(t)}), \ \text{etc.} \end{aligned}$$

What's the acceptance rate of Gibbs sampling?

PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

In general, for a system of K variables, a single iteration involves sampling one parameter at a time:

$$\begin{aligned} x_1^{(t+1)} &\sim P(x_1 \,|\, x_2^{(t)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\ x_2^{(t+1)} &\sim P(x_2 \,|\, x_1^{(t+1)}, x_3^{(t)}, \ldots, x_K^{(t)}) \\ x_3^{(t+1)} &\sim P(x_3 \,|\, x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_K^{(t)}), \ \text{etc.} \end{aligned}$$

What’s the acceptance rate of Gibbs sampling?

PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

What’s the acceptance rate of Gibbs sampling?

PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

What's the acceptance rate of Gibbs sampling?

$$a = \frac{P^*(\mathbf{x}')Q(\mathbf{x}^{(i)}; \mathbf{x}')}{P^*(\mathbf{x}^{(i)})Q(\mathbf{x}'; \mathbf{x}^{(i)})} = \frac{P^*(x_k' \,|\, \mathbf{x}_{-k}^{(i)})P(x_k \,|\, \mathbf{x}'_{-k})}{P^*(x_k \,|\, \mathbf{x}'_{-k})P(x_k' \,|\, \mathbf{x}_{-k}^{(i)})} = \frac{ZP(x_k' \,|\, \mathbf{x}_{-k}^{(i)})P(x_k \,|\, \mathbf{x}'_{-k})}{ZP(x_k \,|\, \mathbf{x}'_{-k})P(x_k' \,|\, \mathbf{x}_{-k}^{(i)})} = 1$$

PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

MCMC algoritms

Gibbs sampling

What’s the acceptance rate of Gibbs sampling?

$$a = \frac{P^*(\mathbf{x}')Q(\mathbf{x}^{(i)}; \mathbf{x}')}{P^*(\mathbf{x}^{(i)})Q(\mathbf{x}'; \mathbf{x}^{(i)})} = \frac{P^*(x_k' \,|\, \mathbf{x}_{-k}^{(i)})P(x_k \,|\, \mathbf{x}'_{-k})}{P^*(x_k \,|\, \mathbf{x}'_{-k})P(x_k' \,|\, \mathbf{x}_{-k}^{(i)})} = \frac{ZP(x_k' \,|\, \mathbf{x}_{-k}^{(i)})P(x_k \,|\, \mathbf{x}'_{-k})}{ZP(x_k \,|\, \mathbf{x}'_{-k})P(x_k' \,|\, \mathbf{x}_{-k}^{(i)})} = 1$$

Gibbs samples are always accepted!

PS: The person who is teaching this class on my behalf will probably say that Gibbs sampling is obsolete. **He himself is obsolete.**

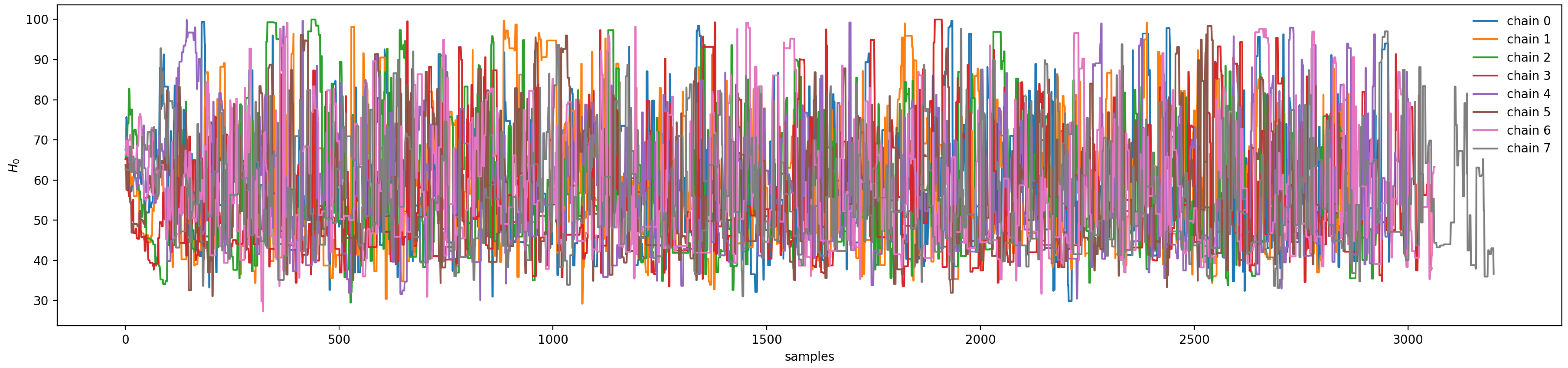
MCMC algoritms

Let’s play

- https://chi-feng.github.io/mcmc-demo/app.html

When to stop sampling?

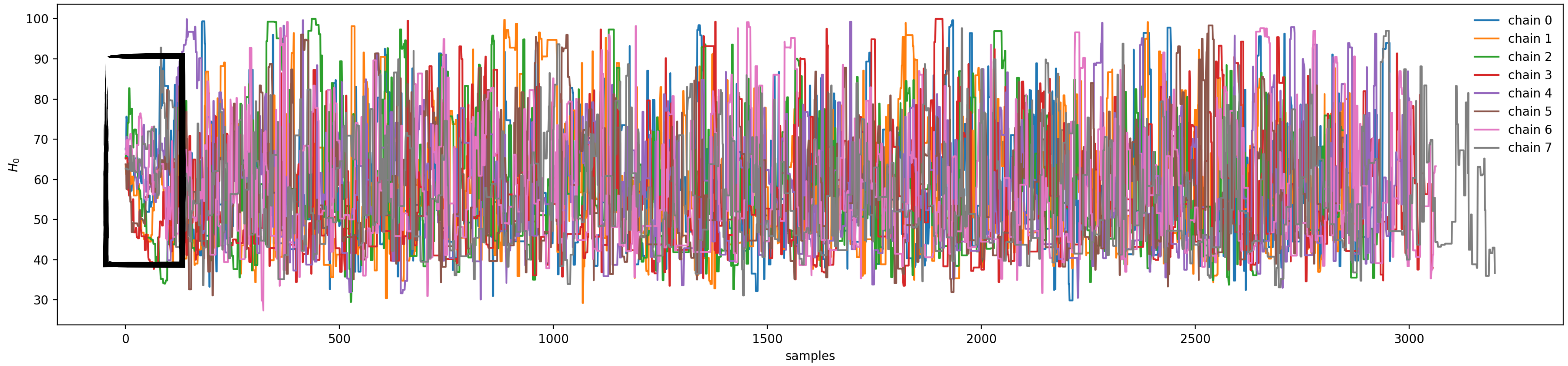
First approach: by eye.



WATERLOO CENTRE FOR ASTROPHYSICS
UNIVERSITY OF WATERLOO

When to stop sampling?

First approach: by eye.



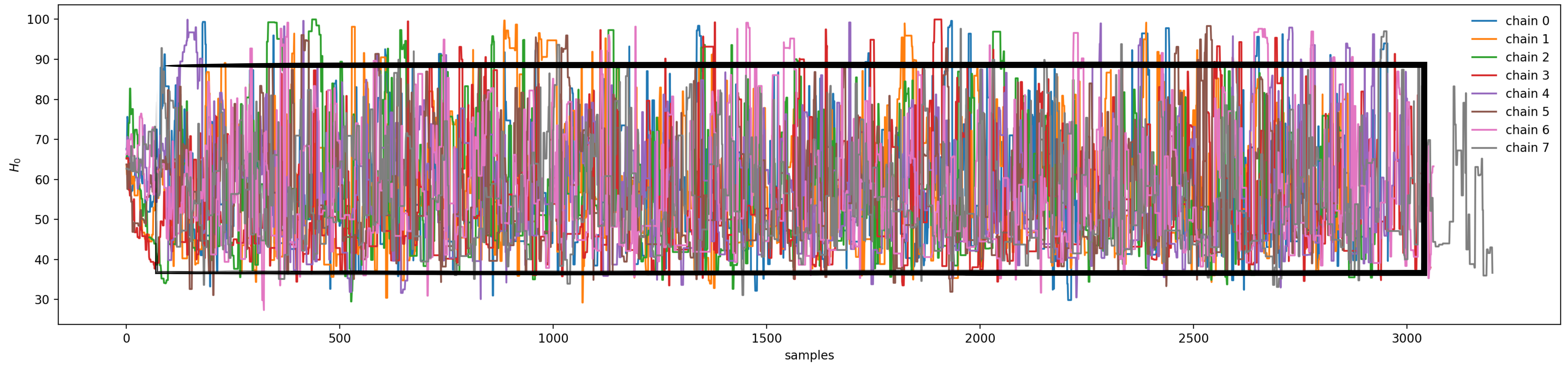
First Burn-in phase: chains leading to a stationary state from initial random points

WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

When to stop sampling?

First approach: by eye.

For a sufficiently long run all chains explore efficiently the parameter space.

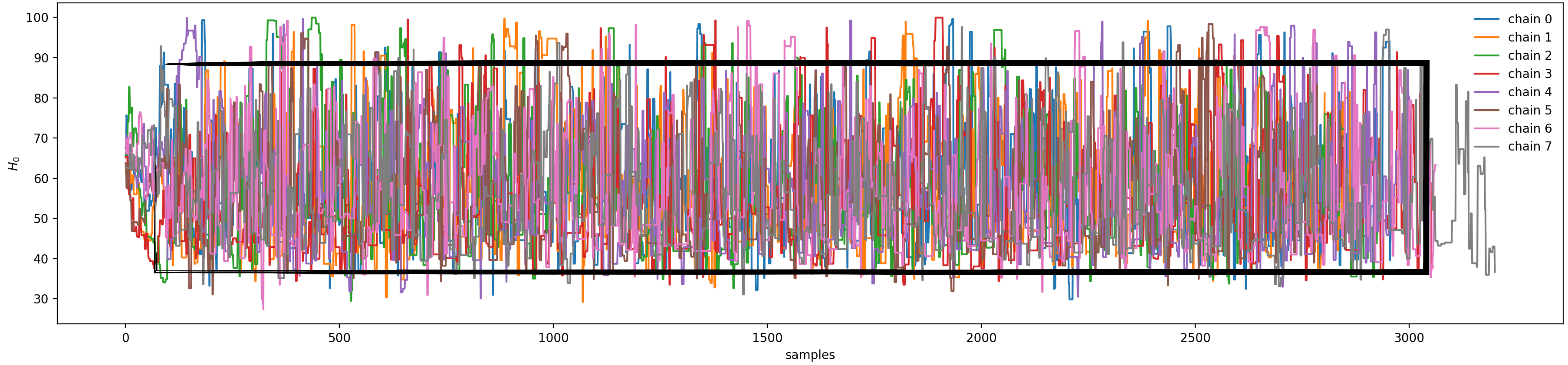


WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

When to stop sampling?

First approach: by eye.

For a sufficiently long run all chains explore efficiently the parameter space.



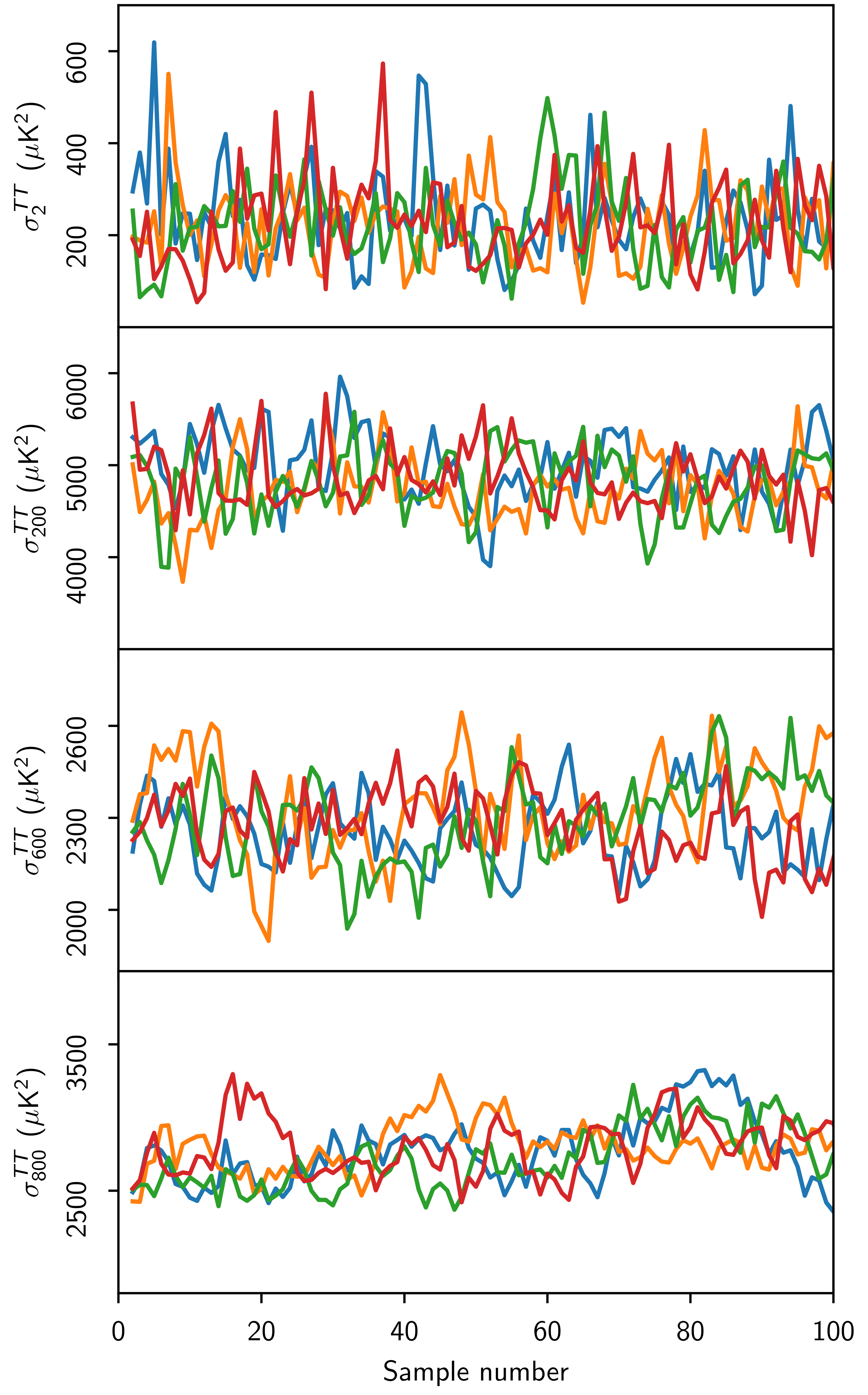
Small correlation length —> good exploration; ergodicity.

When to stop sampling?

First approach: by eye.

Correlation length can be very helpful when it comes to assess quickly the convergency status of chains **individually**.

BeyondPlanck I - The BeyondPlanck collaboration. 2023



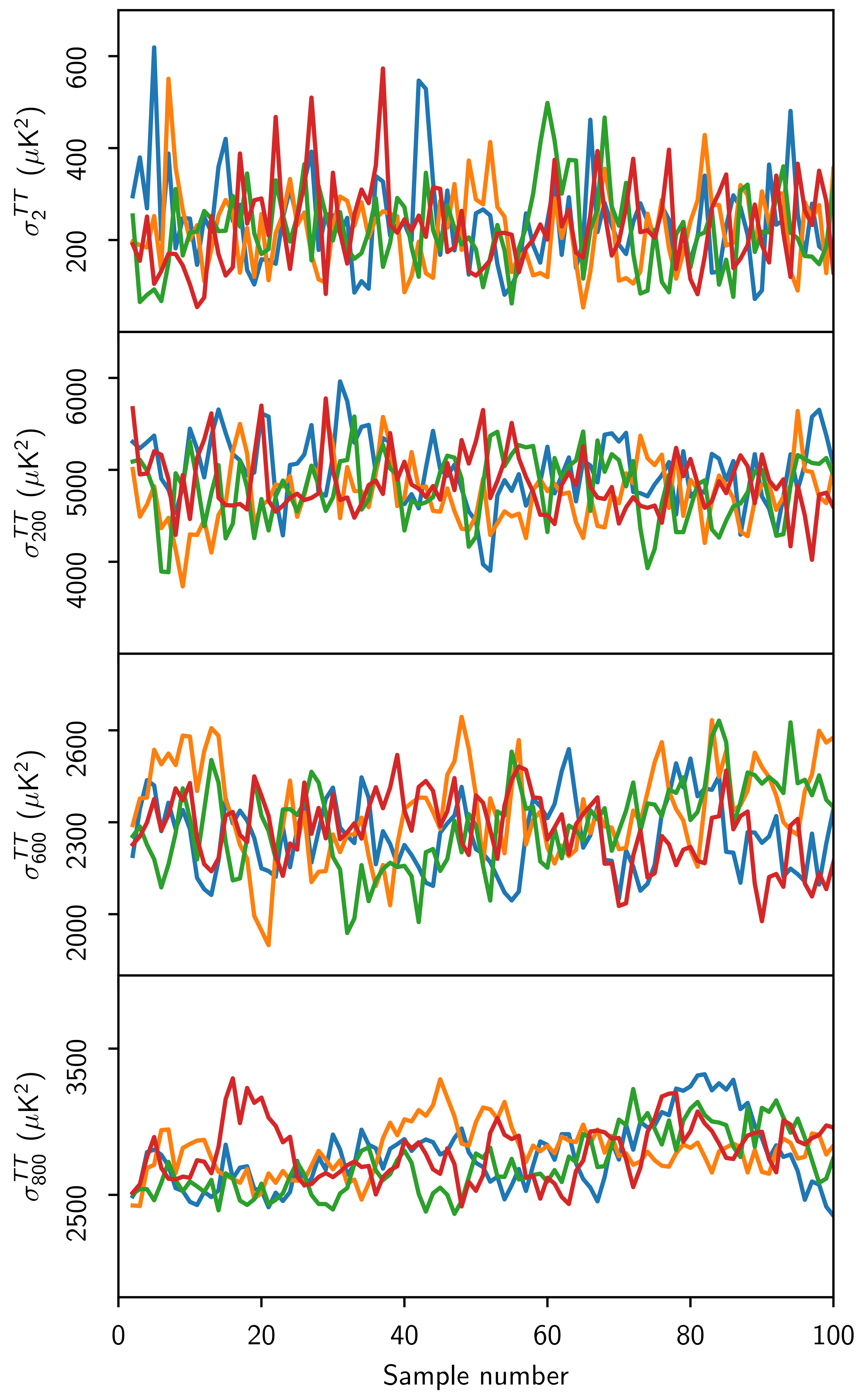
When to stop sampling?

First approach: by eye.

Correlation length can be very helpful when it comes to assess quickly the convergency status of chains **individually**.

- A short correlation length (white noise like samples) indicate a good convergency status.

BeyondPlanck I - The BeyondPlanck collaboration. 2023



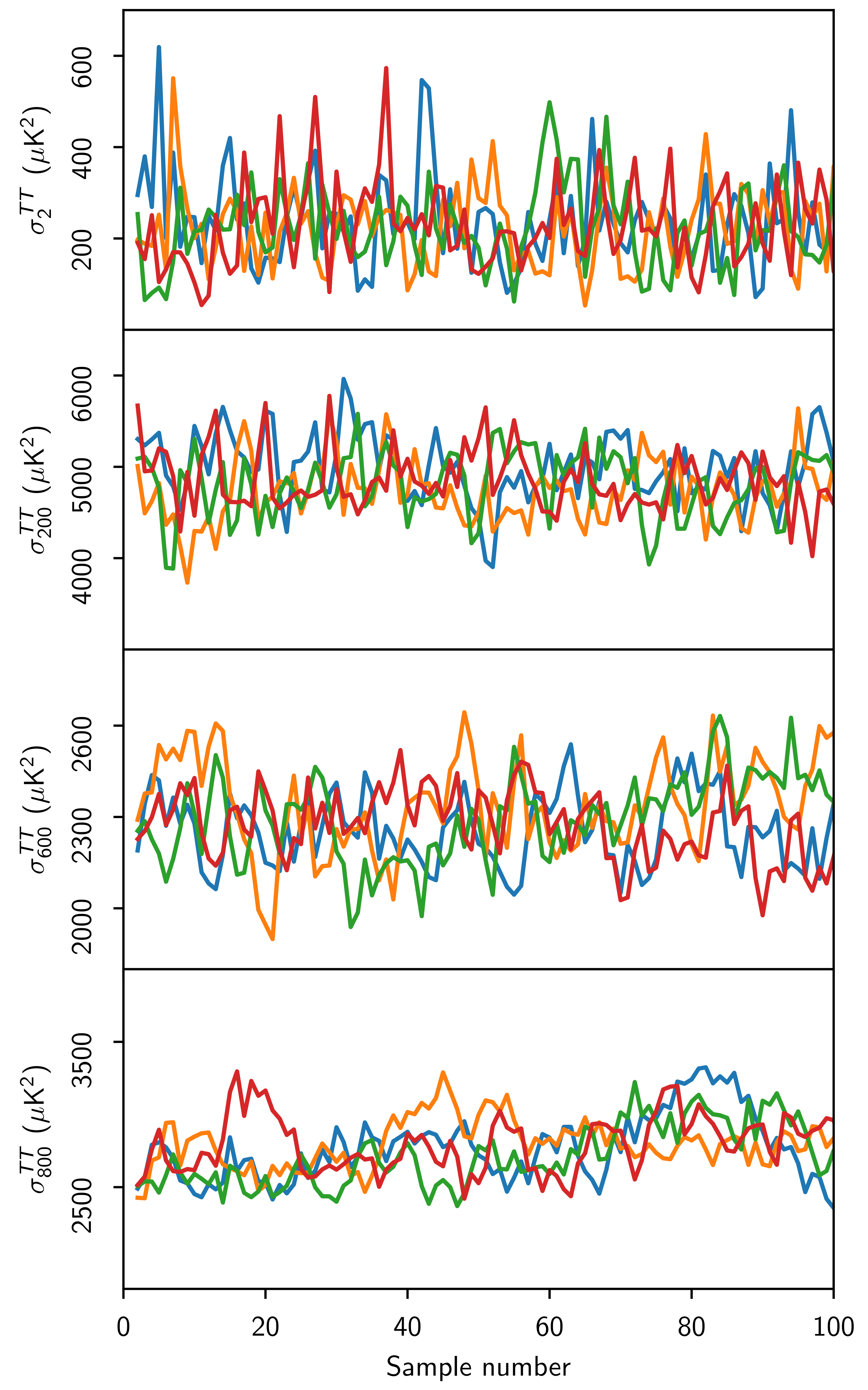
When to stop sampling?

First approach: by eye.

Correlation length can be very helpful when it comes to assess quickly the convergency status of chains **individually**.

- A short correlation length (white noise like samples) indicate a good convergency status.
- A long correlation length, on the other hand, may indicate a poor convergency.

BeyondPlanck I - The BeyondPlanck collaboration. 2023



When to stop sampling?

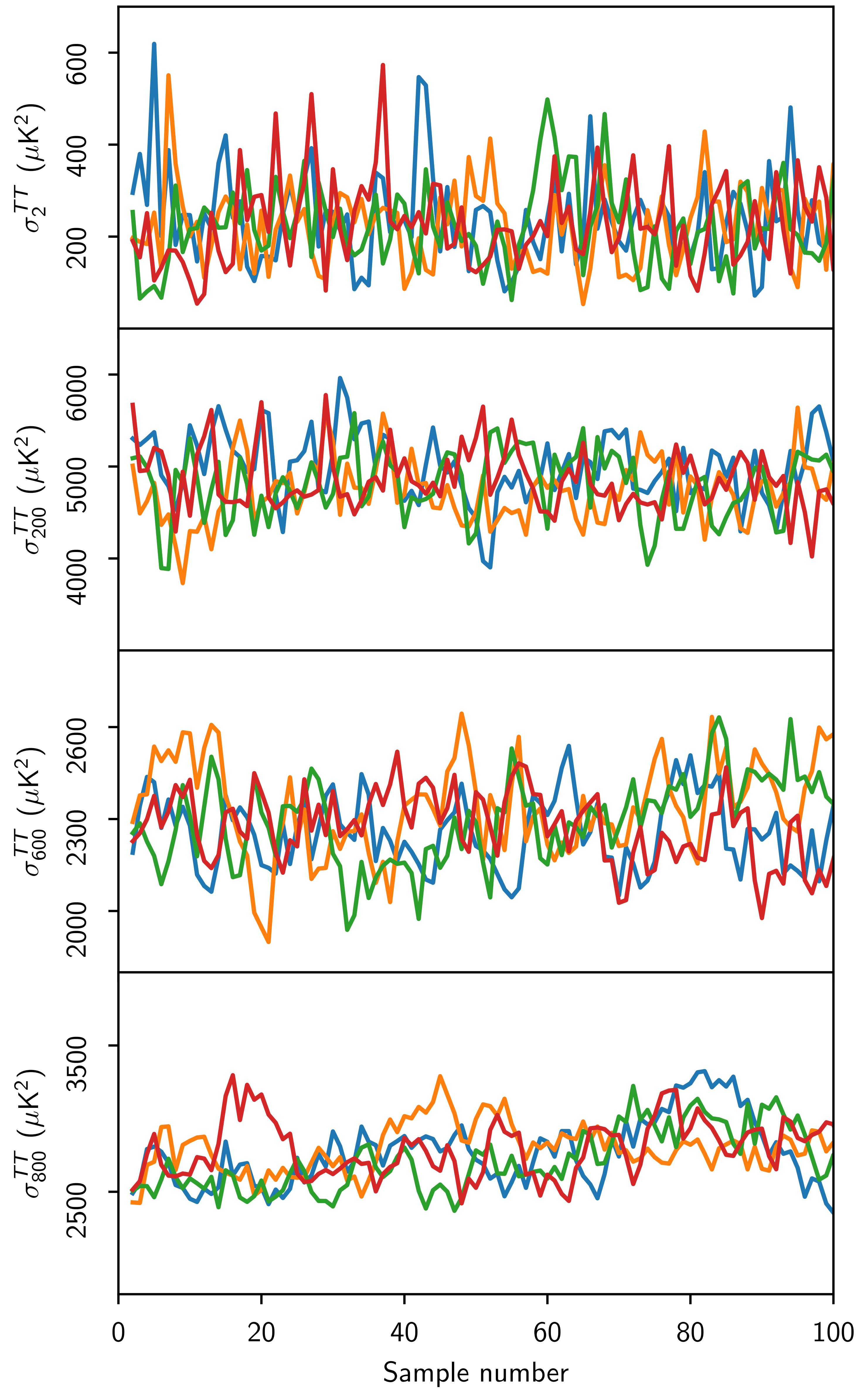
First approach: by eye.

Correlation length can be very helpful when it comes to assess quickly the convergency status of chains **individually**.

- A short correlation length (white noise like samples) indicate a good convergency status.
- A long correlation length, on the other hand, may indicate a poor convergency.

However, parameters can be highly degenerate, and correlation length can just stay long —> **need much more samples** to properly explore the parameter space. This also **depends on the sampling algorithm**.

BeyondPlanck I - The BeyondPlanck collaboration. 2023



When to stop sampling?

Gelman-Rubin statistics

We usually run our MCMC with a handful of chains in parallel.

When to stop sampling?

Gelman-Rubin statistics

We usually run our MCMC with a handful of chains in parallel.

We can exploit the independence of different chains, along with the correlation length within each chain to construct a formal estimator for the MCMC run convergency.

When to stop sampling?

Gelman-Rubin statistics

We usually run our MCMC with a handful of chains in parallel.

We can exploit the independence of different chains, along with the correlation length within each chain to construct a formal estimator for the MCMC run convergency.

Heuristically we want to compare the sampled parameters' variance among different chains and the variance within the chains:

When to stop sampling?

Gelman-Rubin statistics

We usually run our MCMC with a handful of chains in parallel.

We can exploit the independence of different chains, along with the correlation length within each chain to construct a formal estimator for the MCMC run convergency.

Heuristically we want to compare the sampled parameters' variance among different chains and the variance within the chains:

- both these quantities **converge to the true parameter's posterior variance**, but

When to stop sampling?

Gelman-Rubin statistics

We usually run our MCMC with a handful of chains in parallel.

We can exploit the independence of different chains, along with the correlation length within each chain to construct a formal estimator for the MCMC run convergency.

Heuristically we want to compare the sampled parameters' variance among different chains and the variance within the chains:

- both these quantities **converge to the true parameter's posterior variance**, but
- the **between chain variance** is initially an over-estimate (over-dispersed initial points);

When to stop sampling?

Gelman-Rubin statistics

We usually run our MCMC with a handful of chains in parallel.

We can exploit the independence of different chains, along with the correlation length within each chain to construct a formal estimator for the MCMC run convergency.

Heuristically we want to compare the sampled parameters' variance among different chains and the variance within the chains:

- both these quantities **converge to the true parameter's posterior variance**, but
- the **between chain variance** is initially an over-estimate (over-dispersed initial points);
- the **within chain variance** is initially an under-estimate (long correlation length with few samples).

When to stop sampling?

Gelman-Rubin statistics

We usually run our MCMC with a handful of chains in parallel.

We can exploit the independence of different chains, along with the correlation length within each chain to construct a formal estimator for the MCMC run convergency.

Heuristically we want to compare the sampled parameters' variance among different chains and the variance within the chains:

- both these quantities **converge to the true parameter's posterior variance**, but
- the **between chain variance** is initially an over-estimate (over-dispersed initial points);
- the **within chain variance** is initially an under-estimate (long correlation length with few samples).

When to stop sampling?

Gelman-Rubin statistics

This is usually referred as the Gelman-Rubin statistics for J chains of length L:

$$\bar{x}_j = \frac{1}{L}\sum_{t=1}^{L} x_t^{(j)} \qquad \text{(chain mean)}$$

$$\bar{x}_\cdot = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_j \qquad \text{(grand mean)}$$

$$B = \frac{L}{J-1}\sum_{j=1}^{J} (\bar{x}_j - \bar{x}_\cdot)^2 \qquad \text{(between chain variance)}$$

$$s_j^2 = \frac{1}{L-1}\sum_{t=1}^{L} (x_t^{(j)} - \bar{x}_j)^2 \qquad \text{(within chain variance)}$$

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2$$

$$R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$$

When to stop sampling?

Gelman-Rubin statistics

This is usually referred as the Gelman-Rubin statistics for J chains of length L:

$$\bar{x}_j = \frac{1}{L} \sum_{t=1}^{L} x_t^{(j)} \qquad \text{(chain mean)}$$

$$\bar{x}_\cdot = \frac{1}{J} \sum_{j=1}^{J} \bar{x}_j \qquad \text{(grand mean)}$$

$$B = \frac{L}{J-1} \sum_{j=1}^{J} (\bar{x}_j - \bar{x}_\cdot)^2 \qquad \text{(between chain variance)}$$

$$s_j^2 = \frac{1}{L-1} \sum_{t=1}^{L} (x_t^{(j)} - \bar{x}_j)^2 \qquad \text{(within chain variance)}$$

$$W = \frac{1}{J} \sum_{j=1}^{J} s_j^2$$

$$R = \frac{\frac{L-1}{L} W + \frac{1}{L} B}{W}$$

When to stop sampling?

Gelman-Rubin statistics

This is usually referred as the Gelman-Rubin statistics for J chains of length L:

$$\bar{x}_j = \frac{1}{L}\sum_{t=1}^{L} x_t^{(j)} \qquad \text{(chain mean)}$$

$$\bar{x}_\cdot = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_j \qquad \text{(grand mean)}$$

$$B = \frac{L}{J-1}\sum_{j=1}^{J} (\bar{x}_j - \bar{x}_\cdot)^2 \qquad \text{(between chain variance)}$$

$$s_j^2 = \frac{1}{L-1}\sum_{t=1}^{L} (x_t^{(j)} - \bar{x}_j)^2 \qquad \text{(within chain variance)}$$

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2$$

$$R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$$

When to stop sampling?

Gelman-Rubin statistics

This is usually referred as the Gelman-Rubin statistics for J chains of length L:

$$\bar{x}_j = \frac{1}{L}\sum_{t=1}^{L} x_t^{(j)} \qquad \text{(chain mean)}$$

$$\bar{x}_\cdot = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_j \qquad \text{(grand mean)}$$

$$B = \frac{L}{J-1}\sum_{j=1}^{J} (\bar{x}_j - \bar{x}_\cdot)^2 \qquad \text{(between chain variance)}$$

$$s_j^2 = \frac{1}{L-1}\sum_{t=1}^{L} (x_t^{(j)} - \bar{x}_j)^2 \qquad \text{(within chain variance)}$$

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2$$

$$R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$$

WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

When to stop sampling?

Gelman-Rubin statistics

This is usually referred as the Gelman-Rubin statistics for J chains of length L:

$$\bar{x}_j = \frac{1}{L}\sum_{t=1}^{L} x_t^{(j)} \qquad \text{(chain mean)}$$

$$\bar{x}_\cdot = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_j \qquad \text{(grand mean)}$$

$$B = \frac{L}{J-1}\sum_{j=1}^{J} (\bar{x}_j - \bar{x}_\cdot)^2 \qquad \text{(between chain variance)}$$

$$s_j^2 = \frac{1}{L-1}\sum_{t=1}^{L} (x_t^{(j)} - \bar{x}_j)^2 \qquad \text{(within chain variance)}$$

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2$$

$$R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$$

WATERLOO CENTRE FOR ASTROPHYSICS UNIVERSITY OF WATERLOO

When to stop sampling?

Gelman-Rubin statistics

This is usually referred as the Gelman-Rubin statistics for J chains of length L:

$$\bar{x}_j = \frac{1}{L}\sum_{t=1}^{L} x_t^{(j)} \qquad \text{(chain mean)}$$

$$\bar{x}_\cdot = \frac{1}{J}\sum_{j=1}^{J} \bar{x}_j \qquad \text{(grand mean)}$$

$$B = \frac{L}{J-1}\sum_{j=1}^{J} (\bar{x}_j - \bar{x}_\cdot)^2 \qquad \text{(between chain variance)}$$

$$s_j^2 =$$

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2$$

$$R = \frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$$

The GR diagnostic suggests that a good convergency is achieved for $R < 1.1$ or equivalently $R - 1 < 0.1$