Bayesian Statistics

Fabio Sigrist

ETH Zurich, Autumn Semester 2019

Today's topics

- Rejection sampling
- Importance sampling
- Basics of Markov chain Monte Carlo

Goal of simulation methods

- Assume that $X \sim \pi$ where π is the **target distribution**. Our goal is to approximate

$$\mu_h = \mathbb{E}_\pi(h(X)) = \int h(x)\pi(x)dx$$

- Sample $X^1, \ldots X^N$ from π and use the following approximation

$$\mu_h \approx \bar{h}_N := \frac{1}{N}\sum_{t=1}^{N} h(X^t)$$

Rejection sampling

Rejection sampling

Key idea

1. Simulate with a different distribution τ (called the **proposal**)
2. Correct to obtain a sample from the **target** π

Assume that $\pi(x) \leq M\tau(x) < \infty$ for all x and thus*

$$a(x) := \frac{\pi(x)}{M\tau(x)} \leq 1$$

- $a(\cdot)$ is called the **acceptance function**
- $M\tau(x)$ is called the **envelope**

*We assume that densities exist and use the same symbols to denote densities of the distributions

Rejection sampling

Algorithm

1. *Generate* (Y, U) *independent with* $Y \sim \tau$ *and* $U \sim Uniform(0, 1)$
2. *If* $U \leq a(Y)$*, set* $X = Y$*, otherwise go back to step 1*

The X generated by this algorithm has the correct distribution:

$$\mathbb{P}(X \in A) \;=\; \mathbb{P}(Y \in A \mid U \leq a(Y)) \;=\; \int_A \pi(x)dx$$

for all measurable sets A

Proof: see blackboard

Clicker question

Illustration of rejection sampling

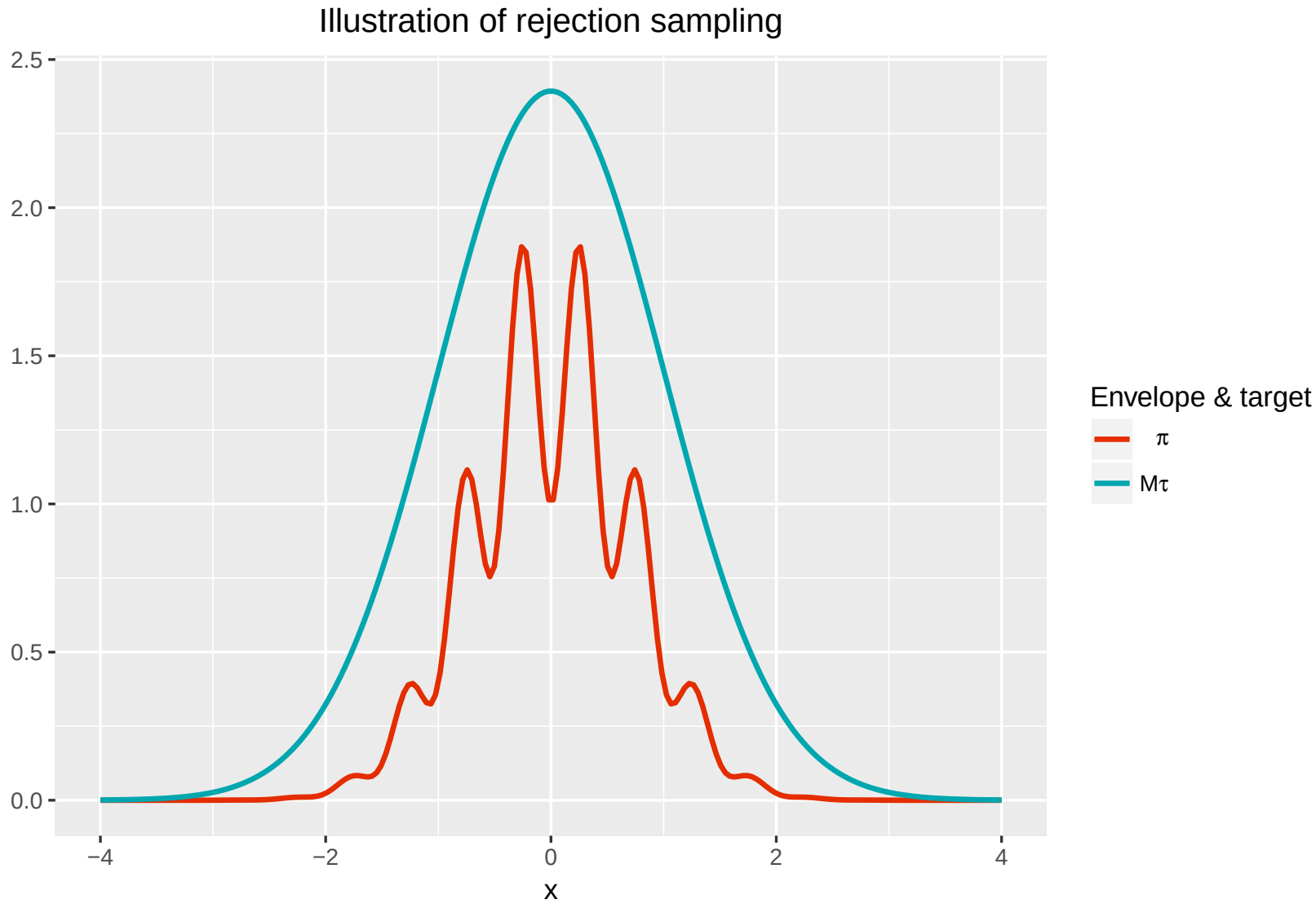
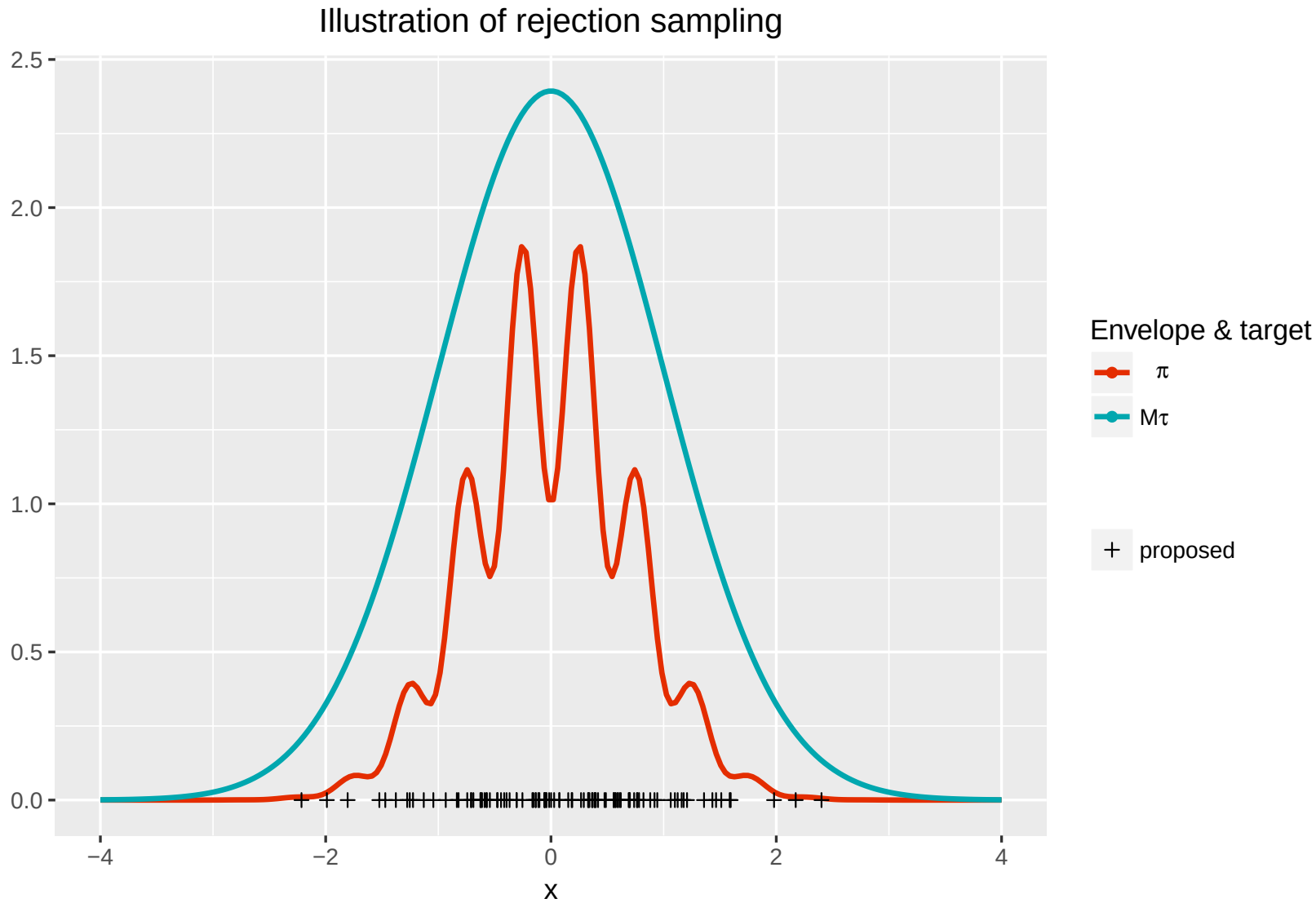
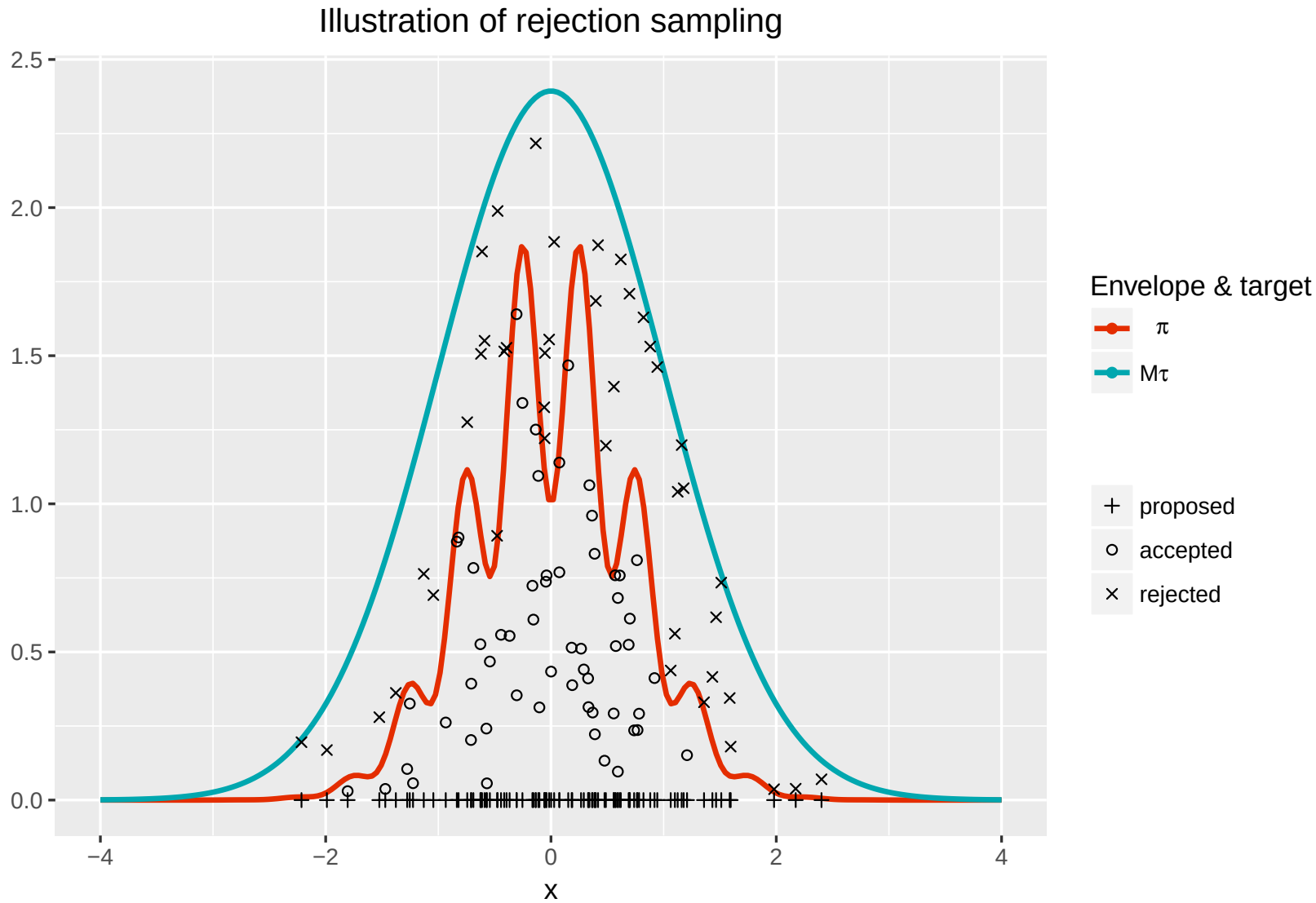


Illustration of rejection sampling



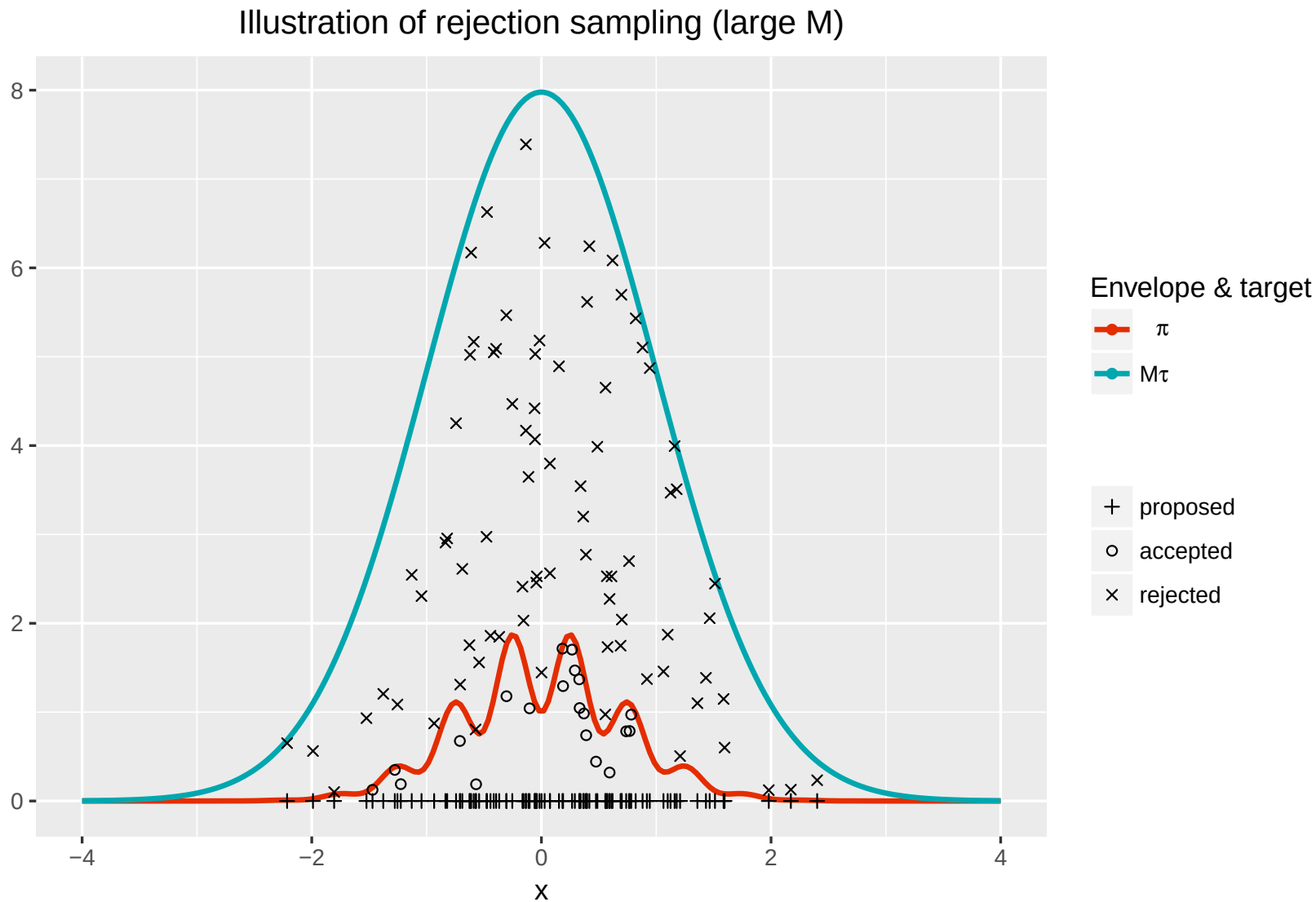
Y accepted if: $U \leq \frac{\pi(Y)}{M\tau(Y)} \Leftrightarrow M\tau(Y) \cdot U \leq \pi(Y)$

Illustration of rejection sampling



Y accepted if: $U \leq \frac{\pi(Y)}{M\tau(Y)} \Leftrightarrow M\tau(Y) \cdot U \leq \pi(Y)$

Illustration of rejection sampling



- *M* controls the number of rejected values

Comments on rejection sampling

- It is sufficient to know π up to a normalizing constant
- The expected number of pairs that need to be sampled until the first Y is accepted is M

  The expected number of rejections is often large unless τ is reasonably close to π
- In high dimensions, it is often difficult to find a proposal which is close to the the target and from which we can simulate

Importance sampling

Importance sampling

- The goal is to calculate

$$\mathbb{E}_\pi(h(X)) = \int h(x)\pi(x)dx$$

- Importance sampling is based on a similar idea as rejection sampling. Instead of rejecting some variables, we weight them with an appropriate weighting function

Importance sampling

If Y^t i.i.d. $\sim \tau$, then

$$\frac{1}{N}\sum_{t=1}^{N} h(Y^t)w(Y^t) \quad \text{where} \quad w(x) = \frac{\pi(x)}{\tau(x)}$$

is an unbiased estimator for $\mathbb{E}_\pi(h(X))$

- Requirement: $h(x)\pi(x) > 0 \Rightarrow \tau(x) > 0$

Derivation: see blackboard

Illustration of importance sampling

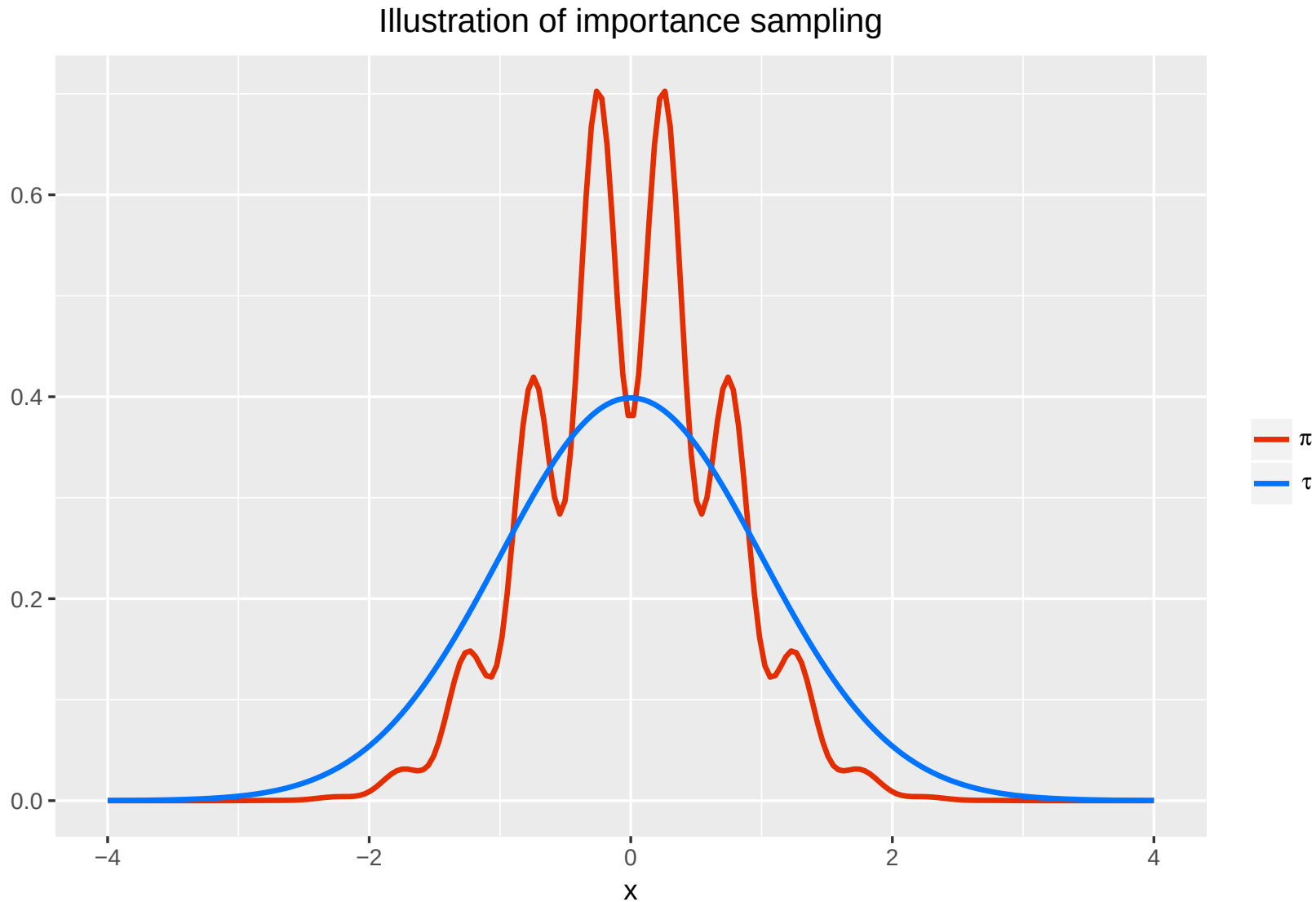


Illustration of importance sampling

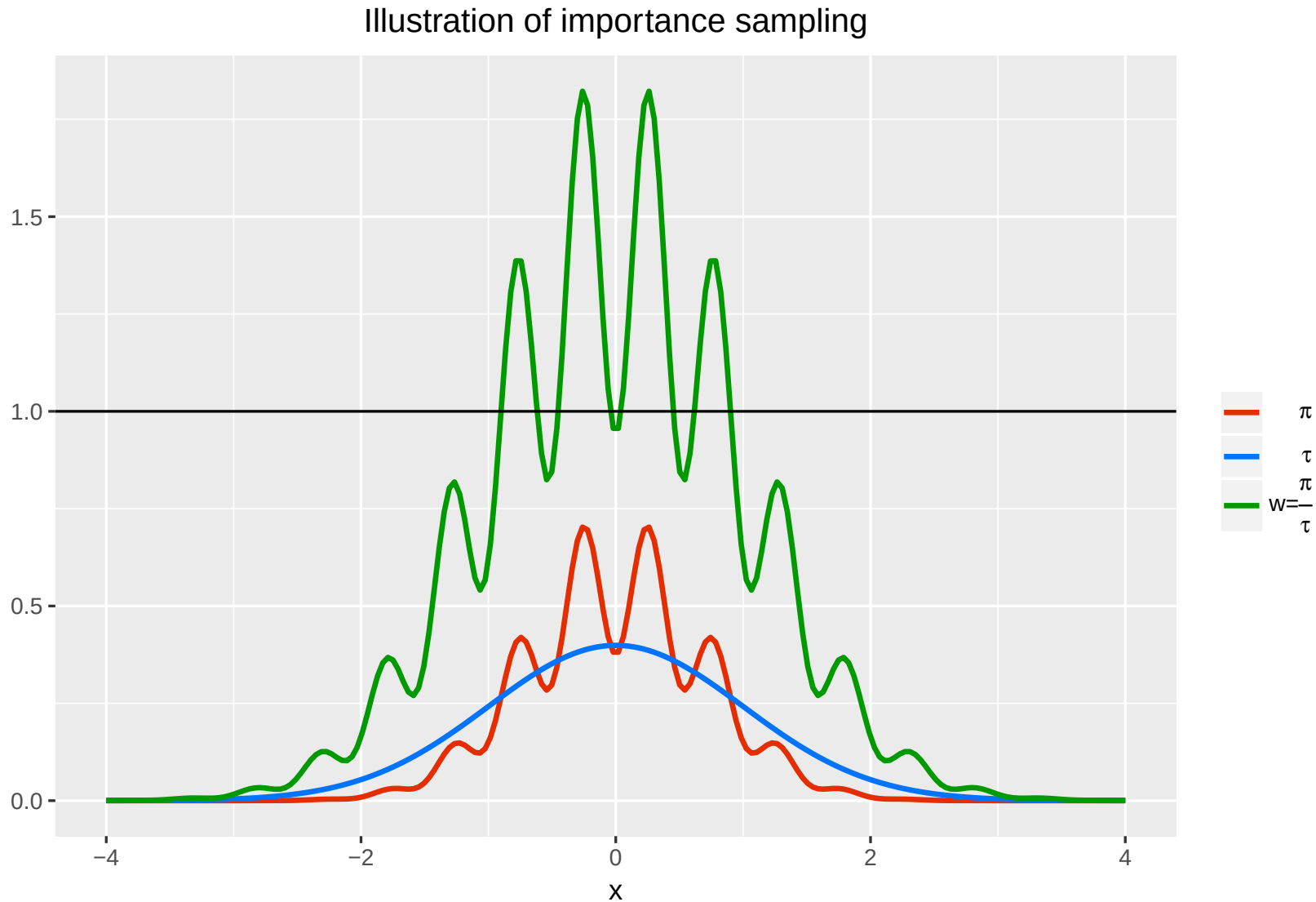


Illustration of importance sampling

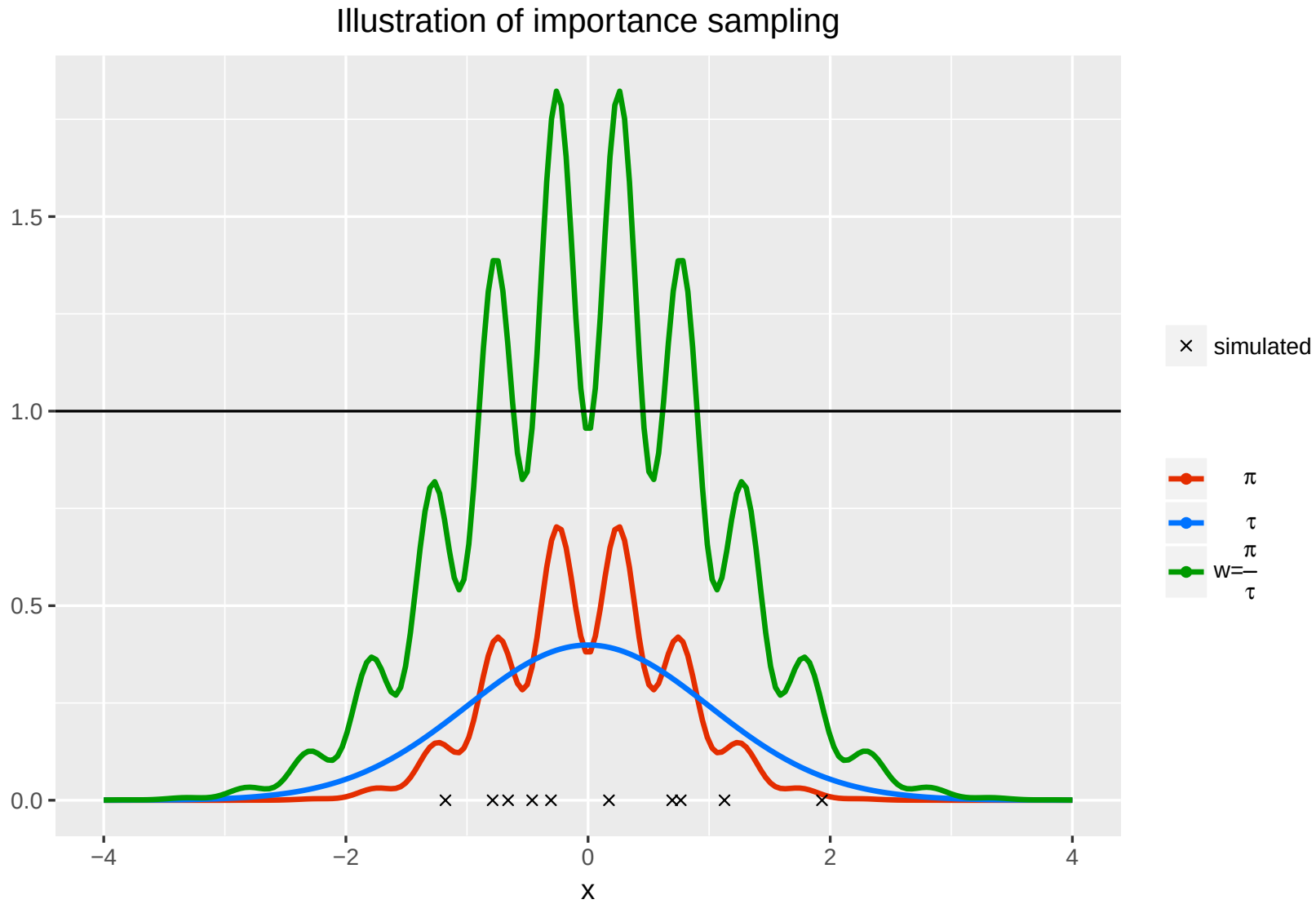
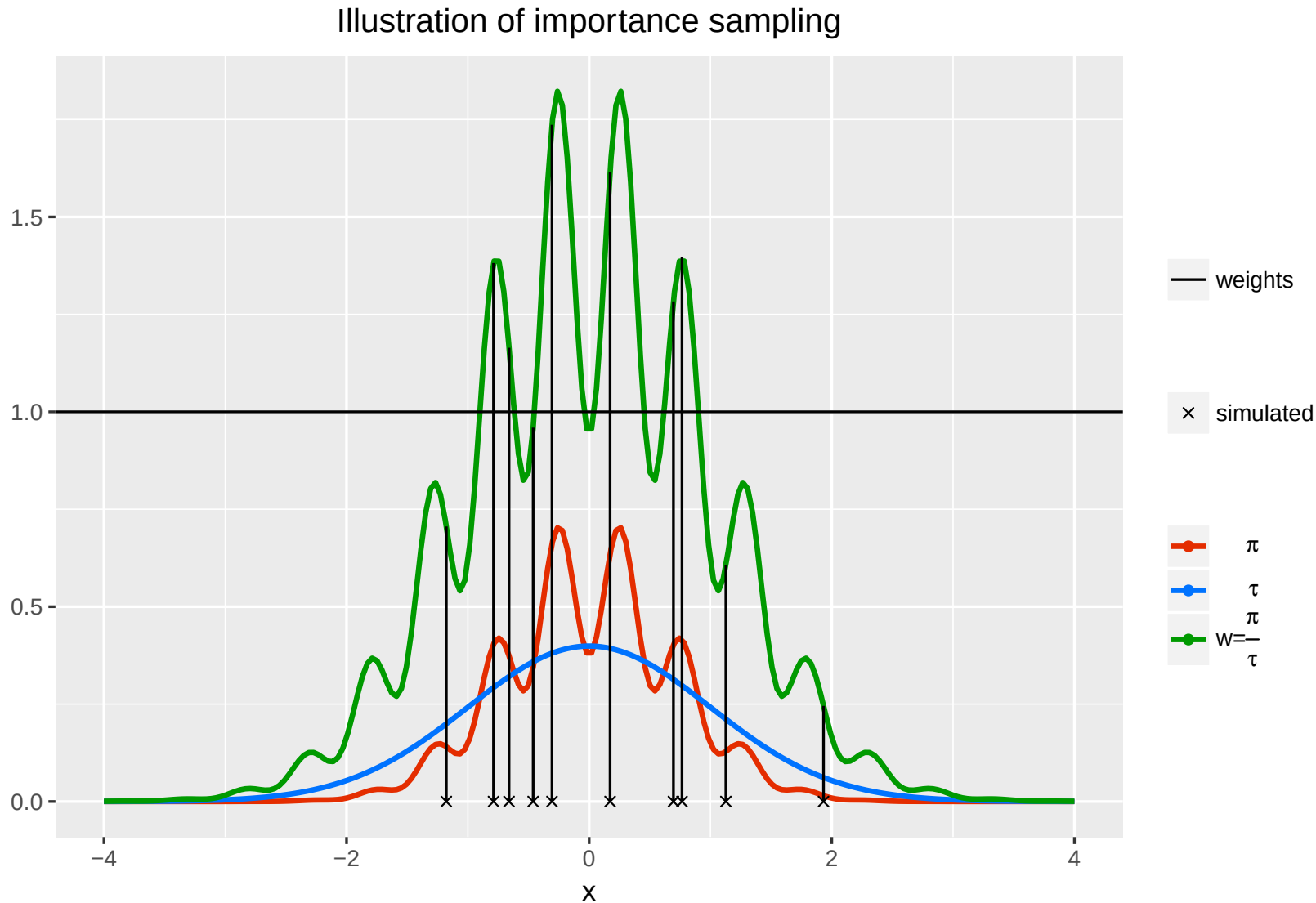


Illustration of importance sampling



Importance sampling for unnormalized densities

If the **normalizing constant for π is not known**, one can use an alternative version of importance sampling:

$$\frac{\frac{1}{N}\sum_{t=1}^{N} h(Y^t)w(Y^t)}{\frac{1}{N}\sum_{t=1}^{N} w(Y^t)} \quad \text{where} \quad w(x) \propto \frac{\pi(x)}{\tau(x)}$$

- This estimator is biased but consistent

Importance sampling

Comments

- No upper bound is needed for the ratio $w = \pi/\tau$
- It is desirable that the estimator has finite variance. This leads to conditions for π and τ
- In order to avoid very large variances, τ must also be similar to π. I.e., the normalized weights $w(Y^t)/\sum_s w(Y^s)$ should not be too far from uniform

  This is difficult in high dimensions (most high dim. distributions tend to differ greatly) $\Rightarrow$ use of importance sampling in high dimensions is limited

Sampling Importance Resampling (SIR)

If we want an **unweighted sample** instead of a weighted one, we can use resampling:

- Fenerate an additional sample (I^t) which takes values in $\{1, 2, \ldots, N\}$ with probabilities proportional to the weights ($w(Y^s)$):

$$\mathbb{P}(I^t = s) = \frac{w(Y^s)}{\sum_{r=1}^{N} w(Y^r)}$$

- Set

$$Z^t = Y^{I^t}$$

See blackboard for justification

Clicker question

Markov chain Monte Carlo

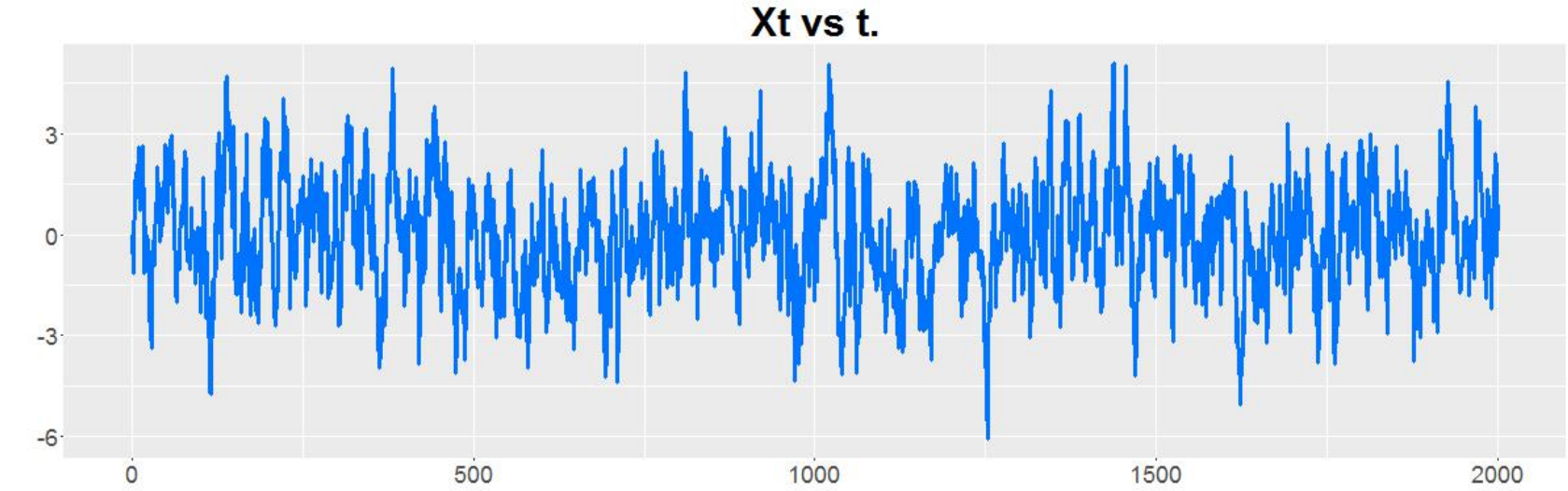
Motivation

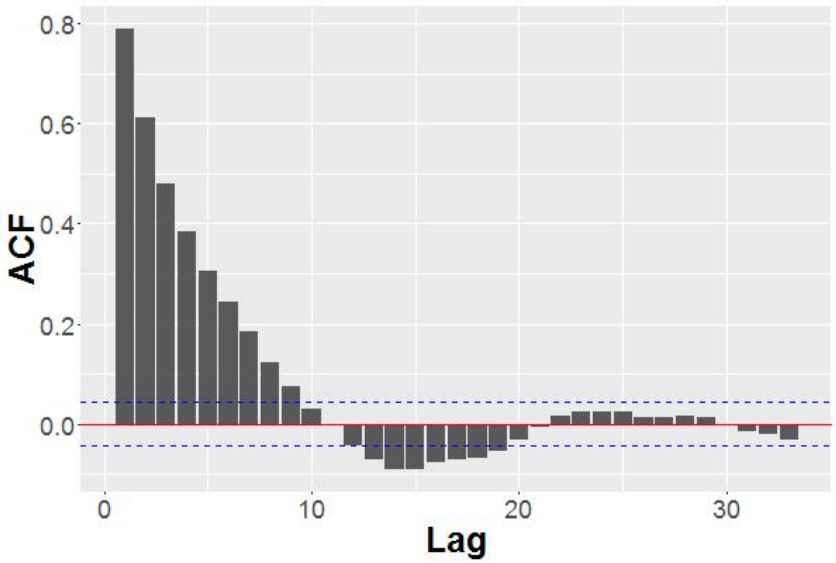
- In many cases, especially in high dimensions, there are no good methods to generate i.i.d. samples from a general target distribution π
  - The rejection algorithm fails because it almost always rejects (the bound for the ratio of the densities is too large)
  - Importance sampling fails because the variance of the weights is too large
- The current standard method for the simulation of distributions in high dimensions is called **Markov chain Monte Carlo (MCMC)**

Idea of Markov chain Monte Carlo

Basic idea: Instead of generating independent samples $X^t \sim \pi$, we generate **dependent samples** X^t such that for large t, X^t has (approximately) the correct distribution π

Illustration of MCMC





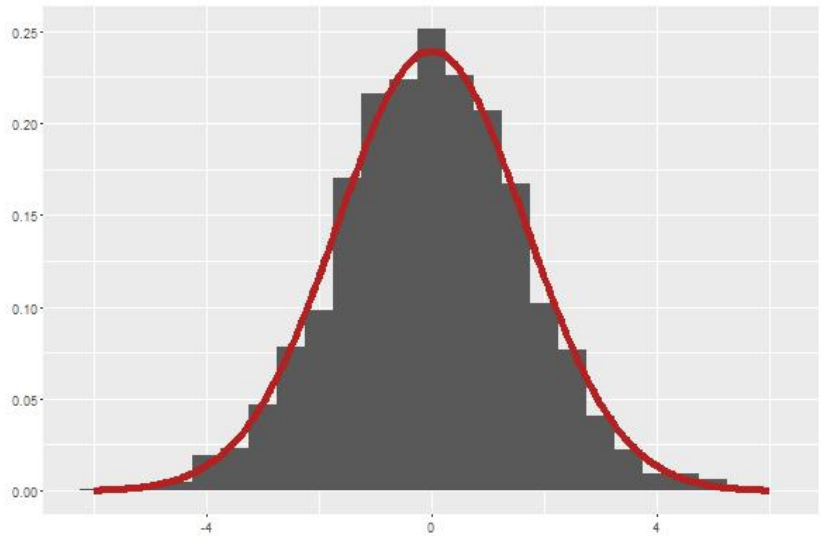
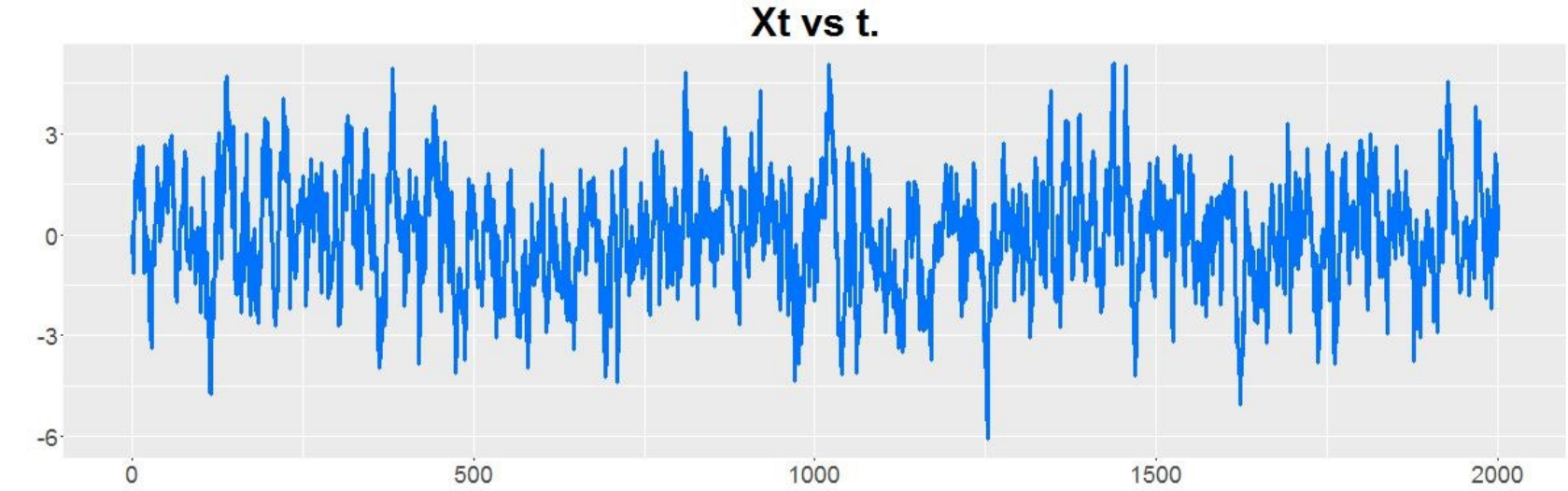
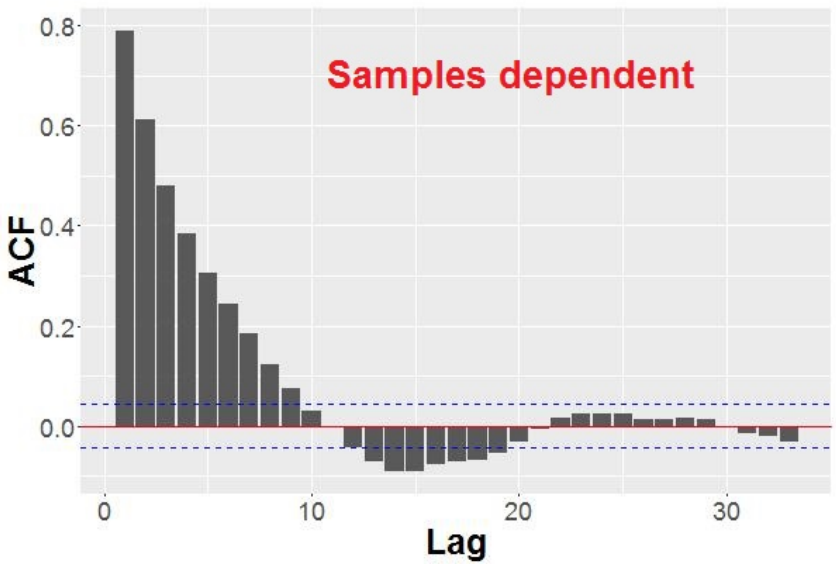
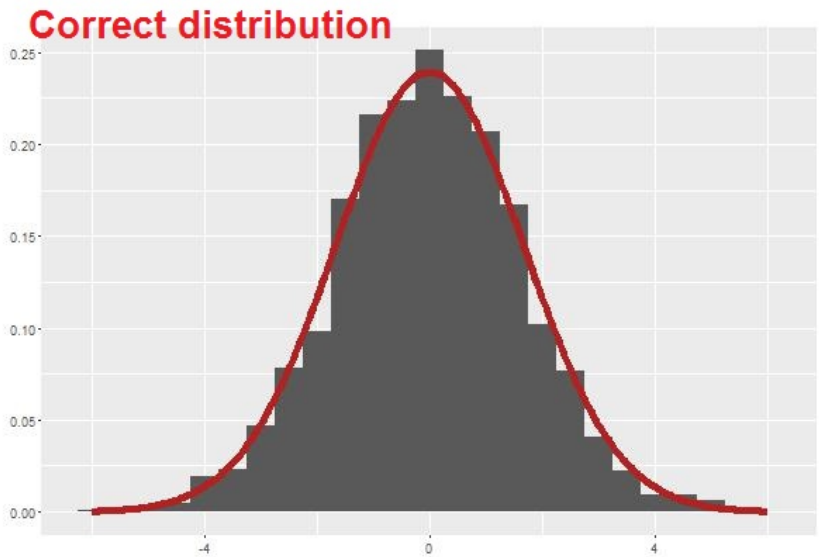


Illustration of MCMC







Basics of Markov chain Monte Carlo

Basics of Markov chain Monte Carlo

1. Start with an arbitrary initial value X^0 and generate a sequence $(X^0, X^1, \ldots)$ **recursively**:

$$X^t = G(X^{t-1}, U^t)$$

   where U^t is a vector of uniform random variables that are independent of $X^0, \ldots, X^{t-1}$
2. Use the **approximation**

$$\int h(x)\pi(x)dx \approx \bar{h}_{N,r} = \frac{1}{N-r}\sum_{t=r+1}^{N} h(X^t)$$

- The first r simulations are discarded until we reach the target π. r is the so-called **burn-in period**
- In contrast to previous methods, the $\boldsymbol{X^t}$**s are not independent**

Basics of Markov chain Monte Carlo

- The random variables $(X^0, X^1, \ldots)$ form a **Markov chain**:
  - X^t depends on X^{t-1} and new (uniform) random variables U_t but not on previous values X^s with $s < t - 1$

- The conditional distribution of X^t given X^{t-1} is called the **transition kernel *P*** of the chain

$$\mathbb{P}(X^t \in A \mid X^0, \ldots, X^{t-1}) = \mathbb{P}(X^t \in A \mid X^{t-1}) = P(X^{t-1}, A)$$

It is determined by the function G through

$$P(x, A) = \mathbb{P}(G(x, U) \in A) = \mathbb{P}(U \in \{u; G(x, u) \in A\})$$

How to specify the transition rule G

How can we specify a transition rule

$$X^t = G(X^{t-1}, U^t)$$

for the Markov chain such that the arithmetic mean $\bar{h}_{N,r} = \frac{1}{N-r} \sum_{t=r+1}^{N} h(X^t)$ converges to $\int h(x)\pi(x)dx$?

The general theory of Markov chains shows that this holds in a wide range of cases if

1. The chain can reach all sets A with $\pi(A) > 0$
2. $X^{t-1} \sim \pi$ implies that $X^t \sim \pi$

Invariance

- We call a distribution π **invariant** or **stationary** for the transition kernel P if $X^{t-1} \sim \pi$ implies that $X^t \sim \pi$. I.e., if

$$\pi(A) = \int \pi(x) P(x, A) dx \quad \forall A$$

- If $P(x, .)$ has the density $p(x, y)$, this equals

$$\pi(y) = \int \pi(x) p(x, y) dx$$

How to specify the transition rule G

There are two widely used recipes for constructing a transition kernel P which has a given target distribution π as stationary distribution:

- The Gibbs sampler
- The Metropolis-Hastings algorithm