Probabilistic Programming

Lecture #4: Markov Chain Monte Carlo

Joost-Pieter Katoen

RWTH Lecture Series on Probabilistic Programming 2022-23

Overview

Overview

Bayes' rule

Sum rule: $Pr(X) = Pr(X, Y) + Pr(X, \neg Y)$

Product rule: $Pr(X, Y) = Pr(X) \cdot Pr(Y \mid X)$

Corollary of the sum and product rule:

$$\underbrace{Pr(X \mid D)}_{\substack{\text{posterior of } X \\ \text{given } D}} = \frac{\overbrace{Pr(D \mid X)}^{\substack{\text{likelihood of } X \\ \text{under } D}} \cdot \overbrace{Pr(X)}^{\text{prior of } X}}{\underbrace{Pr(D)}_{\text{evidence}}}$$

Bayes' rule

$$Pr(X \mid D) = \frac{Pr(D \mid X) \cdot Pr(X)}{Pr(D)}$$

In the discrete setting, this amounts to:

$$Pr(X \mid D) = \frac{Pr(D \mid X) \cdot Pr(X)}{\sum_i Pr(D \mid X_i) \cdot Pr(X_i)}$$

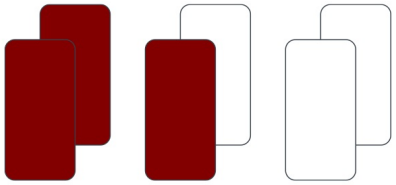
In the continuous setting, this amounts to:

$$Pr(X \mid D) = \frac{Pr(D \mid X) \cdot Pr(X)}{\int Pr(D \mid X) \cdot Pr(X)\, dX}$$

The normalising constant, i.e., the marginal likelihood $Pr(D)$ is important to get an accurate posterior distribution, yet mostly difficult to get.

Example (1)

Pick a card. What is $Pr\{\text{back is red} \mid \text{front is red}\}$?



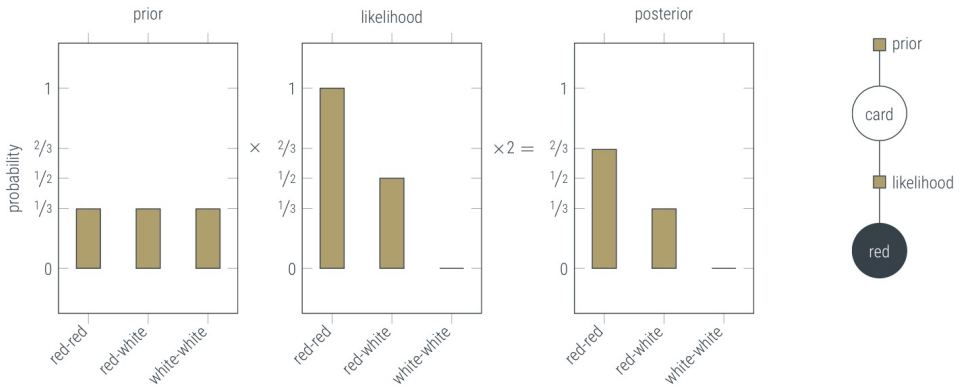
- $\frac{1}{2}$
- $\frac{2}{3}$
- something else
- I don't know yet

$$P(\text{card}|\text{color}) = \frac{P(\text{card}) \cdot P(\text{color} \mid \text{card})}{P(\text{color})} = \frac{P(\text{card}) \cdot P(\text{color} \mid \text{card})}{\sum_{i=1}^{3} P(\text{color} \mid \text{card} = i) P(\text{card} = i)}$$
$$= \frac{1/3 \cdot P(\text{color} \mid \text{card})}{1/2} = \frac{2}{3} \cdot \{1 \quad 1/2 \quad 0\}$$

(Credits: Philipp Hennig, Univ. of Tübingen, 2020)

It is the likelihood that matters, not the prior.

Example (2)



(Credits: Philipp Hennig, Univ. of Tübingen, 2020)

Expected values

A probability distribution Pr on a countable set X is a function

$$Pr : X \to [0, 1] \quad \text{such that} \quad \sum_{x \in X} Pr(x) = 1.$$

The expected value of random variable $f : X \to \mathbb{R}$ under distribution Pr is defined by:

$$E_{Pr}(f) = \sum_{x \in X} f(x) \cdot Pr(x) = \int_X f \, dPr$$

Justification of MCMC

Computing expected values, and posterior distributions can be hard.

Possible remedy: Use Markov chain Monte Carlo sampling

Overview

Monte Carlo Casino

Monte Carlo Sampling

$$E_{Pr}(f) \;=\; \sum_{x \in X} f(x) \cdot Pr(x)$$

As the dimension of X can be huge, or even uncountably infinite, one takes i.i.d. samples $x_1, x_2, \ldots, x_N$ with $x_i \in X$ and $x_i \sim Pr(X)$.

Then:

$$E_{Pr}(f) \;\sim\; \frac{1}{N} \sum_{i=1}^{N} f(x_i)$$

Similarly:

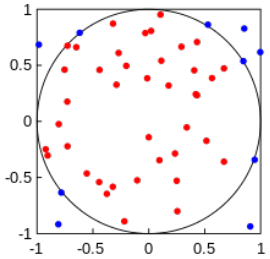
$$\underbrace{\sum_{i} Pr(D \mid A_i)}_{\text{part of marginal likelihood}} \;\sim\; \sum_{i=1}^{N} Pr(D \mid A_i)$$

Computing using samples $x_i \sim Pr(X)$ are called Monte Carlo Methods [1]

[1] due to works in the 1940's by Stanislaw Ulam and John von Neumann

Approximating π by Monte Carlo sampling

- Take a square enclosing a circle
- Sample N points within the square
- Take the ratio of # samples in circle to N
- This approximates area circle to area square



formally: $$\frac{N_{\odot}}{N_{\boxdot}} \sim \frac{Area(\bigcirc)}{Area(\square)},$$

which in turn gives approximation

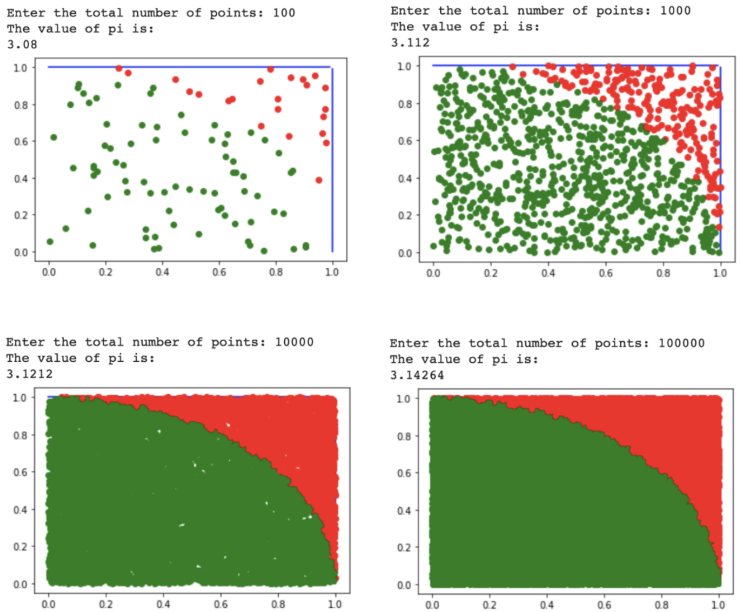
$$\pi \sim \frac{4N_{\odot}}{N_{\boxdot}}.$$

(The approximation becomes more precise as one increases N).

A WebPPL program to approximate π

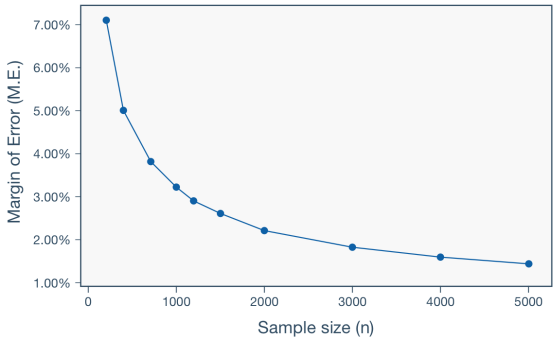
```
var N = 100

var samples_in_circle = function (m) {
if (m == 0) {return 0}
var x = sample(Uniform({a:-1,b:1}))
var y = sample(Uniform({a:-1,b:1}))
var hits = samples_in_circle(m-1)
return (x*x + y*y <= 1) ? 1 + hits : hits
   // count a sample within/on the circle
}

var approx_pi = function() {
  return (4 * samples_in_circle(N)) / N
}

var roughly_pi =
  Infer({method: 'rejection', samples: 20}, approx_pi)
viz.auto(roughly_pi)
```


Increasing the number of samples



Properties of Monte Carlo sampling

1. Yields unbiased estimators for every integrable function
2. The expected square error (variance) drops as $O(N^{-1})$
3. Thus: the expected error drops as $O\left((\sqrt{N})^{-1}\right)$



Continuous distributions

The *distribution function* F_X of random variable X is defined by:

$$F_X(d) \;=\; Pr_X((-\infty, d]) = Pr(\underbrace{\{\, a \in \Omega \mid X(a) \leqslant d \,\}}_{\{\, X \leqslant d \,\}}) \quad \text{for real } d$$

Properties:

- F_X is monotonic and right-continuous
- $0 \leqslant F_X(d) \leqslant 1$
- $\lim_{d \to -\infty} F_X(d) = 0$ and
- $\lim_{d \to \infty} F_X(d) = 1$.

For continuous random variable X, F_X can be written as:

$$F_X(d) = \int_{-\infty}^{d} f_X(u)\, du \quad \text{with } f \text{ the density function}$$

Examples

Obtaining samples from distribution F

- Use sampling from a (continuous) uniform distribution, e.g., $U(0, 1)$
  - Uniformly distributed numbers can be transformed into other distributions
  - Only work numerically efficient in a limited number of cases
- Obstacle: need to know the density $f(x)$ for all x, regions of high density
- Aim of Monte Carlo methods: construct samples from

$$g(x) \;=\; \tilde{f}(x) \;=\; \frac{f(x)}{C}$$

assuming the unnormalised density $\tilde{f}$ can be evaluated at any x

Overview

Rejection sampling

- Aim: take samples from distribution F with density function f
- Key idea: take a proposal distribution[2] G with density $g(x)$
  - such that for all x it holds:

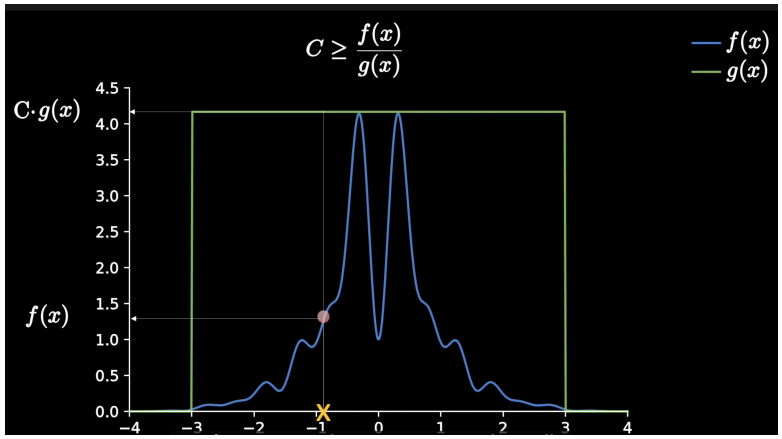
$$f(X) \leqslant C{\cdot}g(x) \quad \text{for some scaling constant} \quad 1 < C < \infty$$

- Draw sample $y \sim g(x)$ and sample $u \sim U(0, C{\cdot}g(y))$
- Accept if $u \leqslant f(y)$, reject if $u > f(y)$

In order to "discover" f, we sample from a different (proposal) function g

[2] e.g., let g be the density of a uniform or Gaussian distribution as sampling from them is easy.

Example



$f(x) = e^{-x^2/2} \cdot (\sin^2(6{+}x) + 3 \cdot \cos^2(x) \cdot \sin^4(4x) + 1)$
proposal $g(x) = U(-3, 3)$, $C = 25$

see: `https://www.youtube.com/watch?v=si76S7QqxTU`

Properties of rejection sampling

- Rejection sampling yields unbiased samples
- For distributions F over d variables, $C \in O(2^d)$
- # samples needed from G to obtain an accepted value follows a geometric distribution with probability $\frac{1}{C}$
- Thus: many samples may be needed to get an accepted value[3]
  - especially, when f is highly concentrated in some region (peeks)
  - or if $C > 1$, or f cannot be nicely "fitted" into a proposal g
- Other deficiency: for posterior distributions, we have:

$$f(x) = \frac{D(x)}{C} \quad \text{with} \quad C = \sum_x D(x) \quad \text{and } C \text{ is not known}$$

[3] We get back to this issue in the lecture on Bayesian networks

Overview

Recall: stationary distributions

Stationary distribution

A probability vector ρ satisfying $\rho = \rho \cdot \mathbf{P}$ is called a *stationary* distribution of MC D.

An irreducible, positive recurrent MC has a unique stationary distribution satisfying $\rho_\sigma = \frac{1}{m_\sigma}$ for every state σ.

m_σ is the mean recurrence time of state σ

A sufficient condition

Detailed balance condition

Let MC D with state space Σ. If for all states $\sigma, \tau \in \Sigma$ it holds

$$\underbrace{\rho(\sigma) \cdot \mathbf{P}(\sigma, \sigma') = \rho(\sigma') \cdot \mathbf{P}(\sigma', \sigma)}_{\text{detailed balance condition}},$$

then ρ is a stationary distribution of MC D.

The reverse does not hold: not every stationary distribution ρ satisfies the detailed balance condition.

Markov chains for Bayes' rule

- Idea: Let a Markov chain generate the Monte Carlo samples
- The MC is supposed to be a representative of the posterior

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}$$

- The MC may start at a distribution $\neq$ actual posterior
- But it asymptotically approaches a stationary distribution
- The MC's stationary distribution = our posterior distribution

How to determine the MC and its transition probabilities?

MCMC methods

MCMC techniques basically differ in how to define the Markov chain.

They differ in how to obtain the MCs, and may yield MCs with distinct convergence rates to the stationary distribution

Examples: Metropolis, Metropolis-Hastings, Gibbs, Hamiltonian MC ...

As opposed to rejection sampling, MCMC scales well with the dimensionality of the sample space

An analogy: optimisation

Assume our aim is to find the maximum of f

1. Given current estimate x_i
2. Draw a candidate sample from a proposal distribution: $x^* \sim g(x^* \mid x_i)$
3. Evaluate $r = \dfrac{f(x^*)}{f(x_i)}$
4. If $r \geqslant 1$, accept and let $x_{i+1} := x^*$ "go uphill"
5. Else (i.e., $r < 1$) stay: $x_{i+1} := x_i$ "stay"

The samples $x_0, x_1, x_2, \ldots$ form a Markov chain

And contain some information about the shape of f

The Metropolis-Hastings method

Assume our aim is to draw sample from distribution with density f

1. Given current estimate x_i
2. Draw a candidate sample from a proposal distribution: $x^* \sim g(x^* \mid x_i)$
3. Evaluate $r = \dfrac{f(x^*) \cdot g(x_i \mid x^*)}{f(x_i) \cdot g(x^* \mid x_i)}$
4. If $r \geqslant 1$, accept and let $x_{i+1} := x^*$ "go uphill"
5. Else (i.e., $r < 1$)
   - accept with probability r: $x_{i+1} := x^*$ "go downhill"
   - stay with probability $1-r$: $x_{i+1} := x_i$ "stay"

In the original Metropolis algorithm: $g(x_i \mid x^*) = g(x^* \mid x_i)$ (symmetry)

Acceptance condition in MH sampling

- Start derivation with the condition of detailed balance[4]:

$$\rho(x) \cdot \mathbf{P}(x, x^*) = \rho(x^*) \cdot \mathbf{P}(x^*, x)$$

- This is equivalent to

$$\frac{\mathbf{P}(x, x^*)}{\mathbf{P}(x^*, x)} = \frac{\rho(x^*)}{\rho(x)} \quad \text{that is} \quad \frac{\mathbf{P}(x^* \mid x)}{\mathbf{P}(x \mid x^*)} = \frac{\rho(x^*)}{\rho(x)}$$

- Break down moving from x to x^* into two steps:

$$\mathbf{P}(x^* \mid x) = \underbrace{g(x^* \mid x)}_{\text{proposal}} \cdot \underbrace{A(x, x^*)}_{\text{acceptance}}$$

1. pick a random value x^* based on a proposal pdf g
2. then, move from x to x^* based on an acceptance distribution A

[4] Recall this is a sufficient condition for the existence of a stationary distribution.

Metropolis-Hastings (2)

- Thus

$$\frac{\mathbf{P}(x^* \mid x)}{\mathbf{P}(x \mid x^*)} = \frac{\rho(x^*)}{\rho(x)}$$

- becomes

$$\frac{A(x, x^*)}{A(x^*, x)} = \frac{\rho(x^*) \cdot g(x \mid x^*)}{\rho(x) \cdot g(x^* \mid x)}$$

- The state x^* is the candidate for transiting to

Metropolis-Hastings (3)

- Consider the acceptance ratio r

$$r = \frac{A(x, x^*)}{A(x^*, x)} = \frac{\rho(x^*) \cdot g(x \mid x^*)}{\rho(x) \cdot g(x^* \mid x)}$$

- If $r \geqslant 1$: the move from x to x^* is a transition to a more likely state. Thus accept this move: $x_{i+1} := x^*$.

- If $r < 1$:
  - with probability r move to state x^*, i.e., $x_{i+1} := x^*$.
  - otherwise, ignore the transition and re-sample at state x, i.e., $x_{i+1} := x$.

Rationale: the MC does not get stuck in a local fragment as it always is possible to move to a less likely state.

Convergence of MH sampling

When the MC is converged to its stationary distribution, we effectively sample the posterior distribution, as desired.

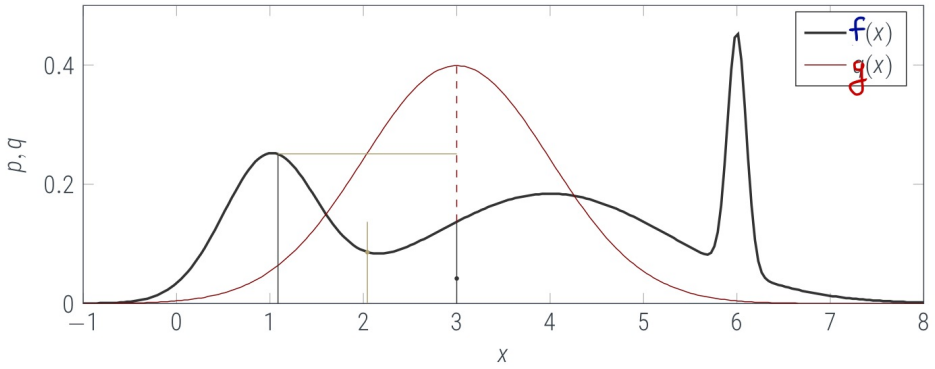
The MC induced by MH sampling is ergodic (aperiodic and positive recurrent)

Recall: ergodic MCs have a unique stationary distribution ρ

Convergence theorem (simplified)

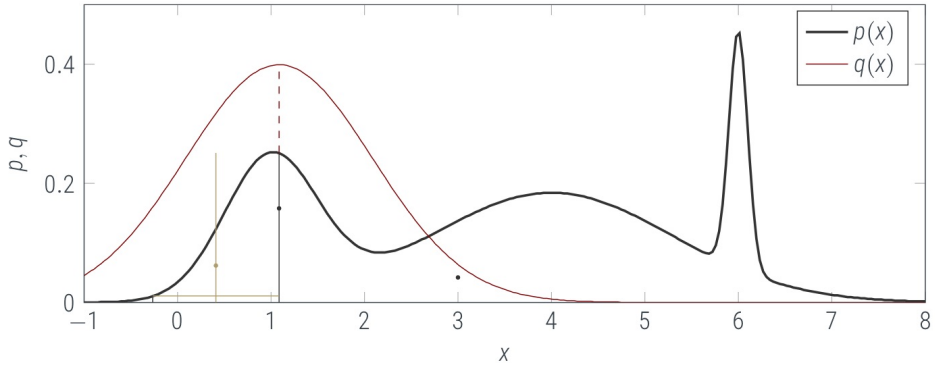
If $g(x^* \mid x_i) > 0$ for all pairs (x^*, x_i), then for any initial sample x_0, the density of $\{ x_i \}_{i \geqslant 0}$ approaches the stationary distribution $f(x)$ as $i \to \infty$.

[5] Example i=1



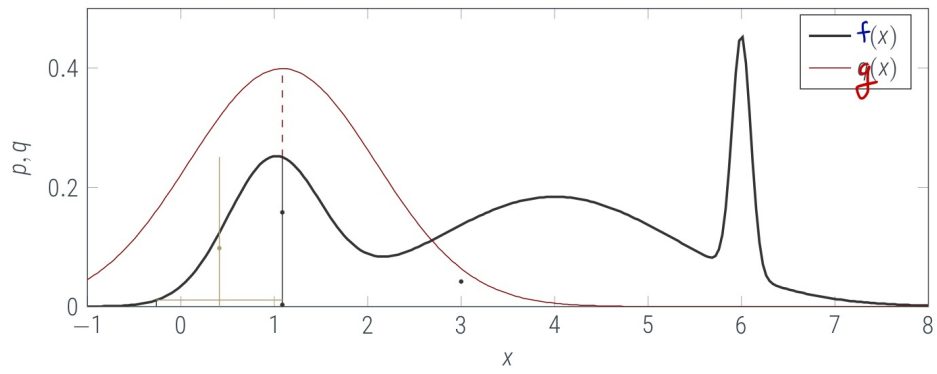
[5] courtesy: Philipp Hennig, 2020.

[6] Example i=2



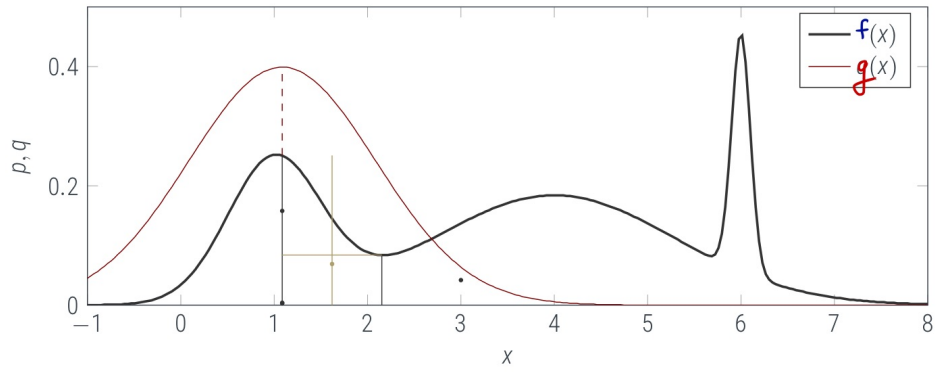
[6] courtesy: Philipp Hennig, 2020.

[7] Example i=3



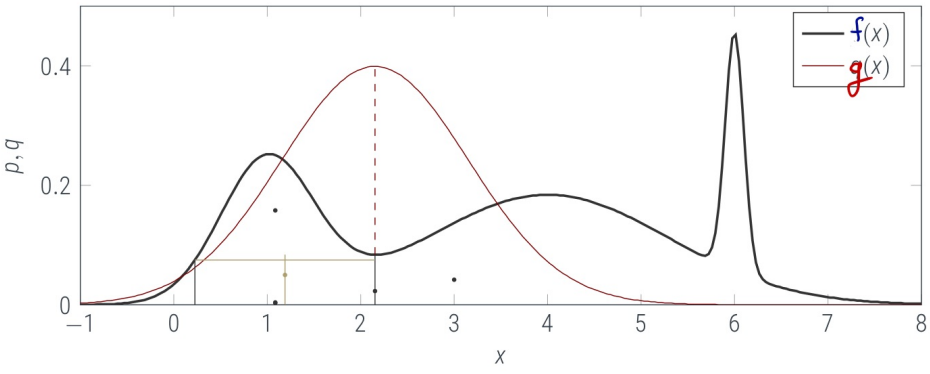
[7] courtesy: Philipp Hennig, 2020.

[8] Example i=4



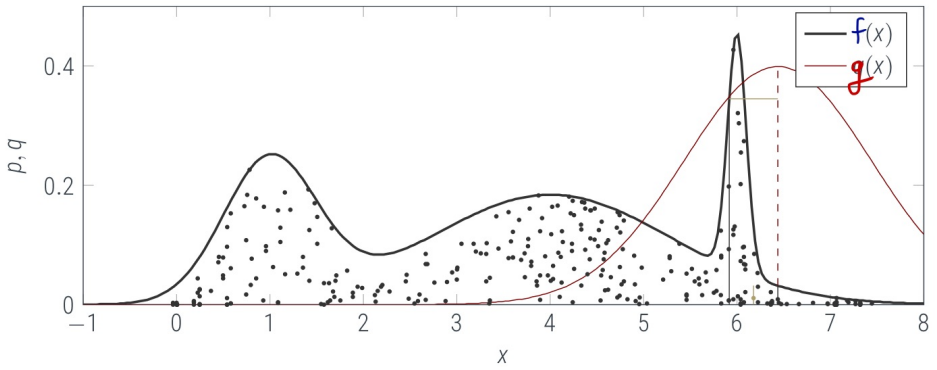
[8] courtesy: Philipp Hennig, 2020.

[9] Example i=5



[9] courtesy: Philipp Hennig, 2020.

[10] Example i=300



[10] courtesy: Philipp Hennig, 2020.

Example

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

Summary

- Bayes rule: posterior $\propto$ likelihood $\cdot$ prior
- Monte Carlo sampling is an effective technique for approximating expected values, marginal likelihoods, etc.
- Expected error with N Monte Carlo samples drops with $O\left((\sqrt{N})^{-1}\right)$
- Rejection sampling samples from a proposal distribution
- Yields unbiased samples but requires many samples[11]
- Metropolis-Hastings: use a random walk converging to a stationary distribution of a MC

[11] In particular for high-dimensional multivariate distributions.

Next lecture

Thursday Oct 27, 16:30

There is no lecture on Tue Oct 25.

Instead, there will be a lecture on Fri Oct 28

No exercise class on Oct 28; next on Nov 4.