昆山杜克大学
DUKE KUNSHAN
UNIVERSITY

Monte Carlo Markov Chain (MCMC)

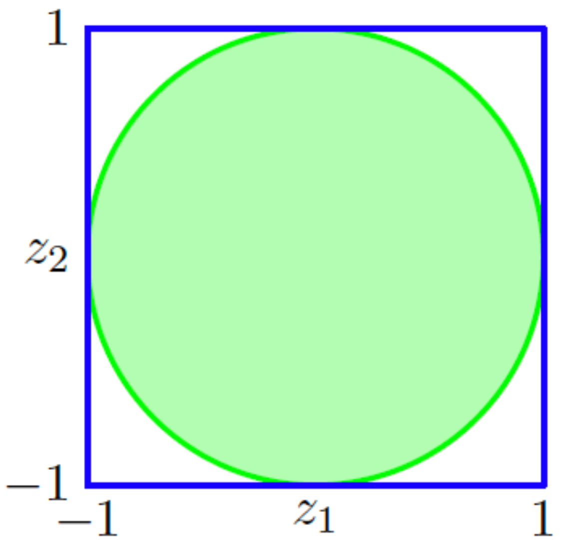
STATS 303 Statistical Machine Learning

Spring 2022

Lecture 14

example: Box-Muller method for Gaussian

The Box-Muller method for generating Gaussian distributed random numbers starts by generating samples from a uniform distribution inside the unit circle.



- First, uniformly sample $(z_1, z_2)^{\mathrm{T}}$ from a unit disk.

How?

1. Sample $\tilde{z}_1 \sim \mathrm{Unif}(0,1)$; take $z_1 = 2\tilde{z}_1 - 1$

   Then $z_1 \sim \mathrm{Unif}(-1,1)$. Then similarly, independently draw $z_2 \sim \mathrm{Unif}(-1,1)$

2. If $z_1^2 + z_2^2 \leq 1$, then accept $(z_1, z_2)^{\mathrm{T}}$ as our sample.

   Otherwise, "reject" the sample and redo step 1.

example: Box-Muller method for Gaussian

- Next, apply the transform: $y_1 = z_1 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$, $y_2 = z_2 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$ where $r^2 = z_1^2 + z_2^2$.

Then it is easy to verify:

$$p(y_1, y_2) = p(z_1, z_2) \left|\frac{\partial(z_1, z_2)}{\partial(y_1, y_2)}\right| = \left[\frac{1}{\sqrt{2\pi}} \exp(-y_1^2/2)\right] \left[\frac{1}{\sqrt{2\pi}} \exp(-y_2^2/2)\right]$$

$$= N(y_1, y_2 \mid 0, I)$$

example: general Gaussian

- If $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, then$\mathbf{y} = \boldsymbol{\mu} + \mathbf{L}\mathbf{z}$ has $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. Therefore, if we can sample from $\mathcal{N}(0, \mathbf{I})$, then we can sample from any Gaussian.
- To sample from $\mathcal{N}(0, \mathbf{I}_D)$, we only need to i.i.d. sample D one-dimensional Gaussians and combine them into a vector.

sample from Unif(0,1) and apply $(\text{cdf})^{-1}$.

rejection sampling

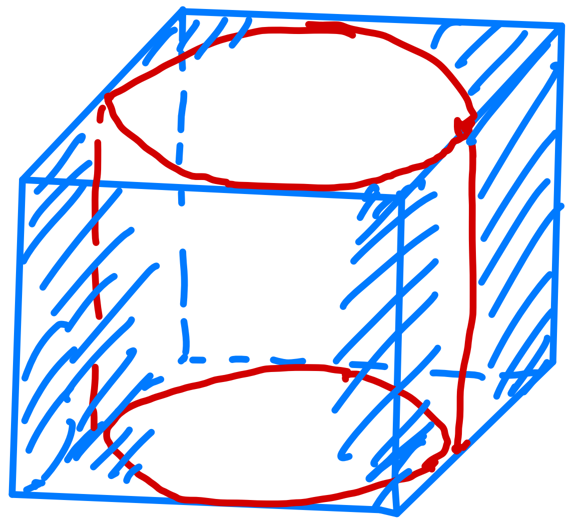
- Suppose it is easy to evaluate $p(\boldsymbol{z})$ up to a (possibly unknown) constant Z

$$p(\boldsymbol{z}) = \frac{1}{Z}\underbrace{\tilde{p}(\boldsymbol{z})}_{\text{easy to evaluate}}$$

- Let $q(\boldsymbol{z})$, called a proposal distribution, be simpler (we can draw samples from q).
- Let k be a constant such that $kq(\boldsymbol{z}) \geq \tilde{p}(\boldsymbol{z})$ for all $\boldsymbol{z}$.

rejection sampling

1. Generate $\boldsymbol{z}_0 \sim q(\boldsymbol{z})$
2. Generate $u_0 \sim \mathrm{Unif}[0, kq(\boldsymbol{z}_0)]$
3. If $u_0 > \tilde{p}(\boldsymbol{z}_0)$, reject! Otherwise accept.



In the rejection sampling method, samples are drawn from a simple distribution $q(z)$ and rejected if they fall in the grey area between the unnormalized distribution $\widetilde{p}(z)$ and the scaled distribution $kq(z)$. The resulting samples are distributed according to $p(z)$, which is the normalized version of $\widetilde{p}(z)$.

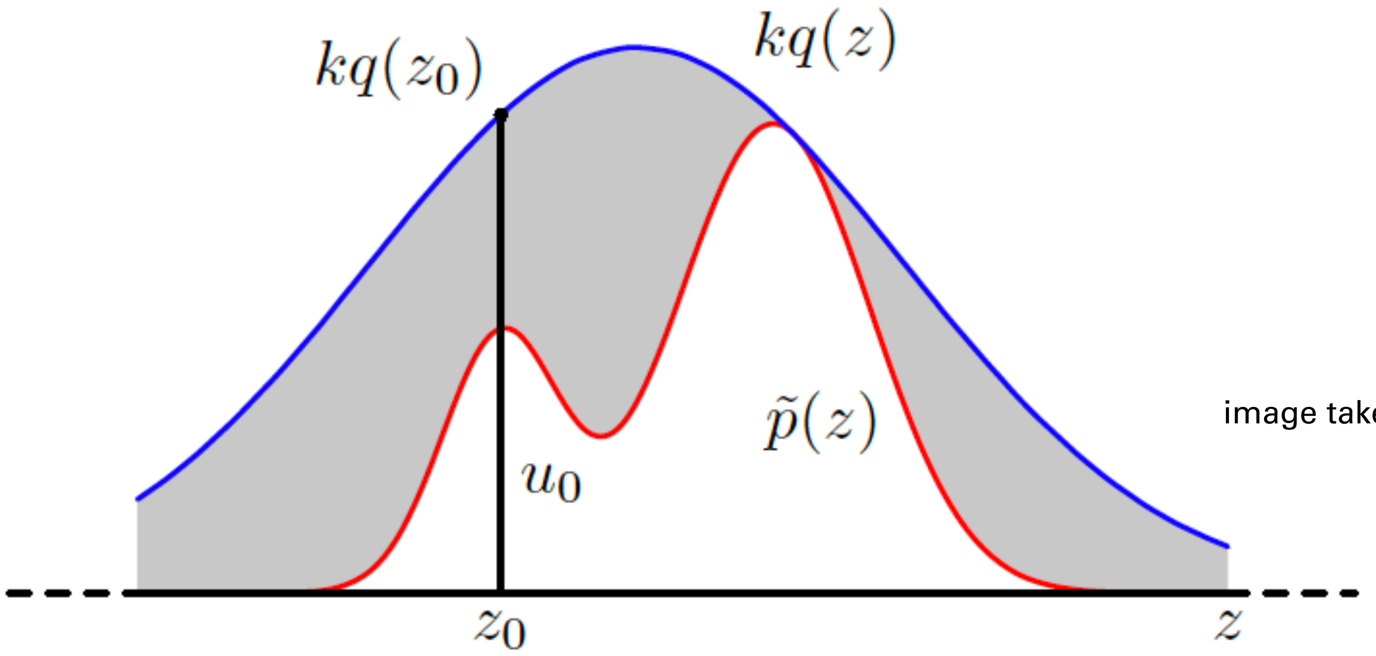


image taken from [Bi]

importance sampling

- In $s = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$, the specific way of decomposing $p(\boldsymbol{x})f(\boldsymbol{x})$ should not matter

$$p(\boldsymbol{x})f(\boldsymbol{x}) = q(\boldsymbol{x})\frac{p(\boldsymbol{x})f(\boldsymbol{x})}{q(\boldsymbol{x})}$$

- We can sample $\frac{pf}{q}$ from q instead of sampling f from p.
- Instead of calculating

$$\hat{s}[p] = \frac{1}{N}\sum_{n=1,\boldsymbol{x}_n\sim p}^{N} f(\boldsymbol{x}_n) \text{ ,}$$

calculate

$$\hat{s}[q] = \frac{1}{N}\sum_{n=1,\boldsymbol{x}_n\sim q}^{N} \frac{p(\boldsymbol{x}_n)f(\boldsymbol{x}_n)}{q(\boldsymbol{x}_n)}$$

importance sampling

Importance sampling addresses the problem of evaluating the expectation of a function $f(z)$ with respect to a distribution $p(z)$ from which it is difficult to draw samples directly. Instead, samples $\{z^{(l)}\}$ are drawn from a simpler distribution $q(z)$, and the corresponding terms in the summation are weighted by the ratios $p(z^{(l)})/q(z^{(l)})$.

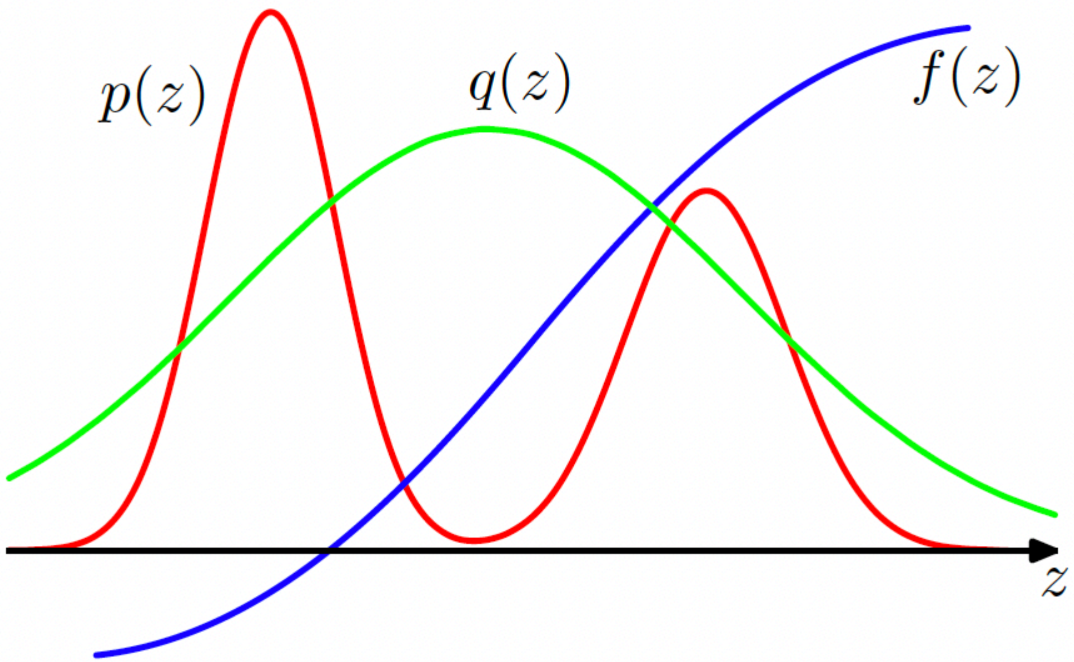


image taken from [Bi]

markov chain

- In $s = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$, we can sample $\frac{pf}{q}$ from q instead of sampling f from p.

- In practice, it is often infeasible to sample directly from p or any good q, due to curse of dimensionality.

- Idea: build a markov chain whose stationary distribution is p

markov chain

- Assume that x has countably many states, say $x \in \mathbb{N}$.
- We initialize some distribution $q^{(0)}$.
- Hope: construct a markov chain $\{q^{(s)}\}_{s \geq 0}$ so that $\{q^{(s)}(x)\}$ converges to $p(x)$

markov chain

- For each probability distribution q, we describe it as a vector $\boldsymbol{v}$ whose i-th entry is given by

$$v_i = q(x = i)$$

- By the Markov property

$$q^{(s+1)}(x') = \sum_x q^{(s)}(x)\, \underbrace{T(x'|x)}$$

We assume homogeneity: the transition probability does not change with s

markov chain

- For the countable case, using the transition matrix $\boldsymbol{A}$

$$A_{i,j} = T(x' = i | x = j)$$

- Then

$$\boldsymbol{v}^{(s)} = \boldsymbol{A}\boldsymbol{v}^{(s-1)}$$

markov chain

- Recursively, $\boldsymbol{v}^{(s)} = \boldsymbol{A}^s \boldsymbol{v}^{(0)}$

- Under some mild conditions (e.g. ergodicity), this process converges to a stationary distribution p represented by a vector $\boldsymbol{v}$:

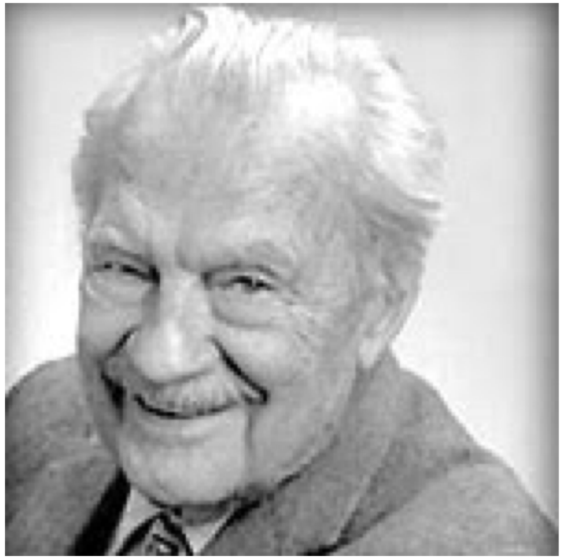
$$\boldsymbol{A}\boldsymbol{v} = \boldsymbol{v}$$

- Then almost surely, suppose $x_1, \cdots, x_N$ are drawn from such a markov chain,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \mathbb{E}_{x\sim p(x)} f(x) = s$$

Metropolis-Hastings (MH)

- A classical MCMC method
  - first proposed by Metropolis in 1953 (for symmetric proposal distributions);
  - then generalized by Hastings in 1970.




Metropolis-Hastings (MH)

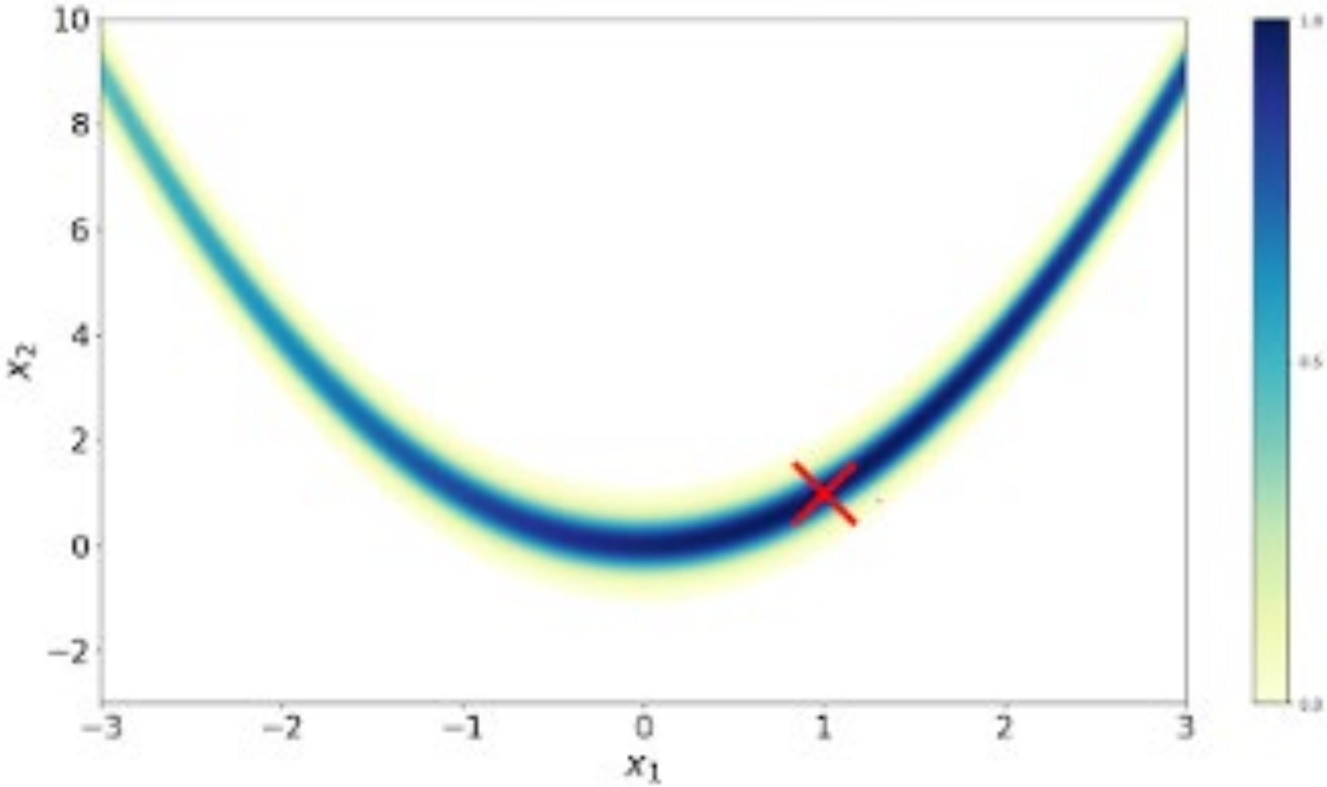
- Assume we can evaluate $\tilde{p}(\boldsymbol{x}) = p(\boldsymbol{x})/Z$ for some (possibly unknown) Z.
- At the beginning, choose a conditional density function, the proposal kernel q.

1. Initialize $\boldsymbol{x}^{(0)}$.
2. Iteration: at step s,
   1) generate $\boldsymbol{y}^{(s+1)} \sim q(\boldsymbol{y} \mid \boldsymbol{x}^{(s)})$
   2) take $\boldsymbol{x}^{(s+1)} = \begin{cases} \boldsymbol{y}^{(s+1)} & \text{with probability } \rho(\boldsymbol{x}^{(s)}, \boldsymbol{y}^{(s+1)}) \\ \boldsymbol{x}^{(s)} & \text{with probability } 1 - \rho(\boldsymbol{x}^{(s)}, \boldsymbol{y}^{(s+1)}) \end{cases}$ where

$$\rho(x, y) = \min \left\{ \frac{\tilde{p}(\boldsymbol{y})}{\tilde{p}(\boldsymbol{x})} \frac{q(\boldsymbol{x}|\boldsymbol{y})}{q(\boldsymbol{y}\,|\boldsymbol{x})}, 1 \right\}$$

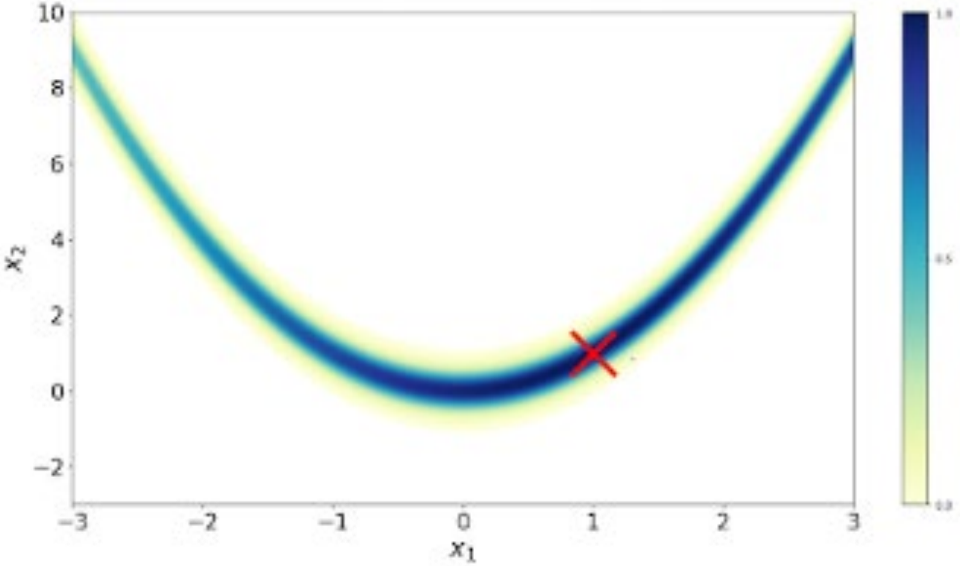
example: Rosenbrock density

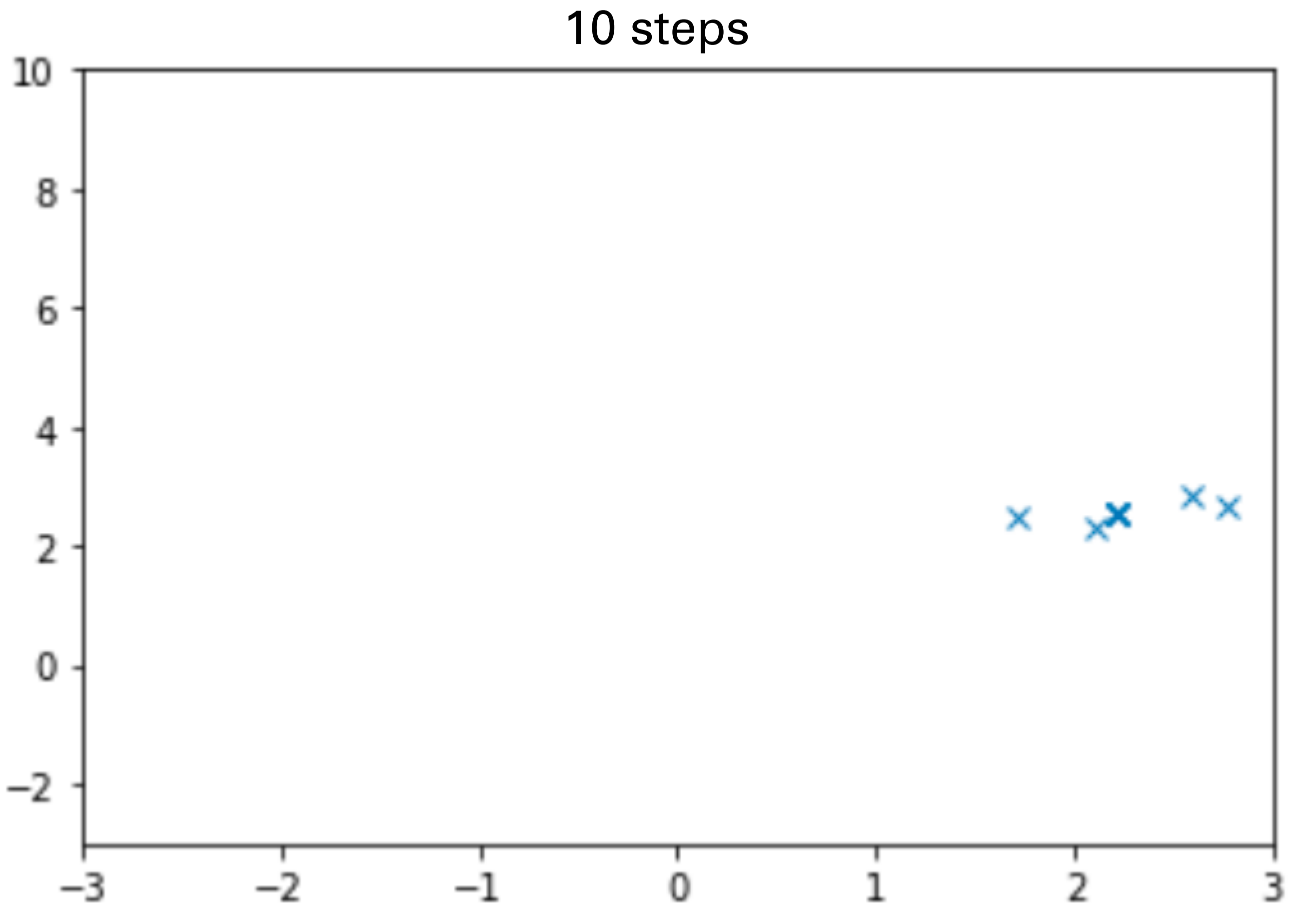
$$p(x_1, x_2) \propto \tilde{p}(x_1, x_2) = \exp\left(-\frac{(1-x_1)^2 + 100\left(x_2 - x_1^2\right)^2}{20}\right)$$



example: Rosenbrock density

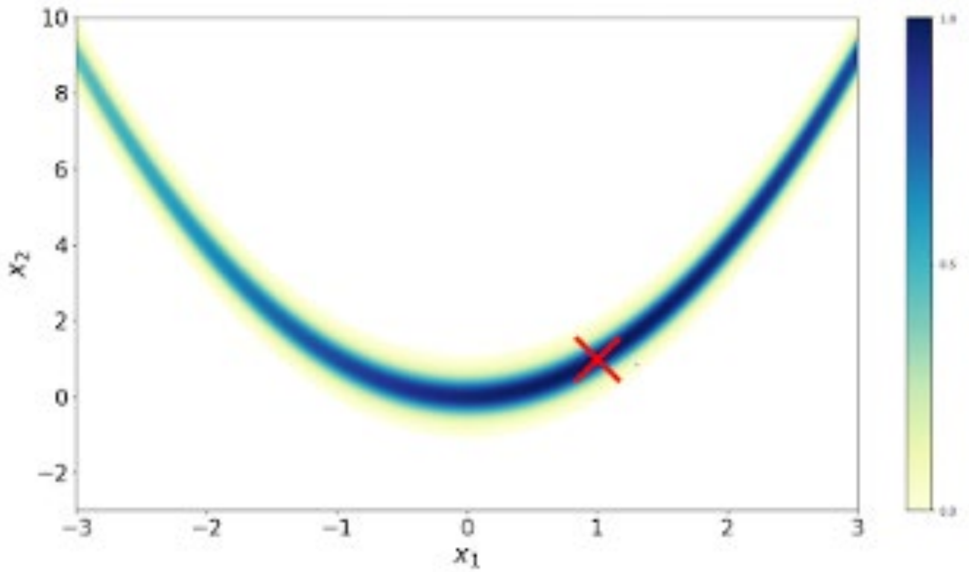
- Initialize $\boldsymbol{x}^{(0)} \sim \mathrm{Unif}[-5,5]^2$.
- Take the proposal kernel to be $q(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x}, \sigma^2 \boldsymbol{I})$, where $\sigma^2 = 0.1$.

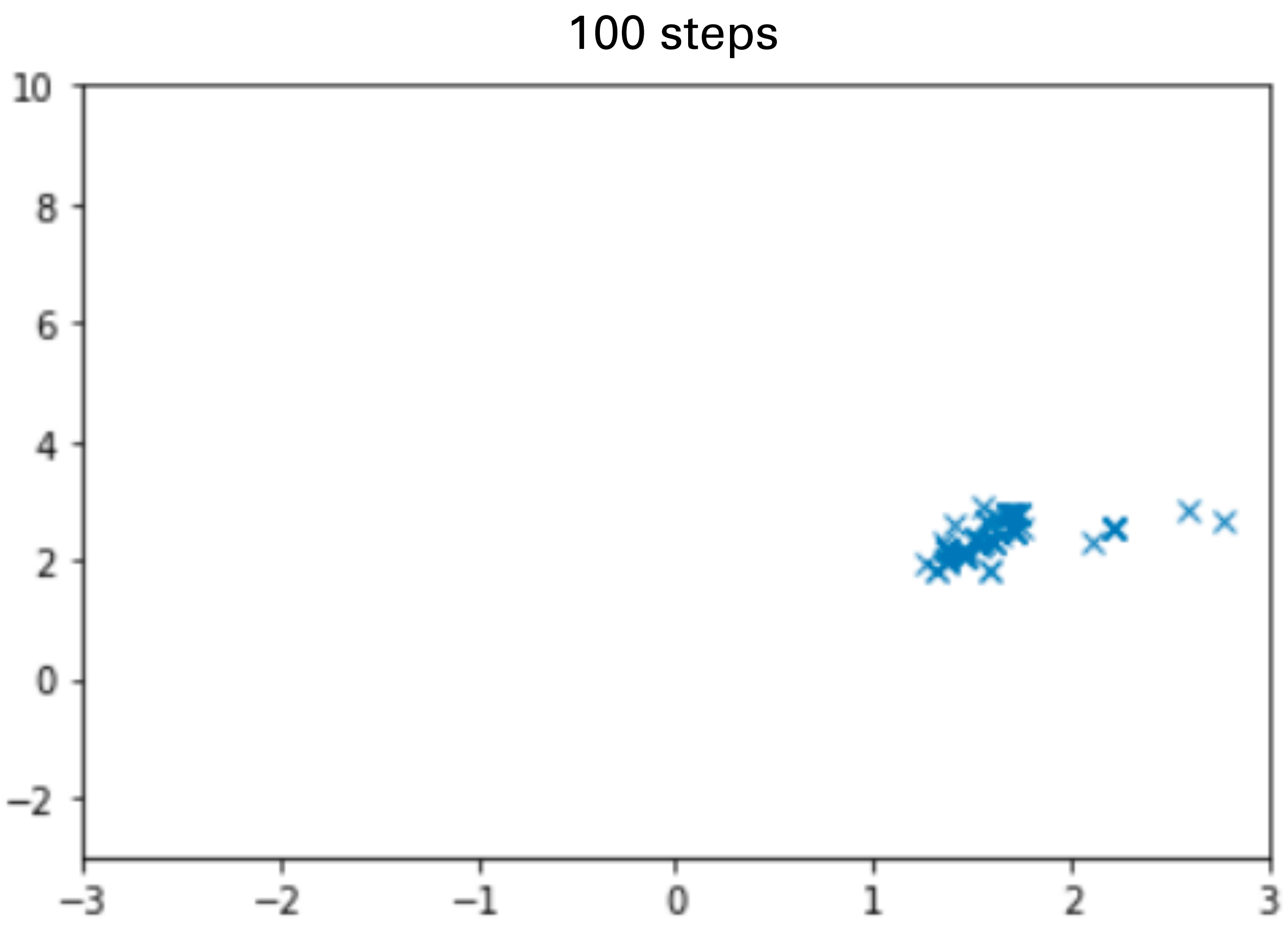




example: Rosenbrock density

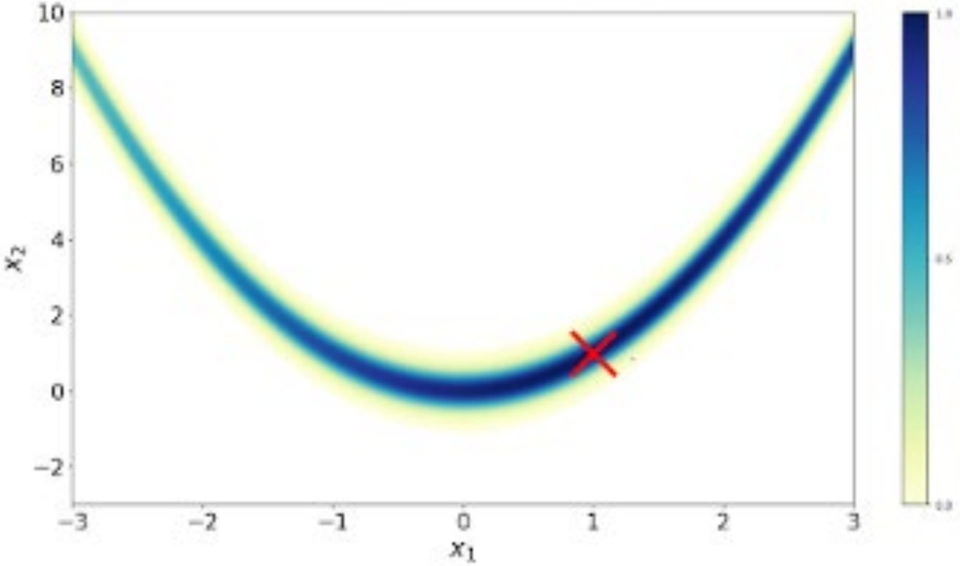
- Initialize $\boldsymbol{x}^{(0)} \sim \text{Unif}[-5, 5]^2$.
- Take the proposal kernel to be $q(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x}, \sigma^2 \boldsymbol{I})$, where $\sigma^2 = 0.1$.



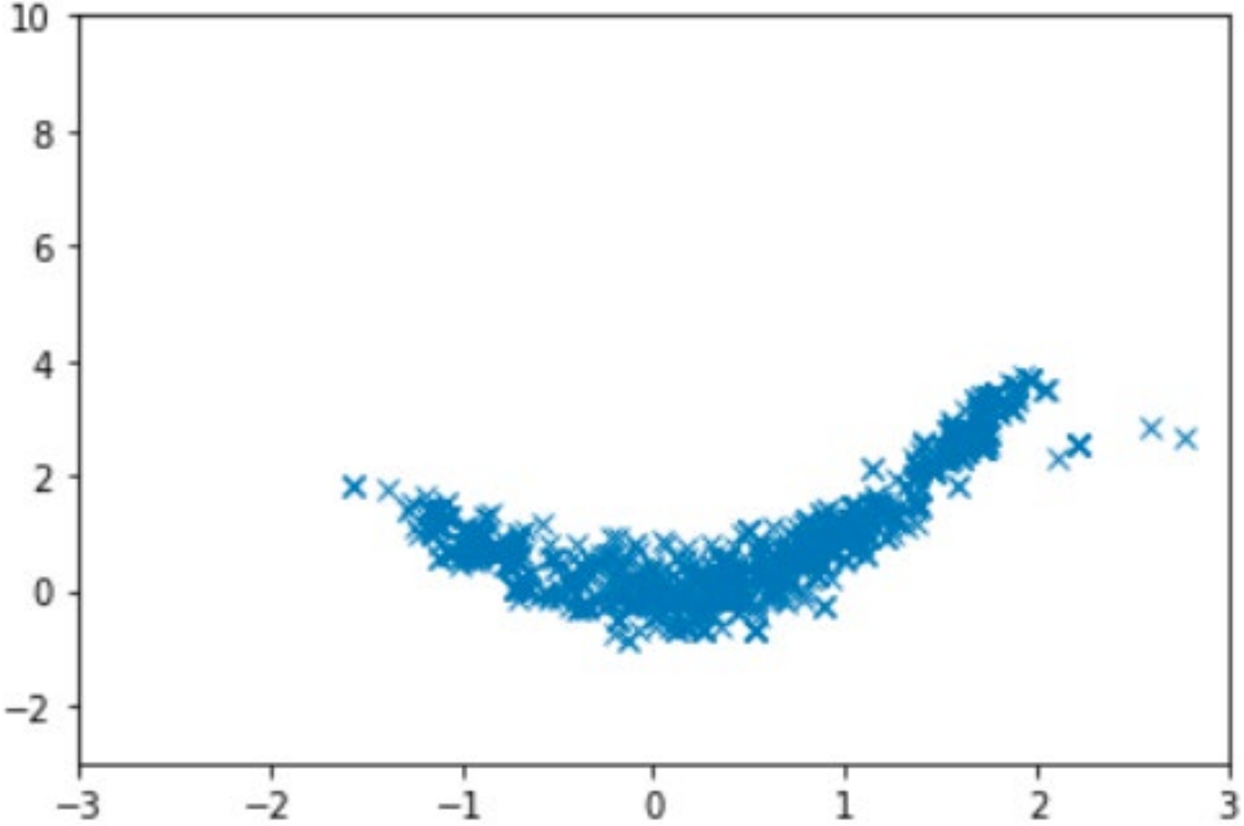


example: Rosenbrock density

- Initialize $\boldsymbol{x}^{(0)} \sim \mathrm{Unif}[-5,5]^2$.
- Take the proposal kernel to be $q(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x}, \sigma^2 \boldsymbol{I})$, where $\sigma^2 = 0.1$.

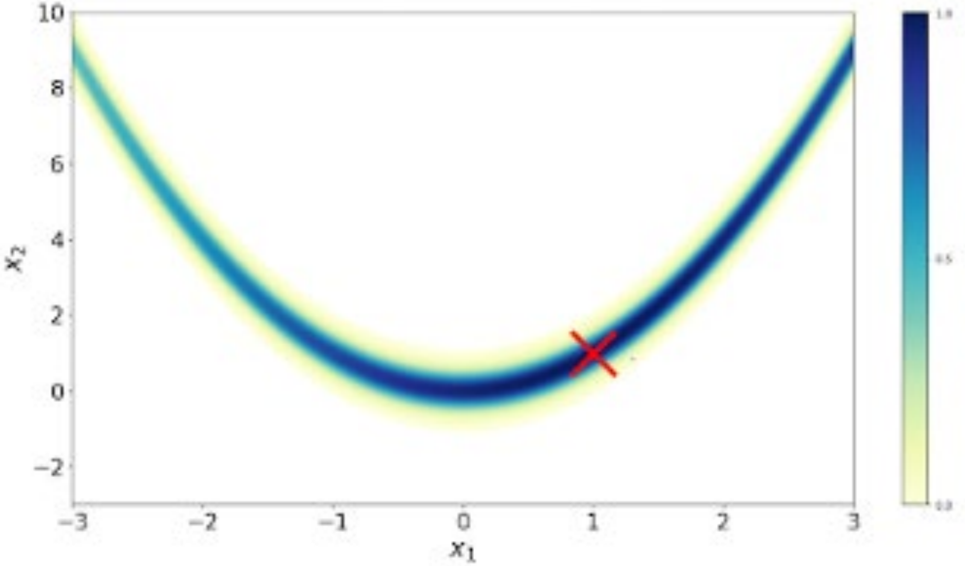


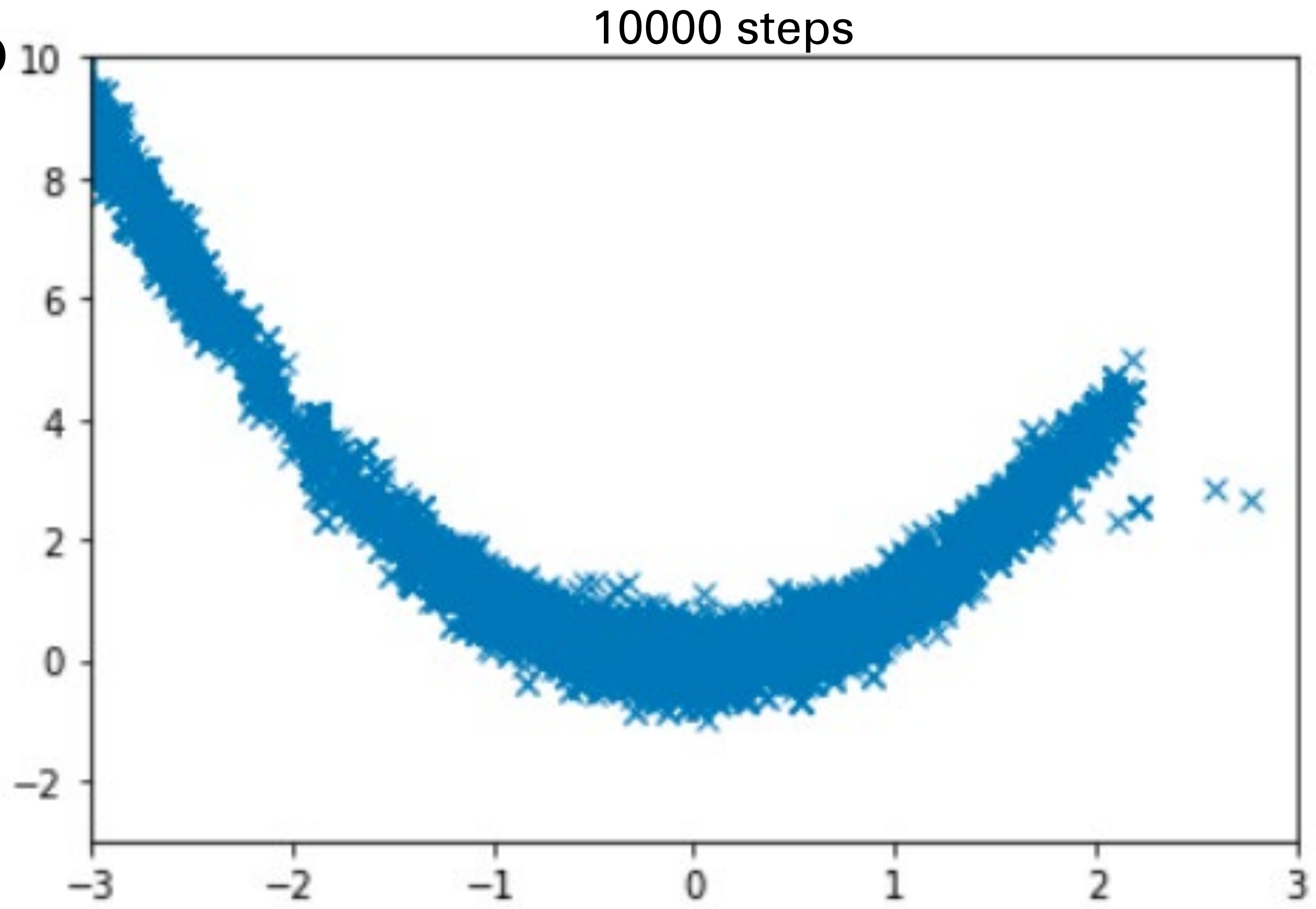
1000 steps



example: Rosenbrock density

- Initialize $\boldsymbol{x}^{(0)} \sim \mathrm{Unif}[-5,5]^2$.
- Take the proposal kernel to be $q(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{x}, \sigma^2 \boldsymbol{I})$, where $\sigma^2 = 0.1$.





Why does MH work?

- A sufficient condition for a markov chain to have a stationary distribution is that it is "reversible":
  - Discrete case: there exists $\boldsymbol{v}$ such that
$$A_{i,j}v_j = A_{j,i}v_i$$
  - Continuous case: there exists a distribution π over the state space such that
$$\pi(\boldsymbol{y})T(\boldsymbol{x}|\boldsymbol{y}) = \pi(\boldsymbol{x})T(\boldsymbol{y}|\boldsymbol{x})$$

Why does MH work?

- Since our proposed q is arbitrary, in general we don't have equality $\pi(\boldsymbol{y})q(\boldsymbol{x}|\boldsymbol{y}) = \pi(\boldsymbol{x})q(\boldsymbol{y}|\boldsymbol{x})$.

- Nevertheless, say for certain $\boldsymbol{x}$ and $\boldsymbol{y}$, without loss of generality,

$$q(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}) \geq q(\boldsymbol{x}|\boldsymbol{y})\pi(\boldsymbol{y})$$

- We need to revise it

$$q(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x})\rho'(\boldsymbol{x},\boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y})\pi(\boldsymbol{y})$$

That is, $\rho'(\boldsymbol{x},\boldsymbol{y}) = \dfrac{q(\boldsymbol{x}|\boldsymbol{y})\pi(\boldsymbol{y})}{q(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x})}$.

Why does MH work?

- Now, we have $q(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x})\rho'(\boldsymbol{x},\boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y})\pi(\boldsymbol{y})$ with $\rho'(\boldsymbol{x},\boldsymbol{y}) = \frac{q(\boldsymbol{x}|\boldsymbol{y})\pi(\boldsymbol{y})}{q(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x})}$. ≤ 1

$$\rho'(y,x) = \frac{q(y|x)\,\pi(x)}{q(x|y)\,\pi(y)} \geq 1$$

- Let $\rho(x,y) = min\{1, \rho'(\boldsymbol{x},\boldsymbol{y})\}$. The above is equivalent to

$$q(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x})\rho(\boldsymbol{x},\boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y})\pi(\boldsymbol{y})\rho(\boldsymbol{y},\boldsymbol{x}).$$

Why does MH work?

- Looking at the equation

$$q(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x})\rho(\boldsymbol{x},\boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y})\pi(\boldsymbol{y})\rho(\boldsymbol{y},\boldsymbol{x}),$$

we realize that the essential transition kernel is

$$q(\boldsymbol{y}|\boldsymbol{x})\rho(\boldsymbol{x},\boldsymbol{y}).$$

Gibbs sampling

- The Metropolis-Hastings algorithm does not leverage any structure of $p(\boldsymbol{x})$.
- Consider $p(\boldsymbol{x}) = p(x_1, \cdots, x_M)$.
- Suppose it is easy to sample from $p(x_i|\boldsymbol{x}_{-i})$.

$x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_M$

Gibbs sampling

i

- At each step s with $\boldsymbol{x}^{(s)} = \left(x_1^{(s)}, \cdots, x_M^{(s)}\right)$,
  1. Uniformly sample an index from $1, \cdots, M$
  2. Draw a sample $z \sim p\left(x_i \mid \boldsymbol{x}_{-i}^{(s)}\right)$
  3. Set $x^{(s+1)} = \left(x_1^{(s)}, \cdots, x_{i-1}^{(s)}, z, x_{i-1}^{(s)}, \cdots, x_M^{(s)}\right)$

Gibbs sampling as a special MH

- Proposal kernel $q(\boldsymbol{y}|\boldsymbol{x}) = p(y_i|\boldsymbol{x}_{-i})$
- We can calculate

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{\pi(\boldsymbol{y})q(\boldsymbol{x}|\boldsymbol{y})}{\pi(\boldsymbol{x})q(\boldsymbol{y}|\boldsymbol{x})} = \frac{p(y_i|\boldsymbol{y}_{-i})\pi(\boldsymbol{y}_{-i})p(x_i|\boldsymbol{y}_{-i})}{p(x_i|\boldsymbol{x}_{-i})\pi(\boldsymbol{x}_{-i})p(y_i|\boldsymbol{x}_{-i})} = 1$$

$\pi(x) = \pi(x_i, x_{-i})$

where the last equality follows the fact that $\boldsymbol{x}_{-i} = \boldsymbol{y}_{-i}$, which is obvious from the algorithm.

Questions?

Reference

- *Sampling-simple methods:*
  - *[Bi] Ch.11.1*
- *MCMC:*
  - *[Bi] Ch.11.2*