Statistics for Data Science

Wintersemester 2023/24

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Thirteenth lecture, January 29, 2024

Why is sampling needed?

Recall that the CM estimator and the conditional covariance require solving integration problems involving the posterior density:

$$\hat{x}_{\mathrm{CM}} = \mathbb{E}[x|y] = \int_{\mathbb{R}^d} x\, f(x|y)\, \mathrm{d}x$$

$$\mathrm{Cov}(x|y) = \int_{\mathbb{R}^d} (x - \hat{x}_{\mathrm{CM}})(x - \hat{x}_{\mathrm{CM}})^{\mathrm{T}} f(x|y)\, \mathrm{d}x.$$

In a non-Gaussian case, these integrals cannot typically be expressed in closed form. Therefore one must resort to numerical integration.

Suppose that our goal is to estimate some quantity of the form

$$\mathcal{I} = \mathbb{E}[G(X)] = \int_{\mathbb{R}^d} G(x)p(x)\,\mathrm{d}x,$$

where $p\colon \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is a probability density function and G is a quantity of interest.

For example, if p is a posterior density and $G(x) = x$, then $\mathcal{I}$ would be precisely the CM estimator.

In principle, we could use a quadrature rule

$$\mathcal{I} = \int_{\mathbb{R}^d} G(x)p(x)\,\mathrm{d}x \approx \sum_{j=1}^{N} w_j G(x_j)p(x_j)$$

with some suitable weights $\{w_j\}_{j=1}^N$ and nodes $\{x_j\}_{j=1}^N$. However, the design of efficient quadrature rules for high-dimensional problems is challenging. Moreover, the implementation of a quadrature rule would require reliable information about the location of the *support* of the probability density p.

Often it is more advisable to resort to sampling: draw a large enough sample $\{x_j\}_{j=1}^N$ from the probability distribution corresponding to p, and use these points to approximate the integral as

$$\mathcal{I} = \int_{\mathbb{R}^d} G(x)p(x)\,\mathrm{d}x = \mathbb{E}[G(x)] \approx \frac{1}{N}\sum_{j=1}^{N} G(x_j) = \mathcal{I}_N.$$

According to the Law of Large Numbers, for any integrable G there holds

$$\lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} G(x_j) = \mathcal{I} \quad \text{almost surely.}$$

Furthermore, if G is square-integrable, then the Central Limit Theorem states that

$$\mathrm{Var}(\mathcal{I} - \mathcal{I}_N) \approx \frac{\mathrm{Var}(G(X))}{N},$$

i.e., the discrepancy between $\mathcal{I}$ and $\mathcal{I}_N$ should go to zero like $1/\sqrt{N}$.

Markov Chain Monte Carlo

Discrete time Markov chains

A sequence $\{X_k\}_{k=0}^{\infty}$ of random variables is called a *discrete time Markov chain* if the probability distribution of any X_{k+1} depends only on the previous state X_k:

$$\pi(x_{k+1} \mid x_0, \ldots, x_k) = \pi(x_{k+1} \mid x_k).$$

Here, $\pi(x_{k+1}|x_0, \ldots, x_k)$ (resp. $\pi(x_{k+1}|x_k)$) denotes the PDF of X_{k+1} conditioned on the previous states $X_0, \ldots, X_k$ (resp. X_k). Suppose in addition that there exists a *probability transition kernel* $q(x, y)$ such that

$$\pi(x_{k+1} \mid x_k) = q(x_k, x_{k+1}).$$

Then the Markov chain is called *time invariant* (or *time homogeneous*) since the kernel q is independent of the time k.

Remark. We assume that transition kernels satisfy the following:

- for each fixed $x \in \mathbb{R}^d$, the function $y \mapsto q(x, y)$ is a probability density. In particular, $\mathbb{P}(Y \in B \mid X = x) = \int_B q(x, y)\,\mathrm{d}y$ and $\int_{\mathbb{R}^d} q(x, y)\,\mathrm{d}y = 1$.

Example (Random walk in $\mathbb{R}^d$)

A *random walk* in $\mathbb{R}^d$ is a process of moving around by taking random steps. Elementary random walk:

1. Choose a starting point $x_0 \in \mathbb{R}^d$ and a "step size" $\sigma > 0$. Set $k = 0$.
2. Draw a random vector $w_{k+1} \sim \mathcal{N}(0, I)$ and set $x_{k+1} = x_k + \sigma w_{k+1}$.
3. Set $k \leftarrow k + 1$ and return to step 2, unless your stopping criterion is satisfied.

The location of a random walk at time k is a realization of the random variable X_k, and we have an evolution model

$$X_{k+1} = X_k + \sigma W_{k+1}, \quad W_{k+1} \sim \mathcal{N}(0, I).$$

The conditional density of X_{k+1}, given $X_k = x_k$, is

$$\pi(x_{k+1}|x_k) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2} \|x_k - x_{x+1}\|^2 \right) = q(x_k, x_{k+1}),$$

where q is the (time invariant) transition kernel.

Let X be a random variable with probability density $p(x)$.

Let $q(x, y)$ be an arbitrary transition kernel used to generate a new random variable Y given $X = x$, i.e.,

$$\pi(y \mid x) = q(x, y).$$

By the law of total probability, the probability density of Y is

$$\pi(y) = \int_{\mathbb{R}^d} \pi(y \mid x) p(x) \, \mathrm{d}x = \int_{\mathbb{R}^d} q(x, y) p(x) \, \mathrm{d}x.$$

If the probability density of Y is equal to the probability density of X,

$$\int_{\mathbb{R}^d} q(x, y) p(x) \, \mathrm{d}x = p(y),$$

then we call p an *invariant density* of the transition kernel q.

Definition (Irreducible transition kernel)

The transition kernel q is *irreducible* if, regardless of the starting point, the Markov chain generated by q can visit any set of positive measure with positive probability.

Definition (Periodic transition kernel)

The transition kernel q is *periodic* if, for some integer $m \geq 2$, there is a set of disjoint nonempty sets $\{E_1, \ldots, E_m\} \subset \mathbb{R}^d$ such that for all $j \in \{1, \ldots, m\}$ and for all $x \in E_j$:

$$\mathbb{P}(Y \in E_{\mathrm{mod}(j,m)+1} | X = x) = \int_{E_{\mathrm{mod}(j,m)+1}} q(x, y)\, \mathrm{d}y = 1.$$

That is, the Markov chain generated by q remains in a periodic loop forever.

Definition (Aperiodic transition kernel)

The transition kernel q is *aperiodic* if it is not periodic.

Theorem

Let $\{X_k\}_{k=0}^{\infty}$ be a time invariant Markov chain with the transition kernel q, i.e.,

$$\pi(x_{k+1} \mid x_k) = q(x_k, x_{k+1}).$$

Assume that p is an invariant density of q and the following technical conditions hold:

- *q is* irreducible*;*
- *q is* aperiodic*.*

Then for all $x_0 \in \mathbb{R}^d$ and any (measurable) $B \subseteq \mathbb{R}^d$, there holds

$$\lim_{N\to\infty} \mathbb{P}(X_N \in B \mid X_0 = x_0) = \int_B p(x)\,\mathrm{d}x.$$

Moreover, for any integrable $G : \mathbb{R}^d \to \mathbb{R}$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} G(X_j) = \int_{\mathbb{R}^d} G(x)p(x)\,\mathrm{d}x \quad \text{a.s.}$$

Suppose we want to sample some probability density p and we know that it is invariant with respect to transition kernel q. Then we can proceed as follows:

1. Select starting point x_0 and set $k = 0$.
2. Draw x_{k+1} from $q(x_k, x_{k+1})$.
3. Set $k \leftarrow k + 1$ and return to step 2.

The previous theorem implies that the sample $\{x_k\}_{k=0}^{N}$ is asymptotically distributed according to p as $N \to \infty$.

This raises the question: *given a probability density p, how do you find a kernel q such that p is its invariant density?*

The *Metropolis–Hastings algorithm* is a method to construct such a kernel!

Derivation of the Metropolis–Hastings algorithm

We are interested in obtaining samples from the probability density p. Consider the following Markov process: if you are currently situated at some $x \in \mathbb{R}^d$, either

1. stay put at x with the probability $r(x)$, $0 \leq r(x) \leq 1$, or
2. move away from x using a transition kernel $R(x, y)$ otherwise.

Here, both $R(x, y)$ and $r(x)$ are as yet undetermined—the trick will be to calibrate these in order to find a kernel such that p is its invariant density as discussed on the previous slide.

Since R is a transition kernel, $y \mapsto R(x, y)$ is a probability density and hence

$$\int_{\mathbb{R}^d} R(x, y) \, \mathrm{d}y = 1 \quad \text{for all } x \in \mathbb{R}^d.$$

Denote by $\mathcal{A}$ the event of moving away from x and by $\neg\mathcal{A}$ the event of not moving. Clearly

$$\mathbb{P}(\mathcal{A}) = 1 - r(x) \quad \text{and} \quad \mathbb{P}(\neg\mathcal{A}) = r(x).$$

Given a current state $X = x$, we want to know what is the probability density of Y generated by the aforementioned strategy. Let $B \subseteq \mathbb{R}^d$ and consider the probability of the event $Y \in B$. Then

$$\begin{aligned}\mathbb{P}(Y \in B \mid X = x) = \mathbb{P}(Y \in B \mid X = x, \mathcal{A})\mathbb{P}(\mathcal{A}) \quad &\text{(move away from } x\text{)} \\ + \mathbb{P}(Y \in B \mid X = x, \neg\mathcal{A})\mathbb{P}(\neg\mathcal{A}). \quad &\text{(stay put at } x\text{)}\end{aligned}$$

The probability of arriving in B through a move is

$$\mathbb{P}(Y \in B \mid X = x, \mathcal{A}) = \int_B R(x, y)\,\mathrm{d}y.$$

The only way to arrive in B without moving is if x is already in B:

$$\mathbb{P}(Y \in B \mid X = x, \neg\mathcal{A}) = \mathbf{1}_B(x) = \begin{cases} 1 & \text{if } x \in B, \\ 0 & \text{if } x \notin B. \end{cases}$$

Hence

$$\begin{aligned}\mathbb{P}(Y \in B \mid X = x) &= \int_B \overbrace{(1 - r(x))R(x, y)}^{=:K(x,y)}\,\mathrm{d}y + r(x)\mathbf{1}_B(x) \\ &= \int_B K(x, y)\,\mathrm{d}y + r(x)\mathbf{1}_B(x).\end{aligned}$$

The probability of $Y \in B$ can be obtained by marginalizing over x:

$$\begin{aligned}
\mathbb{P}(Y \in B) &= \int_{\mathbb{R}^d} \mathbb{P}(Y \in B \mid X = x) p(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \left(\int_B K(x,y)\,\mathrm{d}y \right) p(x)\,\mathrm{d}x + \int_{\mathbb{R}^d} r(x)\mathbf{1}_B(x) p(x)\,\mathrm{d}x \\
&= \int_B \left(\int_{\mathbb{R}^d} K(x,y) p(x)\,\mathrm{d}x \right) \mathrm{d}y + \int_B r(x) p(x)\,\mathrm{d}x \\
&= \int_B \left(\int_{\mathbb{R}^d} K(x,y) p(x)\,\mathrm{d}x + r(y) p(y) \right) \mathrm{d}y \\
&= \int_B \left(\int_{\mathbb{R}^d} K(x,y) p(x)\,\mathrm{d}x - \int_{\mathbb{R}^d} K(y,x) p(y)\,\mathrm{d}x + p(y) \right) \mathrm{d}y,
\end{aligned}$$

where we used $\int_{\mathbb{R}^d} K(y,x)\,\mathrm{d}x = (1 - r(y)) \int_{\mathbb{R}^d} R(y,x)\,\mathrm{d}x = 1 - r(y)$.
If the *balance equation*

$$\int_{\mathbb{R}^d} p(y) K(y,x)\,\mathrm{d}x = \int_{\mathbb{R}^d} p(x) K(x,y)\,\mathrm{d}x \tag{1}$$

holds, then

$$\mathbb{P}(Y \in B) = \int_B p(y)\,\mathrm{d}y \quad \text{as desired.}$$

The Metropolis–Hastings algorithm is a technique for finding a kernel K that satisfies the *detailed balance equation*

$$p(y)K(y,x) = p(x)K(x,y),$$

which implies (1). Let us start with a *proposal density* $q(x,y)$, chosen so that generating a Markov chain with it is easy. (For this reason, a Gaussian kernel is a very popular choice.) There are three separate cases:

1. If $p(y)q(y,x) = p(x)q(x,y)$, then set $r(x) = 0$, $R(x,y) = K(x,y) = q(x,y)$ and the previous analysis ensures that p is an invariant density for kernel q.
2. If $p(y)q(y,x) < p(x)q(x,y)$, then define the kernel K to be

   $$K(x,y) = \alpha(x,y)q(x,y),$$

   where α is chosen s.t. $p(y)\alpha(y,x)q(y,x) = p(x)\alpha(x,y)q(x,y)$. We can make the selection

   $$\alpha(y,x) = 1 \quad \text{and} \quad \alpha(x,y) = \frac{p(y)q(y,x)}{p(x)q(x,y)} < 1.$$
3. If $p(y)q(y,x) > p(x)q(x,y)$, then in complete analogy to the above:

   $$\alpha(x,y) = 1 \quad \text{and} \quad \alpha(y,x) = \frac{p(x)q(x,y)}{p(y)q(y,x)} < 1.$$

In summary, we define K as

$$K(x, y) = \alpha(x, y)q(x, y), \quad \alpha(x, y) = \min\left\{1, \frac{p(y)q(y, x)}{p(x)q(x, y)}\right\}.$$

Even though the expression for K seems complicated, it turns out that the drawing can be performed according to the following procedure.

Metropolis–Hastings algorithm

1. Choose $x^{(0)} \in \mathbb{R}^d$ and set $k = 0$.
2. Given $x = x^{(k)}$, draw y using the transition kernel $q(x, y)$ of your choosing.
3. Calculate the acceptance ratio

$$\alpha(x, y) = \min\left\{1, \frac{p(y)q(y, x)}{p(x)q(x, y)}\right\}.$$

4. Flip the α-coin: draw $t \sim \mathcal{U}([0, 1])$. If $\alpha > t$, set $x^{(k+1)} = y$, otherwise stay put at x and set $x^{(k+1)} = x^{(k)}$.
5. Set $k \leftarrow k + 1$ and return to step 2.

Remark. Note that due to the form of α, both the target p and the proposal density q can be *unnormalized* within the Metropolis–Hastings algorithm.

Why does this work?

Let us focus on the main loop of the Metropolis–Hastings algorithm:

- Given x, draw y using the transition kernel $q(x, y)$.
- Calculate the acceptance ratio $\alpha(x, y) = \min\left\{1, \frac{p(y)q(x,y)}{p(x)q(x,y)}\right\}$.
- Draw $t \sim \mathcal{U}([0, 1])$. If $\alpha > t$, accept y, otherwise stay put at x.

Recall that $\mathcal{A}$ was the event of moving in the Markov chain. Then

$$\mathbb{P}(\mathcal{A}|y, x) = \text{“probability of accepting transition”} = \alpha(x, y),$$
$$\mathbb{P}(y|x) = \text{“probability of drawing } y\text{”} = q(x, y).$$

Then

$$\begin{aligned}\text{“probability of accepted } y\text{”} &= \mathbb{P}(\mathcal{A}, y|x)\\ &= \mathbb{P}(\mathcal{A}|y, x)\mathbb{P}(y|x)\\ &= \alpha(x, y)q(x, y) = K(x, y),\end{aligned}$$

as desired.

Example

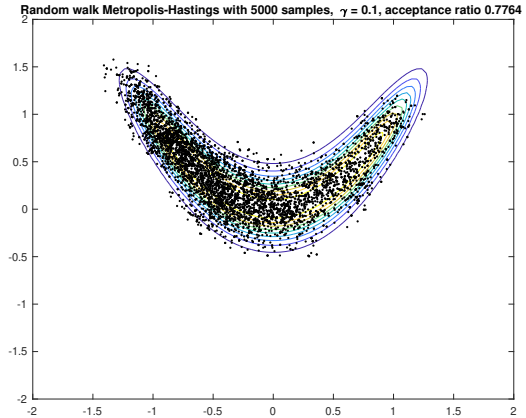
Let us consider sampling from the density

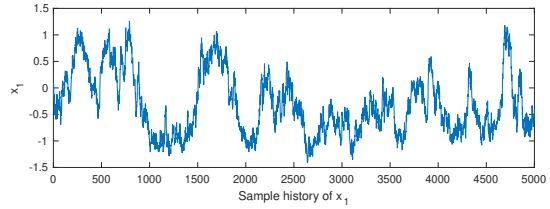
$$p(x_1, x_2) \propto \exp(-10(x_1^2 - x_2)^2 - (x_2 - \tfrac{1}{4})^4).$$

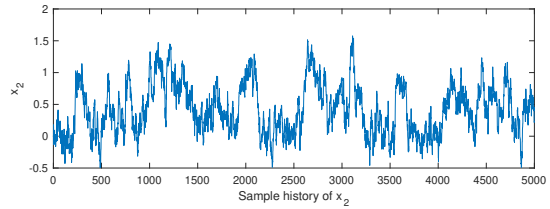
As the proposal distribution, we use the random walk model $Y = X + W$, $W \sim \mathcal{N}(0, \gamma^2 I)$, with the kernel

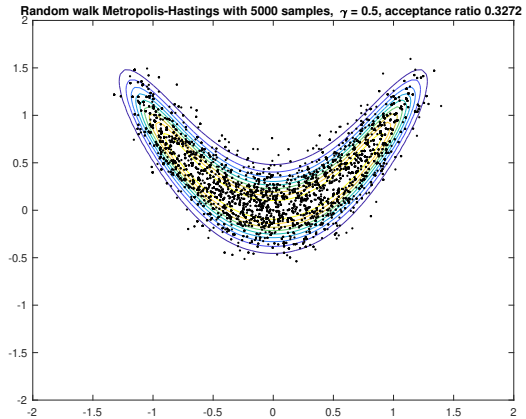
$$q(x, y) \propto \exp\left(-\frac{1}{2\gamma^2} \|x - y\|^2 \right).$$

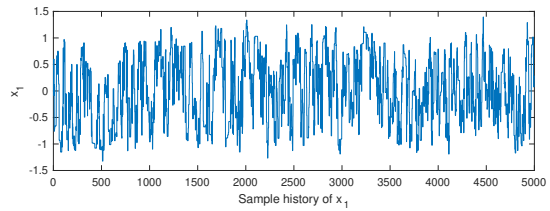
We draw 5000 samples from the probability distribution with density p using three different step sizes: $\gamma = 0.1$, $\gamma = 0.5$, and $\gamma = 2$.

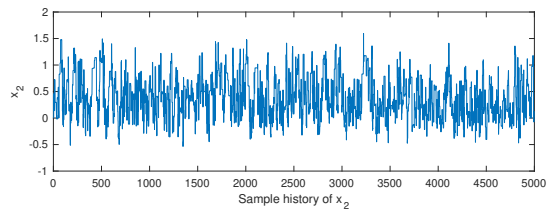

Random walk Metropolis-Hastings with 5000 samples, γ = 0.1, acceptance ratio 0.7764

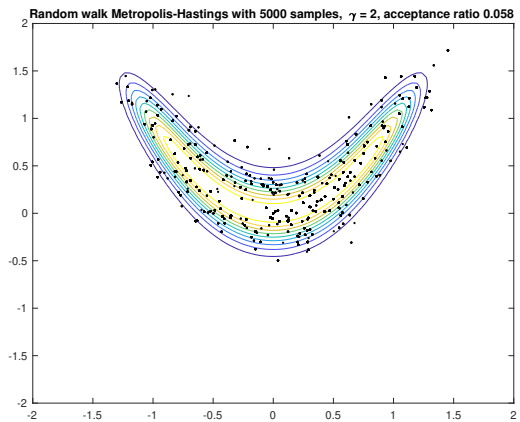

x1
Sample history of x1

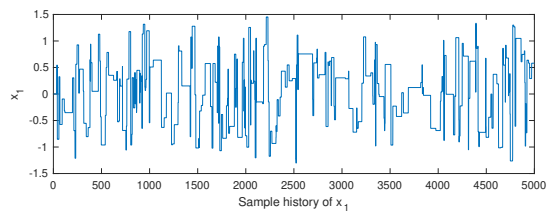

x2
Sample history of x2

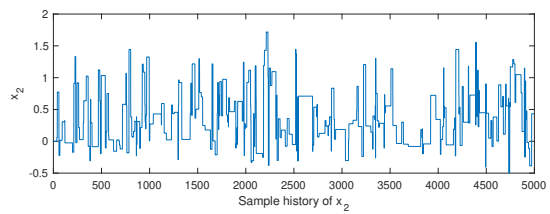

Random walk Metropolis-Hastings with 5000 samples, γ = 0.5, acceptance ratio 0.3272


x1
Sample history of x1


x2
Sample history of x2


Random walk Metropolis-Hastings with 5000 samples, γ = 2, acceptance ratio 0.058


x1
Sample history of x1


x2
Sample history of x2

Derivation of the single component Gibbs sampler

We continue to be interested in sampling the distribution with density $p(x)$. The single component Gibbs sampler is based on the same Markov process that was introduced in the derivation of Metropolis–Hastings: if you are currently situated at some $x \in \mathbb{R}^d$, either

1. stay put at x with the probability $r(x)$, $0 \leq r(x) \leq 1$, or
2. move away from x using a transition kernel $R(x, y)$ otherwise.

Recall also the definition we made in the Metropolis–Hastings derivation:

$$K(x, y) = (1 - r(x))R(x, y).$$

Suppose that x is a d-variate random variable. For the single component Gibbs sampler, we set $r(x) = 0$ (moving is obligatory) and define the transition kernel

$$K(x, y) = R(x, y) = \prod_{i=1}^{d} p(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d),$$

where $p(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d) = \dfrac{p(y_1, \ldots, y_i, x_{i+1}, \ldots, x_d)}{\int_{\mathbb{R}} p(y_1, \ldots, y_i, x_{i+1}, \ldots, x_d)\, \mathrm{d}y_i}$.

This transition kernel K does not in general satisfy the detailed balance equation, but it does satisfy the standard balance equation, which is sufficient to ensure that p is the invariant density of the Markov chain (see derivation of the Metropolis–Hastings method).

Theorem

The transition kernel

$$K(x, y) = \prod_{i=1}^{d} p(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d),$$

where $p(y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_d) = \dfrac{p(y_1, \ldots, y_i, x_{i+1}, \ldots, x_d)}{\int_{\mathbb{R}} p(y_1, \ldots, y_i, x_{i+1}, \ldots, x_d)\,\mathrm{d}y_i}$,

satisfies

$$\int_{\mathbb{R}^d} p(y)K(y, x)\,\mathrm{d}x = \int_{\mathbb{R}^d} p(x)K(x, y)\,\mathrm{d}x.$$

Proof. We begin with the left-hand side of the balance equation and consider $\int_{\mathbb{R}^d} K(y,x)\,\mathrm{d}x$. We integrate inductively over the variables in the order $x_d, x_{d-1}, \ldots, x_1$:

$$\begin{aligned}
\int_{\mathbb{R}} K(y,x)\,\mathrm{d}x_d &= \int_{\mathbb{R}} \left(\prod_{i=1}^{d} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \right) \mathrm{d}x_d \\
&= \left(\prod_{i=1}^{d-1} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \right) \underbrace{\int_{\mathbb{R}} p(x_d|x_1,\ldots,x_{d-1})\,\mathrm{d}x_d}_{=1} \\
&= \prod_{i=1}^{d-1} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \\
\Rightarrow \int_{\mathbb{R}} \int_{\mathbb{R}} K(y,x)\,\mathrm{d}x_d\,\mathrm{d}x_{d-1} &= \int_{\mathbb{R}} \left(\prod_{i=1}^{d-1} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \right) \mathrm{d}x_{d-1} \\
&= \left(\prod_{i=1}^{d-2} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \right) \underbrace{\int_{\mathbb{R}} p(x_{d-1}|x_1,\ldots,x_{d-1},y_d)\,\mathrm{d}x_{d-1}}_{=1} \\
&= \prod_{i=1}^{d-2} p(x_i|x_1,\ldots,x_{i-1},y_{i+1},\ldots,y_d) \quad \Rightarrow \quad \ldots
\end{aligned}$$

Proceeding by inductively integrating over $x_{d-2}, x_{d-3}, \ldots, x_1$, we obtain $\int_{\mathbb{R}^d} K(y,x)\,\mathrm{d}x = 1$ and thus $\int_{\mathbb{R}^d} p(y)K(y,x)\,\mathrm{d}x = p(y)\int_{\mathbb{R}^d} K(y,x)\,\mathrm{d}x = p(y)$.

Next we consider the right-hand side of the balance equation. Recall that $K(x,y) = \prod_{i=1}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1}\ldots,x_d)$. We integrate inductively over the variables, this time in the order $x_1,\ldots,x_d$:

$$\int_{\mathbb{R}} p(x)K(x,y)\,\mathrm{d}x_1 = K(x,y)\int_{\mathbb{R}} p(x_1,x_2,\ldots,x_d)\,\mathrm{d}x_1 \qquad (K \text{ is independent of } x_1)$$

$$= \left(\prod_{i=2}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) \underbrace{p(y_1|x_2,\ldots,x_d)}_{=\frac{p(y_1,x_2,\ldots,x_d)}{\int_{\mathbb{R}} p(x_1,x_2,\ldots,x_d)\,\mathrm{d}x_1}} \int_{\mathbb{R}} p(x_1,x_2,\ldots,x_d)\,\mathrm{d}x_1$$

$$= \left(\prod_{i=2}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) p(y_1,x_2,\ldots,x_d)$$

$$\Rightarrow \int_{\mathbb{R}}\int_{\mathbb{R}} p(x)K(x,y)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \int_{\mathbb{R}} \left(\prod_{i=2}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) p(y_1,x_2,\ldots,x_d)\,\mathrm{d}x_2$$

$$= \left(\prod_{i=3}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) \underbrace{p(y_2|y_1,x_3,\ldots,x_d)}_{=\frac{p(y_1,y_2,x_3,\ldots,x_d)}{\int_{\mathbb{R}} p(y_1,x_2,x_3,\ldots,x_d)\,\mathrm{d}x_2}} \int_{\mathbb{R}} p(y_1,x_2,\ldots,x_d)\,\mathrm{d}x_2$$

$$= \left(\prod_{i=3}^{d} p(y_i|y_1,\ldots,y_{i-1},x_{i+1},\ldots,x_d)\right) p(y_1,y_2,x_3,\ldots,x_d) \quad \Rightarrow \quad \ldots$$

Proceeding by inductively integrating over $x_3,\ldots,x_d$, we eventually obtain $\int_{\mathbb{R}^d} p(x)K(x,y)\,\mathrm{d}x = p(y)$. Therefore the balance equation holds. □

Single component Gibbs sampler

1. Choose the initial value $x^{(0)} \in \mathbb{R}^d$ and set $k = 0$.
2. Draw the next sample as follows:
   (i) Set $x = x^{(k)}$ and $j = 1$.
   (ii) Draw $t \in \mathbb{R}$ from the one-dimensional distribution

$$p(t \mid y_1, \ldots, y_{j-1}, x_{j+1}, \ldots, x_d) \propto p(y_1, \ldots, y_{j-1}, t, x_{j+1}, \ldots, x_d)$$

   and set $y_j = t$.
   (iii) If $j = d$, set $y = (y_1, \ldots, y_d)$ and terminate the inner loop. Otherwise, set $j \leftarrow j + 1$ and return to step (ii).
3. Set $x^{(k+1)} = y$, increase $k \leftarrow k + 1$ and return to step 2.

Example

Let us consider the density from before

$$p(x_1, x_2) = \frac{1}{Z} \exp(-10(x_1^2 - x_2)^2 - (x_2 - \tfrac{1}{4})^4),$$

where the normalizing constant is $Z = 1.1813\ldots$

This time we use the Gibbs sampler. To sample the univariate densities that arise in the process, we use inverse transform sampling. In this case, the explicit algorithm we use is written below.

Fix $x^{(0)} \in \mathbb{R}^2$ and set $x = x^{(0)}$;

For $k = 1, \ldots, N$, do

Calculate $\Phi_1(t) = \int_{-\infty}^{t} p(x_1, x_2)\, \mathrm{d}x_1$;

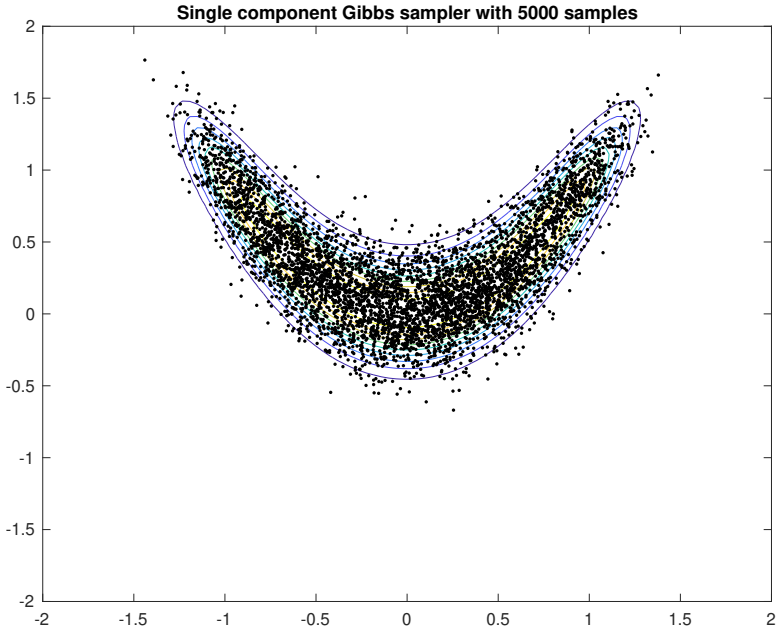
Draw $u \sim \mathcal{U}([0,1])$, set $x_1 = \Phi_1^{-1}(u)$;

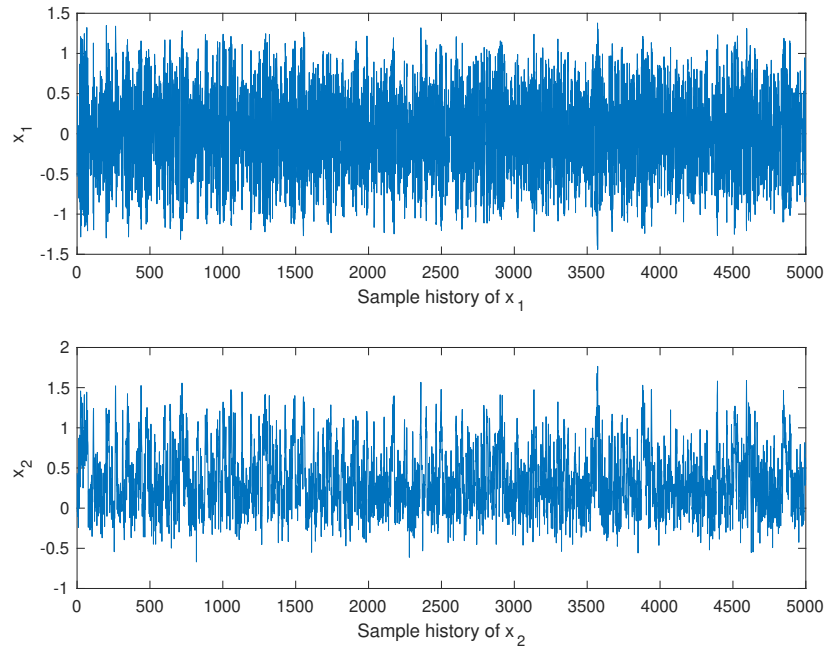
Calculate $\Phi_2(t) = \int_{-\infty}^{t} p(x_1, x_2)\, \mathrm{d}x_2$;

Draw $u \sim \mathcal{U}([0,1])$, set $x_2 = \Phi_2^{-1}(u)$;

Set $x^{(k)} = x$.

End


Single component Gibbs sampler with 5000 samples
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

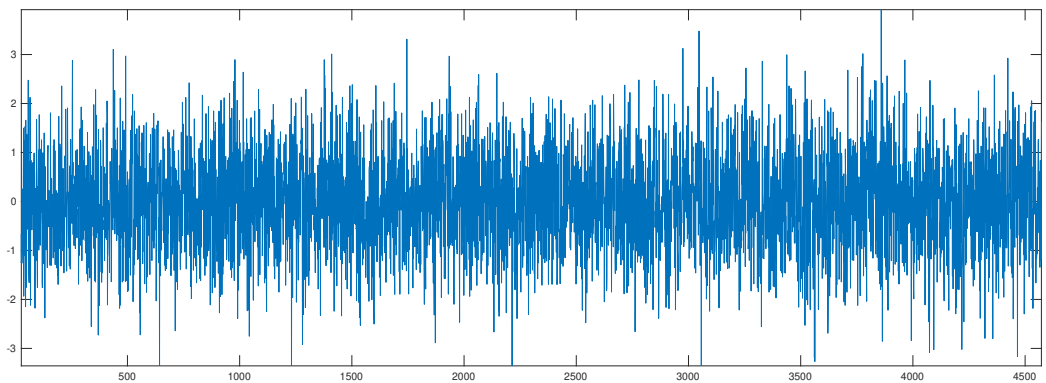

1.5
1
0.5
0
-0.5
-1
-1.5
x1
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Sample history of x 1
2
1.5
1
0.5
0
-0.5
-1
x2
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Sample history of x 2

Computational remarks about MCMC

- As a general rule of thumb, one should aim at roughly 30% acceptance rates when using Gaussian (or close to Gaussian) proposal and target densities with MH.
- It usually takes the Markov chain a number of iterations to reach the steady state. To this end, it is usually advisable to discard the first N_0 obtained samples since they may not be representative of the target distribution—this is the so-called "burn-in" period. The length of the burn-in period varies depending on the application, but one might consider throwing away the first $\sim 5 - 10\%$ steps for a sufficiently large sample size as an example.
- In MH, using a Gaussian kernel (e.g., random walk Metropolis–Hastings) is a popular choice due to the ease of implementation. While it is a safe choice, it does not take into account the form of the posterior density. To increase efficiency, it is advisable to take the shape of the density into account when designing the proposal density. In the high-dimensional setting, this is especially useful if the posterior density is *anisotropic* (stretched in some directions).

Computational remarks about MCMC

- The proposal distribution in MH can also be updated while the sampling algorithm moves around the posterior density. This process is called *adaptation*.
- Visual assessment: we are aiming for independent sample points, where the sample histories look like a "fuzzy worm". One could aim at something like the Gaussian white noise signal below:



- More quantitatively, the independence of consecutive draws can be estimated from the sample itself by computing its (sample-based) autocovariance.

A note on convergence

The convergence of the Metropolis–Hastings and Gibbs sampler algorithms depends on whether they satisfy the ergodicity conditions from before. There are known sufficient conditions concerning the density p that guarantee the ergodicity of these methods. For example, the following proposition gives some relatively general conditions.

Proposition

(a) Let $p\colon \mathbb{R}^d \to \mathbb{R}_+$ and let $q\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ be a candidate-generating kernel. If the Markov chain corresponding to q is aperiodic, then the Metropolis–Hastings chain is also aperiodic. Further, if the Markov chain corresponding to q is irreducible and $\alpha(x, y) > 0$ for all $(x, y) \in E_+ \times E_+$, where $E_+ := \{x \in \mathbb{R}^d \mid p(x) > 0\}$, then the Metropolis–Hastings chain is irreducible.
(b) Let p be a lower semicontinuous density and E_+ as above. The Gibbs sampler defines an irreducible and aperiodic transition kernel if E_+ is connected and each $(d-1)$-dimensional marginal $p(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d) = \int_{\mathbb{R}} p(x)\,\mathrm{d}x_j$ is locally bounded.

Autocovariance and correlation length

The independence of consecutive draws can be estimated from the sample itself. Suppose that we are interested in the convergence of the integral of $G(x)$ with respect to the probability density $p(x)$. Let us denote $z_j = G(x_j)$, where $\{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ is a MCMC sample and let $\overline{z} = N^{-1} \sum_{j=1}^{N} z_j$. Then we define the normalized autocovariance of the sample as

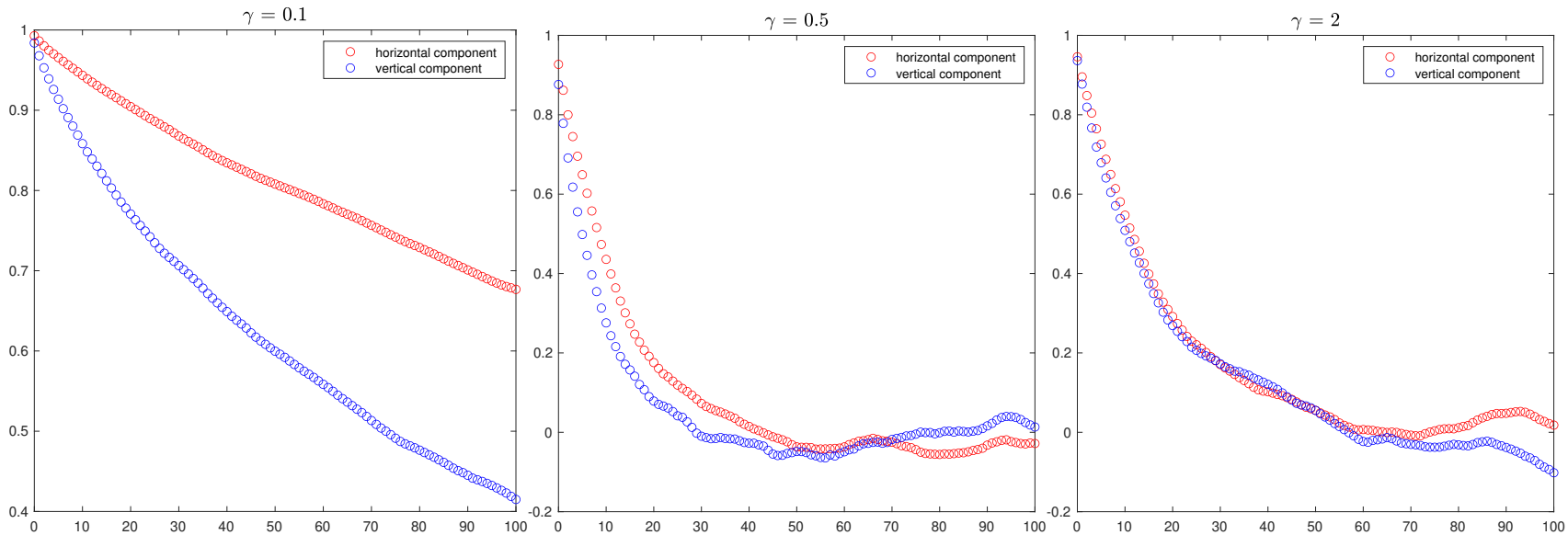
$$\gamma_k = \frac{1}{(N-k)\gamma_0} \sum_{j=1}^{N-k} (z_j - \overline{z})(z_{j+k} - \overline{z}), \quad k \geq 1,$$

where $\gamma_0 = N^{-1} \sum_{j=1}^{N} z_j^2$.

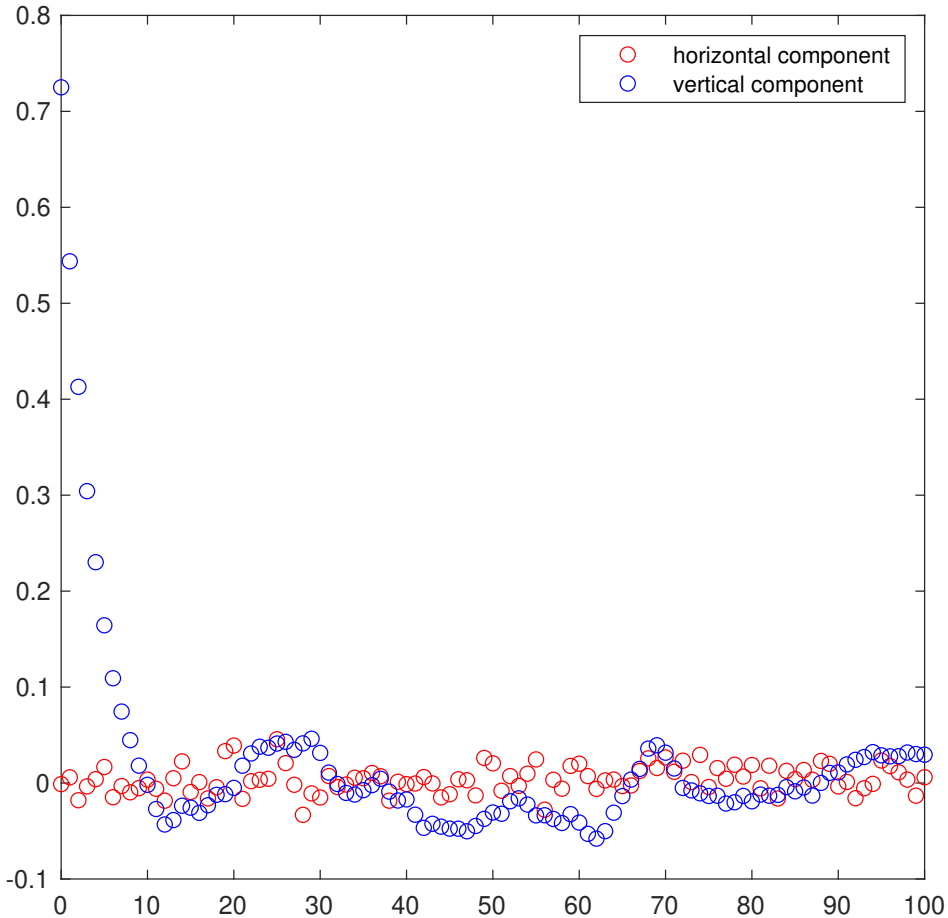
The correlation length of the sample $\{z_j\}_{j=1}^{N}$ can be estimated based on the decay of the normalized autocovariance sequence of the sample.

If every k^{th} sample point is independent, one might expect the discrepancy to behave as $1/\sqrt{N/k} = \sqrt{k/N}$ instead of $1/\sqrt{N}$. In consequence, one should try to choose the proposal distribution so that the *correlation length* is as small as possible.

Normalized autocovariance sequences for the Metropolis–Hastings example



Normalized autocovariance sequences for the Gibbs example



Preconditioned Crank–Nicolson algorithm

- The preconditioned Crank–Nicolson (pCN) algorithm is an instance of the Metropolis–Hastings algorithm with a specially chosen proposal density.
- The proposal is drawn using the model $Y = \sqrt{1-\beta^2}X + \beta W$, where $W \sim \mathcal{N}(0, C_0)$, C_0 is a symmetric and positive definite matrix, with the (*non-symmetric!*) kernel

$$q(x, y) \propto \exp\left(-\frac{1}{2\beta^2}(y - \sqrt{1-\beta^2}x)^{\mathrm{T}} C_0^{-1}(y - \sqrt{1-\beta^2}x) \right).$$

  Here, the step size $0 < \beta < 1$ is a free parameter (which can be optimized for statistical efficiency).
- The pCN method is *dimension robust*: the acceptance probability does not degenerate to zero as the dimension $d \to \infty$. Contrast this with, e.g., random walk Metropolis, whose acceptance probability degenerates to zero as the dimension $d \to \infty$.

Further variations of MCMC

We have only scratched the surface of some basic ideas surrounding MCMC methods. In the literature and practical applications, one can find many variations of these ideas to boost the performance of MCMC for "difficult" / "high-dimensional" problems. To list a couple of notable ones:

- Adaptive Metropolis: as the proposal density $q(x, y)$, use a random walk model $Y = X + W$ with $W \sim \mathcal{N}(0, \Gamma)$, where the covariance Γ is replaced by the *sample covariance* (plus some small perturbation of identity) computed using the sample history. The updating can happen either at every step or after every M steps of the Metropolis iteration. The main theoretical challenge is proving the ergodicity of the chain—this was proved by Haario, Saksman, and Tamminen (2001). Computationally, stable updating formulae for the sample means and covariances are needed in practice.
- Independence Metropolis: as the proposal density $q(x, y)$, use a probability density that is independent of the previous sample x, i.e., $q(x, y) = q(y)$. The proposal density should be similar to the target density.
- Metropolis-within-Gibbs, Delayed rejection adaptive Metropolis, ...

Software: https://mjlaine.github.io/mcmcstat/
https://mc-stan.org/