Stochastic Methods for Optimization and Sampling

M1, TSE

Sampling: MCMC methods

Markov Chains

Markov chains

Let $E = \{1, \dots, d\}$ be a finite set.

Definition (Markov chain)

A sequence of random variables $(X_n)_n$ taking its values in E is a homogeneous Markov chain (or random walk) if for any $n \geq 1$ and any $x_0, \dots, x_n, y \in E$, such that $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) > 0$ we have

$$\mathbb{P}(X_{n+1} = y \,|\, X_0 = x_0, \dots, X_n = x_n) \overset{(1)}{=} \mathbb{P}(X_{n+1} = y \,|\, X_n = x_n) \overset{(2)}{=} \mathbb{P}(X_1 = y \,|\, X_0 = x_n).$$

Remarks: A *Markov chain* (MC) only has to satisfy (1).
A *homogeneous* Markov chain needs to satisfy (1) and (2).

Markov chains

Interpretation: To predict X_{n+1} from $(X_1, \ldots, X_n)$, only the current state X_n is informative, not the past $(X_0, \ldots, X_{n-1})$.

Transition matrix

For any homogeneous Markov chain $(X_n)_n$, we can define a transition matrix:

$$\big(P(x,y)\big)_{x,y\in E}, \qquad \text{where} \qquad P(x,y) = \mathbb{P}(X_1 = y \,|\, X_0 = x).$$

It has the following properties:

1. $\forall x, y \in E : P(x,y) \in [0,1]$.
2. $\forall x \in E : \sum_{y\in E} P(x,y) = 1$.
3. $\forall n \in \mathbb{N}$, $\forall x_0, \ldots, x_n, y \in E$:

$$\begin{aligned}\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) &= \mathbb{P}(X_0 = x_0) \prod_{i=0}^{n-1} \mathbb{P}(X_{i+1} = x_{i+1} \,|\, X_i = x_i) \\ &= \mathbb{P}(X_0 = x_0) \prod_{i=0}^{n-1} P(x_i, x_{i+1}).\end{aligned}$$

Graph of a Markov chain

Any Markov chain on a finite set E can be represented as a graph.

Example: Consider the following transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

We can define the following graph from matrix P

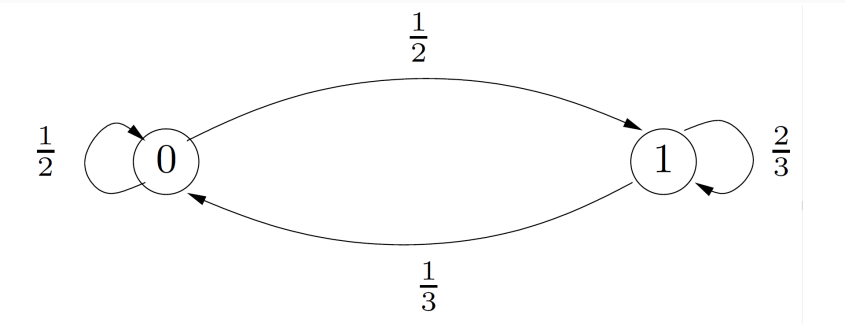

Figure 1: Graph of a Markov chain.

Higher order transition matrices

Let $n \in \mathbb{N}$. The probability to go from x to y in n steps is defined as

$$P^{(n)}(x, y) = \mathbb{P}\left(X_n = y \,|\, X_0 = x\right).$$

We also let $P^{(0)} = I_d$ (identity of size d).

Proposition (Chapman-Kolmogorov equations)

It holds that $P^{(n)} = P^n$ for any $n \in \mathbb{N} \cup \{0\}$.

Chapman-Kolmogorov: Proof

Proof (by induction). It is clear for $n = 0, 1$.

Suppose we have $P^{(n)} = P^n$ for some $n \in \mathbb{N}$. Then for any $x, y \in E$

$$\begin{aligned} P^{(n+1)}(x, y) &= \mathbb{P}(X_{n+1} = y \,|\, X_0 = x) \\ &= \sum_{z \in E} \mathbb{P}(X_{n+1} = y \,|\, X_n = z, X_0 = x) \mathbb{P}(X_n = z \,|\, X_0 = x) \\ &= \sum_{z \in E} P(z, y) P^n(x, z) \\ &= P^{n+1}(x, y), \end{aligned}$$

so $P^{(n+1)} = P^{n+1}$. By induction, we have $P^{(n)} = P^n$ for any $n \in \mathbb{N}$.

Irreducible Markov Chain

Definition

A Markov chain $(X_n)_n$ is irreducible if

$$\forall x, y \in E,\ \exists n \in \mathbb{N} : P^n(x, y) > 0.$$

("We can go from any x to any y in finitely many steps".)

Law of a Markov chain

We define the initial condition

$$\mu_0 = \Big[\mathbb{P}(X_0 = 1), \ldots, \mathbb{P}(X_0 = d)\Big].$$

We also define the law of X_n at time n:

$$\mu_n = \Big[\mathbb{P}(X_n = 1), \ldots, \mathbb{P}(X_n = d)\Big].$$

Corollary

The law of X_n satisfies $\mu_n = \mu_0 P^n$.

Invariant measure

Let $(X_n)_n$ be a Markov chain on E and P be its transition matrix.

Definition (Invariant probability measure)

A probability measure π over E is invariant or stationary for $(X_n)_n$ if

$$\pi P = \pi.$$

Proposition

If (X_n) is irreducible and E is finite, there exists a unique invariant probability measure π, that is, such that $\pi P = \pi$.

Moreover, π is such that $\pi(x) > 0$ for all $x \in E$.

Convergence of irreducible Markov Chains

Theorem (Simplified ergodic theorem)

Let (X_n) be irreducible on a finite set E, and π be its invariant probability measure: $\pi = \pi P$. For any function $\phi : E \to \mathbb{R}$, we have

$$\frac{1}{n}\sum_{k=1}^{n} \phi(X_k) \xrightarrow{\text{a.s.}} \pi\phi \overset{\text{def}}{=} \sum_{x\in E} \pi(x)\phi(x),$$

irrespective of the law of X_0.

Consequence: To estimate π, we can simulate a trajectory $(X_n)_n$:

For any $z \in E$, we can approximate $\pi(z)$ by taking $\phi(x) = \mathbf{1}_{x=z}$

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{X_k=z} \xrightarrow{\text{a.s.}} \sum_{x\in E} \pi(x)\mathbf{1}_{x=z} = \pi(z).$$

Aperiodic Markov chain

Definition

A Markov chain $(X_n)_n$ is *aperiodic* if

$$\forall x \in E : \mathrm{GCD}\left\{n \geq 1 : P^n(x,x) > 0\right\} = 1.$$

where GCD stands for the Greatest Common Divisor.

In contrast, $(X_n)_n$ is *periodic* if for some $x \in E$, going from x back to x must take a multiple of m steps, for some integer $m > 1$.

Convergence of irreducible and aperiodic Markov Chains

Theorem

If $(X_n)_n$ is irreducible and aperiodic on a finite state space E, then there exists a constant C such that, for any initial law μ_0,

$$\left\|\mu_n - \pi\right\|_1 \stackrel{\text{def}}{=} \sum_{j=1}^{d} \left|\mu_n(j) - \pi(j)\right| \leq Cn^{-1}|\lambda|^n,$$

where π denotes the invariant measure of (X_n).

In particular, μ_n converges to π in law (much weaker statement).

Important consequence: If $(X_n)_n$ is irreducible and aperiodic, the law μ_n is always *attracted to* the invariant measure π.

Checking that π satisfies $\pi = \pi P$ is equivalent to finding $\pi = \lim_{n\to\infty} \mu_n$!

Reversible measure

Definition

A probability measure π over E is said to be *reversible* for $(X_n)_n$ if

$$\forall x, y \in E : \quad \pi(x)P(x,y) = \pi(y)P(y,x).$$

Interpretation: Suppose π is an initial measure over E.

- $\pi(x) =$ amount of mass located at state x at time 0.
- $\pi(x)P(x,y) =$ how much mass goes from x to y in one time step.

The measure π is reversible if, starting from π, as much mass goes from x to y as from y to x in one time step, $\forall x, y \in E$.

Reversible = "Time-Reversible"!

Proposition

Let π be an initial reversible measure (i.e. the law of X_0 is π). Then, for any $x_0, \dots, x_n \in E$, we have

$$\mathbb{P}\Big(X_0 = x_0, \dots, X_n = x_n\Big) = \mathbb{P}\Big(X_n = x_0, \dots, X_0 = x_n\Big).$$

In other words, the forward and backward trajectories, $(X_0, \dots, X_n)$ and $(X_n, \dots, X_0)$, have the same law.

Interpretation: When observing a trajectory $X_0, \dots, X_n$, one cannot tell in which direction time goes.

Proof: Time-reversibility

Proof:

$$\mathbb{P}\Big(X_0 = x_0, \ldots, X_n = x_n\Big) = \underbrace{\pi(x_0)P(x_0, x_1)}_{=P(x_1,x_0)\pi(x_1)} \ldots P(x_{n-1}, x_n)$$

$$= P(x_1, x_0)\underbrace{\pi(x_1)P(x_1, x_2)}_{=P(x_2,x_1)\pi(x_2)} \ldots P(x_{n-1}, x_n)$$

$$= \ldots \text{ (induction)}$$

$$= P(x_1, x_0) \ldots P(x_n, x_{n-1})\pi(x_n)$$

$$= \mathbb{P}\Big(X_0 = x_n, \ldots, X_n = x_0\Big).$$

Reversibility $\Longrightarrow$ Stationarity

Theorem

If π is reversible, then π is stationary, i.e. $\pi = \pi P$.

Proof: For any $y \in E$:

$$\begin{aligned} \pi P(y) &= \sum_{x \in E} \pi(x) P(x, y) \\ &= \sum_{x \in E} P(y, x) \pi(y) \qquad \text{(stationarity)} \\ &= 1 \cdot \pi(y) = \pi(y). \end{aligned}$$

Remark: This is useful to show that a *given* measure π is stationary.

(It suffices to check that it is reversible).

Markov chain Monte Carlo (MCMC): Motivation

Framework of MCMC methods

Framework: We aim to sample from a probability density π over $\mathbb{R}^d$

$$\pi(x) = \frac{f(x)}{\displaystyle\int_{\mathbb{R}^d} f(y)dy}.$$

The function $f : \mathbb{R}^d \to \mathbb{R}$ is known, but the normalizing factor

$$\int_{\mathbb{R}^d} f(y)dy$$

is intractable to compute. Thus, we can only evaluate $f(x)$, not $\pi(x)$ (we say that π can only be accessed up to a multiplicative constant).

Goal of MCMC methods

Produce a trajectory $(X_n)_n$ approximating π "well", even if π is only accessible up to an unknown multiplicative constant.

Motivation: Bayesian statistics

Assume we observe a data set $Z_1, \ldots, Z_n \sim \mathbb{P}_{\theta^*}$ *iid* where $\theta^* \in \Theta$.

Goal: Estimate $\theta^* \in \Theta$ in some set of parameters Θ.

Assumptions of the Bayesian framework

- The true parameter θ^* is a random variable itself.
- Its distribution μ, called "prior", is chosen by the statistician.
- $\forall \theta \in \Theta$, the quantity $\mu(\theta)$ is easy to evaluate.
- $\forall\, \theta, z_1, \ldots, z_n$, the likelihood $\mathcal{L}(z_1, \ldots, z_n, \theta)$ is easy to evaluate.

The prior μ is a probability measure over Θ. It represents the statistician's rough initial guess of where θ^* likely is in Θ.

Clever choices of μ can also considerably simplify calculations.

Motivation: Bayesian statistics

Warning:

- μ is a probability measure, not a point in Θ. It takes large values in regions where the statistician thinks θ^* is likely to be located.
- It has to be defined before observing the data $Z_1, \ldots, Z_n$.

Bayesian estimation proceeds as follows:

1. We start from an initial guess μ of the location of θ^*.
2. Observing the data $Z_1, \ldots, Z_n$ reveals information about θ^*: We can incorporate it to refine our initial assessment μ.
3. We thus define a new density $p(\cdot\,|Z_1, \ldots, Z_n)$ over Θ, called "posterior". It represents our updated (and much more precise) guess of the location of θ^* in Θ.

Motivation: Bayesian statistics

The posterior $p(\theta|X_1, \ldots, X_n)$ is defined using the Bayes rule.

Expression of the posterior distribution

$$p(\theta|X_1, \ldots, X_n) = \frac{\mathcal{L}(X_1, \ldots, X_n, \theta)\mu(\theta)}{\int_\Theta \mathcal{L}(X_1, \ldots, X_n, \theta')\mu(\theta')d\theta'}$$

The posterior is crucial: It is our final assessment of where θ is.

The numerator is easy to evaluate by assumption. Unfortunately, the denominator is often intractable.

We often only know the posterior up to an uncomputable constant. We typically need to approximate it using MCMC methods.

MCMC algorithms

Framework: Sample from π known up to a multiplicative constant (i.e. for any x, we can only evaluate $C\pi(x)$ for an unknown C).

Algorithms: We will analyze two important MCMC algorithms:

1. Metropolis-Hastings
2. Gibbs sampler.

Output: They both output a sequence $(X_n)_n$ approximating π well.

⚠ $(X_n)_n$ are non-iid, but form a M.C. with invariant distribution π.

By the ergodic theorem, π can be approximated using $(X_n)_n$ if it is irreducible, regardless of the initialization X_0.

Metropolis-Hastings (MH) algorithm

Metropolis-Hastings (M.H.) algorithm

Let π be a density over E known up to a multiplicative constant.

Strategy: Construct a M.C. $(X_n)_n$ with invariant measure π.

Remark: The end-goal is to handle densities π defined on $E = \mathbb{R}^d$.

However, the M.H. algorithm is easier to understand if E is finite.

For simplicity, we first consider a finite E and will generalize the M.H. algorithm to $E = \mathbb{R}^d$ afterward.

Metropolis-Hastings (M.H.) algorithm on a finite space

In this section, we fix a finite space E.

Goal: Construct a Markov chain $(X_n)_n$ with invariant measure π.

We need an auxiliary transition matrix $Q = (Q(x,y))_{x,y\in E}$ satisfying

$$\forall x, y \in E: \quad \Big[Q(x,y) > 0 \iff Q(y,x) > 0 \Big].$$

We use Q to generate samples, and accept or reject them using $C\pi$.

Remarks:

1. Q can be arbitrary and is chosen by the statistician.
2. Simulating a M.C. with any transition matrix Q is easy (why?).

Metropolis-Hastings algorithm

1. Simulate $X_0 \in E$ according to any initial distribution.
2. Until termination condition, iterate:

   Draw $Y \sim Q(X_k, \cdot)$

   Compute $a = \min\left(\frac{\pi(Y)\, Q(Y, X_k)}{\pi(X_k) Q(X_k, Y)}, 1\right)$.

   Draw $U_k \sim \text{Unif}[0,1]$ independent of the past.

   Update : $X_{k+1} = \begin{cases} Y & \text{if } U_k \leq a \\ X_k & \text{if } U_k > a. \end{cases}$
3. Ouput $(X_0, \ldots, X_n)$.

Remarks

Remarks:

1. The algorithm only uses the ratio $\frac{\pi(x)}{\pi(y)}$. Hence, it is applicable even if π is known up to a multiplicative constant (as this constant cancels out)!

2. Using matrix Q, we generate a new candidate Y for the next step. We choose to move to Y or stay on X_k based on π and Q.

3. If Q is symmetric, $a = \min\left(\frac{\pi(Y)}{\pi(X_k)}, 1\right)$. Thus, we jump to Y w.p. $a = 1$ if $\pi(y) > \pi(X_k)$, i.e. if Y is more likely than X_k under π.

4. The sequence $(X_n)_n$ can be piecewise constant, for instance

$$(X_0, X_0, X_0, X_3, X_4, X_4, X_4, X_7, \ldots, X_n).$$

Convergence of the MH algorithm

Theorem

Suppose $\pi(x) > 0$, $\forall x$, and $Q(x,y) > 0 \iff Q(y,x) > 0$, $\forall x, y$.

Let $(X_n)_n$ be the output of M.H. with inputs Q and $C\pi$.

1. The sequence of random variables $(X_n)_n$ forms a Markov chain.

   Its transition matrix $P = \big(P(x,y)\big)_{x,y\in E}$ is given by

$$P(x,y) = \begin{cases} Q(x,y)\min\left(\dfrac{\pi(y)\,Q(y,x)}{\pi(x)\,Q(x,y)}, 1\right) & \text{if } x \neq y \\ 1 - \sum_{z\neq x} P(x,z) & \text{if } x = y. \end{cases}$$

2. π is an invariant probability measure for $(X_n)_n$.

3. If Q is irreducible, then so is P.

Convergence of the MH algorithm

Item 2 and 3 above ensure the convergence of the empirical measure (histogram if E is finite) of $(X_n)_n$ toward π by the ergodic theorem.

Proof of 1.

$$
\begin{aligned}
&\mathbb{P}(X_{n+1} = y | X_n = x) \\
&= \mathbb{P}\Big(Y_{n+1} = y \text{ and } U_{n+1} \leq \left(\frac{\pi(Y_{n+1})Q(Y_{n+1}, X_n)}{\pi(X_n)Q(X_n, Y_{n+1})}\right) \Big| X_n = x\Big) \\
&= \mathbb{P}\Big(Y_{n+1} = y \text{ and } U_{n+1} \leq \left(\frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}\right) \Big| X_n = x\Big) \\
&= \min\left(1, \left(\frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}\right)\right) Q(x, y).
\end{aligned}
$$

where we used the fact that U_{n+1} is independent of X_n and Y_{n+1}.

Convergence of the MH algorithm

Proof of 2.

- Suppose first that $Q(x, y) = Q(y, x) = 0$. Then for $x \neq y$, we have $\pi(x)P(x, y) = 0 = \pi(y)P(y, x)$
- Now, if $Q(x, y), Q(y, x) > 0$, we have

$$\begin{aligned}
\pi(x)P(x, y) &= \pi(x)Q(x, y) \min\left(1, \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)}\right) \\
&= \min\Big(\pi(x)Q(x, y), \pi(y)Q(y, x)\Big) \\
&= \pi(y)Q(y, x) \min\left(\frac{\pi(x)Q(x, y)}{\pi(y)Q(y, x)}, 1\right) \\
&= \pi(y)P(y, x).
\end{aligned}$$

In both cases, π is reversible, hence invariant.

Convergence of the MH algorithm

Proof of 3.

By definition of P, we have $P(x, y) > 0$ as soon as $Q(x, y) > 0$.

Any possible path under Q is possible under P.

Therefore, is Q is irreducible, then so is P.

MCMC methods on infinite state spaces

Transition kernels

Let $E = \mathbb{R}^d$. Transition matrices are replaced with *transition kernels*.

For simplicity, we restrict to transition kernel with a density over $\mathbb{R}^d$.

Definition (Transition kernel with a density over $\mathbb{R}^d$)

A function $P : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a *transition kernel with a density* if

1. $\forall x \in \mathbb{R}^d$, $P(x, \cdot)$ is a probability density function over $\mathbb{R}^d$,
2. For any $y \subseteq \mathbb{R}^d$, $x \mapsto P(x, y)$ is a measurable function.

Link with transition matrices:

On a finite E, $P(x, \cdot)$ was a probability vector over E for any $x \in E$.

Transition kernels with a density

A transition kernel P defines a Markov chain over $\mathbb{R}^d$:

Markov chain with transition kernel P

1. Draw Z_0 according to some measure μ_0 over $\mathbb{R}^d$.
2. Given the current Z_n, draw Z_{n+1} from the density $P(Z_n, \cdot)$.

Example (Gaussian transition kernel).

For any $x \in \mathbb{R}^d$, take $P(x, \cdot)$ as the density of $N(x, \sigma^2 I_d)$ where $\sigma > 0$:

$$P(x, y) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\Big(- \frac{\|x - y\|^2}{2\sigma^2} \Big).$$

This defines the Markov recurrence

$$Z_{n+1} = Z_n + \xi_{n+1} \qquad \text{where} \quad \xi_i \sim N(0, \sigma^2 I_d) \text{ are iid.}$$

Transition kernel for Metropolis-Hastings

Let π be a probability *density* over $\mathbb{R}^d$

Let Q be an auxiliary density transition kernel over $\mathbb{R}^d$, satisfying:

Conditions on Q

1. $\forall x \in \mathbb{R}^d$, sampling from the density $Q(x, \cdot)$ is easy,
2. $\forall x, y \in \mathbb{R}^d : \Big[Q(x, y) > 0 \iff Q(y, x) > 0\Big]$,
3. $\forall x, y \in \mathbb{R}^d$ it is easy to evaluate

$$a(x, y) = \frac{\pi(x)Q(x, y)}{\pi(y)Q(y, x)}.$$

The transition kernel Q is chosen by the statistician.

We generate proposals from Q and accept/reject based on Q and π.

Metropolis-Hastings algorithm on continuous spaces

The algorithm is identical to the discrete case!

Metropolis-Hastings algorithm

1. Simulate $X_0 \in E$ according to any initial distribution.
2. Until termination condition, iterate:

   Draw $Y \sim Q(X_k, \cdot)$

   Compute $a = \min\left(\frac{\pi(Y)\, Q(Y, X_k)}{\pi(X_k) Q(X_k, Y)}, 1\right)$.

   Draw $U_k \sim \text{Unif}[0,1]$ independent of the past.

   Update : $X_{k+1} = \begin{cases} Y & \text{if } U_k \leq a \\ X_k & \text{if } U_k > a. \end{cases}$
3. Ouput $(X_0, \ldots, X_n)$.

Some important special cases

Special case 1: Random walk Metropolis-Hastings:

Case where there exists a density function $g : \mathbb{R}^d \to \mathbb{R}$ such that

$$Q(x, y) = g(y - x), \qquad \forall x, y \in \mathbb{R}^d.$$

(In words, $Q(x, \cdot)$ is the density g shifted by x).

In this case, a Markov chain associated with Q is defined by

$$Z_{k+1} = Z_k + \xi_{k+1}$$

where $\xi_1, \ldots, \xi_n, \ldots \overset{iid}{\sim} g$ are iid. This is a random walk.

The Metropolis-Hastings algorithm simplifies as follows.

Random walk Metropolis-Hastings algorithm

Input: density g, initial law μ_0, transition kernel Q, target density π.

1. Simulate $X_0 \in E$ according to any initial distribution.
2. Until termination condition, iterate:

   Draw $\xi_{k+1} \sim g$ indep of the past, and set $Y = X_k + \xi_{k+1}$

   Compute $a = \min\left(\dfrac{\pi(Y)g(-\xi_{k+1})}{\pi(X_k)\,g(\xi_{k+1})}, 1\right)$.

   Draw $U_k \sim \mathrm{Unif}[0,1]$ independent of the past.

   Update : $X_{k+1} = \begin{cases} Y & \text{if } U_k \leq a \\ X_k & \text{if } U_k > a. \end{cases}$
3. Ouput $(X_0, \ldots, X_n)$.

Important special cases

Special case 2: Symmetric kernel

Suppose Q is symmetric, i.e. $Q(x, y) = Q(y, x)$ for any $x, y \in \mathbb{R}^d$.

Then the acceptance ratio simplifies as

$$a(x, y) = \frac{\pi(y)Q(y, x)}{\pi(x)Q(x, y)} = \frac{\pi(y)}{\pi(x)}.$$

In this case, the Metropolis-Hastings algorithm is written as follows.

Metropolis-Hastings with symmetric kernel

Suppose Q is symmetric, i.e. $Q(x, y) = Q(y, x)$ for any $x, y \in \mathbb{R}^d$.

Metropolis-Hastings algorithm with symmetric kernel

1. Simulate $X_0 \in E$ according to any initial distribution.
2. Until termination condition, iterate:

   Draw $Y \sim Q(X_k, \cdot)$

   Compute $a = \min\left(\frac{\pi(Y)}{\pi(X_k)}, 1\right)$.

   Draw $U_k \sim \text{Unif}[0, 1]$ independent of the past.

   Update : $X_{k+1} = \begin{cases} Y & \text{if } U_k \leq a \\ X_k & \text{if } U_k > a. \end{cases}$
3. Ouput $(X_0, \ldots, X_n)$.

Gibbs sampler

Gibbs sampler

We will see 2 versions of the Gibbs sampler:

1. Random scan Gibbs sampler,
2. Deterministic scan Gibbs sampler.

The random scan Gibbs sampler is just a special case of the Metropolis-Hastings algorithm where all transitions are accepted!

Gibbs sampler

Notation: For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we define

$$x_{-\ell} = \big(x_1, \ldots, x_{\ell-1}, x_{\ell+1}, \ldots, x_d\big).$$

Suppose we aim to sample from a probability density π. We write

1. $\pi(\,\cdot\,|x_{-\ell})$ for the density conditional on $x_{-\ell}$ (i.e. $x_{-\ell}$ is frozen).
2. $\pi(x_{-\ell})$ for the joint density of the $d-1$ coordinates $x_{-\ell}$.

The Gibbs sampler operates under the following assumption

Condition on π

For any $x \in \mathbb{R}^d$ and any $\ell \in \{1, \ldots, d\}$:

It is easy to sample from the conditional density $\pi(\,\cdot\,|x_{-\ell})$.

Random scan Gibbs sampler

Random scan Gibbs sampler

1. Draw $X_0 \in \mathbb{R}^d$ according to some initial measure μ_0
2. Until termination condition, iterate:

   Set $x = X_k$

   Draw $\ell \sim \text{Unif}(\{1, \dots, d\})$ independent of the past

   Draw $x'_\ell \sim \pi(\,\cdot\,|x_{-\ell})$

   Define $X_{k+1} = (x_1, \dots, x_{\ell-1}, x'_\ell, x_{\ell+1}, \dots, x_d)$

Output $(X_0, \dots, X_n)$.

Remark: The notation x above hides a dependency on k as $x = X_k$.

Systematic scan Gibbs sampler

Systematic scan Gibbs sampler

1. Draw $X_0 \in \mathbb{R}^d$ according to some initial measure μ_0
2. Until termination condition, iterate:

   Set $x = X_k$

   For $\ell = 1, \ldots, d$:

   Draw $x'_\ell \sim \pi\big(\ \cdot\ \big|\ (x'_1, \ldots, x'_{\ell-1}, x_{\ell+1}, \ldots x_d)\big)$

   Define $X_{k+1} = x' = (x'_1, \ldots, x'_d)$.

Output $(X_0, \ldots, X_n)$.

Remark: The notation x above hides a dependency on k as $x = X_k$.

Reversibility of π for the Gibbs sampler

Proposition

Let $(X_n)_n$ be the output of the random scan Gibbs sampler. Then π is reversible, hence stationary, for $(X_n)_n$.

Remark: The stationarity of π is not sufficient to obtain convergence of a Markov chain toward π.

We further need irreducibility and aperiodicity (the formal definitions for infinite state spaces are out of the course's scope).

Proof

With the Metropolis-Hastings notation, we set $Y = (x'_\ell, x_{-\ell})$.

Let also $x \in \mathbb{R}^d$ and $y = (x'_\ell, x_{-\ell})$.

The probability of going to y starting from x is

$$Q(x, y) = \frac{1}{d}\pi(x'_\ell | x_{-\ell}) = \frac{1}{d}\frac{\pi(y)}{\pi(x_{-\ell})}.$$

Similarly $Q(y, x) = \frac{1}{d}\frac{\pi(x)}{\pi(x_{-\ell})}$. The reversibility of π under Q follows:

$$\pi(x)Q(x, y) = \frac{1}{d}\frac{\pi(x)\pi(y)}{\pi(x_{-\ell})} = \pi(y)Q(y, x).$$

It also follows that the acceptance ratio is

$$a(x, y) = \frac{\pi(y)\frac{\pi(x)}{d\cdot\pi(x_{-\ell})}}{\pi(x)\frac{\pi(y)}{d\cdot\pi(x_{-\ell})}} = 1.$$