Advanced Computational Methods in Statistics
Lecture 2

O. Deniz Akyildiz

LTCC Advanced Course

October 14, 2024

IMPERIAL

https://akyildiz.me/

𝕏: @odakyildiz

Recall our basic task:

Recall our basic task:

- We want to sample from a distribution $\pi(x) \propto \gamma(x)$ given only the knowledge of $\gamma(x)$.

Recall our basic task:

- We want to sample from a distribution $\pi(x) \propto \gamma(x)$ given only the knowledge of $\gamma(x)$.
- We want to use these samples to estimate an integral

$$(\varphi, \pi) = \int \varphi(x)\pi(x)\,\mathrm{d}x$$

Last week, we have covered the basic sampling techniques:

Last week, we have covered the basic sampling techniques:

- Uniform random number generation
  - Linear congruential generators

Last week, we have covered the basic sampling techniques:

- Uniform random number generation
  - Linear congruential generators
- Inversion (inverse transform) sampling
  - $U \sim \mathcal{U}(0,1)$
  - $X = F^{-1}(U)$

Last week, we have covered the basic sampling techniques:

- Uniform random number generation
  - Linear congruential generators
- Inversion (inverse transform) sampling
  - $U \sim \mathcal{U}(0, 1)$
  - $X = F^{-1}(U)$
- Rejection sampling
  - $X' \sim q(x)$
  - Accept X' with probability $\gamma(X')/Mq(X')$

Last week, we have covered the basic sampling techniques:

- Uniform random number generation
  - Linear congruential generators
- Inversion (inverse transform) sampling
  - $U \sim \mathcal{U}(0,1)$
  - $X = F^{-1}(U)$
- Rejection sampling
  - $X' \sim q(x)$
  - Accept X' with probability $\gamma(X')/Mq(X')$
- Importance sampling
  - Sample $X_1, \ldots, X_N \sim q(x)$
  - Estimate $(\varphi, \pi) \approx \sum_{i=1}^{N} \varphi(X_i)\mathrm{w}_i$,

Last week, we have covered the basic sampling techniques:

- Uniform random number generation
  - Linear congruential generators
- Inversion (inverse transform) sampling
  - $U \sim \mathcal{U}(0,1)$
  - $X = F^{-1}(U)$
- Rejection sampling
  - $X' \sim q(x)$
  - Accept X' with probability $\gamma(X')/Mq(X')$
- Importance sampling
  - Sample $X_1, \ldots, X_N \sim q(x)$
  - Estimate $(\varphi, \pi) \approx \sum_{i=1}^{N} \varphi(X_i)\mathrm{w}_i$,

The code is also available for these parts:

```
https://akyildiz.me/advanced-computational-statistics
```

OK, so what is wrong with these methods?

The curse of dimensionality

Rejection sampling as $d \to \infty$

Let us exemplify a few issues. Consider the following target distribution on $\mathbb{R}^d$:

$$\pi(x) = \frac{1}{\sigma_\pi^d (2\pi)^{d/2}} \exp\left(-\frac{1}{2\sigma_\pi^2} \|x\|^2 \right)$$

and the following proposal distribution:

$$q(x) = \frac{1}{\sigma_q^d (2\pi)^{d/2}} \exp\left(-\frac{1}{2\sigma_q^2} \|x\|^2 \right)$$

where $\sigma_q > \sigma_\pi$.

The curse of dimensionality

Rejection sampling as $d \to \infty$

We know that the acceptance probability is

$$\alpha(x) = \frac{\pi(x)}{Mq(x)}.$$

Mini-quiz: How do we choose M?

The curse of dimensionality

Rejection sampling as $d \to \infty$

We know that the acceptance probability is

$$\alpha(x) = \frac{\pi(x)}{Mq(x)}.$$

Mini-quiz: How do we choose M?

$$M = \sup_{x \in \mathbb{R}^d} \frac{\pi(x)}{q(x)}.$$

The curse of dimensionality

Rejection sampling as $d \to \infty$

We know that the acceptance probability is

$$\alpha(x) = \frac{\pi(x)}{Mq(x)}.$$

Mini-quiz: How do we choose M?

$$M = \sup_{x \in \mathbb{R}^d} \frac{\pi(x)}{q(x)}.$$

Then, we can write

$$\begin{aligned} M &= \sup_{x \in \mathbb{R}^d} \frac{\sigma_q}{\sigma_\pi} \exp\left(-\frac{1}{2\sigma_\pi^2} \|x\|^2 + \frac{1}{2\sigma_q^2} \|x\|^2 \right) \\ &= \frac{\sigma_q^d}{\sigma_\pi^d} \sup_{x \in \mathbb{R}^d} \exp\left(\frac{\sigma_\pi^2 - \sigma_q^2}{2\sigma_q^2 \sigma_\pi^2} \|x\|^2 \right) = \frac{\sigma_q^d}{\sigma_\pi^d}. \end{aligned}$$

The curse of dimensionality

Rejection sampling as $d \to \infty$

Mini-quiz: Given M, what is the acceptance rate?

The curse of dimensionality

Rejection sampling as $d \to \infty$

Mini-quiz: Given M, what is the acceptance rate?

$$\hat{a} = \frac{1}{M} = \frac{\sigma_\pi^d}{\sigma_q^d}.$$

This means that as $d \to \infty$, given $\sigma_q > \sigma_\pi$, $\hat{a} \to 0$.

The curse of dimensionality for rejection samplers.

The curse of dimensionality

Importance sampling as $d \to \infty$

In standard Monte Carlo methods course, you would hear things like

The curse of dimensionality

Importance sampling as $d \to \infty$

In standard Monte Carlo methods course, you would hear things like

- Monte Carlo estimators are independent of the dimension of the problem.

The curse of dimensionality

Importance sampling as $d \to \infty$

In standard Monte Carlo methods course, you would hear things like

- Monte Carlo estimators are independent of the dimension of the problem.
- Importance sampling estimators are also independent of the dimension of the problem.

The curse of dimensionality

Importance sampling as $d \to \infty$

In standard Monte Carlo methods course, you would hear things like

- Monte Carlo estimators are independent of the dimension of the problem.
- Importance sampling estimators are also independent of the dimension of the problem.

These are false statements.

Importance sampling estimators also suffer badly as $d \to \infty$ (Li et al., 2005).

This motivates us to move on to our next topic: Markov chain Monte Carlo methods.

- In both high-dimensional sampling and more generally generative modelling, techniques based on MCMC and similar ideas are the state-of-the-art.

This motivates us to move on to our next topic: Markov chain Monte Carlo methods.

- In both high-dimensional sampling and more generally generative modelling, techniques based on MCMC and similar ideas are the state-of-the-art.
- Of course, there are many other techniques that are used in practice, but MCMC is the most popular one.

This motivates us to move on to our next topic: Markov chain Monte Carlo methods.

- In both high-dimensional sampling and more generally generative modelling, techniques based on MCMC and similar ideas are the state-of-the-art.
- Of course, there are many other techniques that are used in practice, but MCMC is the most popular one.

Next up: Introducing Markov chains.

What is a Markov chain?

A discrete-time Markov chain is a sequence of random variables $X_0, X_1, \ldots$ such that:

What is a Markov chain?

A discrete-time Markov chain is a sequence of random variables $X_0, X_1, \ldots$ such that:

- X_t depends only on X_{t-1}

What is a Markov chain?

A discrete-time Markov chain is a sequence of random variables $X_0, X_1, \ldots$ such that:

- X_t depends only on X_{t-1}
- In other words, X_t is conditionally independent of $X_0, \ldots, X_{t-2}$ given X_{t-1}.

What is a Markov chain?

A discrete-time Markov chain is a sequence of random variables $X_0, X_1, \ldots$ such that:

- X_t depends only on X_{t-1}
- In other words, X_t is conditionally independent of $X_0, \ldots, X_{t-2}$ given X_{t-1}.

The evolution of the chain is governed by:

What is a Markov chain?

A discrete-time Markov chain is a sequence of random variables $X_0, X_1, \ldots$ such that:

- X_t depends only on X_{t-1}
- In other words, X_t is conditionally independent of $X_0, \ldots, X_{t-2}$ given X_{t-1}.

The evolution of the chain is governed by:

- A transition matrix M (discrete case)

What is a Markov chain?

A discrete-time Markov chain is a sequence of random variables $X_0, X_1, \ldots$ such that:

- X_t depends only on X_{t-1}
- In other words, X_t is conditionally independent of $X_0, \ldots, X_{t-2}$ given X_{t-1}.

The evolution of the chain is governed by:

- A transition matrix M (discrete case)
- A transition kernel K (continuous case)

What is a Markov chain?

A discrete-time Markov chain is a sequence of random variables $X_0, X_1, \ldots$ such that:

- X_t depends only on X_{t-1}
- In other words, X_t is conditionally independent of $X_0, \ldots, X_{t-2}$ given X_{t-1}.

The evolution of the chain is governed by:

- A transition matrix M (discrete case)
- A transition kernel K (continuous case)

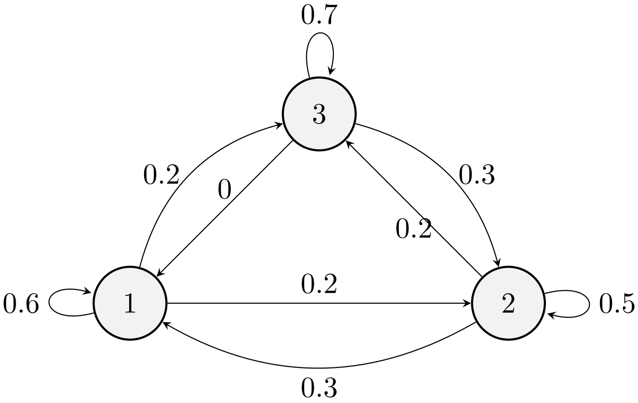
Let us denote our state-space with X.

What is a Markov chain?

Example 1: Simulate a discrete Markov chain

Consider the transition matrix:

$$M = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0.3 & 0.7 \end{bmatrix}, \qquad \text{where } \mathrm{X} = \{1, 2, 3\}.$$



What is a Markov chain?

Example 1: Simulate a discrete Markov chain – What does the matrix M mean?

M	$X_t = 1$	$X_t = 2$	$X_t = 3$
$X_{t-1} = 1$	0.6	0.2	0.2
$X_{t-1} = 2$	0.3	0.5	0.2
$X_{t-1} = 3$	0	0.3	0.7

Example: Given $X_0 = 1$, how to simulate this chain?

Sample:

$$X_t | X_t = x_{t-1} \sim \text{Discrete}(M_{x_{t-1},\cdot}).$$

What is a Markov chain?

Example 1: Simulate a discrete Markov chain – What does the matrix M mean?

M	$X_t = 1$	$X_t = 2$	$X_t = 3$
$X_{t-1} = 1$	0.6	0.2	0.2
$X_{t-1} = 2$	0.3	0.5	0.2
$X_{t-1} = 3$	0	0.3	0.7

Example: Given $X_0 = 1$, how to simulate this chain?

Sample:

$$X_t | X_t = x_{t-1} \sim \text{Discrete}(M_{x_{t-1},\cdot}).$$

Simulation!

What is a Markov chain?

The discrete case: The evolution of the density of the chain

Let $p_0(i) = \mathbb{P}(X_0 = i)$ for $i \in \mathrm{X}$.

What is a Markov chain?

The discrete case: The evolution of the density of the chain

Let $p_0(i) = \mathbb{P}(X_0 = i)$ for $i \in \mathrm{X}$. Then, the density of the chain at time n is given by:

$$\begin{aligned} p_n(i) &= \mathbb{P}(X_n = i) \\ &= \sum_k \mathbb{P}(X_n = i, X_{n-1} = k) \\ &= \sum_k \mathbb{P}(X_n = i | X_{n-1} = k)\mathbb{P}(X_{n-1} = k) \\ &= \sum_k M_{ki} p_{n-1}(k). \end{aligned}$$

What is a Markov chain?

The discrete case: The evolution of the density of the chain

Let $p_0(i) = \mathbb{P}(X_0 = i)$ for $i \in \mathrm{X}$. Then, the density of the chain at time n is given by:

$$\begin{aligned} p_n(i) &= \mathbb{P}(X_n = i) \\ &= \sum_k \mathbb{P}(X_n = i, X_{n-1} = k) \\ &= \sum_k \mathbb{P}(X_n = i | X_{n-1} = k)\mathbb{P}(X_{n-1} = k) \\ &= \sum_k M_{ki} p_{n-1}(k). \end{aligned}$$

This implies that

$$p_n = p_{n-1} M.$$

What is a Markov chain?

The discrete case: The evolution of the density of the chain

Let $p_0(i) = \mathbb{P}(X_0 = i)$ for $i \in \mathrm{X}$. Then, the density of the chain at time n is given by:

$$\begin{aligned} p_n(i) &= \mathbb{P}(X_n = i) \\ &= \sum_k \mathbb{P}(X_n = i, X_{n-1} = k) \\ &= \sum_k \mathbb{P}(X_n = i | X_{n-1} = k)\mathbb{P}(X_{n-1} = k) \\ &= \sum_k M_{ki} p_{n-1}(k). \end{aligned}$$

This implies that

$$p_n = p_{n-1} M.$$

Therefore,

$$p_n = p_0 M^n.$$

Properties of Markov chains

What do we need?

We need Markov chains

Properties of Markov chains

What do we need?

We need Markov chains

- With invariant distributions
- Their convergence is ensured
- Their invariant distribution is unique

Properties of Markov chains

What do we need?

We need Markov chains

- With invariant distributions
- Their convergence is ensured
- Their invariant distribution is unique

We will now review the properties which ensure these in discrete space case.

Properties of Markov chains

Irreducibility

For two states, $x, x' \in \mathrm{X}$, we write $x \rightsquigarrow x'$ if there is a path from x to x':

$$\exists n > 0, \text{ s.t. }, \mathbb{P}(X_n = x' | X_0 = x) > 0.$$

Properties of Markov chains

Irreducibility

For two states, $x, x' \in \mathrm{X}$, we write $x \rightsquigarrow x'$ if there is a path from x to x':

$$\exists n > 0, \text{ s.t. }, \mathbb{P}(X_n = x' | X_0 = x) > 0.$$

If $x \rightsquigarrow x'$ and $x' \rightsquigarrow x$, then we say that x and x' *communicate*.

Properties of Markov chains

Irreducibility

For two states, $x, x' \in \mathrm{X}$, we write $x \rightsquigarrow x'$ if there is a path from x to x':

$$\exists n > 0, \text{ s.t. }, \mathbb{P}(X_n = x' | X_0 = x) > 0.$$

If $x \rightsquigarrow x'$ and $x' \rightsquigarrow x$, then we say that x and x' *communicate*.

A communication class $C \subset \mathrm{X}$ is a set of states such that $x \in C$ and $x' \in C$ if and only if $x \rightsquigarrow x'$ and $x' \rightsquigarrow x$.

Properties of Markov chains

Irreducibility

For two states, $x, x' \in \mathsf{X}$, we write $x \rightsquigarrow x'$ if there is a path from x to x':

$$\exists n > 0, \text{ s.t. }, \mathbb{P}(X_n = x' | X_0 = x) > 0.$$

If $x \rightsquigarrow x'$ and $x' \rightsquigarrow x$, then we say that x and x' *communicate*.

A communication class $C \subset \mathsf{X}$ is a set of states such that $x \in C$ and $x' \in C$ if and only if $x \rightsquigarrow x'$ and $x' \rightsquigarrow x$.

A chain is irreducible if X is a single communication class.

Properties of Markov chains

Recurrence and transience

A Markov chain is recurrent if every state is to be visited infinitely often.

Properties of Markov chains

Recurrence and transience

A Markov chain is recurrent if every state is to be visited infinitely often.

Define the return time:

$$\tau_i = \inf\{n \geq 0 : X_n = i\}.$$

We say that the state is recurrent if

$$\mathbb{P}(\tau_i < \infty | X_1 = i) = 1.$$

If a state is not recurrent, it is transient.

Properties of Markov chains

Positive and null recurrence

We say that a state *i* is positively recurrent if

$$\mathbb{E}[\tau_i | X_1 = i] < \infty.$$

Properties of Markov chains

Positive and null recurrence

We say that a state *i* is positively recurrent if

$$\mathbb{E}[\tau_i | X_1 = i] < \infty.$$

If a recurrent state is not positive recurrent, it is null recurrent.

Properties of Markov chains

Invariant distribution

A probability mass function π is called M-invariant if

$$\pi(i) = \sum_j M_{ij}\pi(j).$$

Properties of Markov chains

Invariant distribution

A probability mass function π is called M-invariant if

$$\pi(i) = \sum_j M_{ij}\pi(j).$$

Equivalently

$$\pi = \pi M.$$

Properties of Markov chains

Existence and uniqueness of the invariant distribution

Theorem 1

If M is irreducible, then M has a unique invariant distribution if and only if it is positive recurrent.

Properties of Markov chains

Existence and uniqueness of the invariant distribution

Theorem 1

If M is irreducible, then M has a unique invariant distribution if and only if it is positive recurrent.

This is existence, we do not talk about convergence yet.

Properties of Markov chains

Reversibility and detailed balance

We will define reversibility through detailed balance condition.

Properties of Markov chains

Reversibility and detailed balance

We will define reversibility through detailed balance condition.

A Markov transition matrix M is reversible w.r.t. π if and only if for all $i, j \in \mathrm{X}$,

$$\pi(i) M_{ij} = \pi(j) M_{ji}.$$

Properties of Markov chains

Reversibility and detailed balance

We will define reversibility through detailed balance condition.

A Markov transition matrix M is reversible w.r.t. π if and only if for all $i, j \in \mathrm{X}$,

$$\pi(i)M_{ij} = \pi(j)M_{ji}.$$

This is called the detailed balance condition (we will discuss the continuous version)

Constructing a chain with stationary distribution π is ensured if detailed balance is satisfied since it implies $\pi = \pi M$.

Properties of Markov chains

Ergodicity

We have seen how to construct chains with invariant distributions π.

Properties of Markov chains

Ergodicity

We have seen how to construct chains with invariant distributions π.

However, the convergence of the chain $p_n \to \pi$ requires one more ingredient: ergodicity.

Properties of Markov chains

Ergodicity

We have seen how to construct chains with invariant distributions π.

However, the convergence of the chain $p_n \to \pi$ requires one more ingredient: ergodicity.

For this, we need a final ingredient: aperiodicity.

Properties of Markov chains

Aperiodicity

A state *i* is called aperiodic if

$$\{n > 0 : \mathbb{P}(X_{n+1} = i | X_1 = i) > 0\}$$

has no common divisor other than 1.

Properties of Markov chains

Ergodicity

Definition 2

An irreducible Markov chain is called ergodic if it is positive recurrent and aperiodic.

Properties of Markov chains

Ergodicity

Definition 2

An irreducible Markov chain is called ergodic if it is positive recurrent and aperiodic.

Ergodicity brings us the missing ingredient for the convergence: We can now ensure p_n to converge to π.

Properties of Markov chains

Ergodicity

If $(X_n)_{n\in\mathbb{N}}$ is an ergodic Markov chain with any initial p_0 and a Markov transition matrix M with π as its invariant distribution, then

$$\lim_{n\to\infty} p_n(i) = \pi(i).$$

Properties of Markov chains

Ergodicity

If $(X_n)_{n\in\mathbb{N}}$ is an ergodic Markov chain with any initial p_0 and a Markov transition matrix M with π as its invariant distribution, then

$$\lim_{n\to\infty} p_n(i) = \pi(i).$$

Moreover, for $i, j \in \mathrm{X}$

$$\lim_{n\to\infty} \mathbb{P}(X_n = i | X_1 = j) = \pi(i).$$

Properties of Markov chains

What about continuous state-space Markov chains, i.e., where X is uncountable, e.g., $\mathsf{X} = \mathbb{R}$?

Properties of Markov chains

What about continuous state-space Markov chains, i.e., where X is uncountable, e.g., $X = \mathbb{R}$?

We will be mainly interested in the continuous case, however, the analogous concepts are defined in a much more complicated way.

Properties of Markov chains

What about continuous state-space Markov chains, i.e., where X is uncountable, e.g., $\mathrm{X} = \mathbb{R}$?

We will be mainly interested in the continuous case, however, the analogous concepts are defined in a much more complicated way.

We will not go into the details here, we will just now introduce the continuous state-space notation.

What is a Markov chain?

The continuous case case

We assume now our state-space is uncountable, e.g., $X = \mathbb{R}$.

What is a Markov chain?

The continuous case case

We assume now our state-space is uncountable, e.g., $\mathsf{X} = \mathbb{R}$.

We denote the initial *density* of the chain by $p_0(x)$.

What is a Markov chain?

The continuous case case

We assume now our state-space is uncountable, e.g., $\mathrm{X} = \mathbb{R}$.

We denote the initial *density* of the chain by $p_0(x)$.

The transition kernel is denoted $K(x_n|x_{n-1})$.

What is a Markov chain?

The continuous case case

We assume now our state-space is uncountable, e.g., $\mathrm{X} = \mathbb{R}$.

We denote the initial *density* of the chain by $p_0(x)$.

The transition kernel is denoted $K(x_n|x_{n-1})$.

The density of the chain at time n is denoted by $p_n(x_n)$.

What is a Markov chain?

The continuous case

A discrete-time Markov chain is a process $(X_n)_{n\in\mathbb{N}}$, when X is uncountable, satisfies:

What is a Markov chain?

The continuous case

A discrete-time Markov chain is a process $(X_n)_{n\in\mathbb{N}}$, when X is uncountable, satisfies:

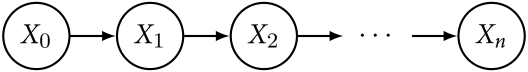
$$p(x_n|x_{1:n-1}) = p(x_n|x_{n-1}) = K(x_n|x_{n-1}).$$

What is a Markov chain?

The continuous case

A discrete-time Markov chain is a process $(X_n)_{n\in\mathbb{N}}$, when X is uncountable, satisfies:

$$p(x_n|x_{1:n-1}) = p(x_n|x_{n-1}) = K(x_n|x_{n-1}).$$



What is a Markov chain?

The continuous case

We will again consider the time-homogeneous case, i.e. the transition kernel is time-independent.

What is a Markov chain?

The continuous case

We will again consider the time-homogeneous case, i.e. the transition kernel is time-independent. A Markov chain therefore can be defined entirely by its:

- Initial state (or initial distribution)
- Transition kernel

What is a Markov chain?

The continuous case

The transition kernel is a density function $K(x_n|x_{n-1})$ for fixed x_{n-1}, i.e.,

$$\int_{\mathrm{X}} K(x_n|x_{n-1})\,\mathrm{d}x_n = 1.$$

Otherwise, it is a function of (x_n, x_{n-1}).

What is a Markov chain?

Example 1: Simulate a continuous-state Markov chain

Consider the following Markov chain: $X_0 = 0$ and

$$K(x_n|x_{n-1}) = \mathcal{N}(x_n; ax_{n-1}, 1),$$

where $0 < a < 1$.

What is a Markov chain?

Example 1: Simulate a continuous-state Markov chain

Consider the following Markov chain: $X_0 = 0$ and

$$K(x_n | x_{n-1}) = \mathcal{N}(x_n; a x_{n-1}, 1),$$

where $0 < a < 1$.
We can simulate this chain by:

$$\begin{aligned} X_1 &\sim \mathcal{N}(0, 1) \\ X_2 &\sim \mathcal{N}(aX_1, 1) \\ X_3 &\sim \mathcal{N}(aX_2, 1) \\ &\vdots \\ X_n &\sim \mathcal{N}(aX_{n-1}, 1). \end{aligned}$$

Simulation.

What is a Markov chain?

The continuous case: Chapman-Kolmogorov equations

The Chapman-Kolmogorov equation for the continuous case

$$p(x_n|x_{n-k}) = \int_{\mathrm{X}} K(x_n|x_{n-1})p(x_{n-1}|x_{n-k})\,\mathrm{d}x_{n-1},$$

for $k > 1$.

What is a Markov chain?

The continuous case: The evolution of the density of the chain

Let $p_0(x)$ be the initial density such that $X_0 \sim p_0(x)$.

Then, the density of the chain at time n is given by

$$p_n(x_n) = \int_{\mathsf{X}} K(x_n | x_{n-1}) p_{n-1}(x_{n-1}) \, \mathrm{d}x_{n-1}.$$

What is a Markov chain?

The continuous case: m-step transition kernel

It is useful for us to define the m-step transition kernel:

$$\begin{aligned} p(x_{m+n}|x_n) &= K^m(x_{m+n}|x_n), \\ &= \int_{\mathrm{X}} K(x_{m+n}|x_{m+n-1}) \cdots K(x_{n+1}|x_n)\,\mathrm{d}x_{m+n-1} \cdots \mathrm{d}x_{n+1}. \end{aligned}$$

What is a Markov chain?

Properties

We have the similar conditions of aperiodicity and irreducibility as in the discrete case, but,

- These are defined over *sets* rather than states.
- irreducibility is replaced by ϕ-irreducibility.
- aperiodicity is defined for sets

What is a Markov chain?

Properties

We have the similar conditions of aperiodicity and irreducibility as in the discrete case, but,

- These are defined over *sets* rather than states.
- irreducibility is replaced by ϕ-irreducibility.
- aperiodicity is defined for sets

We will not go into the details of these conditions for continuous space case.

What is a Markov chain?

Invariant distribution

A probability distribution π is called K-invariant if

$$\pi(x) = \int_{\mathrm{X}} \pi(x')K(x|x')\,\mathrm{d}x'.$$

Similar to the discrete case.

What is a Markov chain?

Detailed balance and reversibility

The detailed balance condition for the continuous case takes a similar form:

$$\pi(x)K(x'|x) = \pi(x')K(x|x').$$

What is a Markov chain?

Detailed balance and reversibility

The detailed balance condition for the continuous case takes a similar form:

$$\pi(x)K(x'|x) = \pi(x')K(x|x').$$

Note that this is a sufficient condition for stationarity of π:

$$\begin{aligned} \int \pi(x)K(x'|x)\mathrm{d}y &= \int \pi(x')K(x|x')\mathrm{d}x', \\ \implies \pi(x) &= \int K(x|x')\pi(x')\mathrm{d}x', \end{aligned}$$

which implies π is K-invariant.

What is a Markov chain?

Detailed balance and reversibility

A useful formulation of reversibility is the following: A Markov kernel K is π-reversible if

$$\int\int f(x,x')\pi(x)K(x|x')\mathrm{d}x\mathrm{d}x' = \int\int f(x,x')\pi(x')K(x'|x)\mathrm{d}x\mathrm{d}x',$$

for every measurable f, which follows from the detailed balance condition.

What is a Markov chain?

Example: Go back to Gaussian model

Consider the following Markov chain: $X_0 = 0$ and

$$K(x_n|x_{n-1}) = \mathcal{N}(x_n; ax_{n-1}, 1),$$

where $0 < a < 1$.

What is a Markov chain?

Example: Go back to Gaussian model

Consider the following Markov chain: $X_0 = 0$ and

$$K(x_n | x_{n-1}) = \mathcal{N}(x_n; a x_{n-1}, 1),$$

where $0 < a < 1$. Note that we can also write this as

$$X_n = a X_{n-1} + \epsilon_n,$$

where $\epsilon_n \sim \mathcal{N}(0, 1)$.

What is a Markov chain?

Example: Go back to Gaussian model

Prove that for

$$\pi(x) = \mathcal{N}\left(x; 0, \frac{1}{1-a^2}\right),$$

the detailed balance condition is satisfied for the kernel

$$K(x_n|x_{n-1}) = \mathcal{N}(x_n; ax_{n-1}, 1),$$

where $0 < a < 1$.

What is a Markov chain?

Example: Go back to Gaussian model

Prove that $K^m(x_{m+n}|x_n)$ is given by

$$K^m(x_{m+n}|x_n) = \mathcal{N}\left(x_{m+n}; a^m x_n, \frac{1-a^{2m}}{1-a^2}\right).$$

Then prove that

$$\pi(x) = \lim_{m\to\infty} K^m(x|x'),$$

independent of x'.

Why Markov chains?

Since we want i.i.d samples

Theorem 3

If K is an irreducible, π-invariant kernel, then for any integrable function φ

$$\lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{T}\varphi(X_i) = \int \varphi(x)\pi(x)\mathrm{d}x = (\varphi, \pi),$$

almost surely, for almost all initial points x_0.

Why Markov chains?

Since we want i.i.d samples

Theorem 3

If K is an irreducible, π-invariant kernel, then for any integrable function φ

$$\lim_{T\to\infty} \frac{1}{T}\sum_{i=1}^{T} \varphi(X_i) = \int \varphi(x)\pi(x)\mathrm{d}x = (\varphi, \pi),$$

almost surely, for almost all initial points x_0.

Therefore, we can use these samples to estimate our integrals.

Why Markov chains?

Since we want i.i.d samples

Theorem 4

If K is irreducible, aperiodic, and π-invariant, then

$$\lim_{T\to\infty} \int_{\mathbb{X}} |K^T(y|x) - \pi(y)| \mathrm{d}y = 0,$$

for π-almost all starting values x.

Markov chain Monte Carlo

How to design good kernels?

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

Markov chain Monte Carlo

How to design good kernels?

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

- We can sample from a proposal $q(x|x')$ (that is a Markov kernel)

Markov chain Monte Carlo

How to design good kernels?

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

- We can sample from a proposal $q(x|x')$ (that is a Markov kernel)
- We can use accept/reject

Markov chain Monte Carlo

How to design good kernels?

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

- We can sample from a proposal $q(x|x')$ (that is a Markov kernel)
- We can use accept/reject

We can design the process so that the stationary distribution of the chain is the target distribution.

Markov chain Monte Carlo

How to design good kernels?

We will first look at a surprisingly simple approach: the Metropolis-Hastings algorithm.

This approach relies on the following idea:

- We can sample from a proposal $q(x|x')$ (that is a Markov kernel)
- We can use accept/reject

We can design the process so that the stationary distribution of the chain is the target distribution.

This is however very different from the rejection sampling approach.

Metropolis-Hastings

A first look

Consider the following method:

- Sample $X' \sim q(x'|X_{n-1})$
- Set $X_n = X'$ with probability

$$\alpha(X'|X_{n-1}) = \min\left\{1, \frac{\pi(X')q(X_{n-1}|X')}{\pi(X_{n-1})q(X'|X_{n-1})}\right\}.$$

- Otherwise, set $X_n = X_{n-1}$.

Metropolis-Hastings

A first look

Consider the following method:

- Sample $X' \sim q(x'|X_{n-1})$
- Set $X_n = X'$ with probability

$$\alpha(X'|X_{n-1}) = \min\left\{1, \frac{\pi(X')q(X_{n-1}|X')}{\pi(X_{n-1})q(X'|X_{n-1})}\right\}.$$

- Otherwise, set $X_n = X_{n-1}$.

Note the last step: we discard the sample X' if rejected BUT set $X_n = X_{n-1}$.

Metropolis-Hastings

Metropolis-Hastings Algorithm

The ratio

$$\mathrm{r}(x, x') = \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)},$$

is called acceptance ratio.

Metropolis-Hastings

Metropolis-Hastings Algorithm

We have discussed explicit kernels in the discrete and continuous cases.

Metropolis-Hastings

Metropolis-Hastings Algorithm

We have discussed explicit kernels in the discrete and continuous cases.

But the MH algorithm automatically gives us a kernel.

We have discussed explicit kernels in the discrete and continuous cases.

But the MH algorithm automatically gives us a kernel.

How to prove that the stationary distribution is the target distribution?

Metropolis-Hastings

Metropolis-Hastings Algorithm

Let us figure out the kernel.

Metropolis-Hastings

Metropolis-Hastings Algorithm

Let us figure out the kernel.

Let us say, we have the sample from the proposal x'. Fixing this sample, the acceptance step samples from the mixture (*intuitively*):

$$\alpha(x'|x)\delta_{x'}(y) + (1 - \alpha(x'|x))\delta_x(y).$$

To get the full kernel, we need to integrate over x':

$$\begin{aligned} K(y|x) &= \int q(x'|x)\left(\alpha(x'|x)\delta_{x'}(y) + (1 - \alpha(x'|x))\delta_x(y)\right) \mathrm{d}x', \\ &= \alpha(y|x)q(y|x) + (1 - a(x))\delta_x(y) \end{aligned}$$

where

$$a(x) = \int \alpha(x'|x)q(x'|x)\mathrm{d}x'.$$

More intuition in terms of x_n and x_{n-1}:

- What is the probability of being at x_{n-1} and getting accepted?

$$a(x_{n-1}) = \int_{\mathsf{X}} \alpha(x|x_{n-1}) q(x|x_{n-1}) \mathrm{d}x.$$

- Therefore, the probability of being at x_{n-1} and getting rejected is $1 - a(x_{n-1})$.

We can see that the kernel is

$$K(x_n|x_{n-1}) = \alpha(x_n|x_{n-1}) q(x_n|x_{n-1}) + (1 - a(x_{n-1}))\delta_{x_{n-1}}(x_n).$$

Metropolis-Hastings

Metropolis-Hastings Algorithm: Detailed Balance

We can now prove that the kernel satisfies the detailed balance condition:

$$K(x'|x)\pi(x) = K(x|x')\pi(x').$$

Metropolis-Hastings

Metropolis-Hastings Algorithm: Detailed Balance

$$
\begin{aligned}
\pi(x)K(x'|x) &= \pi(x)q(x'|x)\alpha(x',x) + \pi(x)(1-a(x))\delta_x(x') \\
&= \pi(x)q(x'|x)\min\left\{1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right\} + \pi(x)(1-a(x))\delta_x(x') \\
&= \min\left\{\pi(x)q(x'|x), \pi(x')q(x|x')\right\} + \pi(x)(1-a(x))\delta_x(x') \\
&= \min\left\{\frac{\pi(x)q(x'|x)}{\pi(x')q(x|x')}, 1\right\}\pi(x')q(x|x') + \pi(x')(1-a(x'))\delta_{x'}(x) \\
&= K(x|x')\pi(x').
\end{aligned}
$$

Metropolis-Hastings

Unnormalised density

Assume we are given an unnormalised density to sample γ where

$$\pi(x) = \frac{\gamma(x)}{Z},$$

where Z is the normalisation constant.

Metropolis-Hastings

Unnormalised density

- Sample $X' \sim q(x'|X_{n-1})$
- Set $X_n = X'$ with probability

$$\alpha(X'|X_{n-1}) = \min\left\{1, \frac{\gamma(X')q(X_{n-1}|X')}{\gamma(X_{n-1})q(X'|X_{n-1})}\right\}.$$

- Otherwise, set $X_n = X_{n-1}$.

as the normalising constants of π would cancel out.

Metropolis-Hastings

How do we choose proposals?

- Independent proposals
- Symmetric (random walk) proposals
- Gradient-based proposals
- Adaptive proposals

Metropolis-Hastings

Independent proposals

Choose the proposal $q(x)$ independently of the current state X_{n-1}. Leads to

- $X' \sim q(x')$
- Accept with probability

$$\alpha(X'|X_{n-1}) = \min\left\{1, \frac{\pi(X')q(X_{n-1})}{\pi(X_{n-1})q(X')}\right\}.$$

- Otherwise, set $X_n = X_{n-1}$.

Metropolis-Hastings

Independent proposals

Let us say

$$\pi(x) = \mathcal{N}(x; \mu, \sigma^2)$$

For the example, assume we want to use MH to sample from it. Choose a proposal

$$q(x) = \mathcal{N}(x; \mu_q, \sigma_q^2).$$

How to compute the acceptance ratio?

Metropolis-Hastings

Independent proposals

$$
\begin{aligned}
\mathrm{r}(x, x') &= \frac{\pi(x')q(x)}{\pi(x)q(x')} \\
&= \frac{\mathcal{N}(x'; \mu, \sigma^2)\mathcal{N}(x; \mu_q, \sigma_q^2)}{\mathcal{N}(x; \mu, \sigma^2)\mathcal{N}(x'; \mu_q, \sigma_q^2)} \\
&= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x'-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x'-\mu_q)^2}{2\sigma_q^2}\right)} \\
&= \frac{\exp\left(-\frac{(x'-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)}{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(x'-\mu_q)^2}{2\sigma_q^2}\right)} \\
&= e^{\left(-\frac{1}{2\sigma^2}\left[(x'-\mu)^2-(x-\mu)^2\right]\right)} e^{\left(-\frac{1}{2\sigma_q^2}\left[(x-\mu_q)^2-(x'-\mu_q)^2\right]\right)}
\end{aligned}
$$

Metropolis-Hastings

Random walk proposal

We can choose:

$$q(x'|x) = \mathcal{N}(x'; x, \sigma_q^2)$$

The proposal looks at where we are and take a random step (random walk).

Metropolis-Hastings

Random walk proposal

We can choose:

$$q(x'|x) = \mathcal{N}(x'; x, \sigma_q^2)$$

The proposal looks at where we are and take a random step (random walk).

Note that $q(x'|x)$ is symmetric, i.e. $q(x|x') = q(x'|x)$.

Metropolis-Hastings

Random walk proposal

Acceptance ratio:

$$\begin{aligned} \mathrm{r}(x, x') &= \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)} \\ &= \frac{\pi(x')}{\pi(x)}, \\ &= \frac{\mathcal{N}(x'; \mu, \sigma^2)}{\mathcal{N}(x; \mu, \sigma^2)} \\ &= e^{\left(-\frac{1}{2\sigma^2}\left[(x'-\mu)^2-(x-\mu)^2\right]\right)}. \end{aligned}$$

Metropolis-Hastings

Random walk proposal

Set a burnin period:

- Run the sampler for fixed number of iterations and discard the first n samples.
- This accounts for the convergence to the stationary measure.

Metropolis-Hastings

Gradient-based proposal

We can *inform* the proposal by using the gradient of the target distribution.

$$q(x'|x) = \mathcal{N}(x'; x + \gamma \nabla \log \pi(x), 2\gamma I),$$

This tends to behave really well.

Metropolis-Hastings

Gradient-based proposal

We can *inform* the proposal by using the gradient of the target distribution.

$$q(x'|x) = \mathcal{N}(x'; x + \gamma\nabla \log \pi(x), 2\gamma I),$$

This tends to behave really well.

This approach is called *Metropolis adjusted Langevin algorithm* (MALA).
(more on these later)

Metropolis-Hastings

Caveats, design rules

- One has to be careful that $p/q < \infty$ (while no theoretical reason, the performance tends to be quite bad).

Metropolis-Hastings

Caveats, design rules

- One has to be careful that $p/q < \infty$ (while no theoretical reason, the performance tends to be quite bad).
- The proposal should attain a balance of acceptance rate and efficiency.

Metropolis-Hastings

Caveats, design rules

- One has to be careful that $p/q < \infty$ (while no theoretical reason, the performance tends to be quite bad).
- The proposal should attain a balance of acceptance rate and efficiency.
- Too high acceptance rate is **not** necessarily good: You might be taking too small steps and getting stuck in some regions

Metropolis-Hastings

Bayesian inference with MH

Let us look at now the Bayesian inference problem.

We can solve it in full generality (in theory) using MH.

Recall the general formulation

$$p(x|y_{1:n}) = \frac{p(y_{1:n}|x)p(x)}{p(y_{1:n})} = \frac{\prod_{i=1}^{n} p(y_i|x)p(x)}{p(y_{1:n})},$$

when $y_1, \ldots, y_n$ are conditionally independent given x.

Metropolis-Hastings

Bayesian inference with MH

We write

$$p(x|y_{1:n}) \propto \prod_{i=1}^{n} p(y_i|x)p(x),$$

and set

$$\gamma(x) = \prod_{i=1}^{n} p(y_i|x)p(x),$$

as our unnormalised posterior.

Metropolis-Hastings

Bayesian inference with MH

The generic MH for Bayesian inference, given x_{n-1}

- Sample $X' \sim q(x'|x_{n-1})$.
- Accept $x_n = x'$ with probability

$$\alpha(x_{n-1}, x') = \min\left\{1, \frac{\gamma(x')q(x_{n-1}|x')}{\gamma(x_{n-1})q(x'|x_{n-1})}\right\}.$$

- Otherwise, $X_n = x_{n-1}$.

Metropolis-Hastings

Example: Source localisation

Recall our example about localising a source using observations from a sensor network.

We can now formalise this problem. Assume that the source is located at $x \in \mathbb{R}^2$ and the sensor network is located at $s_1, \ldots, s_3 \in \mathbb{R}^2$ (3 sensors).

Assume that these three sensors ”observe” the source according to:

$$p(y_i|x, s_i) = \mathcal{N}(y_i; \|x - s_i\|, R),$$

where y_i is the observation from sensor i.

Metropolis-Hastings

Example: Source localisation

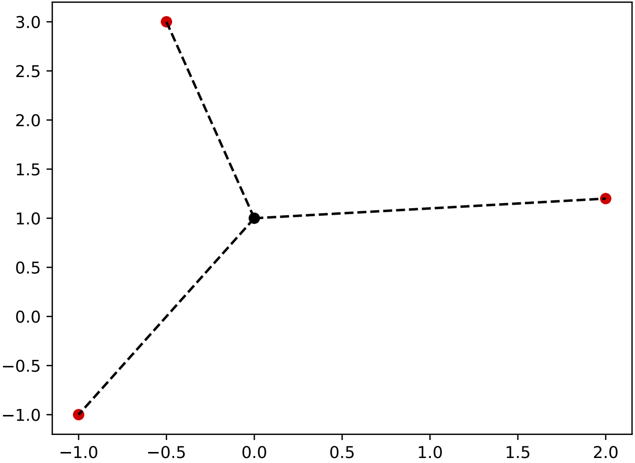


Figure: Source localisation

Metropolis-Hastings

Example: Source localisation

Assume that you are asked to estimate the location of the source given the observations y_1, y_2, y_3. What is the model?

Metropolis-Hastings

Example: Source localisation

Assume that you are asked to estimate the location of the source given the observations y_1, y_2, y_3. What is the model?

We first need a prior on the source location:

$$p(x) = \mathcal{N}(x; \mu, \Sigma),$$

where μ is the prior mean and Σ is the prior covariance. We already have the likelihoods for each y_i.

Metropolis-Hastings

Example: Source localisation

The posterior is given by

$$p(x|y_1, y_2, y_3, s_1, s_2, s_3) \propto p(x) \prod_{i=1}^{3} p(y_i|x, s_i).$$

Metropolis-Hastings

Example: Source localisation

We choose a random walk proposal:

$$q(x'|x) = \mathcal{N}(x'; x, \sigma^2 I).$$

This is symmetric so the acceptance ratio is:

$$\mathrm{r}(x, x') = \frac{p(x')p(y_1|x', s_1)p(y_2|x', s_2)p(y_3|x', s_3)}{p(x)p(y_1|x, s_1)p(y_2|x, s_2)p(y_3|x, s_3)}.$$

Metropolis-Hastings

The banana density

Consider the 2D density

$$p(x,y) \propto \exp\left(-\frac{x^2}{10} - \frac{y^4}{10} - 2(y - x^2)^2\right).$$

Assume we would like to sample from it.

Metropolis-Hastings

The banana density

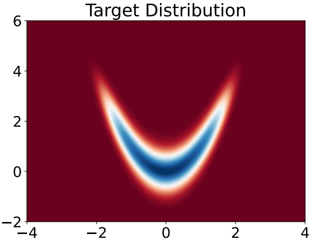


Figure: The banana density (unnormalised)

Metropolis-Hastings

The banana density

We have

$$\gamma(x, y) = \exp\left(-\frac{x^2}{10} - \frac{y^4}{10} - 2(y - x^2)^2\right).$$

and let us choose two alternative proposals

- The random walk proposal:

$$q(x', y'|x, y) = \mathcal{N}(x'; x, \sigma_q^2)\mathcal{N}(y'; y, \sigma_q^2).$$

- and the gradient-based proposal (MALA):

$$q(x', y'|x, y) = \mathcal{N}(z; z + \gamma\nabla \log \gamma(z), \sqrt{2\gamma}\mathbf{I}).$$

where $z = (x, y)$ and γ is a step size.

We have seen Metropolis-Hastings sampler.

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:
    - High-dimensional problems

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:
  - High-dimensional problems
  - Complex models

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:
    - High-dimensional problems
    - Complex models
    - Large datasets

We have seen Metropolis-Hastings sampler.

- Unfortunately, it may not be very efficient.
- Acceptance ratios are very tricky to compute in a variety of settings:
  - High-dimensional problems
  - Complex models
  - Large datasets
- We will now look at a different approach: Langevin MCMC.

Langevin-based approaches

Crash course on Langevin SDE - I

Consider the Langevin SDE for a generic drift ∇V:

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

where $(B_t)_{t\geq 0}$ is a Brownian motion.

Langevin-based approaches

Crash course on Langevin SDE - I

Consider the Langevin SDE for a generic drift ∇V:

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

where $(B_t)_{t\geq 0}$ is a Brownian motion. This SDE has a stationary measure

$$\pi \propto e^{-V(x)}.$$

Therefore, for a classical *sampling* problem for, say $\pi(x)$, we could set $V(x) = -\log \pi(x)$ (negative density).

Langevin-based approaches

Crash course on Langevin SDE - I

Consider the Langevin SDE for a generic drift ∇V:

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

where $(B_t)_{t\geq 0}$ is a Brownian motion. This SDE has a stationary measure

$$\pi \propto e^{-V(x)}.$$

Therefore, for a classical *sampling* problem for, say $\pi(x)$, we could set $V(x) = -\log \pi(x)$ (negative density).

This diffusion converges to its stationary measure exponentially fast if V is μ-strongly-convex.

Langevin-based approaches

Crash course on Langevin SDE - II – Optimisation

Consider the Langevin SDE for a generic drift ∇V:

$$dX_t = -\nabla V(X_t)dt + \sqrt{\frac{2}{\beta}}dB_t,$$

where $(B_t)_{t\geq 0}$ is a Brownian motion.

Langevin-based approaches

Crash course on Langevin SDE - II – Optimisation

Consider the Langevin SDE for a generic drift ∇V:

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}B_t,$$

where $(B_t)_{t\geq 0}$ is a Brownian motion. This SDE has a stationary measure

$$\pi \propto e^{-\beta V(x)}.$$

Consider the Langevin SDE for a generic drift ∇V:

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}B_t,$$

where $(B_t)_{t\geq 0}$ is a Brownian motion. This SDE has a stationary measure

$$\pi \propto e^{-\beta V(x)}.$$

This stationary measure concentrates on the minima of V as $\beta \to \infty$ (Hwang, 1980).

Langevin-based approaches

Crash course on Langevin SDE - II – Optimisation

Consider the Langevin SDE for a generic drift ∇V:

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}B_t,$$

where $(B_t)_{t\geq 0}$ is a Brownian motion. This SDE has a stationary measure

$$\pi \propto e^{-\beta V(x)}.$$

This stationary measure concentrates on the minima of V as $\beta \to \infty$ (Hwang, 1980).

Langevin diffusion is a global optimiser.

Langevin-based approaches

Crash course on Langevin SDE - III: Numerical discretisation

The Euler discretisation is the *unadjusted Langevin algorithm* (ULA):

$$X_{t+1}^{\gamma} = X_t^{\gamma} - \gamma \nabla V(X_t^{\gamma}) + \sqrt{2\gamma} W_{t+1}$$

where $(W_t)_{t \geq 0}$ are i.i.d standard Normal random variables.

Langevin-based approaches

Crash course on Langevin SDE - III: Numerical discretisation

The Euler discretisation is the *unadjusted Langevin algorithm* (ULA):

$$X_{t+1}^{\gamma} = X_t^{\gamma} - \gamma \nabla V(X_t^{\gamma}) + \sqrt{2\gamma} W_{t+1}$$

where $(W_t)_{t \geq 0}$ are i.i.d standard Normal random variables.

This chain has a *different* stationary measure π^{γ} but a number of guarantees can be derived for its convergence.

Theorem 1 (Durmus and Moulines, 2019)

Let $\mathcal{L}(X_t)$ be the law of the iterates of ULA, then

$$W_2^2(\mathcal{L}(X_t^{\gamma}), \pi) \lesssim \left(1 - \frac{\gamma\kappa}{2}\right)^{t+1} (d/m + \|x - x^{\star}\|^2) + \gamma,$$

under suitable regularity conditions for V, restriction on γ where $\kappa := \kappa(m, L)$.

An important note here is that, we can sample from the posterior $p(x|y)$ using ULA as

$$p(x|y) \propto p(x, y),$$

and

$$X_{n+1}^{\gamma} = X_n^{\gamma} + \gamma \nabla \log p(X_n^{\gamma}, y) + \sqrt{2}\gamma W_{n+1}.$$

Langevin-based approaches

ULA for Bayesian inference

An important note here is that, we can sample from the posterior $p(x|y)$ using ULA as

$$p(x|y) \propto p(x, y),$$

and

$$X_{n+1}^{\gamma} = X_n^{\gamma} + \gamma \nabla \log p(X_n^{\gamma}, y) + \sqrt{2}\gamma W_{n+1}.$$

We can see that this algorithm would approximately sample from $p(x|y)$.

Langevin-based approaches

ULA for Bayesian inference

Let us say we have data $y_1, \dots, y_M$ for M large. We can write the posterior as

$$p(x|y_{1:M}) \propto p(x) \prod_{i=1}^{M} p(y_i|x).$$

therefore, our potential becomes

$$V(x) = -\log p(x) - \sum_{i=1}^{M} \log p(y_i|x).$$

Langevin-based approaches

ULA for Bayesian inference

Let us say we have data $y_1, \ldots, y_M$ for M large. We can write the posterior as

$$p(x|y_{1:M}) \propto p(x) \prod_{i=1}^{M} p(y_i|x).$$

therefore, our potential becomes

$$V(x) = -\log p(x) - \sum_{i=1}^{M} \log p(y_i|x).$$

Mini-quiz: What is the problem with MALA (or MH in general) in this case?

Langevin-based approaches

ULA for Bayesian inference

A similar problem of course would be for ULA.

A similar problem of course would be for ULA.

However, we can resolve this, as we can approximate the gradient using subsampling:

$$\begin{aligned}\nabla V(x) &= \nabla \log p(x) + \sum_{i=1}^{M} \nabla \log p(y_i|x), \\ &\approx \nabla \log p(x) + \frac{M}{m} \sum_{j=1}^{m} \nabla \log p(y_{k_j}|x) = \widehat{\nabla V(x)},\end{aligned}$$

where $k_j \sim \text{Unif}\{1, \ldots, M\}$, for $j = 1, \ldots, m$ for $m \ll M$.

Langevin-based approaches

ULA for Bayesian inference

A similar problem of course would be for ULA.

However, we can resolve this, as we can approximate the gradient using subsampling:

$$\begin{aligned}\nabla V(x) &= \nabla \log p(x) + \sum_{i=1}^{M} \nabla \log p(y_i|x), \\ &\approx \nabla \log p(x) + \frac{M}{m} \sum_{j=1}^{m} \nabla \log p(y_{k_j}|x) = \widehat{\nabla V(x)},\end{aligned}$$

where $k_j \sim \text{Unif}\{1, \ldots, M\}$, for $j = 1, \ldots, m$ for $m \ll M$.

Stochastic gradients.

Langevin-based approaches

ULA for Bayesian inference

One can run ULA with stochastic gradients:

$$X_{n+1}^{\gamma} = X_n^{\gamma} - \gamma \widehat{\nabla V(X_n^{\gamma})} + \sqrt{2}\gamma W_{n+1}.$$

Langevin-based approaches

ULA for Bayesian inference

One can run ULA with stochastic gradients:

$$X_{n+1}^{\gamma} = X_n^{\gamma} - \gamma \widehat{\nabla V(X_n^{\gamma})} + \sqrt{2}\gamma W_{n+1}.$$

The resulting method is called *stochastic gradient Langevin dynamics* (SGLD) (Welling and Teh, 2011).

One can run ULA with stochastic gradients:

$$X_{n+1}^{\gamma} = X_n^{\gamma} - \gamma \widehat{\nabla V(X_n^{\gamma})} + \sqrt{2\gamma} W_{n+1}.$$

The resulting method is called *stochastic gradient Langevin dynamics* (SGLD) (Welling and Teh, 2011).

- Widely used for large-scale datasets.

Langevin-based approaches

ULA for Bayesian inference

One can run ULA with stochastic gradients:

$$X_{n+1}^{\gamma} = X_n^{\gamma} - \gamma\nabla\widehat{V(X_n^{\gamma})} + \sqrt{2}\gamma W_{n+1}.$$

The resulting method is called *stochastic gradient Langevin dynamics* (SGLD) (Welling and Teh, 2011).

- Widely used for large-scale datasets.
- It has similar guarantees to ULA in Wasserstein-2 distance for strongly convex V.

Langevin-based approaches

ULA for Bayesian inference

One can run ULA with stochastic gradients:

$$X_{n+1}^{\gamma} = X_n^{\gamma} - \gamma \widehat{\nabla V(X_n^{\gamma})} + \sqrt{2\gamma} W_{n+1}.$$

The resulting method is called *stochastic gradient Langevin dynamics* (SGLD) (Welling and Teh, 2011).

- Widely used for large-scale datasets.
- It has similar guarantees to ULA in Wasserstein-2 distance for strongly convex V.
- Also used to model and analyse the behaviour of stochastic gradient descent methods (SGD) in deep learning.

Langevin-based approaches

ULA for Bayesian inference

One can run ULA with stochastic gradients:

$$X_{n+1}^{\gamma} = X_n^{\gamma} - \gamma \widehat{\nabla V(X_n^{\gamma})} + \sqrt{2\gamma} W_{n+1}.$$

The resulting method is called *stochastic gradient Langevin dynamics* (SGLD) (Welling and Teh, 2011).

- Widely used for large-scale datasets.
- It has similar guarantees to ULA in Wasserstein-2 distance for strongly convex V.
- Also used to model and analyse the behaviour of stochastic gradient descent methods (SGD) in deep learning.

Web based simulations if time permits.

References I

- Durmus, Alain and Eric Moulines (2019). "High-dimensional Bayesian inference via the unadjusted Langevin algorithm". In: *Bernoulli* 25.4A, pp. 2854–2882.
- Hwang, Chii-Ruey (1980). "Laplace's method revisited: weak convergence of probability measures". In: *The Annals of Probability*, pp. 1177–1182.
- Li, Bo, Thomas Bengtsson, and Peter Bickel (2005). "Curse-of-dimensionalit revisited: Collapse of importance sampling in very large scale systems". In: *Rapport technique* 85, p. 205.
- Welling, Max and Yee W Teh (2011). "Bayesian learning via stochastic gradient Langevin dynamics". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, pp. 681–688.