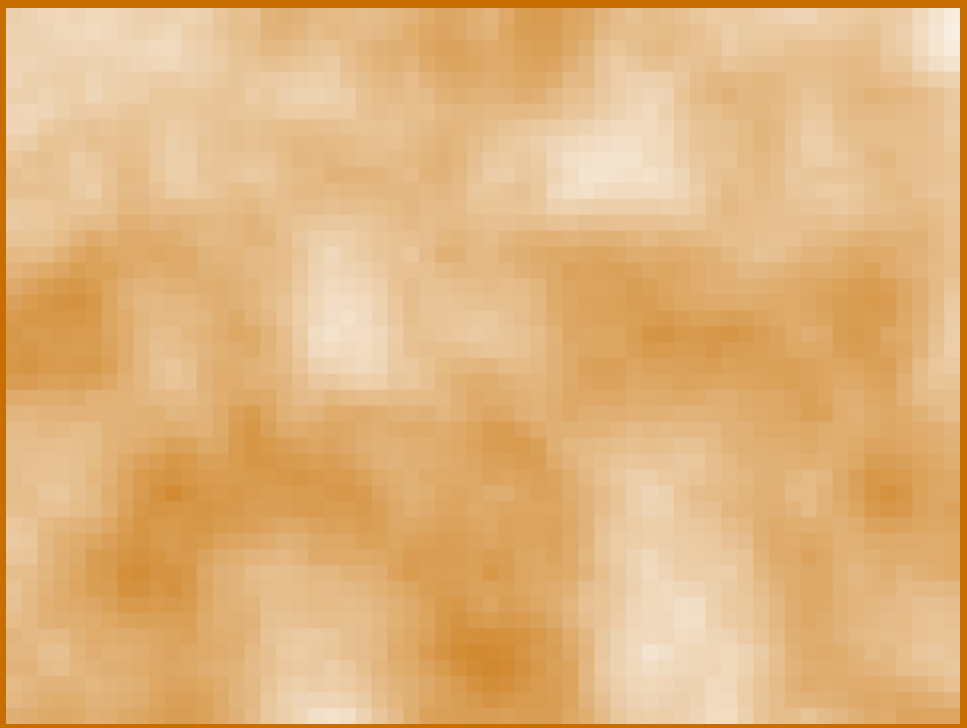
Lecture notes on Mixing times and Markov Chain Monte Carlo methods

Written by
Manuel Hinz
mh@mssh.dev or s6mlhinz@uni-bonn.de

Lecturer
Prof. Dr. Andreas Eberle
eberle (at) uni-bonn.de



University of Bonn
Winter semester 2024/2025
Last update: October 9, 2024

Contents

Chapter 0:
Manuel's notes

Warning

These are unofficial lecture notes written by a student. They are messy, will almost surely contain errors, typos and misunderstandings and may not be kept up to date! I do however try my best and use these notes to prepare for my exams. Feel free to email me any corrections to mh@mssh.dev or s6mlhinz@uni-bonn.de.
Happy learning!

General Information

- Basis: Basis
- Website: https://wt.iam.uni-bonn.de/faculty-staff/eberle/teaching/mcmc-2022-1-1
- Time slot(s): **Wednesday: 16:30-18:05** N0.008
- Exams: Oral

0.1 Organization

Start of lecture 01 (09.10.2024)

- Second week: No lecture
- Notes rearranged and not complete

Chapter 1:
Sampling and Monte Carlo

The state space $S \subseteq \mathbb{R}^d$ measurable with some $\mu(dx) = \mu(x)dx$ absolutely continuous probability measure on S with $\mu(x) > 0, \mu(x) = \frac{1}{z}e^{-U(x)}$ for some normalizing constant $z \in (0, \infty)$, usually unknown.
Many examples: Boltzmann-Gibs, in statistics exponential families.
In physics $U : S \to \mathbb{R}$ is called the **energy**, which we adopt. This is usually know explicitly.
Goals:

1. **Sampling:** Simulate a approximate sample from μ
2. **Integral estimation:** Compute $\vartheta = \int f d\mu$ approximately. For $f = 1_B, \vartheta = \mu(B)$

f is sometimes called a observable

There is a connection between 1. and 2.: For samples $X_1, \dots, X_n$ i.i.d. samples of μ, then

$$\hat{\vartheta}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

is an unbiased estimator of ϑ.

1.1 Sampling (without Markov chains)

1.1.1 Direct simulation only for special models

Example. *For $S = \mathbb{R}^1$, $F(c) = \mu((-\infty, c])$. Consider the generalized inverse $F^{-1}(u) = \inf\{c \in \mathbb{R} : f(c) \geq u\}$. Then draw a uniform variable $u \sim Unif((0,1))$, then $F^{-1}(u) \sim \mu$. This is called the **inversion method**.*

Remark. *$\mu = \mathcal{N}(0,1) \implies F(c) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{c} e^{-\frac{-x^2}{2}}dx$, therefore we can't use F, since it has no explicit representation.*

Example. *$S = \mathbb{R}^2, \mu = \mathcal{N}(0, I_2)$ with $\mu(dx) = \frac{1}{2\pi}e^{\frac{-|x|^2}{2}}dx \overset{polar}{=} \underbrace{\frac{1}{2\pi}e^{-\frac{r^2}{2}}rdrd\phi}_{\nu \otimes Unif(0,2\pi)}$ with*

$$\nu(r) = re^{\frac{-r^2}{2}}, r \in (0, \infty)$$

The distribution function $F(c) = \int_0^c \nu(r)dr = e^{-\frac{c^2}{2}}$, which is explicit!
Algorithm:

1. *Sample (u_1, u_2)*
2. *$r = F^{-1}(u_1), \varphi = 2\pi u_2$*
3. *$x = (r\cos(\varphi), r\sin(\varphi)) \sim \mathcal{N}(0, I_2)$*

4. *Then x_1, x_2 are independent and normally distributed*

This is called the Box-Muller method

Example. *$\mathcal{N}(m, C), m \in \mathbb{R}^d, C \in \mathbb{R}^{d\times d}$ symmetric pos. definite.*

Remark. *$z \sim \mathcal{N}(0, I_d), \sigma \in \mathbb{R}^{d\times d} \implies \sigma Z + m \sim \mathcal{N}(m, \sigma\sigma^d)$*

Algorithm:

This can be used to sample brownian bridges and brownian motions

- Find $\sigma \in \mathbb{R}^{d\times d}$ s.t. $\sigma\sigma^\intercal = C$ by Cholesky.
- Sample $z_1, \dots, z_d \sim \mathcal{N}(0, 1)$ via Box-Muller
- $x = \sigma z + m \sim \mathcal{N}(m, C)$

1.1.2 Acceptance-Rejection

Suppose $\mu(dx) \propto \rho(x)\nu(dx)$ for some *nice* ν, i.e. we can sample from ν.
Assumption: $\exists C \in (0, \infty) : \rho(x) \leq C\nu -$ a.s.
Algorithm: Repeat

1. Sample $x \sim \nu, u \sim \text{Unif}(0, 1)$
2. until $u \leq \rho(x)/C$
3. return x

Model $x_n \sim \nu, U_n \sim \text{Unif}(0, 1)$ all independent. Output $x_T, T = \min\{n \in \mathbb{N} U_n \leq \frac{\rho(X_n)}{C}\}$

Theorem 1.1. 1. *$T \sim Geom(p), p = \frac{1}{C}\int \rho d\nu$*

2. *$X_T \sim \mu$*

Proof. Exercise. □

Problem: $\mathbb{E}(T) = \frac{1}{p} = \frac{C}{\int \rho d\nu}$ will often be very large.

Example. *ν, μ are i.i.d. product measures on $\mathbb{R}^d$*

$$\rho(x_1, \dots, x_d) = \prod_{i=1}^{d} f(x_i)$$

where f is the one-dimensional density. $C \sim A^d$ for some $A > 1$.

1.2 Monte Carlo Methods

We want to approximate $\vartheta = \int f d\mu$
Numerical integration: Curse of dimension, we need some regularity to get good bounds, ...
Classical Monte Carlo:

$$\hat{\vartheta}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

for $X_i \sim \mu$ i.i.d.. We know:

- $\mathbb{E}(\hat{\vartheta}_n) = \vartheta$
- $\text{Var}(\hat{V\vartheta}_n) = \frac{1}{n}\text{Var}_\mu(f)$
- Conecentration inequalies:, like the Hoeffding inequality

$$\mathbb{P}(|\hat{\vartheta}_n - \vartheta| \geq \epsilon) \leq 2\epsilon^{-\frac{n\epsilon^2}{2\sup|f|^2}}$$

But: this requires independent samples from μ, usually not available.
Importance Sampling: $d\mu \propto \rho d\nu$

$$\vartheta = \int f d\mu = \frac{\int f\rho d\nu}{\int \rho d\nu}$$

Now we can approximate ϑ by using Monte Carlo for both integrals.

But then we are estimating a term in the denominator ...

$$\overset{\text{LLN}}{\approx} \frac{\frac{1}{n}\sum_{i=1}^n f(X_i)\rho(X_i)}{\frac{1}{n}\sum_{i=1}^n \rho(X_i)} = \hat{\vartheta}_n$$

where the X_i are independent. $\rho(X_i)$ are also called **importance weights**.

We don't have to choose the same number of samples, but we tipically do

Problems:

- In general $\mathbb{E}(\hat{\vartheta}_n) \neq \vartheta$, which means we have a bias, which might be difficult to control
- Weight degeneracy: often $\rho(X_i) \approx 0$ for most samples (except if ν, μ are close) $\implies$ variance of $\hat{\vartheta}_n$ can be large

1.3 Markov Chain Monte Carlo

- Find a transition kernel $\pi(x, dy)$ on S s.t. $\mu\pi = \mu$, i.e.

  $$\int \mu(dx)\pi(x, B) = \mu(B) \tag{1}$$

  for measurable B

  There are many ways to find such a kernel, but how do we find a kernel that rapidly converges

- Simulate a Markov Chain $X_0, X_1, \ldots, , X_n$ with given initial distribution ν and transition kernel π
- Under weak assumptions:

  $$\text{Law}(X_n) = \nu\pi^n \overset{n\to\infty}{\to} \mu$$

  which means **convergences to stationarity**
- **ergodicty** $\hat{\vartheta}_n = \frac{1}{n}\sum_{i=b}^{b+n} f(X_i) \to \int f d\mu \mathbb{P}$ a.s.

  The b, called **burn-in-time** yields a better estimator

Idea: n sufficiently large $\implies$ X_n is approximately sample from μ and $\hat{\vartheta}_n \approx \vartheta$.
Question: What does *sufficiently large* mean?
Can we get quantitative bounds for fixed n?

That means non-asymptotic bounds

This brings us to mixing times, since they measure how long it takes to converge to the stationary distribution.
How can we find π with invariant measure μ? One possibility is called **Detailed balance**:

Lemma 1.2.

$$\mu(dx)\pi(x, dy) = \mu(dy)\pi(y, dx) \tag{2}$$

i.e. for all measurable A, B :

$$\int_A \mu(dx)\pi(x, B) = \int_B \mu(dy)\pi(y, A)$$

Then μ is invariant for π

This makes it easier, since we only have to make the expression symmetric!

Proof. Choose $A = S$ the whole space. □

Remark. *1. (2) is sufficient, but not necessary, for invariance. E.g. $S = \mathbb{R}^2 \cong \mathbb{C}$ with $X_{n+1} = e^{i\vartheta}X_n$ for $\vartheta \in \mathbb{R}$ $\implies$ $\mu = \mathcal{N}(0, I_2)$ is invariant. (2) holds only for $\vartheta \in \pi\mathbb{Z}$*

2. *Suppose X_n is Markov chain with transition kernel π, $X_0 \sim \mu$. Then (1) $\iff$*

$$Law(X_n, X_{n+1}, \dots) = Law(X_0, X_1, \dots) \forall n \tag{3}$$

and (2) $\iff$

$$Law(X_n, X_{n-1}, \dots) = Law(X_0, X_1, \dots) \forall n \tag{4}$$

3. *(4) is much stronger than (3)*

List of Lectures