Markov Chain Monte Carlo I

Probabilistic Decision Making VU

(Reinforcement Learning VO)

Thomas Wedenig

Dec 10, 2025

Institute of Machine Learning and Neural Computation
Graz University of Technology, Austria

Table of contents

Motivation

- Let π be a **probability distribution** on state space $\mathcal{X}$
  - Discrete $\mathcal{X}$: probability mass function $\pi : \mathcal{X} \to [0, 1]$
  - Continuous $\mathcal{X}$: probability density function $\pi : \mathcal{X} \to \mathbb{R}_{\geq 0}$
- Assume π is either
  - **known**, i.e., we can evaluate $\pi(\mathbf{x}) \; \forall \mathbf{x} \in \mathcal{X}$
  - **known up to normalization**, i.e., we can evaluate $\tilde{\pi}(\mathbf{x}) \propto \pi(\mathbf{x}) \; \forall \mathbf{x} \in \mathcal{X}$

Goal: Draw i.i.d. samples from π.

- In practice, it's **very common** to only know the **unnormalized** distribution $\tilde{\pi}$:

$$\pi(\mathrm{x}) = \frac{\tilde{\pi}(\mathrm{x})}{Z}, \qquad Z := \int_{\mathcal{X}} \tilde{\pi}(\mathrm{x})\,\mathrm{d}\mathrm{x}$$

Bayesian Inference

- Prior $p(\mathrm{x})$
- Likelihood $p(\mathrm{y} \mid \mathrm{x})$
- Posterior:

$$p(\mathrm{x} \mid \mathrm{y}) = \frac{p(\mathrm{y} \mid \mathrm{x})\, p(\mathrm{x})}{\int_{\mathcal{X}} p(\mathrm{y} \mid \mathrm{x}')\, p(\mathrm{x}')\,\mathrm{d}\mathrm{x}'} =: \pi_{\mathrm{y}}(\mathrm{x})$$

- Hence, $\tilde{\pi}_{\mathrm{y}}(\mathrm{x}) = p(\mathrm{y} \mid \mathrm{x})\, p(\mathrm{x}) \propto p(\mathrm{x} \mid \mathrm{y})$

Boltzmann Distributions

- Given an **energy function** $E : \mathcal{X} \to \mathbb{R}$
  - E can be **learned from data**
  - Sometimes, E is **known** (e.g., physics)
- $E(\mathrm{x})$ is small for likely x:

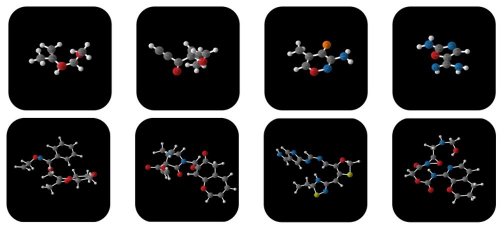
$$\pi(\mathrm{x}) := \frac{\exp(-E(\mathrm{x}))}{\int_{\mathcal{X}} \exp(-E(\mathrm{x}'))\,\mathrm{d}\mathrm{x}'}$$

Why do we care about **samples** from π in the first place? 🤔

Generative Models

- In **Generative Modeling**, **samples** are the quantities of interest ❗
- Generate novel molecules, images, text, audio, video ...

https://www.nytimes.com/2023/04/08/technology/ai-photos-pope-francis.html



Hoogeboom et al., 2023, https://arxiv.org/pdf/2203.17003

Why do we care about **samples** from π in the first place? 🤔

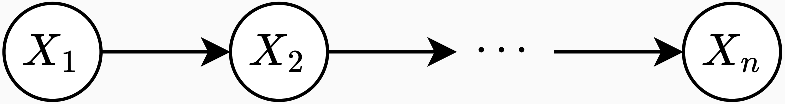
Monte Carlo Estimation

- Often, we care about $\mathbb{E}_{X\sim\pi}[f(X)]$
    - Mean $\mathbb{E}_{X\sim\pi}[X]$, variance $\mathbb{E}_{X\sim\pi}[(X - \mathbb{E}_{X'\sim\pi}[X'])^2]$, higher order moments, ...
- Importantly, in Bayesian methods, we seek the **Posterior Predictive**:

$$p(y_* \mid \mathbf{x}_*, \mathcal{D}) = \int \underbrace{p(y_* \mid \mathbf{x}_*, \boldsymbol{\theta})}_{\text{Likelihood}} \underbrace{p(\boldsymbol{\theta} \mid \mathcal{D})}_{\text{Posterior}} \, \mathrm{d}\boldsymbol{\theta} = \mathbb{E}_{p(\boldsymbol{\theta}\mid\mathcal{D})}[p(y_* \mid \mathbf{x}_*, \boldsymbol{\theta})]$$

- Given i.i.d. samples, we could **estimate all these quantities** with Monte Carlo ❗

Markov Chains

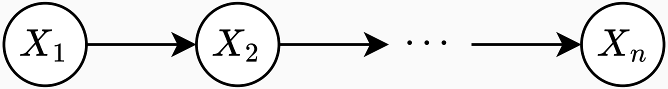


- A **Markov Chain** (MC) consists of n RVs X_i with state space $\mathcal{X}$
- The **joint** factorizes as

$$p(x_1, x_2, \ldots, x_n) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_2) \ldots p(x_n \mid x_{n-1})$$

- **Markov property**: The next state x_{t+1} **only depends on** the current state x_t
  - The **history** $x_1, \ldots, x_{t-1}$ becomes **irrelevant**
- **Ingredients:** *initial* distribution $p(x_1)$, *transition* distributions $p(x_t \mid x_{t-1})$

Markov Chains – Marginals



$$p(x_1, x_2, \ldots, x_n) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_2) \ldots p(x_n \mid x_{n-1})$$

- Assume **finite state space** $\mathcal{X} := \{1, \ldots, s\}$
- Let's think about the **marginal distributions $p_t(x_t)$**:

$$p_1(x_1), \quad p_2(x_2) = \sum_{x_1} p(x_2 \mid x_1)\, p_1(x_1), \quad p_3(x_3) = \sum_{x_2} p(x_3 \mid x_2)\, p_2(x_2),$$

$$\ldots \quad p_n(x_n) = \sum_{x_{n-1}} p(x_n \mid x_{n-1})\, p_{n-1}(x_{n-1})$$

Markov Chains – Sampling

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_n$$

$$p(x_1, x_2, \ldots, x_n) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_2) \ldots p(x_n \mid x_{n-1})$$

How do we sample from $p(x_1, \ldots, x_n)$? 🤔

- Ancestral sampling:

$$x_1 \sim p_1(x_1), \quad x_2 \sim p(x_2 \mid x_1), \quad x_3 \sim p(x_3 \mid x_2), \quad \ldots \quad x_n \sim p(x_n \mid x_{n-1})$$

- The distribution of x_t is $p_t(x_t)$

Goal of MCMC (Informal)

Given π (or $\tilde{\pi}$), construct Markov Chain such that $p_t \approx \pi$ for large t.

Markov Chains – Transition Probability Matrix

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_n$$

- All considered chains will be **time homogeneous**:
- Transition $p(X_t = i \mid X_{t-1} = j)$ will **not depend** on t, only on i, j
- Let's collect them in a **matrix** $P \in \mathbb{R}^{s \times s}$:

$$P_{ij} := p(X_t = i \mid X_{t-1} = j)$$

  i.e., the columns sum to 1.
- We can write the marginal mass function as a vector $\boldsymbol{\mu}_t \in \mathbb{R}^s$:

$$\boldsymbol{\mu}_t := (p_t(X_t = 1), \dots, p_t(X_t = s))^\top$$

Markov Chains – Transition Probability Matrix

$$X_1 \longrightarrow X_2 \longrightarrow \cdots \longrightarrow X_n$$

$$P_{ij} := p(X_t = i \mid X_{t-1} = j)$$

- For all t, we now have:

$$\boldsymbol{\mu}_{t+1} = P\boldsymbol{\mu}_t$$

- Note that

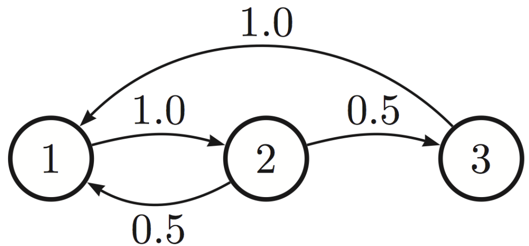
$$\boldsymbol{\mu}_2 = P\boldsymbol{\mu}_1, \quad \boldsymbol{\mu}_3 = P\boldsymbol{\mu}_2 = P(P\boldsymbol{\mu}_1) = P^2\boldsymbol{\mu}_1$$

  where $P^2 = PP$ denotes repeated matrix multiplication

- In general, for all t:

$$\boldsymbol{\mu}_{t+1} = P^t\boldsymbol{\mu}_1 \qquad \left(P^t\right)_{ij} = p(X_t = i \mid X_1 = j)$$

Markov Chain – Example

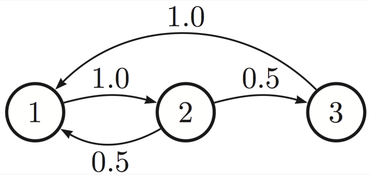


https://ermongroup.github.io/cs228-notes/inference/sampling/

- $\mathcal{X} = \{1, 2, 3\}$
- Transition probability matrix:

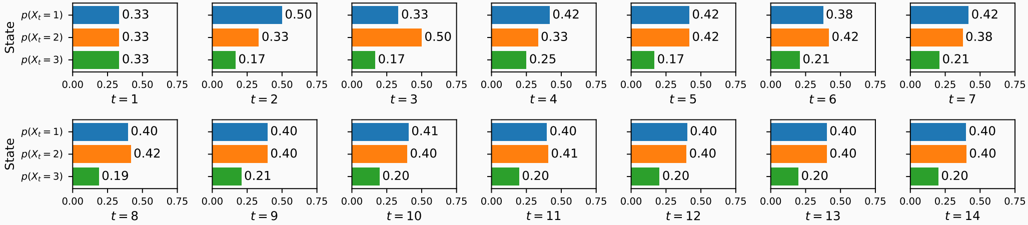
$$P = \begin{bmatrix} 0 & 0.5 & 1 \\ 1 & 0 & 0 \\ 0 & 0.5 & 0 \end{bmatrix}, \quad P_{ij} = p(X_t = i \mid X_{t-1} = j)$$

Markov Chain – Example



https://ermongroup.github.io/cs228-notes/inference/sampling/

- We pick $\boldsymbol{\mu}_1 = (1/3, 1/3, 1/3)^\top$ and iteratively compute $\boldsymbol{\mu}_t = P\boldsymbol{\mu}_{t-1}$



- **Observation 1**: When $\boldsymbol{\mu} = (0.4, 0.4, 0.2)^\top$, we had $\boldsymbol{\mu} = P\boldsymbol{\mu}$ 😮
- Such a $\boldsymbol{\mu}$ is called a **stationary distribution** of the Markov Chain ❗
  - i.e., if $X_t \sim \boldsymbol{\mu}$, then $X_{t+1} \sim \boldsymbol{\mu}$
- **Observation 2**: We actually **converged** to this $\boldsymbol{\mu}$ when starting with our $\boldsymbol{\mu}_1$ 🤯
- In this Markov Chain, it turns out that
  - The **stationary distribution** is **unique**
  - We will converge to it for **any initial distribution** $\boldsymbol{\mu}_1$ 😍

Goal of MCMC in finite state spaces (Less Informal)

Given $\boldsymbol{\pi} \in \mathbb{R}^s$ (or $\tilde{\boldsymbol{\pi}}$), construct P such that $\lim_{t\to\infty} \boldsymbol{\mu}_t = \boldsymbol{\pi}$ for any $\boldsymbol{\mu}_1$.

- For this, $\boldsymbol{\pi}$ needs to be the **unique stationary distribution**

How do we ensure there exists a **unique stationary distribution** ?

Irreducability

A finite-state Markov Chain is **irreducible** if

$$\forall i, j \; \exists t \geq 1 : (P^t)_{ij} > 0.$$

i.e., it's possible to go from any state i to any state j in finite time.

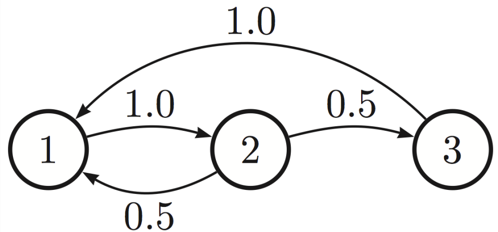
Unique stationary distribution

A finite-state Markov Chain has a **unique stationary distribution**

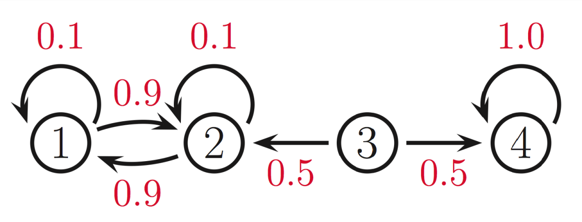
$$\boldsymbol{\mu} = P\boldsymbol{\mu}$$

if and only if it is irreducible.

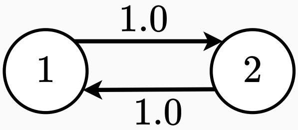
Are the following Markov Chains **irreducible** ?

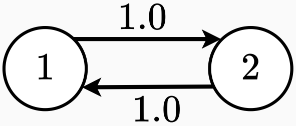


https://ermongroup.github.io/cs228-notes/inference/sampling/



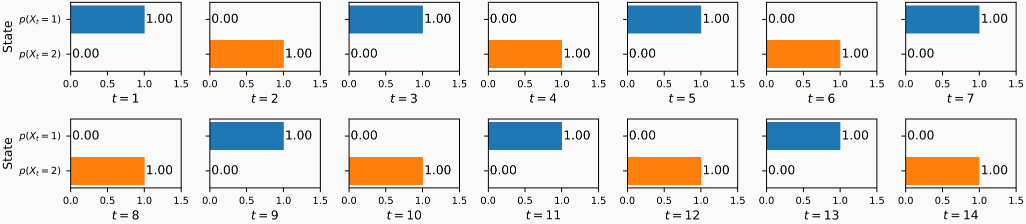
https://ermongroup.github.io/cs228-notes/inference/sampling/





$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad P_{ij} = p(X_t = i \mid X_{t-1} = j)$$

- Since it's irreducible, there exists a unique stationary distribution $\boldsymbol{\mu}$
  - Namely $\boldsymbol{\mu} = (0.5, 0.5)^\top$
- Let's again compute marginal distributions $\boldsymbol{\mu}_t$, starting with $\boldsymbol{\mu}_1 = (1, 0)^\top$



How do we ensure we also **converge to the stationary distribution** ?

Period

Given a finite-state Markov Chain, the **period** of state i is

$$d(i) := \gcd\left(\{t \geq 1 : \left(P^t\right)_{ii} > 0\}\right).$$

Aperiodicity

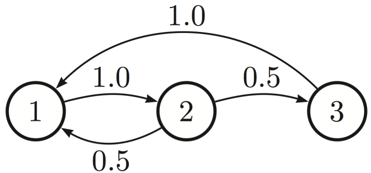
A finite-state Markov Chain is said to be **aperiodic** if $d(i) = 1$ for all i.

Convergence to stationary distribution

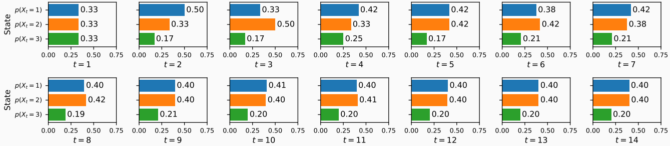
An **irreducible** and **aperiodic** finite-state Markov Chain **converges to its unique stationary distribution** $\boldsymbol{\mu}$:

$$\lim_{t \to \infty} \boldsymbol{\mu}_t = \boldsymbol{\mu}$$

for **any initial distribution** $\boldsymbol{\mu}_1$.



https://ermongroup.github.io/cs228-notes/inference/sampling/



- **Important:** We write down P and (many) $\boldsymbol{\mu}_t$ here for **pedagogical** reasons
  - In practice (e.g. huge s or $\mathcal{X}$ uncountable), this is totally **impossible**
- When **sampling from the chain**, this is **not needed** !
- To **sample** from $\boldsymbol{\mu}_n$
  1. **Sample** x_1 from the **initial distribution**
  2. Repeat: Sample $x_t \sim p(X_t \mid X_{t-1} = x_{t-1})$ for $t = 2, 3, \ldots, n$
  3. Return x_n

- So we wish to design P such that it's **irreducible** and **aperiodic**
- But how do we pick P such that the stationary distribution is $\boldsymbol{\pi}$? 🤔
- It is **nasty** to directly solve for P in

$$\boldsymbol{\pi} = P\boldsymbol{\pi}$$

- Instead, many MCMC algorithms design chains that satisfy **detailed balance**

Detailed Balance

A **sufficient** (*but not necessary*) condition for $\boldsymbol{\pi} = P\boldsymbol{\pi}$ to hold is called **detailed balance**:

$$\forall i, j: \quad P_{ij}\,\pi_j = P_{ji}\,\pi_i$$

i.e., the amount of mass transported from $j \to i$ is the same as from $i \to j$ at stationarity.

Proof.

Assume P satisfies detailed balance at $\boldsymbol{\pi}$. Then, for all i:

$$(P\boldsymbol{\pi})_i = \sum_j P_{ij}\,\pi_j \overset{\text{DB}}{=} \sum_j P_{ji}\,\pi_i = \pi_i \underbrace{\sum_j P_{ji}}_{=1} = \pi_i$$

and hence, $\boldsymbol{\pi} = P\boldsymbol{\pi}$. □

$$\forall i, j : \; P_{ij}\, \pi_j = P_{ji}\, \pi_i$$

$$P = \begin{bmatrix} \frac{7}{12} & \frac{3}{8} & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{6} & \frac{1}{8} & \frac{1}{4} \end{bmatrix}, \quad \boldsymbol{\pi} = \begin{pmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \end{pmatrix}^\top, \quad \boldsymbol{\pi} = P\boldsymbol{\pi}$$

Probability Mass Flow $1 \leftrightarrow 2$

- Flow $1 \rightarrow 2$: $\frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$
- Flow $2 \rightarrow 1$: $\frac{3}{8} \cdot \frac{1}{3} = \frac{1}{8}$

Probability Mass Flow $1 \leftrightarrow 3$

- Flow $1 \rightarrow 3$: $\frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$
- Flow $3 \rightarrow 1$: $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$

Probability Mass Flow $2 \leftrightarrow 3$

- Flow $2 \rightarrow 3$: $\frac{1}{8} \cdot \frac{1}{3} = \frac{1}{24}$
- Flow $3 \rightarrow 2$: $\frac{1}{4} \cdot \frac{1}{6} = \frac{1}{24}$

Metropolis-Hastings

Markov Chain Monte Carlo (MCMC)

Any algorithm that **constructs a Markov Chain**

- that has the target π as the **unique stationary distribution**,
- **converges** to π as $t \to \infty$ (for any initial distribution),
- and **draws samples** from that Markov Chain (**"Monte Carlo Simulation"**)

is called a **Markov Chain Monte Carlo (MCMC)** algorithm.

- We will now look at the **canonical** MCMC algorithm:
- The **Metropolis-Hastings**[1] algorithm

[1]Authorship controversial. Mainly developed by Arianna and Marshall Rosenbluth, while Hastings generalized it.

- For now, again consider finite state space $\mathcal{X} := \{1, \ldots, s\}$
- We wish to construct an **irreducible** and **aperiodic** Markov Chain
- We will use **detailed balance** to enforce that π is the **stationary distribution**
- Of course, the chain should be **easy to sample from**
- **General idea:**
  - When in state X_t, we **propose** a new point $\hat{X}_{t+1}$ by sampling from a **proposal distribution**
  - We will **accept** $\hat{X}_{t+1}$ with a certain probability and **reject** it else, i.e.,

$$X_{t+1} = \begin{cases} \hat{X}_{t+1} & \text{with probability } \alpha \\ X_t & \text{with probability } 1 - \alpha \end{cases}$$

- Consider a **proposal matrix $Q \in \mathbb{R}^{s \times s}$** where

$$Q_{ij} = q(\hat{X}_{t+1} = i \mid X_t = j)$$

  where q is the **proposal distribution**

- We'll pick a q that is **easy to sample from** and induces an **irreducible chain**
  - e.g., q could be the uniform distribution over states "close" to X_t
- When picking $P = Q$, the stationary distribution will in general **not** be $\boldsymbol{\pi}$ **!**

- **Goal**: π should be the **stationary distribution**
- We introduce an **acceptance probability** $\alpha(j \to i) \in [0, 1]$
- For $i \neq j$, we set

$$P_{ij} := Q_{ij}\, \alpha(j \to i)$$

- To ensure the columns of P sum to 1, we set the diagonal to

$$P_{jj} := 1 - \sum_{i \neq j} P_{ij}$$

- Let's pick $\alpha(j \to i)$ such that P satisfies **detailed balance** at $\boldsymbol{\pi}$, i.e., for all i, j:

$$P_{ij}\,\pi_j = P_{ji}\,\pi_i$$

- For $i \neq j$, substitute $P_{ij} = Q_{ij}\,\alpha(j \to i)$:

$$Q_{ij}\,\alpha(j \to i)\,\pi_j = Q_{ji}\,\alpha(i \to j)\,\pi_i$$

- Rearranging yields

$$\frac{\alpha(j \to i)}{\alpha(i \to j)} = \frac{Q_{ji}\,\pi_i}{Q_{ij}\,\pi_j}$$

- There are **infinitely** many choices of $0 \leq \alpha(j \to i), \alpha(i \to j) \leq 1$ that satisfy this
- The **Metropolis-Hastings (MH)** algorithm picks

$$\alpha(j \to i) = \min\left(1, \frac{Q_{ji}\,\pi_i}{Q_{ij}\,\pi_j}\right)$$

which is the largest acceptance rate allowed by detailed balance

$$\alpha(j \to i) = \min\left(1,\ \frac{Q_{ji}\,\pi_i}{Q_{ij}\,\pi_j}\right)$$

- **Note:** We can compute $\alpha(j \to i)$ **even when we only have access to** $\tilde{\pi}$ 😍

$$\frac{Q_{ji}\,\pi_i}{Q_{ij}\,\pi_j} = \frac{Q_{ji}\,\frac{\tilde{\pi}_i}{Z}}{Q_{ij}\,\frac{\tilde{\pi}_j}{Z}} = \frac{Q_{ji}\,\tilde{\pi}_i}{Q_{ij}\,\tilde{\pi}_j}$$

Sample X_1 from **initial distribution** p_1. Repeat for $t = 1, \ldots, n$:

1. Given current state $X_t = j$, **propose**:

$$\hat{X}_{t+1} \sim q(\hat{X}_{t+1} \mid X_t = j)$$

2. Let $\hat{X}_{t+1} = i$. **Compute acceptance probability:**

$$\alpha(j \to i) = \min\left(1, \frac{Q_{ji}\ \tilde{\pi}_i}{Q_{ij}\ \tilde{\pi}_j}\right)$$

3. **Accept or reject**:

$$\text{Sample } U \sim \mathrm{Uniform}(0,1), \qquad X_{t+1} = \begin{cases} \hat{X}_{t+1} & \text{if } U \leq \alpha(j \to i) \\ X_t & \text{else} \end{cases}$$

- For $i \neq j$, detailed balance holds **by construction**, and for $i = j$ it is **trivially true**
- When Q is **irreducible** (on the support of $\boldsymbol{\pi}$), P is **as well**
- P is **aperiodic** because for some j, we have $P_{jj} > 0$[2] since either
  - we pick $Q_{jj} > 0$ for some j, so sometimes we propose to stay in state j, or
  - $\alpha(j \to i) < 1$, i.e., we sometimes reject the proposal and stay in state j
- Thus, $\boldsymbol{\pi}$ is the **unique stationary distribution**, and we will **converge to it** as $t \to \infty$

[2] This implies that an irreducible chain is aperiodic.

Metropolis-Hastings MCMC Demo

Continuous State Spaces

So far, the state space $\mathcal{X}$ was **finite**. What about **continuous** $\mathcal{X}$? 🤔

- The target π is now assumed to be a **probability measure**
  - In practice, this will almost always admit a **density** w.r.t. Lebesgue
- Instead of $P_{ij} = \mathbb{P}(X_{t+1} = i \mid X_t = j)$, we now have a **transition kernel**

$$K(A \mid \mathbf{x}) := \mathbb{P}(X_{t+1} \in A \mid X_t = \mathbf{x})$$

  which is a **probability measure** for every $\mathbf{x} \in \mathcal{X}$
- Instead of evolving $\boldsymbol{\mu}_{t+1} = P\boldsymbol{\mu}_t$, we **integrate against** K:

$$\mu_{t+1}(A) = \int_{\mathcal{X}} K(A \mid \mathbf{x})\, \mu_t(\mathrm{d}\mathbf{x})$$

- **Stationarity** meant $\boldsymbol{\pi} = P\boldsymbol{\pi}$, but now its

$$\pi(A) = \int_{\mathcal{X}} K(A \mid \mathbf{x})\, \pi(\mathrm{d}\mathbf{x})$$

- **Irreducability**, **aperiodicity**, and **detailed balance** are defined **slightly differently**
  - Conceptually the same idea, just with **measures** instead of **mass functions**
- For cont. $\mathcal{X}$, there are irreducible & aperiodic MCs **without stationary distributions**
  - We need an **additional condition** called positive Harris recurrence[3]
  - It makes sure the chain returns infinitely often to sets A with $\pi(A) > 0$
  - Intuitively, we must not "wander off to infinity"
- **Irreducible**, **aperiodic** and **positive Harris recurrent** chains are called **ergodic** !
  - This again guarantees a unique stationary distribution and convergence to it

[3] Finite-state irreducible and aperiodic chains automatically fulfill this.

- **Metropolis-Hastings**: Proposal Q_{ij} becomes **density** $q(y \mid x)$ and

$$\alpha(x \to y) = \min\left(1, \frac{q(x \mid y)\,\tilde{\pi}(y)}{q(y \mid x)\,\tilde{\pi}(x)}\right)$$

  where $\tilde{\pi}$ is the target's unnormalized **density**

- For example, **Gaussian** proposal

$$q(y \mid x) := \mathcal{N}(y; x, \sigma^2 I)$$

- For this choice, we have symmetry $q(y \mid x) = q(x \mid y)$ and thus,

$$\alpha(x \to y) = \min\left(1, \frac{\tilde{\pi}(y)}{\tilde{\pi}(x)}\right)$$

Continuous State Space MCMC Demo

`https://chi-feng.github.io/mcmc-demo/app.html`

Estimating Expectations with MCMC Samples

Recall Goal: Draw i.i.d. samples from π.

Are samples from our Markov Chain **i.i.d.** samples from π? 🤔

No ❗

- If samples were i.i.d., X_t and X_{t+1} would be **independent**
- In the beginning, X_t is not even **marginally** distributed like π (not converged yet)
- Even when the chain has perfectly converged to π, and thus $X_t \sim \pi$, $X_{t+1} \sim \pi$ **marginally**, we still have **possibly large correlation** between X_t and X_{t+1}
- In the finite-state setting:

$$p(X_t = j, X_{t+1} = i) = p(X_t = j)p(X_{t+1} = i \mid X_t = j) = \pi_j P_{ij} \neq \pi_j \pi_i = p(X_t = j)p(X_{t+1} = i)$$

Are we doomed? Monte Carlo estimators need **i.i.d.** samples 😰

Thankfully, no! 🥳

Ergodic Theorem

Assume $X_1, X_2, \ldots$ is an **ergodic** Markov Chain with stationary distribution π. Then,

$$\hat{\mathbb{E}}_T := \frac{1}{T}\sum_{t=1}^{T} f(X_t) \xrightarrow{\text{a.s.}} \mathbb{E}_{X\sim\pi}[f(X)]$$

as $T \to \infty$ for every initial state $X_1 = x$.

- **Almost sure (a.s.) convergence** means
$$\mathbb{P}\left(\lim_{T\to\infty} \hat{\mathbb{E}}_T = \mathbb{E}_{X\sim\pi}[f(X)]\right) = 1$$
- This is the **MCMC counterpart** to classic i.i.d. Monte Carlo

Recap: Properties of MC Estimate

Given **i.i.d.** samples $Y_1, \dots, Y_T \sim \pi$, the **Monte Carlo estimate** of $\mathbb{E}_{X\sim\pi}[f(X)]$ is $\hat{\mathbb{E}}_T^{(MC)} := \frac{1}{T}\sum_{t=1}^T f(Y_t)$.

- **Unbiased** for every T: $\mathbb{E}\left[\hat{\mathbb{E}}_T^{(MC)}\right] = \mathbb{E}_{X\sim\pi}[f(X)]$

Properties of MCMC Estimate

- MCMC estimate is generally[4] **biased** for any finite T
  - We can reduce bias by **discarding the first** B samples (**burn-in phase**)
- Bias goes to 0 as $T \to \infty$, i.e., **asymptotically unbiased**

[4] It would be unbiased if the initial distribution was already π, but we can't do that in practice.

Recap: Properties of MC Estimate

- **Variance** decreases **linearly** with T: $\text{var}\left[\hat{\mathbb{E}}_T^{(\text{MC})}\right] = \frac{\text{var}[f(X)]}{T}$

Properties of MCMC Estimate

- As $T \to \infty$, the **variance** of the MCMC estimator is approximately

$$\text{var}\left[\hat{\mathbb{E}}_T\right] \approx \frac{\sigma_{\text{as}}^2}{T}$$

  with

$$\sigma_{\text{as}}^2 = \text{var}_\pi[f(X)] + 2\sum_{n=1}^{\infty} \text{cov}_\pi(f(X_0), f(X_n)) \,.$$

  where $X_0 \sim \pi, X_1 \sim \pi, \ldots$ (i.e., chain in **stationary** regime)
- **Note:** σ_{as}^2 typically **larger** than $\text{var}_\pi[f(X)]$, but still decreases **linearly** with T

- **Goal**: Sample from **complicated, high-dimensional distribution** π
  - Where typically only $\tilde{\pi}$ is known
- **Idea**: Construct Markov Chain that
  - Has **unique stationary distribution** π
  - Marginal distribution **converges to it** when $T \to \infty$
  - The transition distribution is **easy to sample from**
- **Sampling** from the chain for "long enough" yields
  - Samples with **marginal distribution close to** π
  - That are still **correlated** (even as $T \to \infty$)
  - But that **can still be used to estimate expectations strongly consistently**

Next week, you will hear about...

- More practical, widely used MCMC algorithms
  - Gibb's Sampling
  - Langevin Dynamics
  - Hamiltonian Monte Carlo
- Convergence & Diagnostics
  - Burn-in Phase
  - Autocorrelation
  - Effective Sample Size