

Markov Chain Monte Carlo and variational inference

Chapter 1

Randal Douc and Sholom Schechtman

Télécom SudParis, Institut Polytechnique de Paris
randal.douc@telecom-sudparis.eu



Outline

- ① Introduction
- ② Markov chains and Markov kernels
- ③ Finite dimensional laws
- ④ The canonical space
- ⑤ The Markov property

Outline

- ① Introduction
- ② Markov chains and Markov kernels
- ③ Finite dimensional laws
- ④ The canonical space
- ⑤ The Markov property

Goal:

For a given function f in some class of functions, approximate:

$$\int \pi(\mathrm{d}x) f(x)$$

where the target distribution π is known up to a multiplicative constant:

$$\pi(x) = C\tilde{\pi}(x), \quad \text{where } x \mapsto \tilde{\pi}(x) \text{ is known.}$$

- We use a Markov chain $(X_n)_{n \in \mathbb{N}}$ such that:

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \approx \int \pi(\mathrm{d}x) f(x), \quad n \text{ large.}$$

- Theory of Markov chains: General definitions, invariant measures, ergodicity, Law of Large Numbers, geometric ergodicity, Central Limit theorems. 3 weeks.
- Practice of Markov chains: Metropolis-Hastings Markov chains and variants, Pseudo marginal methods, Hamiltonian MCMC. Alternative methods (Sequential MC, Variational Inference). 3 weeks.

Goal:

For a given function f in some class of functions, approximate:

$$\int \pi(\mathrm{d}x) f(x)$$

where the target distribution π is known up to a multiplicative constant:

$$\pi(x) = C\tilde{\pi}(x), \quad \text{where } x \mapsto \tilde{\pi}(x) \text{ is known.}$$

- We use a **Markov chain** $(X_n)_{n \in \mathbb{N}}$ such that:

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \approx \int \pi(\mathrm{d}x) f(x), \quad n \text{ large.}$$

- **Theory of Markov chains:** General definitions, invariant measures, ergodicity, Law of Large Numbers, geometric ergodicity, Central Limit theorems. **3 weeks.**
- **Practice of Markov chains:** Metropolis-Hastings Markov chains and variants, Pseudo marginal methods, Hamiltonian MCMC. Alternative methods (Sequential MC, Variational Inference). **3 weeks.**

Goal:

For a given function f in some class of functions, approximate:

$$\int \pi(\mathrm{d}x) f(x)$$

where the target distribution π is known up to a multiplicative constant:

$$\pi(x) = C\tilde{\pi}(x), \quad \text{where } x \mapsto \tilde{\pi}(x) \text{ is known.}$$

- We use a **Markov chain** $(X_n)_{n \in \mathbb{N}}$ such that:

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \approx \int \pi(\mathrm{d}x) f(x), \quad n \text{ large.}$$

- **Theory of Markov chains:** General definitions, invariant measures, ergodicity, Law of Large Numbers, geometric ergodicity, Central Limit theorems. **3 weeks.**
- **Practice of Markov chains:** Metropolis-Hastings Markov chains and variants, Pseudo marginal methods, Hamiltonian MCMC. Alternative methods (Sequential MC, Variational Inference). **3 weeks.**

Goal:

For a given function f in some class of functions, approximate:

$$\int \pi(\mathrm{d}x) f(x)$$

where the target distribution π is known up to a multiplicative constant:

$$\pi(x) = C\tilde{\pi}(x), \quad \text{where } x \mapsto \tilde{\pi}(x) \text{ is known.}$$

- We use a **Markov chain** $(X_n)_{n \in \mathbb{N}}$ such that:

$$\frac{1}{n} \sum_{i=0}^{n-1} f(X_i) \approx \int \pi(\mathrm{d}x) f(x), \quad n \text{ large.}$$

- **Theory of Markov chains:** General definitions, invariant measures, ergodicity, Law of Large Numbers, geometric ergodicity, Central Limit theorems. **3 weeks.**
- **Practice of Markov chains:** Metropolis-Hastings Markov chains and variants, Pseudo marginal methods, Hamiltonian MCMC. Alternative methods (Sequential MC, Variational Inference). **3 weeks.**

Outline

- 1 Introduction
- 2 Markov chains and Markov kernels**
- 3 Finite dimensional laws
- 4 The canonical space
- 5 The Markov property

Definitions

Let $(X, \mathcal{X})$ be a measurable space.

Definition (of a Markov kernel)

We say that $P : X \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a **Markov kernel**, if for all $(x, A) \in X \times \mathcal{X}$,

- 1 $y \mapsto P(y, A)$ is $\mathcal{X}/\mathcal{B}(\mathbb{R}^+)$ **measurable**,
- 2 $B \mapsto P(x, B)$ is a **probability measure** on $(X, \mathcal{X})$.

- In particular, $P(x, X) = 1$ for all $x \in X$.
- Recall if ν is a measure on $(X, \mathcal{X})$, $A \mapsto \nu(A)$ is well-defined and we can define the integral associated to ν and we use the notation $\nu(f) = \int f(x)\nu(dx)$,
- Since $P(x, \cdot)$ is a measure, we also use the infinitesimal notation: $P(x, dy)$. For example,

$$P(x, A) = \int_X \mathbf{1}_A(y) P(x, dy) = \int_A P(x, dy).$$

Definitions

Let $(X, \mathcal{X})$ be a measurable space.

Definition (of a Markov kernel)

We say that $P : X \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a **Markov kernel**, if for all $(x, A) \in X \times \mathcal{X}$,

- 1 $y \mapsto P(y, A)$ is $\mathcal{X}/\mathcal{B}(\mathbb{R}^+)$ measurable,
- 2 $B \mapsto P(x, B)$ is a **probability measure** on $(X, \mathcal{X})$.

- In particular, $P(x, X) = 1$ for all $x \in X$.
- Recall if ν is a measure on $(X, \mathcal{X})$, $A \mapsto \nu(A)$ is well-defined and we can define the integral associated to ν and we use the notation $\nu(f) = \int f(x)\nu(dx)$,
- Since $P(x, \cdot)$ is a measure, we also use the infinitesimal notation: $P(x, dy)$. For example,

$$P(x, A) = \int_X \mathbf{1}_A(y)P(x, dy) = \int_A P(x, dy).$$

Definitions

Let $(X, \mathcal{X})$ be a measurable space.

Definition (of a Markov kernel)

We say that $P : X \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a **Markov kernel**, if for all $(x, A) \in X \times \mathcal{X}$,

- ① $y \mapsto P(y, A)$ is $\mathcal{X}/\mathcal{B}(\mathbb{R}^+)$ **measurable**,
- ② $B \mapsto P(x, B)$ is a **probability measure** on $(X, \mathcal{X})$.

- In particular, $P(x, X) = 1$ for all $x \in X$.
- Recall if ν is a measure on $(X, \mathcal{X})$, $A \mapsto \nu(A)$ is well-defined and we can define the integral associated to ν and we use the notation $\nu(f) = \int f(x)\nu(dx)$,
- Since $P(x, \cdot)$ is a measure, we also use the infinitesimal notation: $P(x, dy)$. For example,

$$P(x, A) = \int_X \mathbf{1}_A(y)P(x, dy) = \int_A P(x, dy).$$

Definitions

Let $(X, \mathcal{X})$ be a measurable space.

Definition (of a Markov kernel)

We say that $P : X \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a **Markov kernel**, if for all $(x, A) \in X \times \mathcal{X}$,

- ① $y \mapsto P(y, A)$ is $\mathcal{X}/\mathcal{B}(\mathbb{R}^+)$ **measurable**,
- ② $B \mapsto P(x, B)$ is a **probability measure** on $(X, \mathcal{X})$.

- In particular, $P(x, X) = 1$ for all $x \in X$.
- Recall if ν is a measure on $(X, \mathcal{X})$, $A \mapsto \nu(A)$ is well-defined and we can define the integral associated to ν and we use the notation $\nu(f) = \int f(x)\nu(dx)$,
- Since $P(x, \cdot)$ is a measure, we also use the infinitesimal notation: $P(x, dy)$. For example,

$$P(x, A) = \int_X \mathbf{1}_A(y) P(x, dy) = \int_A P(x, dy).$$

Let $\{X_k : k \in \mathbb{N}\}$ be a sequence of random variables on $(\Omega, \mathcal{G}, \mathbb{P})$ and taking values on X .

Definition (of a Markov chain)

We say that $\{X_k : k \in \mathbb{N}\}$ is a **Markov chain** with Markov kernel P and initial distribution $\nu \in \mathcal{M}_1(X)$ if and only if:

- 1 for all $(k, A) \in \mathbb{N} \times \mathcal{X}$, $\mathbb{P}(X_{k+1} \in A | X_{0:k}) = P(X_k, A)$, $\mathbb{P}$ -a.s.
- 2 $\mathbb{P}(X_0 \in A) = \nu(A)$.

Additional notation

For all $\mu \in \mathbb{M}_+(X)$, all Markov kernels P, Q on $X \times \mathcal{X}$, and all measurable non-negative or bounded functions h on X ,

- ① μP is the (positive) measure:

$$A \mapsto \mu P(A) = \int \mu(dx) P(x, A),$$

- ② PQ is the Markov kernel: $(x, A) \mapsto \int_X P(x, dy) Q(y, A)$,

- ③ Ph is the measurable function: $x \mapsto \int_X P(x, dy) h(y)$.

Example:

$$\begin{aligned} \mu(P(Qh)) &= (\mu P)(Qh) = (\mu(PQ))h = \mu((PQ)h) \\ &= \int \cdots \int_{\mathcal{X}^3} \mu(dx) P(x, dy) Q(y, dz) h(z) = \mu P Q h \end{aligned}$$

Iterates of a kernel:

- define $P^0 = I$ where I is the identity kernel: $(x, A) \mapsto \mathbf{1}_A(x)$,
- set for $k \geq 0$: $P^{k+1} = P^k P$.

Additional notation

For all $\mu \in \mathbb{M}_+(X)$, all Markov kernels P, Q on $X \times \mathcal{X}$, and all measurable non-negative or bounded functions h on X ,

- ① μP is the (positive) measure:

$$A \mapsto \mu P(A) = \int \mu(dx) P(x, A),$$

- ② PQ is the Markov kernel: $(x, A) \mapsto \int_X P(x, dy) Q(y, A)$,

- ③ Ph is the measurable function: $x \mapsto \int_X P(x, dy) h(y)$.

Example:

$$\begin{aligned} \mu(P(Qh)) &= (\mu P)(Qh) = (\mu(PQ))h = \mu((PQ)h) \\ &= \int \cdots \int_{\mathcal{X}^3} \mu(dx) P(x, dy) Q(y, dz) h(z) = \mu PQh \end{aligned}$$

Iterates of a kernel:

- define $P^0 = I$ where I is the identity kernel: $(x, A) \mapsto \mathbf{1}_A(x)$,
- set for $k \geq 0$: $P^{k+1} = P^k P$.

Additional notation

For all $\mu \in \mathbb{M}_+(X)$, all Markov kernels P, Q on $X \times \mathcal{X}$, and all measurable non-negative or bounded functions h on X ,

- ① μP is the (positive) measure:

$$A \mapsto \mu P(A) = \int \mu(dx) P(x, A),$$

- ② PQ is the Markov kernel: $(x, A) \mapsto \int_X P(x, dy) Q(y, A)$,

- ③ Ph is the measurable function: $x \mapsto \int_X P(x, dy) h(y)$.

Example:

$$\begin{aligned} \mu(P(Qh)) &= (\mu P)(Qh) = (\mu(PQ))h = \mu((PQ)h) \\ &= \int \cdots \int_{\mathcal{X}^3} \mu(dx) P(x, dy) Q(y, dz) h(z) = \mu PQh \end{aligned}$$

Iterates of a kernel:

- define $P^0 = I$ where I is the identity kernel: $(x, A) \mapsto \mathbf{1}_A(x)$,
- set for $k \geq 0$: $P^{k+1} = P^k P$.

Outline

- 1 Introduction
- 2 Markov chains and Markov kernels
- 3 Finite dimensional laws**
- 4 The canonical space
- 5 The Markov property

Finite dimensional laws

Let $\{X_k : k \in \mathbb{N}\}$ be a Markov chain with Markov kernel P and initial distribution $\nu \in \mathbb{M}_1(\mathbf{X})$.

Lemma (The joint law)

For any $n \in \mathbb{N}$, the joint law of $X_{0:n}$ is:

$$\nu(\mathrm{d}x_0) \prod_{i=1}^n P(x_{i-1}, \mathrm{d}x_i)$$

(with the convention that $\prod_{i=0}^{-1} = 1$). In particular, the law of X_n is νP^n .

Outline

- 1 Introduction
- 2 Markov chains and Markov kernels
- 3 Finite dimensional laws
- 4 The canonical space**
- 5 The Markov property

The canonical space

- ❶ let P be a Markov kernel on $X \times \mathcal{X}$,
- ❷ let $\nu \in M_1(X)$.

Theorem (The canonical space)

Given (1) and (2), there exists a unique probability measure $\mathbb{P}_\nu$ on the canonical space $(X^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}})$ such that:

- under $\mathbb{P}_\nu$, the coordinate process $\{X_n : n \in \mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution ν .

The canonical space

- ❶ let P be a Markov kernel on $X \times \mathcal{X}$,
- ❷ let $\nu \in M_1(X)$.

Theorem (The canonical space)

Given (1) and (2), there **exists a unique** probability measure $\mathbb{P}_\nu$ on the canonical space $(X^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}})$ such that:

- under $\mathbb{P}_\nu$, the coordinate process $\{X_n : n \in \mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution ν .

❶ We use the notation: $\mathbb{P}_x = \mathbb{P}_{\delta_x}$.

❷ For any $A \in \mathcal{X}^{\otimes(n+1)}$:

$$\mathbb{P}_\nu(X_{0:n} \in A) = \int_{\mathbf{X}} \nu(dx_0) \mathbb{P}_{x_0}(X_{0:n} \in A).$$

❸ We can replace n by ∞ : for all $A \in \mathcal{X}^{\otimes \mathbb{N}}$,

$$\begin{aligned} \mathbb{P}_\nu(A) &= \mathbb{P}_\nu(X_{0:\infty} \in A) = \int_{\mathbf{X}} \nu(dx_0) \mathbb{P}_{x_0}(X_{0:\infty} \in A) \\ &= \int_{\mathbf{X}} \nu(dx_0) \mathbb{P}_{x_0}(A). \end{aligned}$$

- ❶ We use the notation: $\mathbb{P}_x = \mathbb{P}_{\delta_x}$.
- ❷ For any $A \in \mathcal{X}^{\otimes(n+1)}$:

$$\mathbb{P}_\nu(X_{0:n} \in A) = \int_{\mathcal{X}} \nu(dx_0) \mathbb{P}_{x_0}(X_{0:n} \in A).$$

- ❸ We can replace n by ∞ : for all $A \in \mathcal{X}^{\otimes \mathbb{N}}$,

$$\begin{aligned} \mathbb{P}_\nu(A) &= \mathbb{P}_\nu(X_{0:\infty} \in A) = \int_{\mathcal{X}} \nu(dx_0) \mathbb{P}_{x_0}(X_{0:\infty} \in A) \\ &= \int_{\mathcal{X}} \nu(dx_0) \mathbb{P}_{x_0}(A). \end{aligned}$$

- ❶ We use the notation: $\mathbb{P}_x = \mathbb{P}_{\delta_x}$.
- ❷ For any $A \in \mathcal{X}^{\otimes(n+1)}$:

$$\mathbb{P}_\nu(X_{0:n} \in A) = \int_{\mathbf{X}} \nu(\mathrm{d}x_0) \mathbb{P}_{x_0}(X_{0:n} \in A).$$

- ❸ We can replace n by ∞ : for all $A \in \mathcal{X}^{\otimes \mathbb{N}}$,

$$\begin{aligned} \mathbb{P}_\nu(A) &= \mathbb{P}_\nu(X_{0:\infty} \in A) = \int_{\mathbf{X}} \nu(\mathrm{d}x_0) \mathbb{P}_{x_0}(X_{0:\infty} \in A) \\ &= \int_{\mathbf{X}} \nu(\mathrm{d}x_0) \mathbb{P}_{x_0}(A). \end{aligned}$$

Outline

- 1 Introduction
- 2 Markov chains and Markov kernels
- 3 Finite dimensional laws
- 4 The canonical space
- 5 The Markov property**

The Markov property

Theorem (The Markov property)

For any $\nu \in \mathcal{M}_1(\mathcal{X})$, any non-negative or bounded function h on $\mathcal{X}^{\mathbb{N}}$, and any $k \in \mathbb{N}$,

$$\mathbb{E}_{\nu} [h(X_{k:\infty}) \mid \mathcal{F}_k] = \mathbb{E}_{X_k} [h(X_{0:\infty})], \quad \mathbb{P}_{\nu} - a.s.$$

where $\mathcal{F}_k = \sigma(X_{0:k})$.