STAT 547E: LECTURE 2

FUNDAMENTALS OF MCMC

Saifuddin Syed

GOAL FOR TODAY

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc
- Provide multiple equivalent views to think about Markov kernels

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc
- Provide multiple equivalent views to think about Markov kernels
  - Probabilistic

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc
- Provide multiple equivalent views to think about Markov kernels
  - Probabilistic
  - Algorithmic

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc
- Provide multiple equivalent views to think about Markov kernels
  - Probabilistic
  - Algorithmic
  - Algebraic

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc
- Provide multiple equivalent views to think about Markov kernels
  - Probabilistic
  - Algorithmic
  - Algebraic
  - Markov chains

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc
- Provide multiple equivalent views to think about Markov kernels
  - Probabilistic
  - Algorithmic
  - Algebraic
  - Markov chains
- Overview of the properties of Markov kernels and chains

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc
- Provide multiple equivalent views to think about Markov kernels
  - Probabilistic
  - Algorithmic
  - Algebraic
  - Markov chains
- Overview of the properties of Markov kernels and chains
  - Fundamental concepts

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc
- Provide multiple equivalent views to think about Markov kernels
  - Probabilistic
  - Algorithmic
  - Algebraic
  - Markov chains
- Overview of the properties of Markov kernels and chains
  - Fundamental concepts
  - Convergence

GOAL FOR TODAY

- Provide the mathematical language for discussing sampling algorithms
- Formally introduce distributions, statistics, Markov Kernels, etc
- Provide multiple equivalent views to think about Markov kernels
  - Probabilistic
  - Algorithmic
  - Algebraic
  - Markov chains
- Overview of the properties of Markov kernels and chains
  - Fundamental concepts
  - Convergence
- Fundamental recipe of MCMC

PROBABILITY DISTRIBUTIONS

PROBABILITY DISTRIBUTIONS

- Suppose $(\mathbb{X}, \mathscr{F})$ is a measureable space

PROBABILITY DISTRIBUTIONS

- Suppose $(\mathbb{X}, \mathscr{F})$ is a measureable space
- $\mathscr{P}(\mathbb{X})$ be the space of probability distributions over $\mathbb{X}$

PROBABILITY DISTRIBUTIONS

- Suppose $(\mathbb{X}, \mathscr{F})$ is a measureable space
- $\mathscr{P}(\mathbb{X})$ be the space of probability distributions over $\mathbb{X}$
- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, we will assume there is a density over $\mathrm{d}x$

$$\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$$

PROBABILITY DISTRIBUTIONS

- Suppose $(\mathbb{X}, \mathscr{F})$ is a measureable space
- $\mathscr{P}(\mathbb{X})$ be the space of probability distributions over $\mathbb{X}$
- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, we will assume there is a density over $\mathrm{d}x$

$$\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$$

- Given $\mu, \mu' \in \mathscr{P}(\mathbb{X})$ with $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$:

PROBABILITY DISTRIBUTIONS

- Suppose $(\mathbb{X}, \mathscr{F})$ is a measureable space
- $\mathscr{P}(\mathbb{X})$ be the space of probability distributions over $\mathbb{X}$
- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, we will assume there is a density over $\mathrm{d}x$

$$\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$$

- Given $\mu, \mu' \in \mathscr{P}(\mathbb{X})$ with $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$:
  - Define the product measure $\mu_1 \otimes \mu_2 \in \mathscr{P}(\mathbb{X} \times \mathbb{X})$

$$\mu \otimes \mu'(\mathrm{d}x, \mathrm{d}x') = \mu(\mathrm{d}x)\mu'(\mathrm{d}x') = \mu(x)\mu'(x')\mathrm{d}x\mathrm{d}x'$$

PROBABILITY DISTRIBUTIONS

- Suppose $(\mathbb{X}, \mathscr{F})$ is a measureable space
- $\mathscr{P}(\mathbb{X})$ be the space of probability distributions over $\mathbb{X}$
- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, we will assume there is a density over $\mathrm{d}x$

$$\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$$

- Given $\mu, \mu' \in \mathscr{P}(\mathbb{X})$ with $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$:
  - Define the product measure $\mu_1 \otimes \mu_2 \in \mathscr{P}(\mathbb{X} \times \mathbb{X})$

$$\mu \otimes \mu'(\mathrm{d}x, \mathrm{d}x') = \mu(\mathrm{d}x)\mu'(\mathrm{d}x') = \mu(x)\mu'(x')\mathrm{d}x\mathrm{d}x'$$

- E.g. when $\mathbb{X} = \{x_1, \ldots, x_n\}$ is discrete we can represent μ as a n-dimensional row vectors with i -th entry $\mu(x_i)$

FUNCTIONS

FUNCTIONS

- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, and function $f : \mathbb{X} \rightarrow \mathbb{R}$ we define:

FUNCTIONS

- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, and function $f : \mathbb{X} \to \mathbb{R}$ we define:

  - **Expectation:**

$$\mu[f] = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) = \mathbb{E}_{\mu}[f] = \mathbb{E}_{X\sim\mu}[f(X)]$$

FUNCTIONS

- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, and function $f : \mathbb{X} \to \mathbb{R}$ we define:
  - **Expectation:**

$$\mu[f] = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) = \mathbb{E}_{\mu}[f] = \mathbb{E}_{X\sim\mu}[f(X)]$$

  - **Variance:**

$$\mathbb{V}_{\mu}[f] = \mu[f^2] - \mu[f]^2$$

FUNCTIONS

- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, and function $f : \mathbb{X} \to \mathbb{R}$ we define:
  - **Expectation:**
$$\mu[f] = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) = \mathbb{E}_\mu[f] = \mathbb{E}_{X\sim\mu}[f(X)]$$
  - **Variance:**
$$\mathbb{V}_\mu[f] = \mu[f^2] - \mu[f]^2$$
  - **Inner-product:**
$$\langle f, f' \rangle_\mu = \int_{\mathbb{X}} f(x)f'(x)\mu(\mathrm{d}x),$$

FUNCTIONS

- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, and function $f : \mathbb{X} \to \mathbb{R}$ we define:
  - **Expectation:**
  $$\mu[f] = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) = \mathbb{E}_\mu[f] = \mathbb{E}_{X\sim\mu}[f(X)]$$
  - **Variance:**
  $$\mathbb{V}_\mu[f] = \mu[f^2] - \mu[f]^2$$
  - **Inner-product:**
  $$\langle f, f' \rangle_\mu = \int_{\mathbb{X}} f(x)f'(x)\mu(\mathrm{d}x),$$
  - **Norm:**
  $$\|f\|_\mu^2 = \langle f, f \rangle_\mu = \mathbb{V}_\mu[f] + \mathbb{E}_\mu[f]^2$$

FUNCTIONS

- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, and function $f : \mathbb{X} \to \mathbb{R}$ we define:
  - **Expectation:**
  $$\mu[f] = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) = \mathbb{E}_\mu[f] = \mathbb{E}_{X\sim\mu}[f(X)]$$
  - **Variance:**
  $$\mathbb{V}_\mu[f] = \mu[f^2] - \mu[f]^2$$
  - **Inner-product:**
  $$\langle f, f' \rangle_\mu = \int_{\mathbb{X}} f(x)f'(x)\mu(\mathrm{d}x),$$
  - **Norm:**
  $$\|f\|_\mu^2 = \langle f, f \rangle_\mu = \mathbb{V}_\mu[f] + \mathbb{E}_\mu[f]^2$$
- Let $L^2(\mu)$ denote the set of functions with finite norm or equivalent finite variance

FUNCTIONS

- Given a probability distribution $\mu \in \mathscr{P}(\mathbb{X})$, and function $f : \mathbb{X} \to \mathbb{R}$ we define:
  - **Expectation:**
  $$\mu[f] = \int_{\mathbb{X}} f(x)\mu(\mathrm{d}x) = \mathbb{E}_\mu[f] = \mathbb{E}_{X\sim\mu}[f(X)]$$
  - **Variance:**
  $$\mathbb{V}_\mu[f] = \mu[f^2] - \mu[f]^2$$
  - **Inner-product:**
  $$\langle f, f' \rangle_\mu = \int_{\mathbb{X}} f(x)f'(x)\mu(\mathrm{d}x),$$
  - **Norm:**
  $$\|f\|_\mu^2 = \langle f, f \rangle_\mu = \mathbb{V}_\mu[f] + \mathbb{E}_\mu[f]^2$$
- Let $L^2(\mu)$ denote the set of functions with finite norm or equivalent finite variance
- E.g. when $\mathbb{X} = \{x_1, \ldots, x_n\}$ is discrete we can represent f as a n-dimensional column vectors with i-th entry $f(x_i)$
$$\mu[f] = \sum_x \mu(x)f(x)$$

MARKOV KERNELS V1: PROBABILITY

MARKOV KERNELS V1: PROBABILITY

- A Markov kernel $K : \mathbb{X} \times \mathscr{F} \to [0,1]$ is a function such that

MARKOV KERNELS V1: PROBABILITY

- A Markov kernel $K : \mathbb{X} \times \mathscr{F} \to [0,1]$ is a function such that

    1. For all $x \in \mathbb{X}, \quad A \mapsto K(x, A) \in \mathscr{P}(\mathbb{X})$,

MARKOV KERNELS V1: PROBABILITY

- A Markov kernel $K : \mathbb{X} \times \mathscr{F} \to [0,1]$ is a function such that

  1. For all $x \in \mathbb{X}, \quad A \mapsto K(x, A) \in \mathscr{P}(\mathbb{X})$,

  2. For all $A \in \mathscr{F}, \quad x \mapsto K(x, A)$ is $\mathscr{F}$-measurable

MARKOV KERNELS V1: PROBABILITY

- A Markov kernel $K : \mathbb{X} \times \mathscr{F} \to [0,1]$ is a function such that
  1. For all $x \in \mathbb{X}, \quad A \mapsto K(x, A) \in \mathscr{P}(\mathbb{X})$,
  2. For all $A \in \mathscr{F}, \quad x \mapsto K(x, A)$ is $\mathscr{F}$-measurable
- In general we will denote $K(x, \mathrm{d}x')$ as the measure may not have a density over $\mathrm{d}x'$

MARKOV KERNELS V1: PROBABILITY

- A Markov kernel $K : \mathbb{X} \times \mathscr{F} \to [0,1]$ is a function such that

  1. For all $x \in \mathbb{X}, \quad A \mapsto K(x, A) \in \mathscr{P}(\mathbb{X})$,

  2. For all $A \in \mathscr{F}, \quad x \mapsto K(x, A)$ is $\mathscr{F}$-measurable

- In general we will denote $K(x, \mathrm{d}x')$ as the measure may not have a density over $\mathrm{d}x'$
  - We will abuse notation and use $K(x, x')$ it interchangeably

MARKOV KERNELS V1: PROBABILITY

- A Markov kernel $K : \mathbb{X} \times \mathscr{F} \to [0,1]$ is a function such that
  1. For all $x \in \mathbb{X}, \quad A \mapsto K(x, A) \in \mathscr{P}(\mathbb{X})$,
  2. For all $A \in \mathscr{F}, \quad x \mapsto K(x, A)$ is $\mathscr{F}$-measurable
- In general we will denote $K(x, \mathrm{d}x')$ as the measure may not have a density over $\mathrm{d}x'$
  - We will abuse notation and use $K(x, x')$ it interchangeably
- E.g. when $\mathbb{X} = \{x_1, \ldots, x_n\}$ is discrete:
  - We can represent K the a $n \times n$-dimensional square matrix with entries $K(x_i, x_j)$
  - Satisfies $K(x, x') \geq 0$ the rows sum to 1

$$\sum_{x'} K(x, x') = 1$$

MARKOV KERNELS V2: ALGORITHMS

MARKOV KERNELS V2: ALGORITHMS

- Kernels mathematically express algorithmic moves:

MARKOV KERNELS V2: ALGORITHMS

- Kernels mathematically express algorithmic moves:
  - A kernel K provides instructions to move samples $X \in \mathbb{X}$ to $X' \in \mathbb{X}$

$$X' \sim K(X, \mathrm{d}x') \tag{1}$$

MARKOV KERNELS V2: ALGORITHMS

- Kernels mathematically express algorithmic moves:
  - A kernel K provides instructions to move samples $X \in \mathbb{X}$ to $X' \in \mathbb{X}$

$$X' \sim K(X, \mathrm{d}x') \tag{1}$$

  - Conversely, given a set of instructions for generating $X' \in \mathbb{X}$ to $X \in \mathbb{X}$, there exists a kernel K such that (1) holds

MARKOV KERNELS V2: ALGORITHMS

- Kernels mathematically express algorithmic moves:
  - A kernel K provides instructions to move samples $X \in \mathbb{X}$ to $X' \in \mathbb{X}$

$$X' \sim K(X, \mathrm{d}x') \tag{1}$$

  - Conversely, given a set of instructions for generating $X' \in \mathbb{X}$ to $X \in \mathbb{X}$, there exists a kernel K such that (1) holds

- Both representations are important!

MARKOV KERNELS V2: ALGORITHMS

- Kernels mathematically express algorithmic moves:
  - A kernel K provides instructions to move samples $X \in \mathbb{X}$ to $X' \in \mathbb{X}$

$$X' \sim K(X, \mathrm{d}x') \tag{1}$$

  - Conversely, given a set of instructions for generating $X' \in \mathbb{X}$ to $X \in \mathbb{X}$, there exists a kernel K such that (1) holds

- Both representations are important!
  - Kernels allow us to formally study the correctness and mathematical properties of a given algorithm

MARKOV KERNELS V2: ALGORITHMS

- Kernels mathematically express algorithmic moves:
  - A kernel K provides instructions to move samples $X \in \mathbb{X}$ to $X' \in \mathbb{X}$

$$X' \sim K(X, \mathrm{d}x') \tag{1}$$

  - Conversely, given a set of instructions for generating $X' \in \mathbb{X}$ to $X \in \mathbb{X}$, there exists a kernel K such that (1) holds

- Both representations are important!
  - Kernels allow us to formally study the correctness and mathematical properties of a given algorithm
  - Instructions (i.e. pseudo-code) provide implementation details and analyse algorithmic complexity.

MARKOV KERNELS V2: ALGORITHMS

- Kernels mathematically express algorithmic moves:
  - A kernel K provides instructions to move samples $X \in \mathbb{X}$ to $X' \in \mathbb{X}$

$$X' \sim K(X, \mathrm{d}x') \tag{1}$$

  - Conversely, given a set of instructions for generating $X' \in \mathbb{X}$ to $X \in \mathbb{X}$, there exists a kernel K such that (1) holds

- Both representations are important!
  - Kernels allow us to formally study the correctness and mathematical properties of a given algorithm
  - Instructions (i.e. pseudo-code) provide implementation details and analyse algorithmic complexity.

- Define $\mu \otimes K \in \mathscr{P}(\mathbb{X} \times \mathbb{X})$ as the joint law $X \sim \mu$ and $X' \sim K(X, \mathrm{d}x')$

$$\mu \otimes K(\mathrm{d}x, \mathrm{d}x') = \mu(\mathrm{dx})K(x, \mathrm{d}x')$$

EXAMPLES:

EXAMPLES:

- **Identity kernel:**

$$K(x, \mathrm{d}x') = \delta_x(\mathrm{d}x') \quad \Longleftrightarrow \quad \begin{array}{l} \texttt{1. Input } X \\ \texttt{2. Return } X' = X \end{array}$$

EXAMPLES:

- **Identity kernel:**

$$K(x, \mathrm{d}x') = \delta_x(\mathrm{d}x') \quad \iff \quad$$

```
1. Input X
2. Return X' = X
```

- **Transport kernel:** Given $T : \mathbb{X} \to \mathbb{X}$

$$K(x, \mathrm{d}x) = \delta_{T(x)}(\mathrm{dx}') \quad \iff \quad$$

```
1. Input X
2. Return X' = T(X)
```

EXAMPLES:

- **Identity kernel:**

$$K(x, \mathrm{d}x') = \delta_x(\mathrm{d}x') \quad \Longleftrightarrow \quad \begin{array}{l} \texttt{1. Input } X \\ \texttt{2. Return } X' = X \end{array}$$

- **Transport kernel:** Given $T : \mathbb{X} \to \mathbb{X}$

$$K(x, \mathrm{d}x) = \delta_{T(x)}(\mathrm{d}x') \quad \Longleftrightarrow \quad \begin{array}{l} \texttt{1. Input } X \\ \texttt{2. Return } X' = T(X) \end{array}$$

- **Independent kernel:** Given $\eta \in \mathscr{P}(\mathbb{X})$

$$K(x, \mathrm{d}x') = \eta(\mathrm{d}x') \quad \Longleftrightarrow \quad \begin{array}{l} \texttt{1. Input } X \\ \texttt{2. Return } X' \sim \eta \end{array}$$

EXAMPLES:

- **Identity kernel:**

$$K(x, \mathrm{d}x') = \delta_x(\mathrm{d}x') \quad \Longleftrightarrow \quad \begin{array}{l} \texttt{1. Input } X \\ \texttt{2. Return } X' = X \end{array}$$

- **Transport kernel:** Given $T : \mathbb{X} \to \mathbb{X}$

$$K(x, \mathrm{d}x) = \delta_{T(x)}(\mathrm{dx}') \quad \Longleftrightarrow \quad \begin{array}{l} \texttt{1. Input } X \\ \texttt{2. Return } X' = T(X) \end{array}$$

- **Independent kernel:** Given $\eta \in \mathscr{P}(\mathbb{X})$

$$K(x, \mathrm{d}x') = \eta(\mathrm{dx}') \quad \Longleftrightarrow \quad \begin{array}{l} \texttt{1. Input } X \\ \texttt{2. Return } X' \sim \eta \end{array}$$

- **Random Walk:** If $\mathbb{X} = \mathbb{R}^d$

$$K(x, \mathrm{d}x') = N(\mu(x), \Sigma(x), \mathrm{d}x') \quad \Longleftrightarrow \quad \begin{array}{l} \texttt{1. Input } X \\ \texttt{2. Return } X' \sim N(\mu(X), \Sigma(X)) \end{array}$$

MARKOV KERNELS V3: OPERATORS

MARKOV KERNELS V3: OPERATORS

- Given $\mu \in \mathscr{P}(\mathbb{X})$ we can view a kernel K as operators in $L^2(\mu)$

MARKOV KERNELS V3: OPERATORS

- Given $\mu \in \mathscr{P}(\mathbb{X})$ we can view a kernel K as operators in $L^2(\mu)$

- **Left multiplication:** given $\mu \in \mathscr{P}(\mathbb{X})$ define $\mu K \in \mathscr{P}(\mathbb{X})$

$$\mu K(\mathrm{d}x') = \int_{\mathbb{X}} \mu(\mathrm{d}x) K(x, \mathrm{d}x')$$

MARKOV KERNELS V3: OPERATORS

- Given $\mu \in \mathscr{P}(\mathbb{X})$ we can view a kernel K as operators in $L^2(\mu)$

- **Left multiplication:** given $\mu \in \mathscr{P}(\mathbb{X})$ define $\mu K \in \mathscr{P}(\mathbb{X})$

$$\mu K(\mathrm{d}x') = \int_{\mathbb{X}} \mu(\mathrm{d}x) K(x, \mathrm{d}x')$$

  - When discrete equivalent to left multiplication by row vector

$$\mu K(x') = \sum_{x} \mu(x) K(x, x')$$

MARKOV KERNELS V3: OPERATORS

- Given $\mu \in \mathscr{P}(\mathbb{X})$ we can view a kernel K as operators in $L^2(\mu)$
- **Left multiplication:** given $\mu \in \mathscr{P}(\mathbb{X})$ define $\mu K \in \mathscr{P}(\mathbb{X})$

$$\mu K(\mathrm{d}x') = \int_{\mathbb{X}} \mu(\mathrm{d}x) K(x, \mathrm{d}x')$$

  - When discrete equivalent to left multiplication by row vector

$$\mu K(x') = \sum_x \mu(x) K(x, x')$$

- **Right multiplication:** given $f : \mathbb{X} \to \mathbb{R}$ define $Kf : \mathbb{X} \to \mathbb{R}$

$$Kf(x) = \int_{\mathbb{X}} K(x, \mathrm{d}x') f(x')$$

MARKOV KERNELS V3: OPERATORS

- Given $\mu \in \mathscr{P}(\mathbb{X})$ we can view a kernel K as operators in $L^2(\mu)$

- **Left multiplication:** given $\mu \in \mathscr{P}(\mathbb{X})$ define $\mu K \in \mathscr{P}(\mathbb{X})$

$$\mu K(\mathrm{d}x') = \int_{\mathbb{X}} \mu(\mathrm{d}x) K(x, \mathrm{d}x')$$

  - When discrete equivalent to left multiplication by row vector

$$\mu K(x') = \sum_{x} \mu(x) K(x, x')$$

- **Right multiplication:** given $f : \mathbb{X} \to \mathbb{R}$ define $Kf : \mathbb{X} \to \mathbb{R}$

$$Kf(x) = \int_{\mathbb{X}} K(x, \mathrm{d}x') f(x')$$

  - When discrete equivalent to right multiplication by column vector

$$Kf(x) = \sum_{x'} K(x, x') f(x')$$

KERNELS AS BUILDING BLOCKS

KERNELS AS BUILDING BLOCKS

- Kernels act as building blocks to construct algorithms:

KERNELS AS BUILDING BLOCKS

- Kernels act as building blocks to construct algorithms:
- **Product:** Given kernels K_1 and K_2 define the product kernel

$$K_1K_2(x, \mathrm{d}x') = \int_{\mathbb{X}} K_1(x, \mathrm{d}y)K_2(y, \mathrm{d}x')$$

KERNELS AS BUILDING BLOCKS

- Kernels act as building blocks to construct algorithms:
- **Product:** Given kernels K_1 and K_2 define the product kernel

$$K_1K_2(x, \mathrm{d}x') = \int_{\mathbb{X}} K_1(x, \mathrm{d}y)K_2(y, \mathrm{d}x')$$

  - Given K we denote $K^t = K \cdots K$ as the t-times composition of K

KERNELS AS BUILDING BLOCKS

- Kernels act as building blocks to construct algorithms:
- **Product:** Given kernels K_1 and K_2 define the product kernel

$$K_1K_2(x, \mathrm{d}x') = \int_{\mathbb{X}} K_1(x, \mathrm{d}y)K_2(y, \mathrm{d}x')$$

  - Given K we denote $K^t = K \cdots K$ as the t-times composition of K
  - In discrete case corresponds to matrix multiplication

$$(K_1K_2)(x, x') = \sum_y K_1(x, y)K_2(y, x')$$

KERNELS AS BUILDING BLOCKS

- Kernels act as building blocks to construct algorithms:
- **Product:** Given kernels K_1 and K_2 define the product kernel

$$K_1K_2(x, \mathrm{d}x') = \int_{\mathbb{X}} K_1(x, \mathrm{d}y)K_2(y, \mathrm{d}x')$$

  - Given K we denote $K^t = K \cdots K$ as the t-times composition of K
  - In discrete case corresponds to matrix multiplication

$$(K_1K_2)(x, x') = \sum_y K_1(x, y)K_2(y, x')$$

  - Algorithmically corresponds to composition of algorithms

1. `Input` X
2. $Y \sim K_1(X, \mathrm{d}y)$
3. `Return` $X' = K_2(Y, \mathrm{dx}')$

KERNELS AS BUILDING BLOCKS

KERNELS AS BUILDING BLOCKS

- Kernels act as building blocks to construct algorithms:

KERNELS AS BUILDING BLOCKS

- Kernels act as building blocks to construct algorithms:

- **Mixture:** Given kernels K_1 and K_2 and $\alpha : \mathbb{X} \to [0,1]$

$$[\alpha K_1 + (1-\alpha)K_2](x, \mathrm{d}x') = \alpha(x)K_1(x, \mathrm{d}x') + (1-\alpha(x))K_2(x, \mathrm{d}x')$$

KERNELS AS BUILDING BLOCKS

- Kernels act as building blocks to construct algorithms:
- **Mixture:** Given kernels K_1 and K_2 and $\alpha : \mathbb{X} \to [0,1]$

$$[\alpha K_1 + (1-\alpha)K_2](x, \mathrm{d}x') = \alpha(x)K_1(x, \mathrm{d}x') + (1-\alpha(x))K_2(x, \mathrm{d}x')$$

  - In discrete case corresponds to convex combination

$$(\alpha K + (1-\alpha)K')(x, x') = \alpha_i K(x, x') + (1-\alpha_i)K'(x, x')$$

KERNELS AS BUILDING BLOCKS

- Kernels act as building blocks to construct algorithms:
- **Mixture:** Given kernels K_1 and K_2 and $\alpha : \mathbb{X} \to [0,1]$

$$[\alpha K_1 + (1-\alpha)K_2](x, \mathrm{d}x') = \alpha(x)K_1(x, \mathrm{d}x') + (1-\alpha(x))K_2(x, \mathrm{d}x')$$

  - In discrete case corresponds to convex combination

$$(\alpha K + (1-\alpha)K')(x, x') = \alpha_i K(x, x') + (1-\alpha_i)K'(x, x')$$

  - Algorithmically corresponds to stochastically choosing algorithm

1. `Input` X
2. `Generate` $U \sim \mathrm{Uniform}([0,1])$
3. `If` $U < \alpha(X)$ `return` $X' \sim K_1(X, \mathrm{d}x')$
4. `Else return` $X' = K_2(X, \mathrm{d}x')$

INVARIANCE

INVARIANCE

- We will say a K is π-invariant or π-stationary if $\pi K = \pi$

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

INVARIANCE

- We will say a K is π-invariant or π-stationary if $\pi K = \pi$

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

  - Dicrete case:

$$\sum_x \pi(x) K(x, x') = \pi(x)$$

INVARIANCE

- We will say a K is π-invariant or π-stationary if $\pi K = \pi$

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

  - Dicrete case: $$\sum_x \pi(x) K(x, x') = \pi(x)$$
  - Corresponds to the left Eigenvector with Eigenvalue of 1

INVARIANCE

- We will say a K is π-invariant or π-stationary if $\pi K = \pi$

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

  - Dicrete case: $$\sum_x \pi(x) K(x, x') = \pi(x)$$

  - Corresponds to the left Eigenvector with Eigenvalue of 1

  - Algorithmically, the distribution of $X \sim \pi$ is unchanged $X' \sim K(X, \mathrm{d}x') = \pi$

INVARIANCE

- We will say a K is π-invariant or π-stationary if $\pi K = \pi$

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

  - Dicrete case: $$\sum_x \pi(x) K(x, x') = \pi(x)$$
  - Corresponds to the left Eigenvector with Eigenvalue of 1
  - Algorithmically, the distribution of $X \sim \pi$ is unchanged $X' \sim K(X, \mathrm{d}x') = \pi$
- Products and mixtures of π-invariant kernels are π-invariant

INVARIANCE

- We will say a K is π-invariant or π-stationary if $\pi K = \pi$

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

  - Dicrete case:

$$\sum_x \pi(x) K(x, x') = \pi(x)$$

  - Corresponds to the left Eigenvector with Eigenvalue of 1
  - Algorithmically, the distribution of $X \sim \pi$ is unchanged $X' \sim K(X, \mathrm{d}x') = \pi$

- Products and mixtures of π-invariant kernels are π-invariant
- Invariant distributions may not be unique

INVARIANCE

- We will say a K is π-invariant or π-stationary if $\pi K = \pi$

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

  - Dicrete case:

$$\sum_x \pi(x) K(x, x') = \pi(x)$$

  - Corresponds to the left Eigenvector with Eigenvalue of 1
  - Algorithmically, the distribution of $X \sim \pi$ is unchanged $X' \sim K(X, \mathrm{d}x') = \pi$
- Products and mixtures of π-invariant kernels are π-invariant
- Invariant distributions may not be unique
  - E.g. identity kernel is invariant to every distribution

$$K(x, \mathrm{d}x') = \delta_x(\mathrm{d}x')$$

$$\pi K(\mathrm{d}x') = \int \pi(\mathrm{d}x) \delta_x(\mathrm{dx}') = \pi(\mathrm{d}x')$$

REVERSIBILITY

REVERSIBILITY

- We will say a K is π-reversible if the **detailed balance condition** hold:

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

REVERSIBILITY

- We will say a K is π-reversible if the **detailed balance condition** hold:

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

  - In descrete setting equivalent to

$$\pi(x)K(x, x') = \pi(x')K(x', x)$$

REVERSIBILITY

- We will say a K is π-reversible if the **detailed balance condition** hold:

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

  - In descrete setting equivalent to

$$\pi(x)K(x, x') = \pi(x')K(x', x)$$

  - **Intuition:** the mass flowing in is equivalent to the mass flowing out

REVERSIBILITY

- We will say a K is π-reversible if the **detailed balance condition** hold:

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

  - In descrete setting equivalent to

$$\pi(x)K(x, x') = \pi(x')K(x', x)$$

  - **Intuition:** the mass flowing in is equivalent to the mass flowing out

- **Proposition:** If K is π-reversible, then it is π-invariant

REVERSIBILITY

- We will say a K is π-reversible if the **detailed balance condition** hold:

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

  - In descrete setting equivalent to

$$\pi(x)K(x, x') = \pi(x')K(x', x)$$

  - **Intuition:** the mass flowing in is equivalent to the mass flowing out

- **Proposition:** If K is π-reversible, then it is π-invariant

- **Proof:** For any bounded $f : \mathbb{X} \to \mathbb{R}$

REVERSIBILITY

- We will say a K is π-reversible if the **detailed balance condition** hold:

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

  - In descrete setting equivalent to

$$\pi(x)K(x, x') = \pi(x')K(x', x)$$

  - **Intuition:** the mass flowing in is equivalent to the mass flowing out

- **Proposition:** If K is π-reversible, then it is π-invariant

- **Proof:** For any bounded $f : \mathbb{X} \to \mathbb{R}$

$$\int_{\mathbb{X}} f(x')\pi K(\mathrm{dx}') = \int_{\mathbb{X}^2} f(x')\pi(\mathrm{d}x)K(x, \mathrm{d}x')$$

REVERSIBILITY

- We will say a K is π-reversible if the **detailed balance condition** hold:

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

  - In descrete setting equivalent to

$$\pi(x)K(x, x') = \pi(x')K(x', x)$$

  - **Intuition:** the mass flowing in is equivalent to the mass flowing out

- **Proposition:** If K is π-reversible, then it is π-invariant

- **Proof:** For any bounded $f : \mathbb{X} \to \mathbb{R}$

$$\begin{aligned}\int_{\mathbb{X}} f(x')\pi K(\mathrm{dx}') &= \int_{\mathbb{X}^2} f(x')\pi(\mathrm{d}x)K(x, \mathrm{d}x') \\ &= \int_{\mathbb{X}^2} f(x')\pi(\mathrm{d}x')K(x', \mathrm{d}x)\end{aligned}$$

REVERSIBILITY

- We will say a K is π-reversible if the **detailed balance condition** hold:

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

  - In descrete setting equivalent to

$$\pi(x)K(x, x') = \pi(x')K(x', x)$$

  - **Intuition:** the mass flowing in is equivalent to the mass flowing out

- **Proposition:** If K is π-reversible, then it is π-invariant

- **Proof:** For any bounded $f : \mathbb{X} \to \mathbb{R}$

$$\begin{aligned} \int_{\mathbb{X}} f(x')\pi K(\mathrm{dx}') &= \int_{\mathbb{X}^2} f(x')\pi(\mathrm{d}x)K(x, \mathrm{d}x') \\ &= \int_{\mathbb{X}^2} f(x')\pi(\mathrm{d}x')K(x', \mathrm{d}x) \\ &= \int_{\mathbb{X}^2} f(x')\pi(\mathrm{d}x') \end{aligned}$$

TRANSPOSE OF KERNEL

TRANSPOSE OF KERNEL

- Given an π-invariant kernel, K in $L^2(\pi)$ we define the adjoint (or transpose) as $K^\top$ if all $f, g \in L^2(\pi)$ we have

$$\langle Kf, g \rangle_\pi = \langle f, K^\top g \rangle_\pi$$

TRANSPOSE OF KERNEL

- Given an π-invariant kernel, K in $L^2(\pi)$ we define the adjoint (or transpose) as $K^\top$ if all $f, g \in L^2(\pi)$ we have

$$\langle Kf, g \rangle_\pi = \langle f, K^\top g \rangle_\pi$$

- **Proposition:** The adjoint kernel satisfies

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \pi(\mathrm{d}x')K^\top(x', \mathrm{d}x)$$

TRANSPOSE OF KERNEL

- Given an π-invariant kernel, K in $L^2(\pi)$ we define the adjoint (or transpose) as $K^\top$ if all $f, g \in L^2(\pi)$ we have

$$\langle Kf, g\rangle_\pi = \langle f, K^\top g\rangle_\pi$$

- **Proposition:** The adjoint kernel satisfies

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \pi(\mathrm{d}x')K^\top(x', \mathrm{d}x)$$

  - K is π-reversible if and only if K is self-adjoint $K = K^\top$

TRANSPOSE OF KERNEL

- Given an π-invariant kernel, K in $L^2(\pi)$ we define the adjoint (or transpose) as $K^\top$ if all $f, g \in L^2(\pi)$ we have

$$\langle Kf, g\rangle_\pi = \langle f, K^\top g\rangle_\pi$$

- **Proposition:** The adjoint kernel satisfies

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \pi(\mathrm{d}x')K^\top(x', \mathrm{d}x)$$

  - K is π-reversible if and only if K is self-adjoint $K = K^\top$

- **Proof:** Using the definition we compute both sides:

TRANSPOSE OF KERNEL

- Given an π-invariant kernel, K in $L^2(\pi)$ we define the adjoint (or transpose) as $K^\top$ if all $f, g \in L^2(\pi)$ we have

$$\langle Kf, g\rangle_\pi = \langle f, K^\top g\rangle_\pi$$

- **Proposition:** The adjoint kernel satisfies

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \pi(\mathrm{d}x')K^\top(x', \mathrm{d}x)$$

  - K is π-reversible if and only if K is self-adjoint $K = K^\top$

- **Proof:** Using the definition we compute both sides:

$$\langle Kf, g\rangle_\pi = \int_{\mathbb{X}} Kf(x)g(x)\pi(\mathrm{d}x)$$

TRANSPOSE OF KERNEL

- Given an π-invariant kernel, K in $L^2(\pi)$ we define the adjoint (or transpose) as $K^\top$ if all $f, g \in L^2(\pi)$ we have

$$\langle Kf, g \rangle_\pi = \langle f, K^\top g \rangle_\pi$$

- **Proposition:** The adjoint kernel satisfies

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \pi(\mathrm{d}x')K^\top(x', \mathrm{d}x)$$

  - K is π-reversible if and only if K is self-adjoint $K = K^\top$

- **Proof:** Using the definition we compute both sides:

$$\langle Kf, g \rangle_\pi = \int_{\mathbb{X}} Kf(x)g(x)\pi(\mathrm{d}x) = \int_{\mathbb{X}} \int_{\mathbb{X}} K(x, \mathrm{d}x')f(x')g(x)\pi(\mathrm{d}x)$$

TRANSPOSE OF KERNEL

- Given an π-invariant kernel, K in $L^2(\pi)$ we define the adjoint (or transpose) as $K^\top$ if all $f, g \in L^2(\pi)$ we have

$$\langle Kf, g \rangle_\pi = \langle f, K^\top g \rangle_\pi$$

- **Proposition:** The adjoint kernel satisfies

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \pi(\mathrm{d}x')K^\top(x', \mathrm{d}x)$$

  - K is π-reversible if and only if K is self-adjoint $K = K^\top$

- **Proof:** Using the definition we compute both sides:

$$\langle Kf, g \rangle_\pi = \int_{\mathbb{X}} Kf(x)g(x)\pi(\mathrm{d}x) = \int_{\mathbb{X}} \int_{\mathbb{X}} K(x, \mathrm{d}x')f(x')g(x)\pi(\mathrm{d}x)$$

$$\langle f, K^\top g \rangle_\pi = \int_{\mathbb{X}} f(x')K^\top g(x')\pi(\mathrm{d}x')$$

TRANSPOSE OF KERNEL

- Given an π-invariant kernel, K in $L^2(\pi)$ we define the adjoint (or transpose) as $K^\top$ if all $f, g \in L^2(\pi)$ we have

$$\langle Kf, g \rangle_\pi = \langle f, K^\top g \rangle_\pi$$

- **Proposition:** The adjoint kernel satisfies

$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \pi(\mathrm{d}x')K^\top(x', \mathrm{d}x)$$

  - K is π-reversible if and only if K is self-adjoint $K = K^\top$

- **Proof:** Using the definition we compute both sides:

$$\langle Kf, g \rangle_\pi = \int_{\mathbb{X}} Kf(x)g(x)\pi(\mathrm{d}x) \quad = \int_{\mathbb{X}} \int_{\mathbb{X}} K(x, \mathrm{d}x')f(x')g(x)\pi(\mathrm{d}x)$$

$$\langle f, K^\top g \rangle_\pi = \int_{\mathbb{X}} f(x')K^\top g(x')\pi(\mathrm{d}x') \quad = \int_{\mathbb{X}} \int_{\mathbb{X}} f(x')K^\top(x', \mathrm{d}x)g(x)\pi(\mathrm{d}x')$$

TRANSPOSE OF KERNEL

- Given an π-invariant kernel, K in $L^2(\pi)$ we define the adjoint (or transpose) as $K^\top$ if all $f, g \in L^2(\pi)$ we have
$$\langle Kf, g\rangle_\pi = \langle f, K^\top g\rangle_\pi$$
- **Proposition:** The adjoint kernel satisfies
$$\pi(\mathrm{d}x)K(x, \mathrm{d}x') = \pi(\mathrm{d}x')K^\top(x', \mathrm{d}x)$$
  - K is π-reversible if and only if K is self-adjoint $K = K^\top$
- **Proof:** Using the definition we compute both sides:
$$\langle Kf, g\rangle_\pi = \int_{\mathbb{X}} Kf(x)g(x)\pi(\mathrm{d}x) = \int_{\mathbb{X}}\int_{\mathbb{X}} K(x, \mathrm{d}x')f(x')g(x)\pi(\mathrm{d}x)$$
$$\langle f, K^\top g\rangle_\pi = \int_{\mathbb{X}} f(x')K^\top g(x')\pi(\mathrm{d}x') = \int_{\mathbb{X}}\int_{\mathbb{X}} f(x')K^\top(x', \mathrm{d}x)g(x)\pi(\mathrm{d}x')$$
  - We see they are equivalent if and only if detailed balance holds

ALGEBRA OF KERNELS

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms
- **Example:** Kernels commute if and only if:

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms
- **Example:** Kernels commute if and only if:
  - The operators commute:

$$K_1 K_2 = K_2 K_1$$

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms
- **Example:** Kernels commute if and only if:
    - The operators commute:

$$K_1 K_2 = K_2 K_1$$

    - The order of algorithms doesn't matter (can parallelise)

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms
- **Example:** Kernels commute if and only if:
  - The operators commute:

$$K_1 K_2 = K_2 K_1$$

  - The order of algorithms doesn't matter (can parallelise)
- **Example:** Kernels is reversible if and only if:

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms
- **Example:** Kernels commute if and only if:
  - The operators commute:

$$K_1 K_2 = K_2 K_1$$

  - The order of algorithms doesn't matter (can parallelise)
- **Example:** Kernels is reversible if and only if:
  - Operators are self-adjoint $$K^\top = K$$

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms
- **Example:** Kernels commute if and only if:
  - The operators commute:
  $$K_1K_2 = K_2K_1$$
  - The order of algorithms doesn't matter (can parallelise)
- **Example:** Kernels is reversible if and only if:
  - Operators are self-adjoint $$K^\top = K$$
  - Product of reversible kernels is not always reversible:
  $$(K_1K_2)^\top = K_2^\top K_1^\top$$

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms
- **Example:** Kernels commute if and only if:
  - The operators commute:
    $$K_1K_2 = K_2K_1$$
  - The order of algorithms doesn't matter (can parallelise)
- **Example:** Kernels is reversible if and only if:
  - Operators are self-adjoint $$K^\top = K$$
  - Product of reversible kernels is not always reversible:
    $$(K_1K_2)^\top = K_2^\top K_1^\top$$
  - Mixutre of reversible kernels is:
    $$(\alpha K_1 + (1-\alpha)K_2)^\top = \alpha K_1^\top + (1-\alpha)K_2^\top$$

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms
- **Example:** Kernels commute if and only if:
  - The operators commute:
  $$K_1K_2 = K_2K_1$$
  - The order of algorithms doesn't matter (can parallelise)
- **Example:** Kernels is reversible if and only if:
  - Operators are self-adjoint $$K^\top = K$$
  - Product of reversible kernels is not always reversible:
  $$(K_1K_2)^\top = K_2^\top K_1^\top$$
  - Mixutre of reversible kernels is:
  $$(\alpha K_1 + (1-\alpha)K_2)^\top = \alpha K_1^\top + (1-\alpha)K_2^\top$$
  - Adjoint corresponds to chain running backward in time

ALGEBRA OF KERNELS

- Can use intuition from linear algebra to analyse kernels and algorithms
- **Example:** Kernels commute if and only if:
  - The operators commute:
$$K_1 K_2 = K_2 K_1$$
  - The order of algorithms doesn't matter (can parallelise)
- **Example:** Kernels is reversible if and only if:
  - Operators are self-adjoint $$K^\top = K$$
  - Product of reversible kernels is not always reversible:
$$(K_1 K_2)^\top = K_2^\top K_1^\top$$
  - Mixutre of reversible kernels is:
$$(\alpha K_1 + (1-\alpha) K_2)^\top = \alpha K_1^\top + (1-\alpha) K_2^\top$$
  - Adjoint corresponds to chain running backward in time
- Exercise: what the interpreations of orthogonality, eigenvalues, eigenvectors, normality, etc

MARKOV KERNELS V4: MARKOV CHAINS

MARKOV KERNELS V4: MARKOV CHAINS

- A Markov Chain $X_0, X_1, \ldots \in \mathbb{X}$ is a stochastic process in $\mathbb{X}$ such that

$$\mathbb{P}[X_t \in A \,|\, X_0, \ldots, X_{t-1}] = \mathbb{P}[X_t \in A \,|\, X_{t-1}]$$

MARKOV KERNELS V4: MARKOV CHAINS

- A Markov Chain $X_0, X_1, \ldots \in \mathbb{X}$ is a stochastic process in $\mathbb{X}$ such that

$$\mathbb{P}[X_t \in A \,|\, X_0, \ldots, X_{t-1}] = \mathbb{P}[X_t \in A \,|\, X_{t-1}]$$

  - Given a Markov chain defines a Markov kernel as the one step transition

$$K(x, A) = \mathbb{P}[X_1 \in A \,|\, X_0 = x] = \mathbb{P}_x[X_1 \in A]$$

MARKOV KERNELS V4: MARKOV CHAINS

- A Markov Chain $X_0, X_1, \ldots \in \mathbb{X}$ is a stochastic process in $\mathbb{X}$ such that

$$\mathbb{P}[X_t \in A \,|\, X_0, \ldots, X_{t-1}] = \mathbb{P}[X_t \in A \,|\, X_{t-1}]$$

  - Given a Markov chain defines a Markov kernel as the one step transition

$$K(x, A) = \mathbb{P}[X_1 \in A \,|\, X_0 = x] = \mathbb{P}_x[X_1 \in A]$$

- Given a Markov kernel K can construct a chain X_t by iterating over $X_0 \sim \mu$

$$X_t \sim K(X_{t-1}, \mathrm{d}x_t)$$

MARKOV KERNELS V4: MARKOV CHAINS

- A Markov Chain $X_0, X_1, \ldots \in \mathbb{X}$ is a stochastic process in $\mathbb{X}$ such that

$$\mathbb{P}[X_t \in A \,|\, X_0, \ldots, X_{t-1}] = \mathbb{P}[X_t \in A \,|\, X_{t-1}]$$

  - Given a Markov chain defines a Markov kernel as the one step transition

$$K(x, A) = \mathbb{P}[X_1 \in A \,|\, X_0 = x] = \mathbb{P}_x[X_1 \in A]$$

- Given a Markov kernel K can construct a chain X_t by iterating over $X_0 \sim \mu$

$$X_t \sim K(X_{t-1}, \mathrm{d}x_t)$$

  - The marginal law μ_t of X_t after t steps satisfies

$$\mu_t = \mathrm{Law}(X_t) = \mu_{t-1}K = \mu K^t$$

MARKOV KERNELS V4: MARKOV CHAINS

- A Markov Chain $X_0, X_1, \ldots \in \mathbb{X}$ is a stochastic process in $\mathbb{X}$ such that

$$\mathbb{P}[X_t \in A \,|\, X_0, \ldots, X_{t-1}] = \mathbb{P}[X_t \in A \,|\, X_{t-1}]$$

  - Given a Markov chain defines a Markov kernel as the one step transition

$$K(x, A) = \mathbb{P}[X_1 \in A \,|\, X_0 = x] = \mathbb{P}_x[X_1 \in A]$$

- Given a Markov kernel K can construct a chain X_t by iterating over $X_0 \sim \mu$

$$X_t \sim K(X_{t-1}, \mathrm{d}x_t)$$

  - The marginal law μ_t of X_t after t steps satisfies

$$\mu_t = \mathrm{Law}(X_t) = \mu_{t-1}K = \mu K^t$$

  - The joint law of $X_{0:t} = (X_0, \ldots, X_t)$

$$\mu \otimes K \cdots \otimes K(\mathrm{d}x_{0:t}) = \mu(\mathrm{d}x_0) \prod_{s=1}^{t} K(x_{t-1}, \mathrm{d}x_t)$$

f-DIVERGENCES

f-DIVERGENCES

- The likelihood ratio or Radon-Nikodym derivative $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$

$$\frac{\mathrm{d}\mu'}{\mathrm{d}\mu} : \mathbb{X} \to \mathbb{R}, \qquad \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x) = \frac{\mu'(x)}{\mu(x)}$$

f-DIVERGENCES

- The likelihood ratio or Radon-Nikodym derivative $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$

$$\frac{\mathrm{d}\mu'}{\mathrm{d}\mu} : \mathbb{X} \to \mathbb{R}, \qquad \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x) = \frac{\mu'(x)}{\mu(x)}$$

- Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an convex function such that $f(1) = 0$, define the f-divergence:

$$D_f(\mu' \| \mu) = \mu\left[f \circ \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right] = \int_{\mathbb{X}} f\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x)\right)\mu(x)\mathrm{d}x$$

f-DIVERGENCES

- The likelihood ratio or Radon-Nikodym derivative $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$

$$\frac{\mathrm{d}\mu'}{\mathrm{d}\mu} : \mathbb{X} \to \mathbb{R}, \qquad \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x) = \frac{\mu'(x)}{\mu(x)}$$

- Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an convex function such that $f(1) = 0$, define the f-divergence:

$$D_f(\mu' \| \mu) = \mu \left[f \circ \frac{\mathrm{d}\mu'}{\mathrm{d}\mu} \right] = \int_{\mathbb{X}} f\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x) \right) \mu(x)\mathrm{d}x$$

- We have $D_f(\mu' \| \mu) \geq 0$ and with equality if and only if $\mu = \mu'$

f-DIVERGENCES

- The likelihood ratio or Radon-Nikodym derivative $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$

$$\frac{\mathrm{d}\mu'}{\mathrm{d}\mu} : \mathbb{X} \to \mathbb{R}, \qquad \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x) = \frac{\mu'(x)}{\mu(x)}$$

- Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an convex function such that $f(1) = 0$, define the f-divergence:

$$D_f(\mu' \| \mu) = \mu \left[f \circ \frac{\mathrm{d}\mu'}{\mathrm{d}\mu} \right] = \int_{\mathbb{X}} f\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x) \right) \mu(x)\mathrm{d}x$$

- We have $D_f(\mu' \| \mu) \geq 0$ and with equality if and only if $\mu = \mu'$

- **Examples:**

f-DIVERGENCES

- The likelihood ratio or Radon-Nikodym derivative $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$

$$\frac{\mathrm{d}\mu'}{\mathrm{d}\mu} : \mathbb{X} \to \mathbb{R}, \qquad \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x) = \frac{\mu'(x)}{\mu(x)}$$

- Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an convex function such that $f(1) = 0$, define the f-divergence:

$$D_f(\mu' \| \mu) = \mu\left[f \circ \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right] = \int_{\mathbb{X}} f\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x)\right)\mu(x)\mathrm{d}x$$

- We have $D_f(\mu' \| \mu) \geq 0$ and with equality if and only if $\mu = \mu'$

- **Examples:**

$$f(r) = \frac{1}{2}|1 - r| \qquad \mathrm{TV}(\mu, \mu') = \frac{1}{2}\int_{\mathbb{X}} |\mu(x) - \mu'(x)|\,\mathrm{d}x = 1 - \int_{\mathbb{X}} \mu(x) \wedge \mu'(x)\mathrm{d}x$$

f-DIVERGENCES

- The likelihood ratio or Radon-Nikodym derivative $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$

$$\frac{\mathrm{d}\mu'}{\mathrm{d}\mu} : \mathbb{X} \to \mathbb{R}, \qquad \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x) = \frac{\mu'(x)}{\mu(x)}$$

- Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an convex function such that $f(1) = 0$, define the f-divergence:

$$D_f(\mu' \| \mu) = \mu\left[f \circ \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right] = \int_{\mathbb{X}} f\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x)\right)\mu(x)\mathrm{d}x$$

- We have $D_f(\mu' \| \mu) \geq 0$ and with equality if and only if $\mu = \mu'$

- **Examples:**

$$f(r) = \frac{1}{2}|1 - r| \qquad \mathrm{TV}(\mu, \mu') = \frac{1}{2}\int_{\mathbb{X}} |\mu(x) - \mu'(x)|\,\mathrm{d}x = 1 - \int_{\mathbb{X}} \mu(x) \wedge \mu'(x)\mathrm{d}x$$

$$f(r) = (1 - r)^2 \qquad \chi^2(\mu' \| \mu) = \mathbb{V}_\mu\left[\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right] = \int_{\mathbb{X}} \left(\frac{\mu'(x)}{\mu(x)} - 1\right)^2 \mu(x)\mathrm{d}x$$

f-DIVERGENCES

- The likelihood ratio or Radon-Nikodym derivative $\mu(\mathrm{d}x) = \mu(x)\mathrm{d}x$ and $\mu'(\mathrm{d}x) = \mu'(x)\mathrm{d}x$

$$\frac{\mathrm{d}\mu'}{\mathrm{d}\mu} : \mathbb{X} \to \mathbb{R}, \qquad \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x) = \frac{\mu'(x)}{\mu(x)}$$

- Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an convex function such that $f(1) = 0$, define the f-divergence:

$$D_f(\mu' \| \mu) = \mu\left[f \circ \frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right] = \int_{\mathbb{X}} f\left(\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}(x)\right)\mu(x)\mathrm{d}x$$

- We have $D_f(\mu' \| \mu) \geq 0$ and with equality if and only if $\mu = \mu'$

- **Examples:**

$$f(r) = \frac{1}{2}|1 - r| \qquad \mathrm{TV}(\mu, \mu') = \frac{1}{2}\int_{\mathbb{X}} |\mu(x) - \mu'(x)|\,\mathrm{d}x = 1 - \int_{\mathbb{X}} \mu(x) \wedge \mu'(x)\mathrm{d}x$$

$$f(r) = (1 - r)^2 \qquad \chi^2(\mu' \| \mu) = \mathbb{V}_\mu\left[\frac{\mathrm{d}\mu'}{\mathrm{d}\mu}\right] = \int_{\mathbb{X}} \left(\frac{\mu'(x)}{\mu(x)} - 1\right)^2 \mu(x)\mathrm{d}x$$

$$f(r) = r \log r \qquad \mathrm{KL}(\mu' \| \mu) = \int_{\mathbb{X}} \mu'(x) \log\frac{\mu'(x)}{\mu(x)}\mathrm{d}x$$

CONVERGENCE

CONVERGENCE

- If $X \sim \mu$ and K is π-invariant, is $X' \sim K(X, \mathrm{d}x')$ closer to π than X?

CONVERGENCE

- If $X \sim \mu$ and K is π-invariant, is $X' \sim K(X, \mathrm{d}x')$ closer to π than X?
- **Data processing inequality:** for any kernel K and $\mu, \mu' \in \mathscr{P}(\mathbb{X})$

$$D_f(\mu' K \| \mu K) \leq D_f(\mu' \| \mu)$$

CONVERGENCE

- If $X \sim \mu$ and K is π-invariant, is $X' \sim K(X, \mathrm{d}x')$ closer to π than X?
- **Data processing inequality:** for any kernel K and $\mu, \mu' \in \mathscr{P}(\mathbb{X})$

$$D_f(\mu' K \| \mu K) \leq D_f(\mu' \| \mu)$$

- Therefore, if K is π-invariant, then $\pi = \pi K$ and,

$$D_f(\pi \| K\mu) \leq D_f(\pi \| \mu)$$

CONVERGENCE

- If $X \sim \mu$ and K is π-invariant, is $X' \sim K(X, \mathrm{d}x')$ closer to π than X?
- **Data processing inequality:** for any kernel K and $\mu, \mu' \in \mathscr{P}(\mathbb{X})$

$$D_f(\mu' K \| \mu K) \leq D_f(\mu' \| \mu)$$

- Therefore, if K is π-invariant, then $\pi = \pi K$ and,

$$D_f(\pi \| K\mu) \leq D_f(\pi \| \mu)$$

- A π-Invariant kernel can not push you away from target!

CONVERGENCE

- If $X \sim \mu$ and K is π-invariant, is $X' \sim K(X, \mathrm{d}x')$ closer to π than X?

- **Data processing inequality:** for any kernel K and $\mu, \mu' \in \mathscr{P}(\mathbb{X})$

$$D_f(\mu' K \| \mu K) \leq D_f(\mu' \| \mu)$$

- Therefore, if K is π-invariant, then $\pi = \pi K$ and,

$$D_f(\pi \| K\mu) \leq D_f(\pi \| \mu)$$

- A π-Invariant kernel can not push you away from target!

- **BUT** invariance alone does not guarantee that it will bring you closer either.

CONVERGENCE

- If $X \sim \mu$ and K is π-invariant, is $X' \sim K(X, \mathrm{d}x')$ closer to π than X?
- **Data processing inequality:** for any kernel K and $\mu, \mu' \in \mathscr{P}(\mathbb{X})$

$$D_f(\mu' K \| \mu K) \leq D_f(\mu' \| \mu)$$

- Therefore, if K is π-invariant, then $\pi = \pi K$ and,

$$D_f(\pi \| K\mu) \leq D_f(\pi \| \mu)$$

- A π-Invariant kernel can not push you away from target!
- **BUT** invariance alone does not guarantee that it will bring you closer either.
  - Example: if $K(x, \mathrm{d}x) = \delta_x(\mathrm{d}x')$ is the identity kernel, then $\mu K = \mu$ and hence,

$$D_f(\pi \| K\mu) = D_f(\pi \| \mu)$$

CONVERGENCE OF MARKOV CHAINS

CONVERGENCE OF MARKOV CHAINS

- A Markov chain is μ-irreducible if $\forall x \in \mathbb{X}$, and $\mu(A) > 0$, there exists $t \geq 0$ such that $K^t(x, A) > 0$

CONVERGENCE OF MARKOV CHAINS

- A Markov chain is μ-irreducible if $\forall x \in \mathbb{X}$, and $\mu(A) > 0$, there exists $t \geq 0$ such that $K^t(x, A) > 0$

- μ-irreducible Markov chain is **aperiodic** if for all $\mu(A) > 0$,

$$\gcd(t : \mathbb{P}[X_t \in A \,|\, X_0 \in A] > 0) = 1$$

CONVERGENCE OF MARKOV CHAINS

- A Markov chain is μ-irreducible if $\forall x \in \mathbb{X}$, and $\mu(A) > 0$, there exists $t \geq 0$ such that $K^t(x, A) > 0$

- μ-irreducible Markov chain is **aperiodic** if for all $\mu(A) > 0$,

$$\gcd(t : \mathbb{P}[X_t \in A \,|\, X_0 \in A] > 0) = 1$$

- **Theorem**: If K is a π-irreducible, π-invariant and aperiodic, then for any μ

$$\lim_{t \to \infty} \|\mu K^t - \pi\|_{\mathrm{TV}} = 0$$

CONVERGENCE OF MARKOV CHAINS

- A Markov chain is μ-irreducible if $\forall x \in \mathbb{X}$, and $\mu(A) > 0$, there exists $t \geq 0$ such that $K^t(x, A) > 0$

- μ-irreducible Markov chain is **aperiodic** if for all $\mu(A) > 0$,

$$\gcd(t : \mathbb{P}[X_t \in A \,|\, X_0 \in A] > 0) = 1$$

- **Theorem**: If K is a π-irreducible, π-invariant and aperiodic, then for any μ

$$\lim_{t \to \infty} \|\mu K^t - \pi\|_{\mathrm{TV}} = 0$$

  - Implies X_t is an approximate sample from π when t is large enough

CONVERGENCE OF MARKOV CHAINS

- A Markov chain is μ-irreducible if $\forall x \in \mathbb{X}$, and $\mu(A) > 0$, there exists $t \geq 0$ such that $K^t(x, A) > 0$

- μ-irreducible Markov chain is **aperiodic** if for all $\mu(A) > 0$,

$$\gcd(t : \mathbb{P}[X_t \in A \mid X_0 \in A] > 0) = 1$$

- **Theorem**: If K is a π-irreducible, π-invariant and aperiodic, then for any μ

$$\lim_{t\to\infty} \|\mu K^t - \pi\|_{\mathrm{TV}} = 0$$

  - Implies X_t is an approximate sample from π when t is large enough

- Irreducible chains always bring your closer to the target, but not always quickly

CONVERGENCE OF MARKOV CHAINS

- A Markov chain is μ-irreducible if $\forall x \in \mathbb{X}$, and $\mu(A) > 0$, there exists $t \geq 0$ such that $K^t(x, A) > 0$

- μ-irreducible Markov chain is **aperiodic** if for all $\mu(A) > 0$,

$$\gcd(t : \mathbb{P}[X_t \in A \,|\, X_0 \in A] > 0) = 1$$

- **Theorem**: If K is a π-irreducible, π-invariant and aperiodic, then for any μ

$$\lim_{t \to \infty} ||\mu K^t - \pi||_{\mathrm{TV}} = 0$$

  - Implies X_t is an approximate sample from π when t is large enough

- Irreducible chains always bring your closer to the target, but not always quickly

  - E.g. multi-modal distribution…

ERGODIC THEOREM

ERGODIC THEOREM

- A μ-irreducible Markov chain X_t is Harris recurrent if $\forall x \in \mathbb{X}$, and $\mu(A) > 0$,

$$\mathbb{P}_x\left[\sum_t 1_A(X_t) = \infty\right] = 1$$

ERGODIC THEOREM

- A μ-irreducible Markov chain X_t is Harris recurrent if $\forall x \in \mathbb{X},$ and $\mu(A) > 0$,

$$\mathbb{P}_x\left[\sum_t 1_A(X_t) = \infty\right] = 1$$

- **Theorem**: If X_t is a π-irreducible, π-invariant, Harris recurrent Markov chain, then for any integrable function $f : \mathbb{X} \to \mathbb{R}$ the following limit a.s. holds

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \pi[f]$$

ERGODIC THEOREM

- A μ-irreducible Markov chain X_t is Harris recurrent if $\forall x \in \mathbb{X},$ and $\mu(A) > 0$,

$$\mathbb{P}_x \left[\sum_t 1_A(X_t) = \infty \right] = 1$$

- **Theorem**: If X_t is a π-irreducible, π-invariant, Harris recurrent Markov chain, then for any integrable function $f : \mathbb{X} \to \mathbb{R}$ the following limit a.s. holds

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \pi[f]$$

- Given a function $f : \mathbb{X} \to \mathbb{R}$ we can approximate expectation using the average

$$\hat{\pi}_T[f] = \frac{1}{T} \sum_{t=1}^{T} f(X_t)$$

ERGODIC THEOREM

- A μ-irreducible Markov chain X_t is Harris recurrent if $\forall x \in \mathbb{X}$, and $\mu(A) > 0$,

$$\mathbb{P}_x\left[\sum_t 1_A(X_t) = \infty\right] = 1$$

- **Theorem**: If X_t is a π-irreducible, π-invariant, Harris recurrent Markov chain, then for any integrable function $f : \mathbb{X} \to \mathbb{R}$ the following limit a.s. holds

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} f(X_t) = \pi[f]$$

- Given a function $f : \mathbb{X} \to \mathbb{R}$ we can approximate expectation using the average

$$\hat{\pi}_T[f] = \frac{1}{T}\sum_{t=1}^{T} f(X_t)$$

- Unlike the Monte Carlo estimator, this is biased for a finite time T

ERGODIC THEOREM

- A μ-irreducible Markov chain X_t is Harris recurrent if $\forall x \in \mathbb{X},$ and $\mu(A) > 0$,

$$\mathbb{P}_x\left[\sum_t 1_A(X_t) = \infty\right] = 1$$

- **Theorem**: If X_t is a π-irreducible, π-invariant, Harris recurrent Markov chain, then for any integrable function $f : \mathbb{X} \to \mathbb{R}$ the following limit a.s. holds

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} f(X_t) = \pi[f]$$

- Given a function $f : \mathbb{X} \to \mathbb{R}$ we can approximate expectation using the average

$$\hat{\pi}_T[f] = \frac{1}{T}\sum_{t=1}^{T} f(X_t)$$

- Unlike the Monte Carlo estimator, this is biased for a finite time T
- How long does it take to forget the initial distribution $X_0 \sim \mu$ and enter the stationary regime approximating the target π?

MIXING-TIME

MIXING-TIME

- For $\epsilon > 0$, we define the **mixing time** for a π-invariant kernel K equals

$$\tau_{\mathrm{mix}}(\epsilon) = \inf \left\{ t : \sup_{\mu \in \mathscr{P}(\mathbb{X})} \mathrm{TV}(\mu K^t, \pi) < \epsilon \right\}$$

MIXING-TIME

- For $\epsilon > 0$, we define the **mixing time** for a π-invariant kernel K equals

$$\tau_{\text{mix}}(\epsilon) = \inf \left\{ t : \sup_{\mu \in \mathscr{P}(\mathbb{X})} \text{TV}(\mu K^t, \pi) < \epsilon \right\}$$

  - τ_{mix} measures how long it takes to achieve stationarity and forget μ

MIXING-TIME

- For $\epsilon > 0$, we define the **mixing time** for a π-invariant kernel K equals

$$\tau_{\text{mix}}(\epsilon) = \inf \left\{ t : \sup_{\mu \in \mathscr{P}(\mathbb{X})} \text{TV}(\mu K^t, \pi) < \epsilon \right\}$$

  - τ_{mix} measures how long it takes to achieve stationarity and forget μ

- A Markov chain is **geometrically ergodic** if there exists a $M(x) > 0$ and $\rho \in [0,1]$

$$\forall x \in \mathbb{X}, \quad \text{TV}(K^t(x, \cdot\,), \pi) \leq M(x)\rho^t$$

MIXING-TIME

- For $\epsilon > 0$, we define the **mixing time** for a π-invariant kernel K equals

$$\tau_{\text{mix}}(\epsilon) = \inf \left\{ t : \sup_{\mu \in \mathscr{P}(\mathbb{X})} \mathrm{TV}(\mu K^t, \pi) < \epsilon \right\}$$

  - τ_{mix} measures how long it takes to achieve stationarity and forget μ

- A Markov chain is **geometrically ergodic** if there exists a $M(x) > 0$ and $\rho \in [0,1]$

$$\forall x \in \mathbb{X}, \quad \mathrm{TV}(K^t(x, \cdot\,), \pi) \leq M(x)\rho^t$$

  - A Markov chain is **uniformly ergodic** if $M(x) \leq M$ for some $M < \infty$

MIXING-TIME

- For $\epsilon > 0$, we define the **mixing time** for a π-invariant kernel K equals

$$\tau_{\text{mix}}(\epsilon) = \inf \left\{ t : \sup_{\mu \in \mathscr{P}(\mathbb{X})} \text{TV}(\mu K^t, \pi) < \epsilon \right\}$$

  - τ_{mix} measures how long it takes to achieve stationarity and forget μ

- A Markov chain is **geometrically ergodic** if there exists a $M(x) > 0$ and $\rho \in [0,1]$

$$\forall x \in \mathbb{X}, \quad \text{TV}(K^t(x, \cdot\,), \pi) \leq M(x)\rho^t$$

  - A Markov chain is **uniformly ergodic** if $M(x) \leq M$ for some $M < \infty$

- The mixing time for a uniformly ergodic chain satisfies:

$$\tau_{\text{mix}}(\epsilon) < \frac{\log \frac{\epsilon}{M}}{\log \rho}$$

CENTRAL LIMIT THEOREM

CENTRAL LIMIT THEOREM

- **Theorem (CLT):** Under certain regularity assumptions, a Harris recurrent, π-invariant Markov chain satisfying enough regularity conditions (e.g. reversible geometrically ergodic), as $T \to \infty$

$$\sqrt{T}(\hat{\pi}_T[f] - \pi[f]) \quad \Longrightarrow \quad N(0, \sigma^2(f))$$

CENTRAL LIMIT THEOREM

- **Theorem (CLT):** Under certain regularity assumptions, a Harris recurrent, π-invariant Markov chain satisfying enough regularity conditions (e.g. reversible geometrically ergodic), as $T \to \infty$

$$\sqrt{T}(\hat{\pi}_T[f] - \pi[f]) \quad \Longrightarrow \quad N(0, \sigma^2(f))$$

  - The asymptotic variance decomposes as:

$$\sigma^2[f] = \lim_{T \to \infty} T\mathbb{V}[\hat{\pi}_T[f]] = \mathbb{V}_\pi[f]\tau_{\mathrm{corr}}[f]$$

CENTRAL LIMIT THEOREM

- **Theorem (CLT):** Under certain regularity assumptions, a Harris recurrent, π-invariant Markov chain satisfying enough regularity conditions (e.g. reversible geometrically ergodic), as $T \to \infty$

$$\sqrt{T}(\hat{\pi}_T[f] - \pi[f]) \quad \Longrightarrow \quad N(0, \sigma^2(f))$$

  - The asymptotic variance decomposes as:

$$\sigma^2[f] = \lim_{T \to \infty} T\,\mathbb{V}[\hat{\pi}_T[f]] = \mathbb{V}_\pi[f]\tau_{\text{corr}}[f]$$

  - τ_{corr} is the **integrated autocorrelation time**:

$$\tau_{\text{corr}}[f] = 1 + 2\sum_{t=1}^{\infty} \rho_t[f]$$

CENTRAL LIMIT THEOREM

- **Theorem (CLT):** Under certain regularity assumptions, a Harris recurrent, π-invariant Markov chain satisfying enough regularity conditions (e.g. reversible geometrically ergodic), as $T \to \infty$

$$\sqrt{T}(\hat{\pi}_T[f] - \pi[f]) \quad \Longrightarrow \quad N(0, \sigma^2(f))$$

  - The asymptotic variance decomposes as:

$$\sigma^2[f] = \lim_{T \to \infty} T\mathbb{V}[\hat{\pi}_T[f]] = \mathbb{V}_\pi[f]\tau_{\mathrm{corr}}[f]$$

  - τ_{corr} is the **integrated autocorrelation time**:

$$\tau_{\mathrm{corr}}[f] = 1 + 2\sum_{t=1}^{\infty} \rho_t[f]$$

  - $\rho_t[f]$ is the t-th lag defined as the autocorrelation coefficient at stationarity

$$\rho_t[f] = \frac{\mathrm{Cov}[f(X_0), f(X_t)]}{\mathbb{V}_\pi[f]}, \qquad (X_0, X_t) \sim \pi \otimes K^t$$

CENTRAL LIMIT THEOREM

- **Theorem (CLT):** Under certain regularity assumptions, a Harris recurrent, π-invariant Markov chain satisfying enough regularity conditions (e.g. reversible geometrically ergodic), as $T \to \infty$

$$\sqrt{T}(\hat{\pi}_T[f] - \pi[f]) \quad \Longrightarrow \quad N(0, \sigma^2(f))$$

  - The asymptotic variance decomposes as:

$$\sigma^2[f] = \lim_{T\to\infty} T\mathbb{V}[\hat{\pi}_T[f]] = \mathbb{V}_\pi[f]\tau_{\text{corr}}[f]$$

  - τ_{corr} is the **integrated autocorrelation time**:

$$\tau_{\text{corr}}[f] = 1 + 2\sum_{t=1}^{\infty} \rho_t[f]$$

  - $\rho_t[f]$ is the t-th lag defined as the autocorrelation coefficient at stationarity

$$\rho_t[f] = \frac{\text{Cov}[f(X_0), f(X_t)]}{\mathbb{V}_\pi[f]}, \qquad (X_0, X_t) \sim \pi \otimes K^t$$

- τ_{corr} measures how long it takes to forget a stationary sample $X_0 \sim \pi$

EFFECTIVE SAMPLE SIZE

EFFECTIVE SAMPLE SIZE

- For large T the CLT implies the variance of the MCMC estimator

$$\mathbb{V}[\hat{\pi}_T[f]] \approx \frac{\sigma^2[f]}{T} = \frac{\mathbb{V}_\pi[f]\tau_{\mathrm{corr}}[f]}{T} = \frac{\mathbb{V}_\pi[f]}{T_{\mathrm{ess}}[f]}$$

EFFECTIVE SAMPLE SIZE

- For large T the CLT implies the variance of the MCMC estimator

$$\mathbb{V}[\hat{\pi}_T[f]] \approx \frac{\sigma^2[f]}{T} = \frac{\mathbb{V}_\pi[f]\tau_{\text{corr}}[f]}{T} = \frac{\mathbb{V}_\pi[f]}{T_{\text{ess}}[f]}$$

- Where we define the **effective sample size (ESS)** defined as

$$T_{\text{ESS}}[f] = \frac{T}{\tau_{\text{corr}}[f]} = \frac{T}{1 + 2\sum_{t=1}^{\infty} \rho_t[f]}$$

EFFECTIVE SAMPLE SIZE

- For large T the CLT implies the variance of the MCMC estimator

$$\mathbb{V}[\hat{\pi}_T[f]] \approx \frac{\sigma^2[f]}{T} = \frac{\mathbb{V}_\pi[f]\tau_{\text{corr}}[f]}{T} = \frac{\mathbb{V}_\pi[f]}{T_{\text{ess}}[f]}$$

- Where we define the **effective sample size (ESS)** defined as

$$T_{\text{ESS}}[f] = \frac{T}{\tau_{\text{corr}}[f]} = \frac{T}{1 + 2\sum_{t=1}^{\infty} \rho_t[f]}$$

- The ESS estimates the number of iid samples from π a Monte Carlo estimator would require to achieve a comparable variance to the MCMC

EFFECTIVE SAMPLE SIZE

- For large T the CLT implies the variance of the MCMC estimator

$$\mathbb{V}[\hat{\pi}_T[f]] \approx \frac{\sigma^2[f]}{T} = \frac{\mathbb{V}_\pi[f]\tau_{\text{corr}}[f]}{T} = \frac{\mathbb{V}_\pi[f]}{T_{\text{ess}}[f]}$$

- Where we define the **effective sample size (ESS)** defined as

$$T_{\text{ESS}}[f] = \frac{T}{\tau_{\text{corr}}[f]} = \frac{T}{1 + 2\sum_{t=1}^{\infty} \rho_t[f]}$$

- The ESS estimates the number of iid samples from π a Monte Carlo estimator would require to achieve a comparable variance to the MCMC
- The goal is to design MCMC kernels to reduce the auto-corellations

EFFECTIVE SAMPLE SIZE

- For large T the CLT implies the variance of the MCMC estimator

$$\mathbb{V}[\hat{\pi}_T[f]] \approx \frac{\sigma^2[f]}{T} = \frac{\mathbb{V}_\pi[f]\tau_{\mathrm{corr}}[f]}{T} = \frac{\mathbb{V}_\pi[f]}{T_{\mathrm{ess}}[f]}$$

- Where we define the **effective sample size (ESS)** defined as

$$T_{\mathrm{ESS}}[f] = \frac{T}{\tau_{\mathrm{corr}}[f]} = \frac{T}{1 + 2\sum_{t=1}^{\infty} \rho_t[f]}$$

- The ESS estimates the number of iid samples from π a Monte Carlo estimator would require to achieve a comparable variance to the MCMC
- The goal is to design MCMC kernels to reduce the auto-corellations
- ESS is a useful tool: but use with caution!!

EFFECTIVE SAMPLE SIZE

- For large T the CLT implies the variance of the MCMC estimator

$$\mathbb{V}[\hat{\pi}_T[f]] \approx \frac{\sigma^2[f]}{T} = \frac{\mathbb{V}_\pi[f]\tau_{\mathrm{corr}}[f]}{T} = \frac{\mathbb{V}_\pi[f]}{T_{\mathrm{ess}}[f]}$$

- Where we define the **effective sample size (ESS)** defined as

$$T_{\mathrm{ESS}}[f] = \frac{T}{\tau_{\mathrm{corr}}[f]} = \frac{T}{1 + 2\sum_{t=1}^{\infty} \rho_t[f]}$$

- The ESS estimates the number of iid samples from π a Monte Carlo estimator would require to achieve a comparable variance to the MCMC
- The goal is to design MCMC kernels to reduce the auto-corellations
- ESS is a useful tool: but use with caution!!
  - Only valid **after burn-in** until chain has converged is meaningless

EFFECTIVE SAMPLE SIZE

- For large T the CLT implies the variance of the MCMC estimator

$$\mathbb{V}[\hat{\pi}_T[f]] \approx \frac{\sigma^2[f]}{T} = \frac{\mathbb{V}_\pi[f]\tau_{\mathrm{corr}}[f]}{T} = \frac{\mathbb{V}_\pi[f]}{T_{\mathrm{ess}}[f]}$$

- Where we define the **effective sample size (ESS)** defined as

$$T_{\mathrm{ESS}}[f] = \frac{T}{\tau_{\mathrm{corr}}[f]} = \frac{T}{1 + 2\sum_{t=1}^{\infty} \rho_t[f]}$$

- The ESS estimates the number of iid samples from π a Monte Carlo estimator would require to achieve a comparable variance to the MCMC
- The goal is to design MCMC kernels to reduce the auto-corellations
- ESS is a useful tool: but use with caution!!
  - Only valid **after burn-in** until chain has converged is meaningless
  - Generally not fun to compute and can be unstable

EFFECTIVE SAMPLE SIZE

- For large T the CLT implies the variance of the MCMC estimator

$$\mathbb{V}[\hat{\pi}_T[f]] \approx \frac{\sigma^2[f]}{T} = \frac{\mathbb{V}_\pi[f]\tau_{\text{corr}}[f]}{T} = \frac{\mathbb{V}_\pi[f]}{T_{\text{ess}}[f]}$$

- Where we define the **effective sample size (ESS)** defined as

$$T_{\text{ESS}}[f] = \frac{T}{\tau_{\text{corr}}[f]} = \frac{T}{1 + 2\sum_{t=1}^{\infty} \rho_t[f]}$$

- The ESS estimates the number of iid samples from π a Monte Carlo estimator would require to achieve a comparable variance to the MCMC
- The goal is to design MCMC kernels to reduce the auto-corellations
- ESS is a useful tool: but use with caution!!
  - Only valid **after burn-in** until chain has converged is meaningless
  - Generally not fun to compute and can be unstable
  - ESS of a Markov chain doesn’t mean anything, it depends on the statistics of interest

EXAMPLE RANDOM WALK ON A CIRLE

EXAMPLE RANDOM WALK ON A CIRLE

- Consider the random walk on a circle $\mathbb{Z}_n$ where n is even.

EXAMPLE RANDOM WALK ON A CIRLE

- Consider the random walk on a circle $\mathbb{Z}_n$ where n is even.
- Suppose $X_0 = 0$ and $\mathbb{P}[X_t = m \,|\, X_{t-1} = n] = K(n, m)$ where

$$K(n,m) = \frac{1}{4}\delta_{n-1}(m) + \frac{1}{2}\delta_n(m) + \frac{1}{4}\delta_{n+1}(m)$$

EXAMPLE RANDOM WALK ON A CIRLE

- Consider the random walk on a circle $\mathbb{Z}_n$ where n is even.
- Suppose $X_0 = 0$ and $\mathbb{P}[X_t = m \,|\, X_{t-1} = n] = K(n, m)$ where

$$K(n, m) = \frac{1}{4}\delta_{n-1}(m) + \frac{1}{2}\delta_n(m) + \frac{1}{4}\delta_{n+1}(m)$$

- X_t is statiationary with respect to ´$\pi = \mathrm{Uniform}[\mathbb{Z}_n]$

EXAMPLE RANDOM WALK ON A CIRLE

- Consider the random walk on a circle $\mathbb{Z}_n$ where n is even.
- Suppose $X_0 = 0$ and $\mathbb{P}[X_t = m \,|\, X_{t-1} = n] = K(n, m)$ where

$$K(n, m) = \frac{1}{4}\delta_{n-1}(m) + \frac{1}{2}\delta_n(m) + \frac{1}{4}\delta_{n+1}(m)$$

- X_t is statiationary with respect to ´$\pi = \text{Uniform}[\mathbb{Z}_n]$
- This is a lazy random walk on a circle, mixes after $O(n^2)$ iterations

EXAMPLE RANDOM WALK ON A CIRLE

- Consider the random walk on a circle $\mathbb{Z}_n$ where n is even.
- Suppose $X_0 = 0$ and $\mathbb{P}[X_t = m \,|\, X_{t-1} = n] = K(n, m)$ where

$$K(n, m) = \frac{1}{4}\delta_{n-1}(m) + \frac{1}{2}\delta_n(m) + \frac{1}{4}\delta_{n+1}(m)$$

- X_t is statiationary with respect to ´$\pi = \mathrm{Uniform}[\mathbb{Z}_n]$
- This is a lazy random walk on a circle, mixes after $O(n^2)$ iterations
- $f(n) = 1$ if n is even and $f(n) = 0$ if n is odd

EXAMPLE RANDOM WALK ON A CIRLE

- Consider the random walk on a circle $\mathbb{Z}_n$ where n is even.
- Suppose $X_0 = 0$ and $\mathbb{P}[X_t = m \,|\, X_{t-1} = n] = K(n, m)$ where

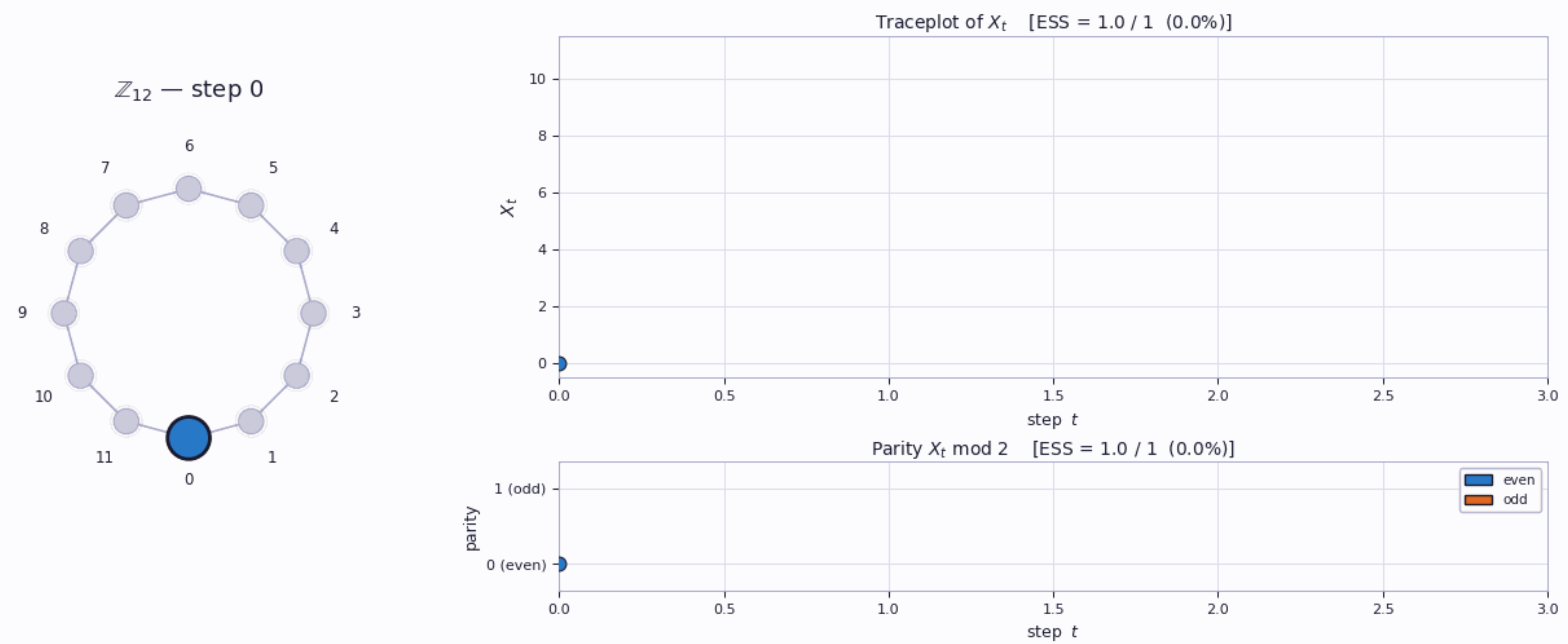
$$K(n,m) = \frac{1}{4}\delta_{n-1}(m) + \frac{1}{2}\delta_n(m) + \frac{1}{4}\delta_{n+1}(m)$$

- X_t is statiationary with respect to ´$\pi = \text{Uniform}[\mathbb{Z}_n]$
- This is a lazy random walk on a circle, mixes after $O(n^2)$ iterations
- $f(n) = 1$ if n is even and $f(n) = 0$ if n is odd
- For all $t > 0$ we have $f(X_t)$ are iid and ESS is $T_{\text{ESS}}[f] = T$

$$\mathbb{P}[f(X_t) = 1] = \frac{1}{2} = \pi[f]$$

EXAMPLE RANDOM WALK ON A CIRLE

EXAMPLE RANDOM WALK ON A CIRLE



MCMC IN PRACTICE

MCMC IN PRACTICE

- The goal of MCMC is to construct a π-stationary Markov chain X_t initialised as $X_0 \sim \mu$ satisfies the ergodic theorem and CLT

MCMC IN PRACTICE

- The goal of MCMC is to construct a π-stationary Markov chain X_t initialised as $X_0 \sim \mu$ satisfies the ergodic theorem and CLT

- Given a budget of T iterations decompose as tuning and sampling budget:

$$T = T_{\text{tune}} + T_{\text{sample}}$$

MCMC IN PRACTICE

- The goal of MCMC is to construct a π-stationary Markov chain X_t initialised as $X_0 \sim \mu$ satisfies the ergodic theorem and CLT

- Given a budget of T iterations decompose as tuning and sampling budget:

$$T = T_{\text{tune}} + T_{\text{sample}}$$

  - It's not uncommon for half the budget to go into tuning

MCMC IN PRACTICE

- The goal of MCMC is to construct a π-stationary Markov chain X_t initialised as $X_0 \sim \mu$ satisfies the ergodic theorem and CLT

- Given a budget of T iterations decompose as tuning and sampling budget:

$$T = T_{\text{tune}} + T_{\text{sample}}$$

  - It's not uncommon for half the budget to go into tuning

- **Tuning phase (aka burn-in):**

MCMC IN PRACTICE

- The goal of MCMC is to construct a π-stationary Markov chain X_t initialised as $X_0 \sim \mu$ satisfies the ergodic theorem and CLT

- Given a budget of T iterations decompose as tuning and sampling budget:

$$T = T_{\text{tune}} + T_{\text{sample}}$$

  - It's not uncommon for half the budget to go into tuning

- **Tuning phase (aka burn-in):**

    - Run long enough to forget the initial distribution μ and exceed $\tau_{\text{mix}}(\epsilon)$

MCMC IN PRACTICE

- The goal of MCMC is to construct a π-stationary Markov chain X_t initialised as $X_0 \sim \mu$ satisfies the ergodic theorem and CLT

- Given a budget of T iterations decompose as tuning and sampling budget:

$$T = T_{\text{tune}} + T_{\text{sample}}$$

  - It's not uncommon for half the budget to go into tuning

- **Tuning phase (aka burn-in):**
    - Run long enough to forget the initial distribution μ and exceed $\tau_{\text{mix}}(\epsilon)$
    - Used to tune hyper parameters of chain

MCMC IN PRACTICE

- The goal of MCMC is to construct a π-stationary Markov chain X_t initialised as $X_0 \sim \mu$ satisfies the ergodic theorem and CLT

- Given a budget of T iterations decompose as tuning and sampling budget:

$$T = T_{\text{tune}} + T_{\text{sample}}$$

  - It's not uncommon for half the budget to go into tuning

- **Tuning phase (aka burn-in):**
    - Run long enough to forget the initial distribution μ and exceed $\tau_{\text{mix}}(\epsilon)$
    - Used to tune hyper parameters of chain
    - Typically discard these samples

MCMC IN PRACTICE

- The goal of MCMC is to construct a π-stationary Markov chain X_t initialised as $X_0 \sim \mu$ satisfies the ergodic theorem and CLT

- Given a budget of T iterations decompose as tuning and sampling budget:

$$T = T_{\text{tune}} + T_{\text{sample}}$$

  - It's not uncommon for half the budget to go into tuning

- **Tuning phase (aka burn-in):**
  - Run long enough to forget the initial distribution μ and exceed $\tau_{\text{mix}}(\epsilon)$
  - Used to tune hyper parameters of chain
  - Typically discard these samples

- **Sampling phase:**

MCMC IN PRACTICE

- The goal of MCMC is to construct a π-stationary Markov chain X_t initialised as $X_0 \sim \mu$ satisfies the ergodic theorem and CLT

- Given a budget of T iterations decompose as tuning and sampling budget:

$$T = T_{\text{tune}} + T_{\text{sample}}$$

  - It's not uncommon for half the budget to go into tuning

- **Tuning phase (aka burn-in):**
    - Run long enough to forget the initial distribution μ and exceed $\tau_{\text{mix}}(\epsilon)$
    - Used to tune hyper parameters of chain
    - Typically discard these samples

- **Sampling phase:**
    - Run long enough to achieve a target ESS or untill estimates stabilise

$$T_{\text{ESS}}[f] = \frac{T_{\text{sample}}}{\tau_{\text{corr}}[f]}$$