

# CPSC 661: Sampling Algorithms in ML

---

Andre Wibisono

February 3, 2021

# Today

1. Markov chain
2. Stationary distribution
3. Reversibility
4. Mixing time

# References

- Aldous and Fill, *Reversible Markov Chains and Random Walks on Graphs*, 2002, available at <https://www.stat.berkeley.edu/users/aldous/RWG/book.html>
- Levin, Peres, and Wilmer, *Markov Chains and Mixing Times*, AMS, 2009
- Lovász and Simonovits, *Random Walks in a Convex Body and an Improved Volume Algorithm*, Random Structures and Algorithms, 1993
- Vempala, *Geometric Random Walk: A Survey*, Combinatorial and Computational Geometry, 2005

# Markov Chain

---

# Random Walk

Let  $\mathcal{X}$  be *state space*.

A **random walk** on  $\mathcal{X}$  is a sequence of random variables

$$X_0, X_1, X_2, \dots \in \mathcal{X}$$

where the **future** depends on the **past** and **present**:

$$\{X_0, \dots, X_{t-1}, X_t\} \mapsto X_{t+1}$$

# Random Walk

Let  $\mathcal{X}$  be *state space*.

A **random walk** on  $\mathcal{X}$  is a sequence of random variables

$$X_0, X_1, X_2, \dots \in \mathcal{X}$$

where the **future** depends on the **past** and **present**:

$$\{X_0, \dots, X_{t-1}, X_t\} \mapsto X_{t+1}$$

A **Markov chain** is when the **future** is independent of the **past** given the **present**:

$$X_{t+1} \perp\!\!\!\perp (X_0, \dots, X_{t-1}) \mid X_t$$

$X_0 \rightarrow \dots \rightarrow X_{t-1} \rightarrow X_t \rightarrow X_{t+1} \rightarrow \dots$

# Measure Theory Formalism

---

# Measure Theory

Let space  $\mathcal{X}$  be a set with a  $\sigma$ -algebra  $\mathcal{A} \subseteq 2^{\mathcal{X}} = \{ \text{all subsets of } \mathcal{X} \}$

- $\sigma$ -alg; •  $\mathcal{X} \in \mathcal{A}$ ,  $\mathcal{A}$  is closed under complement and countable union
- Elements of  $\mathcal{A}$  are the *measurable* subsets of  $\mathcal{X}$



# Measure Theory

Let space  $\mathcal{X}$  be a set with a  $\sigma$ -algebra  $\mathcal{A} \subseteq 2^{\mathcal{X}}$

- $\mathcal{X} \in \mathcal{A}$ ,  $\mathcal{A}$  is closed under complement and countable union
- Elements of  $\mathcal{A}$  are the *measurable* subsets of  $\mathcal{X}$

A **measure** on  $\mathcal{X}$  is a function  $\nu: \mathcal{A} \rightarrow [0, \infty]$

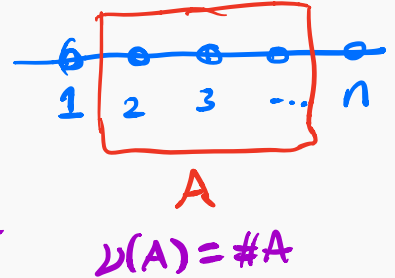
- $\nu(\emptyset) = 0$  and  $\nu$  is countably additive  $\therefore A = \bigcup_{i=1}^{\infty} A_i$  disjoint
- $\nu$  is a probability measure if  $\nu(\mathcal{X}) = 1$   $\Rightarrow \nu(A) = \sum_{i=1}^{\infty} \nu(A_i)$

# Examples

discrete: 1.  $\mathcal{X} = \{1, \dots, n\}$ ,  $\mathcal{A} = 2^{\mathcal{X}}$

- $\nu = \text{Counting}$ :  $\nu(A) = |A|$  = # elements in  $A$

- $\nu = \underline{\text{Uniform}}$ :  $\nu(A) = \frac{|A|}{n}$  ← prob. measure



# Examples

1.  $\mathcal{X} = \{1, \dots, n\}$ ,  $\mathcal{A} = 2^{\mathcal{X}}$

- $\nu = \text{Counting}$ :  $\nu(A) = |A| \rightarrow \nu(A) = n$

- $\nu = \text{Uniform}$ :  $\nu(A) = \frac{|A|}{n}$

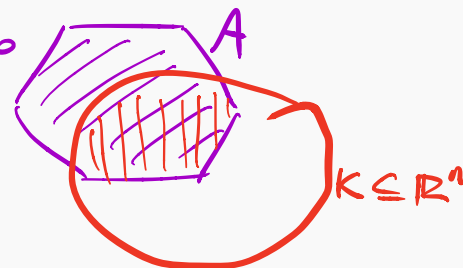
2.  $\mathcal{X} = \mathbb{R}^n$ ,  $\mathcal{A} = \text{Borel } \sigma\text{-algebra}$  *generated by open sets*

- $\nu = \text{Lebesgue}$ :  $\nu(A) = \text{Vol}(A)$ ,  $\nu(\mathbb{R}^n) = \infty$

- $\nu = \text{Lebesgue on } K$ :  $\nu(A) = \text{Vol}(A \cap K)$

- $\nu = \text{Uniform on } K$ :  $\nu(A) = \frac{\text{Vol}(A \cap K)}{\text{Vol}(K)}$

- $\nu = \text{Gaussian}$ :  $\nu(A) = \int_A (2\pi)^{-n/2} e^{-\frac{\|x\|^2}{2}} dx$



# Examples

1.  $\mathcal{X} = \{1, \dots, n\}$ ,  $\mathcal{A} = 2^{\mathcal{X}}$

- $\nu = \text{Counting}$ :  $\nu(A) = |A|$

- $\nu = \text{Uniform}$ :  $\nu(A) = \frac{|A|}{n}$

→ •  $\nu = \text{Point mass at } x^*$ :  $\nu(A) = \mathbf{1}\{x^* \in A\} = \begin{cases} 1 & \text{if } x^* \in A \\ 0 & \text{else} \end{cases}$



2.  $\mathcal{X} = \mathbb{R}^n$ ,  $\mathcal{A} = \text{Borel } \sigma\text{-algebra}$

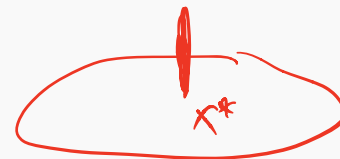
- $\nu = \text{Lebesgue}$ :  $\nu(A) = \text{Vol}(A)$

- $\nu = \text{Lebesgue on } K$ :  $\nu(A) = \text{Vol}(A \cap K)$

- $\nu = \text{Uniform on } K$ :  $\nu(A) = \frac{\text{Vol}(A \cap K)}{\text{Vol}(K)}$

- $\nu = \text{Gaussian}$ :  $\nu(A) = \int_A (2\pi)^{-n/2} e^{-\frac{\|x\|^2}{2}} dx$

→ •  $\nu = \text{Point mass at } x^*$ :  $\nu(A) = \mathbf{1}\{x^* \in A\} = \begin{cases} 1 & \text{if } x^* \in A \\ 0 & \text{else} \end{cases}$



Assume there is a reference measure  $\mu$  on space  $\mathcal{X}$


- $\mathcal{X} = \{1, \dots, n\}$ :  $\mu =$  counting measure
- $\mathcal{X} = \mathbb{R}^n$ :  $\mu =$  Lebesgue measure

# Density

Assume there is a reference measure  $\mu$  on space  $\mathcal{X}$

- $\mathcal{X} = \{1, \dots, n\}$ :  $\mu$  = counting measure
- $\mathcal{X} = \mathbb{R}^n$ :  $\mu$  = Lebesgue measure

Write a measure  $\nu$  as a density (*Radon-Nikodym derivative*) wrt  $\mu$

$$\underline{h} = \frac{d\nu}{d\mu} : \mathcal{X} \rightarrow \mathbb{R} \Leftrightarrow \underline{\nu(A)} = \int_A \underline{h(x)} \, \underline{d\mu(x)}$$


- Requires  $\nu \ll \mu$ : if  $\mu(A) = 0$ , then  $\nu(A) = 0$
- We usually also denote the density by  $\nu$  itself

# Examples

1.  $\mathcal{X} = \{1, \dots, n\}$ ,  $\mu =$  counting measure

- $\nu =$  Uniform:  $\nu(x) = \frac{1}{n}$

# Examples

1.  $\mathcal{X} = \{1, \dots, n\}$ ,  $\mu =$  counting measure

- $\nu =$  Uniform:  $\nu(x) = \frac{1}{n}$

2.  $\mathcal{X} = \mathbb{R}^n$ ,  $\mu =$  Lebesgue measure

- $\nu =$  Uniform on  $K$ :  $\nu(x) = \frac{1}{\text{Vol}(K)}$
- $\nu =$  Gaussian:  $\nu(x) = (2\pi)^{-\frac{n}{2}} e^{-\frac{\|x\|^2}{2}}$   $\mathcal{N}(0, I)$



# Examples

1.  $\mathcal{X} = \{1, \dots, n\}$ ,  $\mu =$  counting measure

- $\nu =$  Uniform:  $\nu(x) = \frac{1}{n}$

→ •  $\nu =$  Point mass:  $\delta_{x^*}(x) = \begin{cases} 1 & \text{if } x = x^* \\ 0 & \text{else} \end{cases}$

2.  $\mathcal{X} = \mathbb{R}^n$ ,  $\mu =$  Lebesgue measure

- $\nu =$  Uniform on  $K$ :  $\nu(x) = \frac{1}{\text{Vol}(K)}$

- $\nu =$  Gaussian:  $\nu(x) = (2\pi)^{-\frac{n}{2}} e^{-\frac{\|x\|^2}{2}}$

→ •  $\nu =$  Point mass:  $\delta_{x^*}(x) = \begin{cases} \infty & \text{if } x = x^* \\ 0 & \text{else} \end{cases}$

Dirac  $\delta$ -function

$$\int_{\mathbb{R}^n} \delta_{x^*}(x) dx = 1$$

$$\int_{\mathbb{R}^n} \delta_{x^*}(x) f(x) dx = f(x^*)$$

# Markov Chain

---

# Markov Chain

A **Markov chain** is a random walk where the **future** is independent of the **past** given the **present**:

$$X_{t+1} \overset{\text{independence}}{\perp\!\!\!\perp} (X_0, \dots, X_{t-1}) \mid X_t$$

$$X_{t+1} \mid \{\cancel{X_0}, \dots, \cancel{X_{t-1}}, X_t\} \stackrel{d}{=} X_{t+1} \mid X_t$$

The distribution of  $X_{t+1}$  is determined only by  $X_t$ :

$$X_{t+1} \mid X_t \sim \underbrace{P_{X_t}(\cdot)}_{P_{t, X_t} : \text{non-time-homogeneous}}$$

(this is a first-order, time-homogeneous Markov chain)

only 1 previous point      transition does not change

# Markov Chain

A **Markov chain** on  $\mathcal{X}$  is a collection

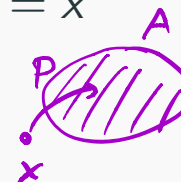
$$P = (P_x : x \in \mathcal{X})$$

where each  $P_x$  is a probability measure on  $\mathcal{X}$

- $\underline{P_x}$  = distribution of next point  $X_1$  if we start at  $X_0 = x$

$P_x = \delta_x$   
when  $\delta_0 = \delta_x$

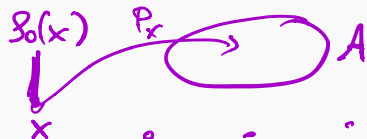
$$\rightarrow \underline{P_x(A) = \Pr(X_1 \in A \mid X_0 = x)}$$



- $P$  maps an initial distribution  $\rho_0$  to the next distribution  $\rho_1$

$X_0 \sim \rho_0$   
 $\Rightarrow X_1 \sim \rho_1$

$$\rightarrow \rho_1(A) = \int_{\mathcal{X}} \underbrace{P_x(A)}_{\leftarrow} d\rho_0(x) \leftarrow$$



$\mathcal{X}$  discrete:  
 $\int = \sum$

# Markov Chain

- Starting from  $X_0 \sim \rho_0$ ,  $P$  defines a Markov chain

$$X_0 \rightarrow \cdots \rightarrow X_k \rightarrow X_{k+1} \rightarrow \cdots$$

The distribution of  $X_k \sim \rho_k$  is

$$\rightarrow \underbrace{\rho_{k+1}(A)}_{\uparrow} = \int_{\mathcal{X}} P_x(A) \underbrace{d\rho_k(x)}_{\uparrow}$$

# Stationary distribution

$\nu$  is a stationary distribution for  $P$  if

$$\rightarrow \left[ \nu(A) = \int_{\mathcal{X}} P_x(A) d\nu(x) \right]$$

- If  $X_0 \sim \nu$ , then  $X_1 \sim \nu$  (and all  $X_k \sim \nu$ : stationary chain)

# Stationary distribution

$\nu$  is a **stationary distribution** for  $P$  if

$$\nu(A) = \int_{\mathcal{X}} P_x(A) d\nu(x)$$

- If  $X_0 \sim \nu$ , then  $X_1 \sim \nu$  (and all  $X_k \sim \nu$ : *stationary chain*)
- Existence and uniqueness follow from irreducibility, recurrence  
[ see [AF], [LPW], [Diaconis & Freedman, *On Markov Chains with Continuous State Space*, 1997, <https://statistics.berkeley.edu/tech-reports/501>]
- Always holds for us, interested in *convergence speed*  $\rho_k \rightarrow \nu$

# Discrete Space

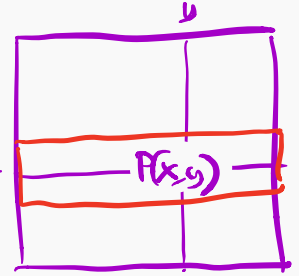
---



# Markov Chain in Discrete Space

A Markov chain  $P$  on  $\mathcal{X} = \{1, \dots, n\}$  can be represented as a stochastic matrix  $P \in \mathbb{R}^{n \times n}$  with:

- non-negative entries:  $P(x, y) = \underline{P_x(y)} \geq 0$   $P =$
- row sums to 1:  $\sum_{y \in \mathcal{X}} P(x, y) = 1 = \underline{P_x(X_1 \in \mathcal{X} \mid X_0 = x)}$



# Markov Chain in Discrete Space

A Markov chain  $P$  on  $\mathcal{X} = \{1, \dots, n\}$  can be represented as a *stochastic matrix*  $P \in \mathbb{R}^{n \times n}$  with:

- non-negative entries:  $P(x, y) = P_x(y) \geq 0$
- row sums to 1:  $\sum_{y \in \mathcal{X}} P(x, y) = 1$

A probability distribution  $\rho$  on  $\mathcal{X}$  is a row vector

$$\rho = \left( \underline{\rho(x_1)} \quad \cdots \quad \underline{\rho(x_n)} \right) \in \mathbb{R}^n$$

$\rho(x_i) \geq 0$   
 $\sum_{i=1}^n \rho(x_i) = 1$

Operation of Markov chain  $X_0 \rightarrow X_1$  is a right multiplication by  $P$ :

$$\rho_1 = \rho_0 P$$

$\rho_1(y) = \sum_x \rho_0(x) \cdot P(x, y)$

# Stationary Distribution

A stationary distribution  $\nu$  is a *left eigenvector* of  $P$ :

$$\nu = \nu P$$

- Existence and uniqueness by Perron-Frobenius theorem

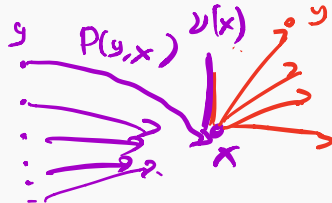
# Stationary Distribution

A stationary distribution  $\nu$  is a *left eigenvector* of  $P$ :

$$\nu = \nu P$$

- Existence and uniqueness by Perron-Frobenius theorem
- Stationary condition  $\Leftrightarrow$  global balance:

$$\forall x: \underbrace{\sum_{y \in \mathcal{X}} \nu(y) P(y, x)}_{\text{LHS: flux from } x} = \nu(x) = \underbrace{\sum_{y \in \mathcal{X}} \nu(x) P(x, y)}_{\text{RHS: flux to } x}$$



# Stationary Distribution

A stationary distribution  $\nu$  is a *left eigenvector* of  $P$ :

$$\nu = \nu P$$

- Existence and uniqueness by Perron-Frobenius theorem
- Stationary condition  $\Leftrightarrow$  global balance:

$$\forall x: \sum_{y \in \mathcal{X}} \nu(x) P(x, y) = \nu(x) = \sum_{y \in \mathcal{X}} \nu(y) P(y, x)$$

- cf. detailed balance (reversibility):  $\Uparrow$

$$\forall x, y: \nu(x) P(x, y) = \nu(y) P(y, x)$$

## Examples: Discrete Space

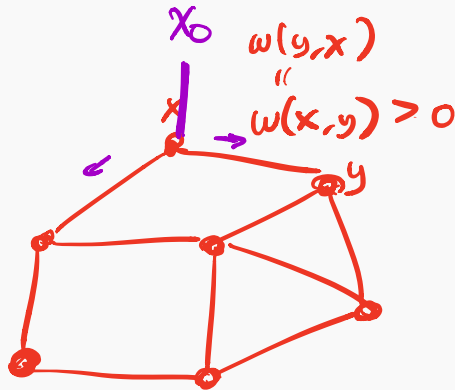
---

# Random Walk on Graph

Let  $G = (V, E, w)$  be an undirected, weighted graph

**Random walk on  $G$ :** Start at any vertex  $x \in V$ .

At each time, jump to a neighbor  $y$  with probability  $\propto w(x, y)$ .



unweighted:  $w(x, y) = 1$

$\Rightarrow$  jump to a neighbor

uniformly at random

# Random Walk on Graph

Let  $G = (V, E, w)$  be an undirected, weighted graph

**Random walk on  $G$ :** Start at any vertex  $x \in V$ .

At each time, jump to a neighbor  $y$  with probability  $\propto w(x, y)$ .

**Markov chain:**  $P_x(y) = \frac{w(x, y)}{\deg(x)}$  ←

**Stationary:** Degree distribution  $\nu(x) \propto \deg(x) = \sum_{y \in \mathcal{X}} w(x, y)$

check:  $\nu = \nu P$

$$\nu(y) = \sum_x \nu(x) \cdot P(x, y)$$



# Random Walk on Graph

Let  $G = (V, E, w)$  be an undirected, weighted graph

**Random walk on  $G$ :** Start at any vertex  $x \in V$ .

At each time, jump to a neighbor  $y$  with probability  $\propto w(x, y)$ .

**Markov chain:**  $P_x(y) = \frac{w(x, y)}{\deg(x)}$

$$D = \sum_x \deg(x) = 2 \sum_{(x,y) \in E} w(x, y)$$

**Stationary:** Degree distribution  $\nu(x) \propto \deg(x) = \sum_{y \in \mathcal{X}} w(x, y)$



Fact: Satisfies detailed balance ( $P$  is reversible with respect to  $\nu$ )

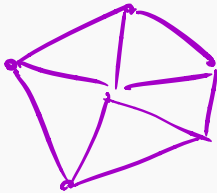
Proof:  $\nu(x) \cdot P(x, y) \stackrel{?}{=} \nu(y) \cdot P(y, x)$

$$\frac{\cancel{\deg(x)}}{D} \cdot \frac{w(x, y)}{\cancel{\deg(x)}} = \frac{w(x, y)}{D} = \frac{w(y, x)}{D} = \frac{\deg(y)}{D} \cdot \frac{w(y, x)}{\deg(y)} \quad D$$

# Random Walk on Graph

Conversely:

- Any Markov chain on a finite space  $\mathcal{X}$  can be written as a random walk on a directed, weighted graph



$$G = (V, E, w)$$

$$V = \mathcal{X}$$

$$w(x, y) = P_x(y) \geq 0$$

# Random Walk on Graph

Conversely:

- Any Markov chain on a finite space  $\mathcal{X}$  can be written as a random walk on a directed, weighted graph
- Any reversible Markov chain on a finite space  $\mathcal{X}$  can be written as a random walk on an undirected, weighted graph

$$\mathcal{V} = \mathcal{X} = \{1, \dots, n\}$$

$$\text{weight: } w(x, y) = \nu(x) \cdot P(x, y) \stackrel{\text{rev.}}{=} \nu(y) \cdot P(y, x) = w(y, x)$$

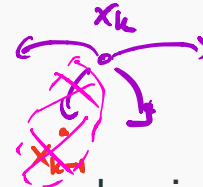
$$\text{degree: } d(x) = \sum_y w(x, y) = \sum_y \nu(x) \cdot P(x, y) = \nu(x)$$

$$\text{Markov chain: } \frac{w(x, y)}{d(x)} = \frac{\cancel{\nu(x)} \cdot P(x, y)}{\cancel{\nu(x)}} = P(x, y)$$

# Non-Example: Non-Backtracking Walk

Non-Backtracking Walk on Graph: At time  $k$  from  $X_k \in V$ :  
Jump to a neighbor  $X_{k+1}$  which is not the previous point  $X_{k-1}$

- This is a second-order Markov chain
- Can write as a first-order Markov chain by enlarging space (but lose reversibility)



$$\chi = V \times V$$
$$(x_k, x_{k-1})$$

# Non-Example: Non-Backtracking Walk

**Non-Backtracking Walk on Graph:** At time  $k$  from  $X_k \in V$ :  
Jump to a neighbor  $X_{k+1}$  which is not the previous point  $X_{k-1}$

- This is a *second-order* Markov chain
- Can write as a first-order Markov chain by enlarging space (but lose reversibility)
- Second-order/non-reversible Markov chains can be faster

Project? :

- ★ Alon et al., *Non-backtracking random walks mix faster*, Communications in Contemporary Mathematics, 2007
- ★ Diaconis & Miclo, *On the spectral analysis of second-order Markov chains*, 2013, available at [http://www.numdam.org/item/AFST\\_2013\\_6\\_22\\_3\\_573\\_0](http://www.numdam.org/item/AFST_2013_6_22_3_573_0)

# Continuous Space

---

# Markov Chain in Continuous Space

A Markov chain  $P$  on  $\mathcal{X} = \mathbb{R}^n$  is written as a transition density  
 $P: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  with:

→ • non-negative entries:  $P(x, y) = P_x(y) = P(y \mid x) \geq 0$

→ • row sums to 1:  $\int_{\mathbb{R}^n} P(x, y) dy = 1$

# Markov Chain in Continuous Space

A Markov chain  $P$  on  $\mathcal{X} = \mathbb{R}^n$  is written as a *transition density*  $P: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  with:

- non-negative entries:  $P(x, y) = P_x(y) = P(y \mid x) \geq 0$
- row sums to 1:  $\int_{\mathbb{R}^n} P(x, y) dy = 1$

A probability distribution  $\rho$  on  $\mathbb{R}^n$  is  $\rho: \mathbb{R}^n \rightarrow [0, \infty]$  with

$$\int_{\mathbb{R}^n} \rho(x) dx = 1$$

Operation of Markov chain  $X_0 \rightarrow X_1$  is a linear functional:

$$\rho_1(y) = \int_{\mathbb{R}^n} \rho_0(x) P(x, y) dx$$



# Stationary Distribution


A stationary distribution  $\nu$  is a *left eigenfunction* of  $P$ :

$$\nu(y) = \int_{\mathbb{R}^n} \nu(x) P(x, y) dx$$

# Stationary Distribution

A stationary distribution  $\nu$  is a *left eigenfunction* of  $P$ :

$$\nu(y) = \int_{\mathbb{R}^n} \nu(x) P(x, y) dx$$

- $\Leftrightarrow$  *Global balance*:  $\int_{\mathbb{R}^n} \nu(y) P(y, x) dx = \int_{\mathbb{R}^n} \nu(x) P(x, y) dy$
- cf. detailed balance (reversibility): 

$$\nu(y) P(y, x) = \nu(x) P(x, y)$$

## Examples: Continuous Space

---

# Brownian Motion

Let  $\eta > 0$  be constant (*step size*),  $\mathcal{X} = \mathbb{R}^n$

**Brownian motion** (Gaussian walk): From  $X$ , move to

$$Y = X + \sqrt{2\eta}Z$$

where  $Z \sim \mathcal{N}(0, I)$  is independent



# Brownian Motion

Let  $\eta > 0$  be constant (*step size*),  $\mathcal{X} = \mathbb{R}^n$

**Brownian motion** (Gaussian walk): From  $X$ , move to

$$Y = X + \sqrt{2\eta}Z$$

where  $Z \sim \mathcal{N}(0, I)$  is independent

- Markov chain:  $P_x = \mathcal{N}(x, 2\eta I)$  with density

$$P_x(y) = \frac{1}{(4\pi\eta)^{n/2}} \exp\left(-\frac{\|y - x\|^2}{4\eta}\right)$$

# Brownian Motion

Let  $\eta > 0$  be constant (*step size*),  $\mathcal{X} = \mathbb{R}^n$

**Brownian motion** (Gaussian walk): From  $X$ , move to

$$Y = X + \sqrt{2\eta}Z$$

where  $Z \sim \mathcal{N}(0, I)$  is independent

- Markov chain:  $P_x = \mathcal{N}(x, 2\eta I)$  with density

$$P_x(y) = \frac{1}{(4\pi\eta)^{n/2}} \exp\left(-\frac{\|y - x\|^2}{4\eta}\right) = \mathcal{P}_y(x)$$

- Stationary: Lebesgue measure  $\nu(x) = 1 = \nu(y)$
  - Reversible (because symmetric:  $P_x(y) = P_y(x)$ )
- }  $\nu(x) \cdot P_x(y) = \nu(y) \cdot P_y(x)$

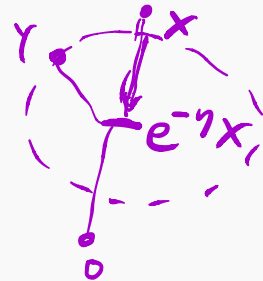
# Ornstein-Uhlenbeck

Let  $\eta > 0$  be constant (*step size*),  $\mathcal{X} = \mathbb{R}^n$

**Ornstein-Uhlenbeck:** From  $X$ , move to

$$Y = e^{-\eta}X + \underbrace{\sqrt{(1 - e^{-2\eta})}}_{\text{step size}} Z$$

where  $Z \sim \mathcal{N}(0, I)$  is independent



# Ornstein-Uhlenbeck

Let  $\eta > 0$  be constant (*step size*),  $\mathcal{X} = \mathbb{R}^n$

**Ornstein-Uhlenbeck:** From  $X$ , move to

$$Y = e^{-\eta}X + \sqrt{(1 - e^{-2\eta})}Z$$

where  $Z \sim \mathcal{N}(0, I)$  is independent

- Markov chain:  $P_x = \mathcal{N}(e^{-\eta}x, (1 - e^{-2\eta})I)$  with density

$$\rightarrow P_x(y) = \frac{1}{(2\pi(1 - e^{-2\eta}))^{n/2}} \exp\left(-\frac{\|y - e^{-\eta}x\|^2}{2(1 - e^{-2\eta})}\right)$$



# Ornstein-Uhlenbeck

Let  $\eta > 0$  be constant (*step size*),  $\mathcal{X} = \mathbb{R}^n$

**Ornstein-Uhlenbeck:** From  $X$ , move to

$$Y = e^{-\eta}X + \sqrt{(1 - e^{-2\eta})}Z$$

where  $Z \sim \mathcal{N}(0, I)$  is independent

- Markov chain:  $P_x = \mathcal{N}(e^{-\eta}x, (1 - e^{-2\eta})I)$  with density

$$P_x(y) = \frac{1}{(2\pi(1 - e^{-2\eta}))^{n/2}} \exp\left(-\frac{\|y - e^{-\eta}x\|^2}{2(1 - e^{-2\eta})}\right)$$

- Reversible wrt stationary Gaussian distribution  $\nu = \mathcal{N}(0, I)$ :

★  
Exercise

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

# Lazy Markov Chain

A Markov chain  $P$  is **lazy** if  $P_x(\{x\}) \geq \frac{1}{2}$  for all  $x \in \mathcal{X}$

Given a Markov chain  $P$ , can make a *lazy* version:

From  $X$ , draw  $U \sim \text{Uniform}(\{0, 1\})$ :

1. If  $U = 0$ , stay at  $X$ ;
2. If  $U = 1$ , move to  $Y \sim P_X$

New Markov chain is lazy:

$$\tilde{P}_x = \frac{1}{2}\delta_x + \frac{1}{2}P_x$$

# Distance between distributions

---

# Distance between distributions

Many ways to measure distance between distributions  $\rho$  and  $\nu$

- 1. Total Variation (TV) distance ( $L^1(\nu)$ )
- 2.  $\chi^2$ -divergence ( $L^2(\nu)$ )
- 3. KL divergence (relative entropy)

(also  $f$ -divergence, Wasserstein distance, Rényi divergence, ...)

TV is good for discrete+continuous;  $\chi^2$ , KL have smoothness

## Total Variation distance

$$\mathrm{TV}(\rho, \nu) = \sup_A |\rho(A) - \nu(A)| = \frac{1}{2} \|\rho - \nu\|_{L^1(dx)}$$

# Total Variation distance

$$\text{TV}(\rho, \nu) = \sup_A |\rho(A) - \nu(A)| = \frac{1}{2} \|\rho - \nu\|_{L^1(dx)}$$

- A metric (symmetric, satisfies triangle inequality)
- Range:  $0 \leq \text{TV}(\rho, \nu) \leq 1$
- Works for any distribution: Discrete, continuous, mixture, ...
- If  $\rho$  has density  $h = \frac{d\rho}{d\nu}$ , then can also write

$$\text{TV}(\rho, \nu) = \frac{1}{2} \|h - 1\|_{L^1(\nu)}$$

## $\chi^2$ -divergence

Assume  $\rho$  has density  $h = \frac{d\rho}{d\nu}$  with respect to  $\nu$

$$\chi_\nu^2(\rho) = \text{Var}_\nu(h) = \|h - 1\|_{L^2(\nu)}^2$$

# $\chi^2$ -divergence

Assume  $\rho$  has density  $h = \frac{d\rho}{d\nu}$  with respect to  $\nu$

$$\chi_\nu^2(\rho) = \text{Var}_\nu(h) = \|h - 1\|_{L^2(\nu)}^2$$

- Not a metric (not symmetric)
- Range can be large ( $\sim$  exponential in  $n$ , or  $\infty$ )
- Inequality:

$$\text{TV}(\rho, \nu) \leq \frac{1}{2} \sqrt{\chi_\nu^2(\rho)}$$



## Example

Let  $\nu = \mathcal{N}(0, I)$ ,  $\rho = \mathcal{N}(0, \alpha I)$  for some  $\alpha > 0$

$$\text{TV}(\rho, \nu) \leq 1$$

$$\chi^2_\nu(\rho) = \begin{cases} \frac{1}{(\alpha(2-\alpha))^{n/2}} - 1 & \text{if } 0 < \alpha < 2 \\ 0 & \text{if } \alpha \geq 2 \end{cases}$$

# Mixing time

---

# Mixing time

$$X_0 \rightarrow X_1 \rightarrow X_2 \rightarrow \dots, \quad X_n \sim \mathcal{P}_n$$

Choose a distance/divergence  $d$  between distributions

**Mixing time:**

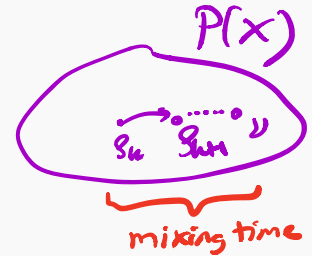
$$\tau(\epsilon) = \inf \{ K : \underbrace{d(\rho_k, \nu)}_{\sim} \leq \epsilon \quad \forall k \geq K \}$$

# Mixing time

Choose a distance/divergence  $d$  between distributions

Mixing time:

$$X_k \sim \rho_k$$



$$\tau(\epsilon) = \inf\{K: d(\rho_k, \nu) \leq \epsilon \quad \forall k \geq K\}$$

$$d(\rho_k, \nu) \leq \alpha^k, \quad \alpha < 1$$

- Exponential convergence  $\Rightarrow$  dependence on  $\epsilon$  is  $\log(1/\epsilon)$
- Want to examine dependence on:

- Problem size/dimension
- Target distribution  $\leftarrow$
- Initial distribution  $\leftarrow$

$$\text{eg.} \quad \chi^2_\nu(\rho_k) \leq \underbrace{(1-r)^k}_{\uparrow} \cdot \underbrace{\chi^2_\nu(\rho_0)}_{\uparrow}$$

THM: for  $\nu$  nice (uniform/log-concave)  
for  $P_x$  nice (....)

mixing time is  $\boxed{\tau(\epsilon) \leq \frac{L}{\alpha} (1-r)^k \log \frac{1}{\epsilon}} \leq 200$

# Mixing Time Example: Card Shuffles

The New York Times

TUESDAY, JANUARY 9, 1990

## In Shuffling Cards, 7 Is Winning Number

Computers confirm gamblers' intuition on how decks mix.

By GINA KOLATA

**I**T takes just seven ordinary, imperfect shuffles to mix a deck of cards thoroughly, researchers have found. Fewer are not enough and more do not significantly improve the mixing.

The mathematical proof, discovered after studies from elaborate computer calculations and careful observation of card games, confirms the intuition of many gamblers, bridge enthusiasts and casual players that most shuffling is inadequate.

The finding has implications for everyone who plays cards and everyone, from casino operators to magicians, who has a stake in knowing whether a shuffle is random.

The mathematical problem was complicated because of the immense number of possible ways the cards in a deck can be arranged; any of 52 could be first in the deck, any of 51 could be second, 50 could be third and so on. Multiplied out, the number of possible permutations, 52 factorial, or 52 times 51 times 50, etc., is  $10^{68}$  or  $10$  with 62 zeros after it.

No one expected that the shuffling problem would have a simple answer, said Dr. Dave Bayer, a mathematician and computer scientist at Columbia who is a co-author of the recent discovery. Other problems in statistics, like analyzing speech patterns to identify speakers, might be



The New York Times/Rick Friedman  
Dr. Persi Diaconis, Harvard mathematician, used a computer to analyze the shuffling patterns.

amenable to similar approaches, he said.

The new result "definitely solves the problem," said Dr. David Aldous, a statistician at the University of California at Berkeley. "All their calculations are right. It's a fascinating result."

Dr. Persi Diaconis, a mathematician and statistician at Harvard University who is the other author of the discovery, said the methods used are already helping mathematicians analyze problems in abstract mathematics that have nothing to do with shuffling or with any known real-world phenomena.

Dr. Diaconis, who is also a magi-

Continued on Page C12

$$X = S_d = \{\text{permutations on } d \text{ cards}\}$$

$$\Rightarrow |X| = d! \sim d^d \leftarrow$$

$\nu$  = uniform

Mixing time in TV distance for  $d$  cards:

- $\rightarrow$  • Top-to-random shuffle  $\sim d \log d$
- $\rightarrow$  • Rifle shuffle  $\sim \frac{3}{2} \log_2 d$

Bayes & Diaconis, *Trailing the Dovetail Shuffle to Its Lair*, Annals of Applied Probability, 1992

$$X = S_d \leftarrow \text{permutation}$$

