CS 263: Counting and Sampling

Nima Anari

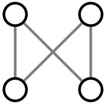
Stanford University

slides for

Markov Chain Mixing

Review

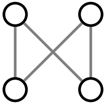
- [Pólya]'s scheme:



$$\det\left(\begin{bmatrix} +1 & -1 \\ +1 & +1 \end{bmatrix}\right) = \operatorname{per}\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)$$

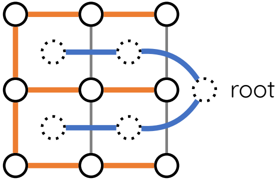
Review

- [Pólya]'s scheme:



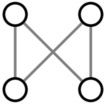
$$\det\left(\begin{bmatrix} +1 & -1 \\ +1 & +1 \end{bmatrix}\right) = \mathrm{per}\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)$$

- Good signing exists for planar graphs [Fisher-Kasteleyn-Temperley].



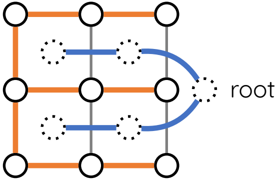
Review

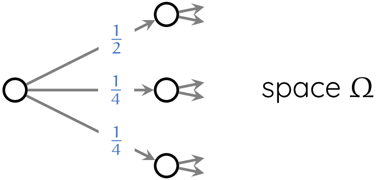
- [Pólya]'s scheme:



$$\det\left(\begin{bmatrix}+1 & -1\\ +1 & +1\end{bmatrix}\right) = \mathrm{per}\left(\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\right)$$

- Good signing exists for planar graphs [Fisher-Kasteleyn-Temperley].

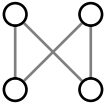




transition matrix: $P \in \mathbb{R}_{\geqslant 0}^{\Omega\times\Omega}$

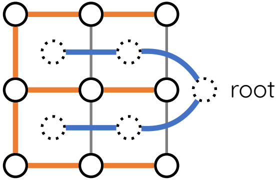
Review

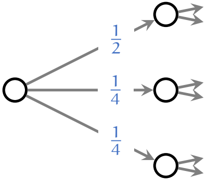
- [Pólya]'s scheme:



$$\det\left(\begin{bmatrix} +1 & -1 \\ +1 & +1 \end{bmatrix}\right) = \operatorname{per}\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)$$

- Good signing exists for planar graphs [Fisher-Kasteleyn-Temperley].





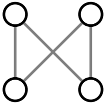
space Ω

transition matrix: $P \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega}$

- Stationary dist: $\mu = \mu P$

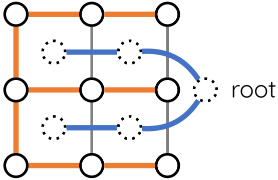
Review

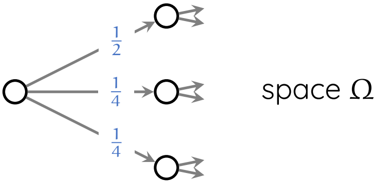
- [Pólya]'s scheme:



$$\det\left(\begin{bmatrix}+1 & -1\\ +1 & +1\end{bmatrix}\right) = \mathrm{per}\left(\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\right)$$

- Good signing exists for planar graphs [Fisher-Kasteleyn-Temperley].





transition matrix: $P \in \mathbb{R}_{\geqslant 0}^{\Omega\times\Omega}$

- Stationary dist: $\mu = \mu P$

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

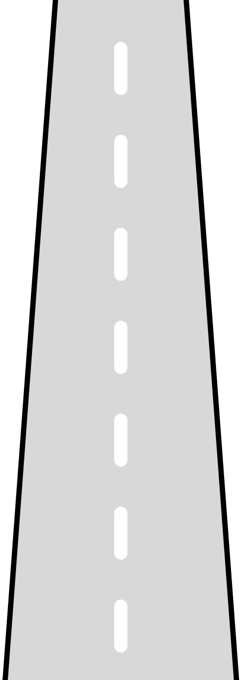
$$\lim_{t\to\infty} \nu P^t = \mu.$$

Markov Chain Mixing

- Fundamental theorem
- Mixing time growth
- Strong stationary time

Designing Markov Chains

- Reversible chains
- Metropolis filter

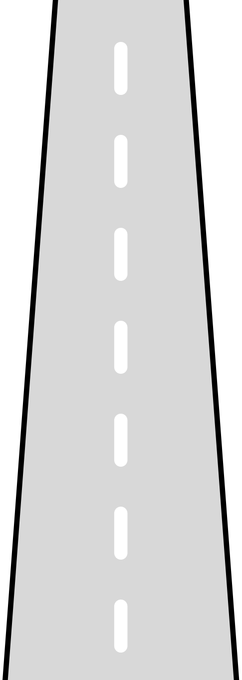


Markov Chain Mixing

- Fundamental theorem
- Mixing time growth
- Strong stationary time

Designing Markov Chains

- Reversible chains
- Metropolis filter



- Irreducible: possible to reach from every x to every y.

- Irreducible: possible to reach from every x to every y.
- Aperiodic: length of cycles from x to x have $\gcd = 1$.

- Irreducible: possible to reach from every x to every y.
- Aperiodic: length of cycles from x to x have $\gcd = 1$.
- Ergodic: irreducible + aperiodic

- Irreducible: possible to reach from every x to every y.
- Aperiodic: length of cycles from x to x have $\gcd = 1$.
- Ergodic: irreducible + aperiodic

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t \to \infty} \nu P^t = \mu.$$

- Irreducible: possible to reach from every x to every y.
- Aperiodic: length of cycles from x to x have $\gcd = 1$.
- Ergodic: irreducible + aperiodic

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

- Stationary dist: $\mu = \mu P$

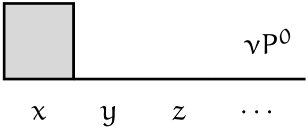
- Irreducible: possible to reach from every x to every y.
- Aperiodic: length of cycles from x to x have $\gcd = 1$.
- Ergodic: irreducible + aperiodic

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

- Stationary dist: $\mu = \mu P$



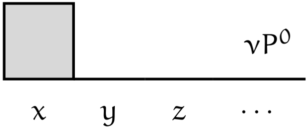
- Irreducible: possible to reach from every x to every y.
- Aperiodic: length of cycles from x to x have $\gcd = 1$.
- Ergodic: irreducible + aperiodic

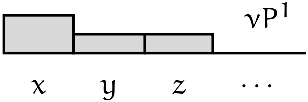
Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

- Stationary dist: $\mu = \mu P$





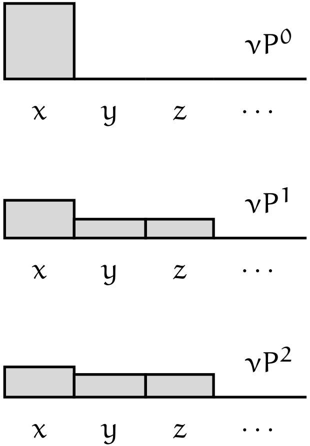
- Irreducible: possible to reach from every x to every y.
- Aperiodic: length of cycles from x to x have $\gcd = 1$.
- Ergodic: irreducible + aperiodic

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

- Stationary dist: $\mu = \mu P$



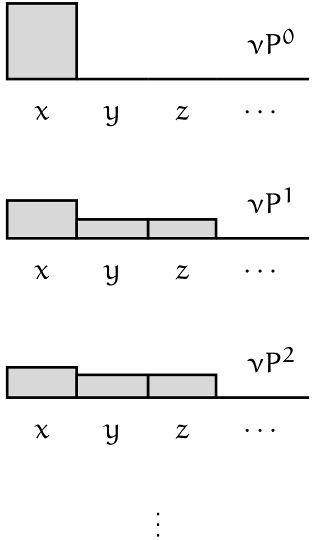
- Irreducible: possible to reach from every x to every y.
- Aperiodic: length of cycles from x to x have $\gcd = 1$.
- Ergodic: irreducible + aperiodic

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

- Stationary dist: $\mu = \mu P$



How to prove convergence?

How to prove convergence?

contraction

Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

if we could sneak in a 0.99 we'd be done

Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

if we could sneak in a 0.99 we'd be done

- Strong contraction $\implies$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:

  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$

  so it converges.

Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

if we could sneak in a 0.99 we'd be done

- Strong contraction $\implies$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:

  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$

  so it converges.
- Stationary is unique, because $d_{TV} = 0$ for any two limits.

Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

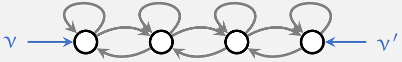
if we could sneak in a 0.99 we'd be done

- Strong contraction $\Longrightarrow$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:

  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$

  so it converges.
- Stationary is unique, because $d_{TV} = 0$ for any two limits.

Example: no strong contraction



Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

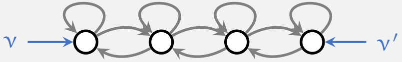
if we could sneak in a 0.99 we'd be done

- Strong contraction $\implies$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:

  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$

  so it converges.
- Stationary is unique, because $d_{TV} = 0$ for any two limits.

Example: no strong contraction



Proof of weak contraction:

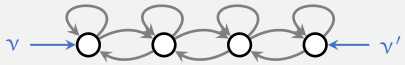
Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

if we could sneak in a 0.99 we'd be done

- Strong contraction $\implies$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:
  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$
  so it converges.
- Stationary is unique, because $d_{TV} = 0$ for any two limits.

Example: no strong contraction



Proof of weak contraction:

- There is an optimal coupling π:
  $$d_{TV}(\nu, \nu') = \mathbb{P}_{(X,X')\sim\pi}[X \neq X'].$$

Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

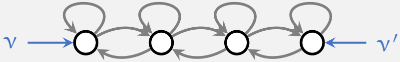
if we could sneak in a 0.99 we'd be done

- Strong contraction $\implies$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:

  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$

  so it converges.
- Stationary is unique, because $d_{TV} = 0$ for any two limits.

Example: no strong contraction



Proof of weak contraction:

- There is an optimal coupling π:

  $$d_{TV}(\nu, \nu') = \mathbb{P}_{(X,X')\sim\pi}[X \neq X'].$$
- Will construct coupling for $\nu P, \nu' P$.

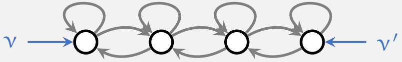
Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

if we could sneak in a 0.99 we'd be done

- Strong contraction $\implies$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:
  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$
  so it converges.
- Stationary is unique, because $d_{TV} = 0$ for any two limits.

Example: no strong contraction



Proof of weak contraction:

- There is an optimal coupling π:
  $$d_{TV}(\nu, \nu') = \mathbb{P}_{(X,X')\sim\pi}[X \neq X'].$$
- Will construct coupling for $\nu P, \nu' P$.
- Sample $(X_0, X_0') \sim \pi$.

Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

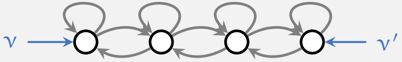
if we could sneak in a 0.99 we'd be done

- Strong contraction $\implies$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:

  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$

  so it converges.
- Stationary is unique, because $d_{TV} = 0$ for any two limits.

Example: no strong contraction



Proof of weak contraction:

- There is an optimal coupling π:

  $$d_{TV}(\nu, \nu') = \mathbb{P}_{(X,X')\sim\pi}[X \neq X'].$$
- Will construct coupling for $\nu P, \nu' P$.
- Sample $(X_0, X_0') \sim \pi$.
- Evolve by P to get X_1, X_1':
  - if $X_0 = X_0'$, use same transition
  - else, evolve arbitrarily

for example, independently

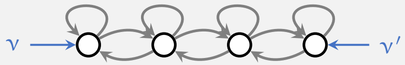
Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

if we could sneak in a 0.99 we'd be done

- Strong contraction $\implies$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:
  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$
  so it converges.
- Stationary is unique, because $d_{TV} = 0$ for any two limits.

Example: no strong contraction



Proof of weak contraction:

- There is an optimal coupling π:
  $$d_{TV}(\nu, \nu') = \mathbb{P}_{(X,X')\sim\pi}[X \neq X'].$$
- Will construct coupling for $\nu P, \nu' P$.
- Sample $(X_0, X_0') \sim \pi$.
- Evolve by P to get X_1, X_1':
  - if $X_0 = X_0'$, use same transition
  - else, evolve arbitrarily

    for example, independently
- We get
  $$\mathbb{P}[X_1 \neq X_1'] \leqslant \mathbb{P}[X_0 \neq X_0']$$

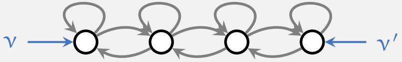
Weak contraction

$$d_{TV}(\nu P, \nu' P) \leqslant d_{TV}(\nu, \nu')$$

if we could sneak in a 0.99 we'd be done

- Strong contraction $\Longrightarrow$ $\nu, \nu P, \nu P^2, \ldots$ is Cauchy:
  $$d_{TV}(\nu P^n, \nu P^m) \leqslant 0.99^{\min\{n,m\}}$$
  so it converges.
- Stationary is unique, because $d_{TV} = 0$ for any two limits.

Example: no strong contraction



Proof of weak contraction:

- There is an optimal coupling π:
  $$d_{TV}(\nu, \nu') = \mathbb{P}_{(X,X')\sim\pi}[X \neq X'].$$
- Will construct coupling for $\nu P, \nu' P$.
- Sample $(X_0, X_0') \sim \pi$.
- Evolve by P to get X_1, X_1':
  - if $X_0 = X_0'$, use same transition
  - else, evolve arbitrarily

    for example, independently
- We get
  $$\mathbb{P}[X_1 \neq X_1'] \leqslant \mathbb{P}[X_0 \neq X_0']$$
- Rhs is $d_{TV}(\nu, \nu')$ and lhs is an upper bound on $d_{TV}(\nu P, \nu' P)$.

no ergodicity needed

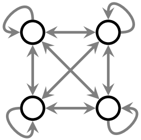
- Weak contraction always holds, but not enough 🙁

no ergodicity needed

- Weak contraction always holds, but not enough ☹
- Idea: what if different starts X_0, X_0' have chance of collision?

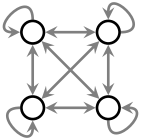
no ergodicity needed

- Weak contraction always holds, but not enough ☹
- Idea: what if different starts X_0, X_0' have chance of collision?
- Suppose $P(x, y) > 0$ for all x, y:



no ergodicity needed

- Weak contraction always holds, but not enough ☹
- Idea: what if different starts X_0, X_0' have chance of collision?
- Suppose $P(x, y) > 0$ for all x, y:

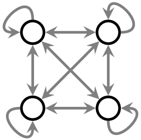


- Then in our coupling

$$\mathbb{P}[X_1 \neq X_1'] \leqslant (1 - \epsilon)\,\mathbb{P}[X_0 \neq X_0']$$

no ergodicity needed ↓

- Weak contraction always holds, but not enough ☹
- Idea: what if different starts X_0, X_0' have chance of collision?
- Suppose $P(x, y) > 0$ for all x, y:



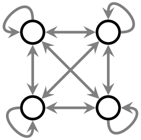
- Then in our coupling

$$\mathbb{P}[X_1 \neq X_1'] \leqslant (1 - \epsilon)\, \mathbb{P}[X_0 \neq X_0']$$

- Strong contraction holds ☺

no ergodicity needed

- Weak contraction always holds, but not enough ☹
- Idea: what if different starts X_0, X'_0 have chance of collision?
- Suppose $P(x, y) > 0$ for all x, y:



- Then in our coupling

$$\mathbb{P}[X_1 \neq X'_1] \leqslant (1 - \epsilon)\,\mathbb{P}[X_0 \neq X'_0]$$

- Strong contraction holds ☺

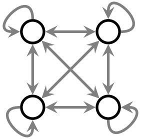
- For general P, no strong contraction ☹, but

ergodic $\implies P^t > 0$ for some t

every entry

no ergodicity needed

- Weak contraction always holds, but not enough ☹
- Idea: what if different starts X_0, X_0' have chance of collision?
- Suppose $P(x,y) > 0$ for all x, y:



- Then in our coupling

$$\mathbb{P}[X_1 \neq X_1'] \leqslant (1-\epsilon)\,\mathbb{P}[X_0 \neq X_0']$$

- Strong contraction holds ☺

- For general P, no strong contraction ☹, but

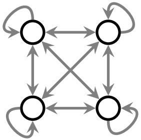
$$\text{ergodic} \implies P^t > 0 \text{ for some } t$$

every entry

- Enough for fundamental theorem.

no ergodicity needed

- Weak contraction always holds, but not enough ☹
- Idea: what if different starts X_0, X_0' have chance of collision?
- Suppose $P(x, y) > 0$ for all x, y:



- Then in our coupling

$$\mathbb{P}[X_1 \neq X_1'] \leqslant (1 - \epsilon)\, \mathbb{P}[X_0 \neq X_0']$$

- Strong contraction holds ☺

- For general P, no strong contraction ☹, but

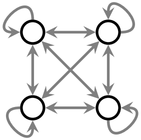
$$\text{ergodic} \implies P^t > 0 \text{ for some } t$$

every entry

- Enough for fundamental theorem.

no ergodicity needed

- Weak contraction always holds, but not enough ☹
- Idea: what if different starts X_0, X_0' have chance of collision?
- Suppose $P(x, y) > 0$ for all x, y:



- Then in our coupling

$$\mathbb{P}[X_1 \neq X_1'] \leqslant (1 - \epsilon)\, \mathbb{P}[X_0 \neq X_0']$$

- Strong contraction holds ☺

- For general P, no strong contraction ☹, but

$$\text{ergodic} \implies P^t > 0 \text{ for some } t$$

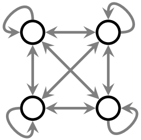
every entry

- Enough for fundamental theorem.

- Aperiodic: there are $x \to x$ loops of every len $\ell \in [\ell_0, \infty)$

no ergodicity needed

- Weak contraction always holds, but not enough 🙁
- Idea: what if different starts X_0, X'_0 have chance of collision?
- Suppose $P(x, y) > 0$ for all x, y:



- Then in our coupling

$$\mathbb{P}[X_1 \neq X'_1] \leqslant (1 - \epsilon)\,\mathbb{P}[X_0 \neq X'_0]$$

- Strong contraction holds 🙂

- For general P, no strong contraction 🙁, but

$$\text{ergodic} \implies P^t > 0 \text{ for some } t$$

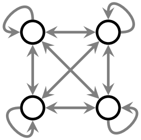
every entry

- Enough for fundamental theorem.

- Aperiodic: there are $x \to x$ loops of every len $\ell \in [\ell_0, \infty)$
- Irreducible: there is one $x \to y$ path (of len $\leqslant |\Omega|$)

no ergodicity needed

- Weak contraction always holds, but not enough ☹
- Idea: what if different starts X_0, X'_0 have chance of collision?
- Suppose $P(x, y) > 0$ for all x, y:



- Then in our coupling

$$\mathbb{P}[X_1 \neq X'_1] \leqslant (1 - \epsilon)\,\mathbb{P}[X_0 \neq X'_0]$$

- Strong contraction holds ☺

- For general P, no strong contraction ☹, but

$$\text{ergodic} \implies P^t > 0 \text{ for some } t$$

every entry

- Enough for fundamental theorem.

- Aperiodic: there are $x \to x$ loops of every len $\ell \in [\ell_0, \infty)$
- Irreducible: there is one $x \to y$ path (of len $\leqslant |\Omega|$)
- So $P^{\ell_0 + |\Omega|} > 0$

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

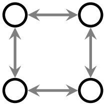
- When irreducible, all $x \in \Omega$ have same period.

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

- When irreducible, all $x \in \Omega$ have same period.
- Aperiodicity is easy to get via lazification:

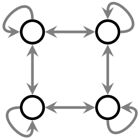


Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

- When irreducible, all $x \in \Omega$ have same period.
- Aperiodicity is easy to get via lazification:



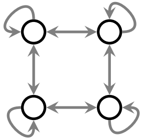
$$P \mapsto \lambda I + (1-\lambda)P$$

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

- When irreducible, all $x \in \Omega$ have same period.
- Aperiodicity is easy to get via lazification:



$$P \mapsto \lambda I + (1-\lambda)P$$

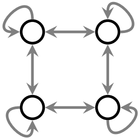
- Lazy chain has the same stationary dist.

Fundamental theorem

Every ergodic chain has a unique stationary dist μ, and for any dist ν

$$\lim_{t\to\infty} \nu P^t = \mu.$$

- When irreducible, all $x \in \Omega$ have same period.
- Aperiodicity is easy to get via lazification:



$$P \mapsto \lambda I + (1-\lambda)P$$

- Lazy chain has the same stationary dist.
- Corollary: irreducible chains have unique stationary.

Mixing time

Mixing time

For chain P with stationary μ:

$$t_{\text{mix}}(P, \epsilon, \nu) = \min\left\{t \;\middle|\; d_{\text{TV}}(\mu, \nu P^t) \leqslant \epsilon\right\}$$
$$t_{\text{mix}}(P, \epsilon) = \max\{t_{\text{mix}}(P, \epsilon, \nu) \mid \nu\}$$

Mixing time

Mixing time

For chain P with stationary μ:

$$t_{mix}(P, \epsilon, \nu) = \min\left\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\right\}$$
$$t_{mix}(P, \epsilon) = \max\{t_{mix}(P, \epsilon, \nu) \mid \nu\}$$

- Worst case ν: $\mathbb{1}_x$ for some x

Mixing time

Mixing time

For chain P with stationary μ:

$$t_{mix}(P, \epsilon, \nu) = \min\left\{t \;\middle|\; d_{TV}(\mu, \nu P^t) \leqslant \epsilon\right\}$$
$$t_{mix}(P, \epsilon) = \max\{t_{mix}(P, \epsilon, \nu) \mid \nu\}$$

- Worst case ν: $\mathbb{1}_x$ for some x
- We often talk about t_{mix} without specifying ϵ. This is justified by

Mixing time

Mixing time

For chain P with stationary μ:

$$t_{\text{mix}}(P, \epsilon, \nu) = \min\left\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\right\}$$
$$t_{\text{mix}}(P, \epsilon) = \max\{t_{\text{mix}}(P, \epsilon, \nu) \mid \nu\}$$

- Worst case ν: $\mathbb{1}_x$ for some x
- We often talk about t_{mix} without specifying ϵ. This is justified by

Lemma: mixing time growth

$$t_{\text{mix}}(P, \epsilon) \leqslant t_{\text{mix}}(P, 1/4) \cdot O(\log(1/\epsilon))$$

Mixing time

Mixing time

For chain P with stationary μ:

$$t_{\text{mix}}(P,\epsilon,\nu) = \min\{t \mid d_{TV}(\mu,\nu P^t) \leqslant \epsilon\}$$
$$t_{\text{mix}}(P,\epsilon) = \max\{t_{\text{mix}}(P,\epsilon,\nu) \mid \nu\}$$
$$t_{\text{mix}}(P) = t_{\text{mix}}(P,1/4)$$

- Worst case ν: $\mathbb{1}_x$ for some x
- We often talk about t_{mix} without specifying ϵ. This is justified by

Lemma: mixing time growth

$$t_{\text{mix}}(P,\epsilon) \leqslant t_{\text{mix}}(P,1/4) \cdot O(\log(1/\epsilon))$$

Mixing time

Mixing time

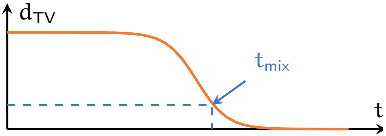
For chain P with stationary μ:

$$t_{\text{mix}}(P, \epsilon, \nu) = \min\left\{t \;\middle|\; d_{TV}(\mu, \nu P^t) \leqslant \epsilon\right\}$$
$$t_{\text{mix}}(P, \epsilon) = \max\{t_{\text{mix}}(P, \epsilon, \nu) \mid \nu\}$$
$$t_{\text{mix}}(P) = t_{\text{mix}}(P, 1/4)$$

- Worst case ν: $\mathbb{1}_x$ for some x
- We often talk about t_{mix} without specifying ϵ. This is justified by

Lemma: mixing time growth

$$t_{\text{mix}}(P, \epsilon) \leqslant t_{\text{mix}}(P, 1/4) \cdot O(\log(1/\epsilon))$$



Mixing time

Mixing time

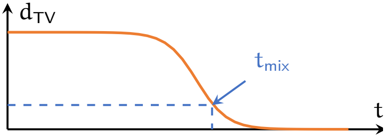
For chain P with stationary μ:

$$t_{\mathsf{mix}}(P, \epsilon, \nu) = \min\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\}$$
$$t_{\mathsf{mix}}(P, \epsilon) = \max\{t_{\mathsf{mix}}(P, \epsilon, \nu) \mid \nu\}$$
$$t_{\mathsf{mix}}(P) = t_{\mathsf{mix}}(P, 1/4)$$

- Worst case ν: $\mathbb{1}_x$ for some x
- We often talk about t_{mix} without specifying ϵ. This is justified by

Lemma: mixing time growth

$$t_{\mathsf{mix}}(P, \epsilon) \leqslant t_{\mathsf{mix}}(P, 1/4) \cdot O(\log(1/\epsilon))$$



Proof:

Mixing time

Mixing time

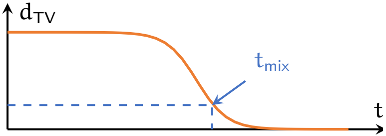
For chain P with stationary μ:

$$t_{\mathrm{mix}}(P,\epsilon,\nu) = \min\left\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\right\}$$
$$t_{\mathrm{mix}}(P,\epsilon) = \max\{t_{\mathrm{mix}}(P,\epsilon,\nu) \mid \nu\}$$
$$t_{\mathrm{mix}}(P) = t_{\mathrm{mix}}(P, 1/4)$$

- Worst case ν: $\mathbb{1}_x$ for some x
- We often talk about t_{mix} without specifying ϵ. This is justified by

Lemma: mixing time growth

$$t_{\mathrm{mix}}(P,\epsilon) \leqslant t_{\mathrm{mix}}(P, 1/4) \cdot O(\log(1/\epsilon))$$



Proof:

- Sample $X_0 \sim \nu, X_0' \sim \nu'$

Mixing time

Mixing time

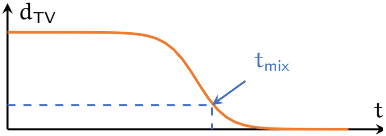
For chain P with stationary μ:

$$t_{mix}(P, \epsilon, \nu) = \min\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\}$$
$$t_{mix}(P, \epsilon) = \max\{t_{mix}(P, \epsilon, \nu) \mid \nu\}$$
$$t_{mix}(P) = t_{mix}(P, 1/4)$$

- Worst case ν: $\mathbb{1}_x$ for some x
- We often talk about t_{mix} without specifying ϵ. This is justified by

Lemma: mixing time growth

$$t_{mix}(P, \epsilon) \leqslant t_{mix}(P, 1/4) \cdot O(\log(1/\epsilon))$$



Proof:

- Sample $X_0 \sim \nu, X'_0 \sim \nu'$
- Couple X_t, X'_t for $t = t_{mix}(P)$:
  - if $X_0 = X'_0$, use same transition
  - else, couple them so $\mathbb{P}[X_t \neq X'_t \mid X_0, X'_0] \leqslant 2 \cdot 1/4 = 1/2$

Mixing time

Mixing time

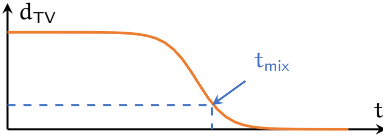
For chain P with stationary μ:

$$t_{mix}(P, \epsilon, \nu) = \min\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\}$$
$$t_{mix}(P, \epsilon) = \max\{t_{mix}(P, \epsilon, \nu) \mid \nu\}$$
$$t_{mix}(P) = t_{mix}(P, 1/4)$$

- Worst case ν: $\mathbb{1}_x$ for some x
- We often talk about t_{mix} without specifying ϵ. This is justified by

Lemma: mixing time growth

$$t_{mix}(P, \epsilon) \leqslant t_{mix}(P, 1/4) \cdot O(\log(1/\epsilon))$$



Proof:

- Sample $X_0 \sim \nu$, $X_0' \sim \nu'$
- Couple X_t, X_t' for $t = t_{mix}(P)$:
  - if $X_0 = X_0'$, use same transition
  - else, couple them so $\mathbb{P}[X_t \neq X_t' \mid X_0, X_0'] \leqslant 2 \cdot 1/4 = 1/2$
- Possible because

$$d_{TV}(P^t(x,\cdot), P^t(y,\cdot)) \leqslant d_{TV}(P^t(x,\cdot), \mu) + d_{TV}(P^t(y,\cdot), \mu)$$

- So we proved

$$d_{TV}(\nu P^{t_{mix}}, \nu' P^{t_{mix}}) \leqslant d_{TV}(\nu, \nu')/2$$

- So we proved

$$d_{TV}(\nu P^{t_{mix}}, \nu' P^{t_{mix}}) \leqslant d_{TV}(\nu, \nu')/2$$

- This finishes the proof: 🙂

$$d_{TV}(\nu P^{k\, t_{mix}}, \mu) \leqslant 1/2^k$$

- So we proved

$$d_{TV}(\nu P^{t_{mix}}, \nu' P^{t_{mix}}) \leqslant d_{TV}(\nu, \nu')/2$$

- This finishes the proof: 🙂

$$d_{TV}(\nu P^{k\, t_{mix}}, \mu) \leqslant 1/2^k$$

- Note: usually

$$t_{mix}(P, \epsilon) \ll t_{mix}(P) \log(1/\epsilon)$$

- So we proved

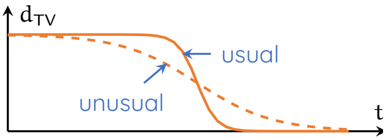
$$d_{TV}(\nu P^{t_{mix}}, \nu' P^{t_{mix}}) \leqslant d_{TV}(\nu, \nu')/2$$

- This finishes the proof: 🙂

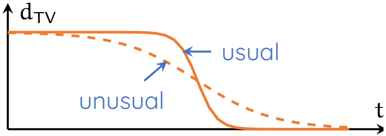
$$d_{TV}(\nu P^{k\, t_{mix}}, \mu) \leqslant 1/2^k$$

- Note: usually

$$t_{mix}(P, \epsilon) \ll t_{mix}(P) \log(1/\epsilon)$$



- So we proved
$$d_{TV}(\nu P^{t_{mix}}, \nu' P^{t_{mix}}) \leqslant d_{TV}(\nu, \nu')/2$$
- This finishes the proof: 🙂
$$d_{TV}(\nu P^{k\, t_{mix}}, \mu) \leqslant 1/2^k$$
- Note: usually
$$t_{mix}(P, \epsilon) \ll t_{mix}(P) \log(1/\epsilon)$$



- Cutoff phenomenon:
(asymptotically) plot becomes a step function

- So we proved

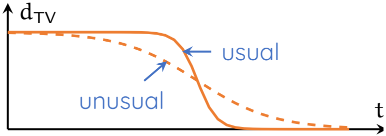
$$d_{TV}(\nu P^{t_{mix}}, \nu' P^{t_{mix}}) \leqslant d_{TV}(\nu, \nu')/2$$

- This finishes the proof: 🙂

$$d_{TV}(\nu P^{k\, t_{mix}}, \mu) \leqslant 1/2^k$$

- Note: usually

$$t_{mix}(P, \epsilon) \ll t_{mix}(P) \log(1/\epsilon)$$



- Cutoff phenomenon:
(asymptotically) plot becomes a step function

How to bound mixing time?

- So we proved

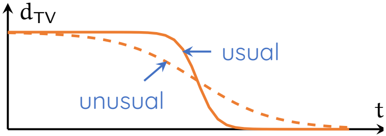
$$d_{TV}(\nu P^{t_{mix}}, \nu' P^{t_{mix}}) \leqslant d_{TV}(\nu, \nu')/2$$

- This finishes the proof: ☺

$$d_{TV}(\nu P^{k\, t_{mix}}, \mu) \leqslant 1/2^k$$

- Note: usually

$$t_{mix}(P, \epsilon) \ll t_{mix}(P) \log(1/\epsilon)$$

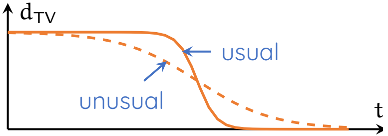


- Cutoff phenomenon:
(asymptotically) plot becomes a step function

How to bound mixing time?

- Stationary time ← today
- Coupling
- Functional analysis
- Fourier analysis
- Canonical paths
- Comparison
- Localization
- ...

- So we proved
$$d_{TV}(\nu P^{t_{mix}}, \nu' P^{t_{mix}}) \leqslant d_{TV}(\nu, \nu')/2$$
- This finishes the proof: 🙂
$$d_{TV}(\nu P^{k\, t_{mix}}, \mu) \leqslant 1/2^k$$
- Note: usually
$$t_{mix}(P, \epsilon) \ll t_{mix}(P) \log(1/\epsilon)$$



- Cutoff phenomenon:
(asymptotically) plot becomes a step function

How to bound mixing time?

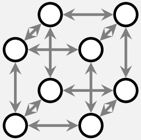
- Stationary time ← today
- Coupling
- Functional analysis
- Fourier analysis
- Canonical paths
- Comparison
- Localization
- …

prevalent idea: contraction of some proxy for d_{TV}

Strong stationary time

Example: hypercube

- $\Omega = \{0,1\}^n$
- Pick u.r. $i \in [n]$
- Replace coord i with Ber($\frac{1}{2}$)



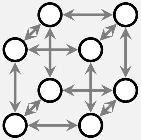
stationary: uniform

$$X_0 \to X_1 \to X_2 \to \dots$$

Strong stationary time

Example: hypercube

- $\Omega = \{0, 1\}^n$
- Pick u.r. $i \in [n]$
- Replace coord i with $\mathrm{Ber}(\frac{1}{2})$



stationary: uniform

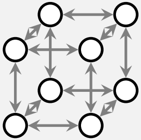
$$X_0 \to X_1 \to X_2 \to \dots$$

- Define τ: first time we have replaced every coordinate $\in [n]$

Strong stationary time

Example: hypercube

- $\Omega = \{0,1\}^n$
- Pick u.r. $i \in [n]$
- Replace coord i with $\mathrm{Ber}(\frac{1}{2})$



stationary: uniform

$$X_0 \to X_1 \to X_2 \to \dots$$

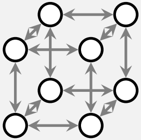
- Define τ: first time we have replaced every coordinate $\in [n]$
- We have for every k:

$$\mathrm{dist}(X_\tau \mid \tau = k) = \text{uniform}$$

Strong stationary time

Example: hypercube

- $\Omega = \{0,1\}^n$
- Pick u.r. $i \in [n]$
- Replace coord i with Ber$(\frac{1}{2})$



stationary: uniform

$$X_0 \to X_1 \to X_2 \to \dots$$

- Define τ: first time we have replaced every coordinate $\in [n]$
- We have for every k:

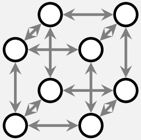
$$\text{dist}(X_\tau \mid \tau = k) = \text{uniform}$$

- Such a τ is called a strong stationary time [Aldous-Diaconis]

Strong stationary time

Example: hypercube

- $\Omega = \{0, 1\}^n$
- Pick u.r. $i \in [n]$
- Replace coord i with $\text{Ber}(\frac{1}{2})$



stationary: uniform

$$X_0 \to X_1 \to X_2 \to \dots$$

- Define τ: first time we have replaced every coordinate $\in [n]$
- We have for every k:

$$\text{dist}(X_\tau \mid \tau = k) = \text{uniform}$$

- Such a τ is called a strong stationary time [Aldous-Diaconis]

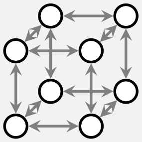
Lemma

$$\mathbb{P}[\tau > t] \leqslant \epsilon \implies t_{\text{mix}}(\epsilon) \leqslant t$$

Strong stationary time

Example: hypercube

- $\Omega = \{0,1\}^n$
- Pick u.r. $i \in [n]$
- Replace coord i with Ber$(\frac{1}{2})$



stationary: uniform

$$X_0 \to X_1 \to X_2 \to \dots$$

- Define τ: first time we have replaced every coordinate $\in [n]$
- We have for every k:

$$\mathrm{dist}(X_\tau \mid \tau = k) = \text{uniform}$$

- Such a τ is called a strong stationary time [Aldous-Diaconis]

Lemma

$$\mathbb{P}[\tau > t] \leqslant \epsilon \implies t_{\mathrm{mix}}(\epsilon) \leqslant t$$

Proof: we can write $\mathrm{dist}(X_t)$ as

$$\mathbb{P}[\tau = 0]\,\mathrm{dist}(X_0 \mid \tau = 0)P^t +$$
$$\mathbb{P}[\tau = 1]\,\mathrm{dist}(X_1 \mid \tau = 1)P^{t-1} +$$
$$\dots +$$
$$\mathbb{P}[\tau = t]\,\mathrm{dist}(X_t \mid \tau = t) +$$
$$\mathbb{P}[\tau > t]\,\mathrm{dist}(X_t \mid \tau > t)$$

and every $\mathrm{dist}(X_i \mid \tau = i)$ is μ.

- For the hypercube, we have

$$\mathbb{P}[\tau > t] \leqslant n(1 - 1/n)^t \leqslant ne^{-t/n}$$

- For the hypercube, we have

$$\mathbb{P}[\tau > t] \leqslant n(1 - 1/n)^t \leqslant ne^{-t/n}$$

- So we can bound the mixing time 🙂

$$t_{\mathsf{mix}}(\epsilon) \leqslant n \log n + n \log(1/\epsilon)$$

- For the hypercube, we have

$$\mathbb{P}[\tau > t] \leqslant n(1 - 1/n)^t \leqslant ne^{-t/n}$$

- So we can bound the mixing time 🙂

$$t_{\text{mix}}(\epsilon) \leqslant n \log n + n \log(1/\epsilon)$$

- Poll: is this tight?

- For the hypercube, we have

$$\mathbb{P}[\tau > t] \leqslant n(1 - 1/n)^t \leqslant ne^{-t/n}$$

- So we can bound the mixing time 🙂

$$t_{\mathsf{mix}}(\epsilon) \leqslant n \log n + n \log(1/\epsilon)$$

- Poll: is this tight?
- Exercise: show $t_{\mathsf{mix}} \geqslant \Omega(n \log n)$

- For the hypercube, we have

$$\mathbb{P}[\tau > t] \leqslant n(1 - 1/n)^t \leqslant ne^{-t/n}$$

- So we can bound the mixing time 🙂

$$t_{\text{mix}}(\epsilon) \leqslant n \log n + n \log(1/\epsilon)$$

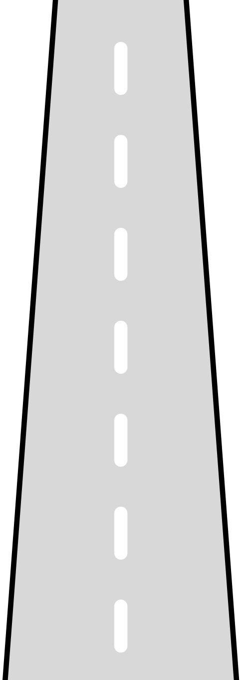
- Poll: is this tight?
- Exercise: show $t_{\text{mix}} \geqslant \Omega(n \log n)$
- Note: we have NOT proved cutoff, even though cutoff does hold for this chain.

Markov Chain Mixing

- Fundamental theorem
- Mixing time growth
- Strong stationary time

Designing Markov Chains

- Reversible chains
- Metropolis filter

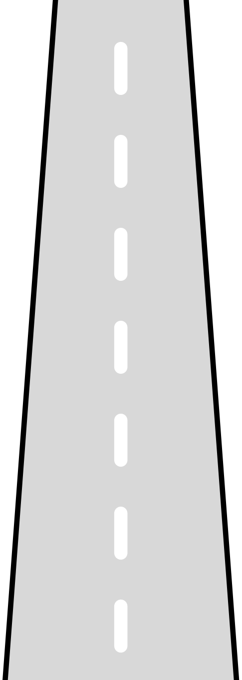


Markov Chain Mixing

- Fundamental theorem
- Mixing time growth
- Strong stationary time

Designing Markov Chains

- Reversible chains
- Metropolis filter



How to design chains?

Criteria: correct stationary dist

How to design chains?

Criteria: correct stationary dist

Definition: ergodic flow

For dist μ and chain P we define ergodic flow $Q \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega}$ as

$$Q(x, y) = \mu(x) P(x, y)$$

prob flow from x to y

How to design chains?

Criteria: correct stationary dist

Definition: ergodic flow

For dist μ and chain P we define ergodic flow $Q \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega}$ as

$$Q(x, y) = \mu(x) P(x, y)$$

prob flow from x to y

- Lemma:

  μ stationary $\leftrightarrow$ Q proper flow

  incoming=outgoing

How to design chains?

Criteria: correct stationary dist

Definition: ergodic flow

For dist μ and chain P we define ergodic flow $Q \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega}$ as

$$Q(x,y) = \mu(x)P(x,y)$$

prob flow from x to y

- Lemma:

  μ stationary $\leftrightarrow$ Q proper flow

  incoming=outgoing

- Proof:

  $$\sum_x \mu(x)P(x,y) = \sum_z \mu(y)P(y,z)$$

  $(\mu P)(y)$ $\qquad$ $\mu(y)$

How to design chains?

Criteria: correct stationary dist

Definition: ergodic flow

For dist μ and chain P we define ergodic flow $Q \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega}$ as

$$Q(x,y) = \mu(x)P(x,y)$$

prob flow from x to y

- Lemma:

  μ stationary $\leftrightarrow$ Q proper flow

  incoming=outgoing

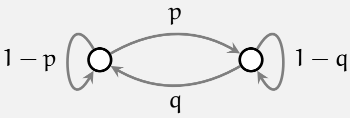
- Proof:

  $$\sum_x \mu(x)P(x,y) = \sum_z \mu(y)P(y,z)$$

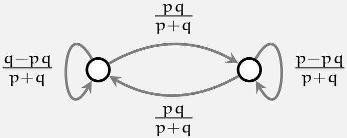
  $(\mu P)(y)$ $\qquad$ $\mu(y)$

Example

The two-state chain



has stationary $\mu = [q/(p+q), p/(p+q)]$ and ergodic flow



Time-reversible

$$Q(x, y) = Q(y, x)$$

Time-reversible

$$Q(x, y) = Q(y, x)$$

- This is also called detailed balance.

Time-reversible

$$Q(x, y) = Q(y, x)$$

- This is also called detailed balance.
- Meaning of time-reversibility: if we watch a movie of the chain at stationarity we cannot tell the direction of time.

Time-reversible

$$Q(x, y) = Q(y, x)$$

- This is also called detailed balance.
- Meaning of time-reversibility: if we watch a movie of the chain at stationarity we cannot tell the direction of time.
- Very useful design tool: make sure P is chosen so that

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

Time-reversible

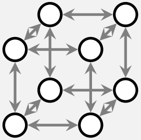
$$Q(x,y) = Q(y,x)$$

- This is also called detailed balance.
- Meaning of time-reversibility: if we watch a movie of the chain at stationarity we cannot tell the direction of time.
- Very useful design tool: make sure P is chosen so that

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

Example: hypercube

- $\Omega = \{0,1\}^n$
- Pick u.r. $i \in [n]$
- Replace coord i with Ber$(\frac{1}{2})$



Time-reversible

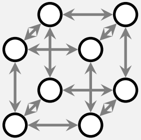
$$Q(x, y) = Q(y, x)$$

- This is also called detailed balance.
- Meaning of time-reversibility: if we watch a movie of the chain at stationarity we cannot tell the direction of time.
- Very useful design tool: make sure P is chosen so that

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

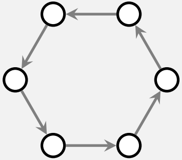
Example: hypercube

- $\Omega = \{0, 1\}^n$
- Pick u.r. $i \in [n]$
- Replace coord i with $\text{Ber}(\frac{1}{2})$



Non-example: cycle

- $\Omega = \mathbb{Z}_n$
- Go from x to $x + 1$



Time-reversal

A Markov chain P and stationary dist μ defines process

$$\cdots \to X_{-1} \to X_0 \to X_1 \to \ldots$$

Time-reversal

A Markov chain P and stationary dist μ defines process

$$\cdots \to X_{-1} \to X_0 \to X_1 \to \dots$$

Properties:

- Conditioned on X_i, left and right are independent.
- Time-homogeneous: every (X_i, X_{i+1}) is distributed $\sim Q$

Time-reversal

A Markov chain P and stationary dist μ defines process

$$\dots \to X_{-1} \to X_0 \to X_1 \to \dots$$

Properties:

- Conditioned on X_i, left and right are independent.
- Time-homogeneous: every (X_i, X_{i+1}) is distributed $\sim Q$

If we reverse the arrows, we still get a similar process

$$\dots \to X_1 \to X_0 \to X_{-1} \to \dots$$

Time-reversal

A Markov chain P and stationary dist μ defines process

$$\dots \to X_{-1} \to X_0 \to X_1 \to \dots$$

Properties:

- Conditioned on X_i, left and right are independent.
- Time-homogeneous: every (X_i, X_{i+1}) is distributed $\sim Q$

If we reverse the arrows, we still get a similar process

$$\dots \to X_1 \to X_0 \to X_{-1} \to \dots$$

Time-reversal

Given P, μ, time-reversal P° is the chain whose ergodic flow is the reversal of P's.

$$\mu(x)P(x, y) = \mu(y)P^\circ(y, x)$$

Time-reversal

A Markov chain P and stationary dist μ defines process

$$\cdots \to X_{-1} \to X_0 \to X_1 \to \dots$$

Properties:

- Conditioned on X_i, left and right are independent.
- Time-homogeneous: every (X_i, X_{i+1}) is distributed $\sim Q$

If we reverse the arrows, we still get a similar process

$$\cdots \to X_1 \to X_0 \to X_{-1} \to \dots$$

Time-reversal

Given P, μ, time-reversal P° is the chain whose ergodic flow is the reversal of P's.

$$\mu(x)P(x,y) = \mu(y)P^\circ(y,x)$$

- Time-reversible: $P = P^\circ$

Time-reversal

A Markov chain P and stationary dist μ defines process

$$\cdots \to X_{-1} \to X_0 \to X_1 \to \dots$$

Properties:

- Conditioned on X_i, left and right are independent.
- Time-homogeneous: every (X_i, X_{i+1}) is distributed $\sim Q$

If we reverse the arrows, we still get a similar process

$$\cdots \to X_1 \to X_0 \to X_{-1} \to \dots$$

Time-reversal

Given P, μ, time-reversal P° is the chain whose ergodic flow is the reversal of P's.

$$\mu(x)P(x,y) = \mu(y)P^\circ(y,x)$$

- Time-reversible: $P = P^\circ$
- Time-reversal is more generally defined for Markov kernels $P \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega'}$:

$$\mu(x)P(x,y) = \mu^\circ(y)P^\circ(y,x)$$

where $\mu^\circ = \mu P$.

How to design time-reversible chains?

Metropolis filter

- Suppose P doesn't have μ as stationary.

Metropolis filter

- Suppose P doesn't have μ as stationary.
- Define $P'(x, y)$ for $x \neq y$:

$$P(x, y) \min\left\{1, \frac{\mu(y)P(y, x)}{\mu(x)P(x, y)}\right\}$$

Metropolis filter

- Suppose P doesn't have μ as stationary.
- Define $P'(x, y)$ for $x \neq y$:

$$P(x, y) \min\left\{1, \frac{\mu(y)P(y, x)}{\mu(x)P(x, y)}\right\}$$

- Put remaining prob as $x \to x$ transition.

Metropolis filter

- Suppose P doesn't have μ as stationary.
- Define $P'(x, y)$ for $x \neq y$:

$$P(x, y) \min\left\{1, \frac{\mu(y)P(y, x)}{\mu(x)P(x, y)}\right\}$$

- Put remaining prob as $x \to x$ transition.
- Only need to know μ up to proportionality.

Metropolis filter

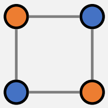
- Suppose P doesn't have μ as stationary.
- Define $P'(x, y)$ for $x \neq y$:

$$P(x, y) \min\left\{1, \frac{\mu(y)P(y, x)}{\mu(x)P(x, y)}\right\}$$

- Put remaining prob as $x \to x$ transition.
- Only need to know μ up to proportionality.

Example: coloring

- $\Omega =$ colorings
- Pick u.r. vert v
- Pick u.r. color c
- Replace v's color with c



stationary: not proper colorings

Metropolis filter

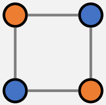
- Suppose P doesn't have μ as stationary.
- Define $P'(x, y)$ for $x \neq y$:

$$P(x, y) \min\left\{1, \frac{\mu(y)P(y, x)}{\mu(x)P(x, y)}\right\}$$

- Put remaining prob as $x \to x$ transition.
- Only need to know μ up to proportionality.

Example: coloring

- $\Omega =$ colorings
- Pick u.r. vert v
- Pick u.r. color c
- Replace v's color with c



stationary: not proper colorings

- Metropolis filter: reject transitions to invalid colorings.