CS 263: Counting and Sampling

Nima Anari

Stanford University

slides for

Mixing via Transport

Review

Fundamental theorem

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

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

Review

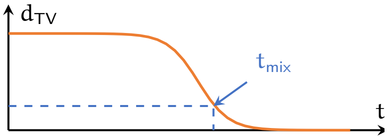
Fundamental theorem

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

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

- Mixing time $t_{\mathrm{mix}}(P, \epsilon, \nu)$:

$$\min\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\}$$



Review

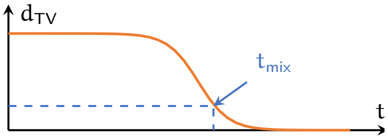
Fundamental theorem

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

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

- Mixing time $t_{\mathrm{mix}}(P, \epsilon, \nu)$:

$$\min\{t \mid d_{\mathrm{TV}}(\mu, \nu P^t) \leqslant \epsilon\}$$



- $t_{\mathrm{mix}}(P, \epsilon) = O\left(t_{\mathrm{mix}}\left(P, \frac{1}{4}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right)$

Review

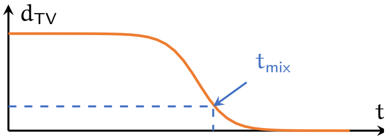
Fundamental theorem

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

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

- Mixing time $t_{\text{mix}}(P, \epsilon, \nu)$:

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

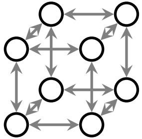


- $t_{\text{mix}}(P, \epsilon) = O\left(t_{\text{mix}}\left(P, \frac{1}{4}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right)$

- Strong stationary time:

$$\text{dist}(X_t \mid \tau = k) = \text{stationary}$$

  - τ: all coords replaced
  - $t_{\text{mix}}(\epsilon) \leqslant n \log(n/\epsilon)$



Review

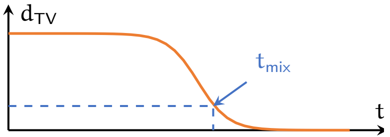
Fundamental theorem

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

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

- Mixing time $t_{mix}(P, \epsilon, \nu)$:

$$\min\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\}$$

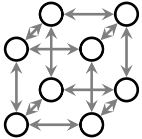


- $t_{mix}(P, \epsilon) = O\left(t_{mix}\left(P, \frac{1}{4}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right)$

- Strong stationary time:

$$\text{dist}(X_t \mid \tau = k) = \text{stationary}$$

  - τ: all coords replaced
  - $t_{mix}(\epsilon) \leqslant n \log(n/\epsilon)$



- Ergodic flow: $Q(x, y) = \mu(x) P(x, y)$

Review

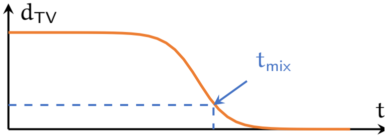
Fundamental theorem

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

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

- Mixing time $t_{\text{mix}}(P, \epsilon, \nu)$:

$$\min\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\}$$

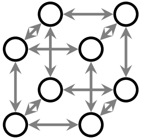


- $t_{\text{mix}}(P, \epsilon) = O\left(t_{\text{mix}}\left(P, \frac{1}{4}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right)$

- Strong stationary time:

$$\text{dist}(X_t \mid \tau = k) = \text{stationary}$$

  - τ: all coords replaced
  - $t_{\text{mix}}(\epsilon) \leqslant n \log(n/\epsilon)$



- Ergodic flow: $Q(x, y) = \mu(x) P(x, y)$
- Lemma: stationary $\leftrightarrow$ proper flow

Review

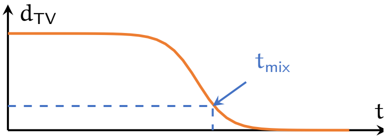
Fundamental theorem

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

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

- Mixing time $t_{\text{mix}}(P, \epsilon, \nu)$:

$$\min\{t \mid d_{TV}(\mu, \nu P^t) \leq \epsilon\}$$

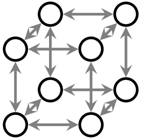


- $t_{\text{mix}}(P, \epsilon) = O\left(t_{\text{mix}}\left(P, \frac{1}{4}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right)$

- Strong stationary time:

$$\text{dist}(X_t \mid \tau = k) = \text{stationary}$$

  - τ: all coords replaced
  - $t_{\text{mix}}(\epsilon) \leq n \log(n/\epsilon)$



- Ergodic flow: $Q(x, y) = \mu(x) P(x, y)$
- Lemma: stationary $\leftrightarrow$ proper flow
- Detailed balance/time-reversible:

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

Review

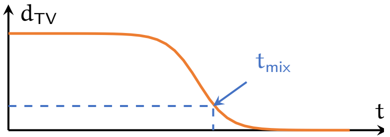
Fundamental theorem

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

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

- Mixing time $t_{\text{mix}}(P, \epsilon, \nu)$:

$$\min\{t \mid d_{TV}(\mu, \nu P^t) \leqslant \epsilon\}$$

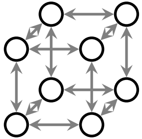


- $t_{\text{mix}}(P, \epsilon) = O\left(t_{\text{mix}}\left(P, \frac{1}{4}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right)$

- Strong stationary time:

$$\text{dist}(X_t \mid \tau = k) = \text{stationary}$$

  - τ: all coords replaced
  - $t_{\text{mix}}(\epsilon) \leqslant n \log(n/\epsilon)$



- Ergodic flow: $Q(x, y) = \mu(x)P(x, y)$
- Lemma: stationary $\leftrightarrow$ proper flow
- Detailed balance/time-reversible:

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

- Metropolis filter: $P(x, y) \mapsto$

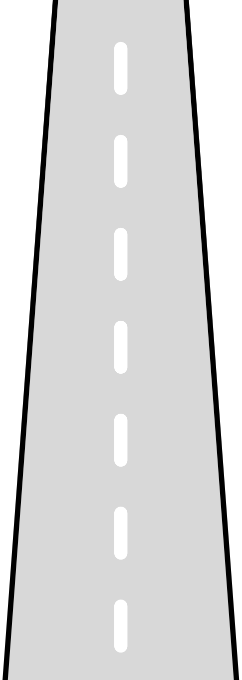
$$P(x, y) \min\left\{1, \frac{\mu(y)P(y,x)}{\mu(x)P(x,y)}\right\}$$

Designing Markov Chains

- Markov kernels
- Combination with time-reversal

Mixing via Transport

- Wasserstein distance
- Path coupling

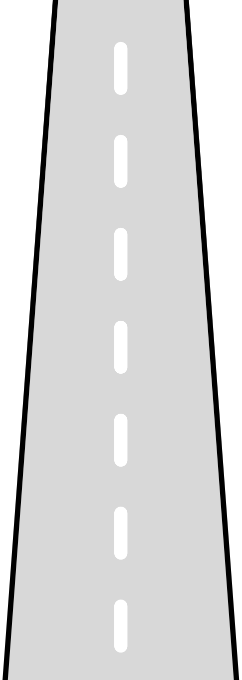


Designing Markov Chains

- Markov kernels
- Combination with time-reversal

Mixing via Transport

- Wasserstein distance
- Path coupling



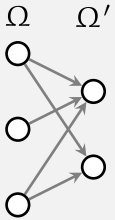
Markov kernel

We can generalize time-reversal to

Markov kernel

$$P \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega'}$$

$$\sum_y P(x, y) = 1$$



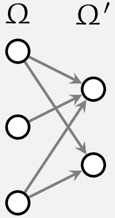
Markov kernel

We can generalize time-reversal to

Markov kernel

$$P \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega'}$$

$$\sum_y P(x, y) = 1$$



- Markov kernels are conditional dists. Combined with dist μ on Ω, they give joint dist/ergodic flow:

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

Markov kernel

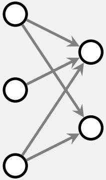
We can generalize time-reversal to

Markov kernel

$$P \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega'}$$

$$\sum_y P(x,y) = 1$$

Ω Ω'



- Markov kernels are conditional dists. Combined with dist μ on Ω, they give joint dist/ergodic flow:

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

- Time-reversal:

$$Q^{\circ}(y,x) = Q(x,y)$$

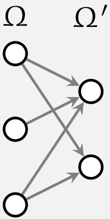
Markov kernel

We can generalize time-reversal to

Markov kernel

$P \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega'}$

$\sum_y P(x,y) = 1$



- Markov kernels are conditional dists. Combined with dist μ on Ω, they give joint dist/ergodic flow:

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

- Time-reversal:

$$Q^\circ(y,x) = Q(x,y)$$

- Dist on Ω': $\mu^\circ = \mu P$ is marginal of y in Q° or Q.

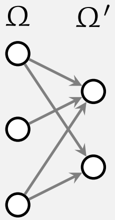
Markov kernel

We can generalize time-reversal to

Markov kernel

$$P \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega'}$$

$$\sum_y P(x, y) = 1$$



- Markov kernels are conditional dists. Combined with dist μ on Ω, they give joint dist/ergodic flow:

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

- Time-reversal:

$$Q^\circ(y, x) = Q(x, y)$$

- Dist on Ω': $\mu^\circ = \mu P$ is marginal of y in Q° or Q.
- The time-reversal Markov kernel is the conditional dist of x given y:

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

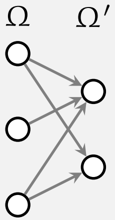
Markov kernel

We can generalize time-reversal to

Markov kernel

$$P \in \mathbb{R}_{\geqslant 0}^{\Omega \times \Omega'}$$

$$\sum_y P(x,y) = 1$$



- Markov kernels are conditional dists. Combined with dist μ on Ω, they give joint dist/ergodic flow:

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

- Time-reversal:

$$Q^\circ(y,x) = Q(x,y)$$

- Dist on Ω': $\mu^\circ = \mu P$ is marginal of y in Q° or Q.
- The time-reversal Markov kernel is the conditional dist of x given y:

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

- Note the detailed balance equation:

$$\underbrace{\mu(x)P(x,y)}_{Q(x,y)} = \underbrace{\mu^\circ(y)P^\circ(y,x)}_{Q^\circ(y,x)}$$

Combination with time-reversal

Design recipe:

1. Target dist μ on Ω

Combination with time-reversal

Design recipe:

1. Target dist μ on Ω
2. Markov kernel N from Ω to Ω'

Combination with time-reversal

Design recipe:

1. Target dist μ on Ω
2. Markov kernel N from Ω to Ω'
3. Let $P = NN^{\circ}$

Combination with time-reversal

Design recipe:

1. Target dist μ on Ω
2. Markov kernel N from Ω to Ω'
3. Let $P = NN^{\circ}$

Lemma

$\mu P = \mu$ and P is time-reversible

Combination with time-reversal

Design recipe:

1. Target dist μ on Ω
2. Markov kernel N from Ω to Ω'
3. Let $P = NN^\circ$

Lemma

$\mu P = \mu$ and P is time-reversible

Proof: we have $\mu(x)P(x,z) =$

$$\sum_y \mu(x)N(x,y)N^\circ(y,z) =$$

$$\sum_y \frac{\mu(x)N(x,y)\mu(z)N(z,y)}{\mu^\circ(y)}$$

symmetric in x, z

Combination with time-reversal

Design recipe:

1. Target dist μ on Ω
2. Markov kernel N from Ω to Ω'
3. Let $P = NN^\circ$

Lemma

$\mu P = \mu$ and P is time-reversible

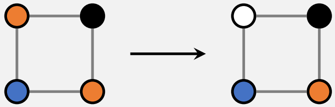
Proof: we have $\mu(x)P(x,z) =$

$$\sum_y \mu(x)N(x,y)N^\circ(y,z) =$$

$$\sum_y \frac{\mu(x)N(x,y)\mu(z)N(z,y)}{\mu^\circ(y)}$$

symmetric in x, z

Example: Glauber dynamics



Combination with time-reversal

Design recipe:

1. Target dist μ on Ω
2. Markov kernel N from Ω to Ω'
3. Let $P = NN^\circ$

Lemma

$\mu P = \mu$ and P is time-reversible

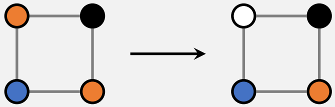
Proof: we have $\mu(x)P(x,z) =$

$$\sum_y \mu(x)N(x,y)N^\circ(y,z) =$$

$$\sum_y \frac{\mu(x)N(x,y)\mu(z)N(z,y)}{\mu^\circ(y)}$$

symmetric in x, z

Example: Glauber dynamics



- N: erase u.r. vertex

Combination with time-reversal

Design recipe:

1. Target dist μ on Ω
2. Markov kernel N from Ω to Ω'
3. Let $P = NN^\circ$

Lemma

$\mu P = \mu$ and P is time-reversible

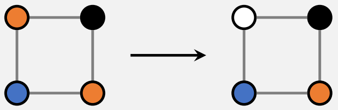
Proof: we have $\mu(x)P(x,z) =$

$$\sum_y \mu(x)N(x,y)N^\circ(y,z) =$$

$$\sum_y \frac{\mu(x)N(x,y)\mu(z)N(z,y)}{\mu^\circ(y)}$$

symmetric in x, z

Example: Glauber dynamics



- N: erase u.r. vertex
- N°: recolor with prob $\propto$

$$\mu(\text{result})N(\text{result}, \text{partial})$$

cancels out

Combination with time-reversal

Design recipe:

1. Target dist μ on Ω
2. Markov kernel N from Ω to Ω'
3. Let $P = NN^\circ$

Lemma

$\mu P = \mu$ and P is time-reversible

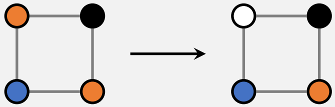
Proof: we have $\mu(x)P(x,z) =$

$$\sum_y \mu(x)N(x,y)N^\circ(y,z) =$$

$$\sum_y \frac{\mu(x)N(x,y)\mu(z)N(z,y)}{\mu^\circ(y)}$$

symmetric in x, z

Example: Glauber dynamics



- N: erase u.r. vertex
- N°: recolor with prob $\propto$

  $\mu(\text{result})N(\text{result}, \text{partial})$

  cancels out

- P: pick u.r. valid color for u.r. vert

Combination with time-reversal

Design recipe:

1. Target dist μ on Ω
2. Markov kernel N from Ω to Ω'
3. Let $P = NN^\circ$

Lemma

$\mu P = \mu$ and P is time-reversible

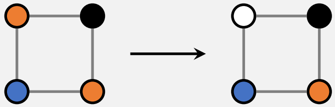
Proof: we have $\mu(x)P(x,z) =$

$$\sum_y \mu(x)N(x,y)N^\circ(y,z) =$$

$$\sum_y \underbrace{\frac{\mu(x)N(x,y)\mu(z)N(z,y)}{\mu^\circ(y)}}_{\text{symmetric in } x,z}$$

symmetric in x, z

Example: Glauber dynamics



- N: erase u.r. vertex
- N°: recolor with prob $\propto$

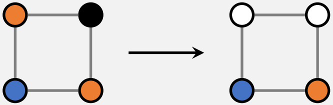
$$\mu(\text{result})N(\text{result}, \text{partial})$$

cancels out

- P: pick u.r. valid color for u.r. vert

- Note: different from Metropolis.

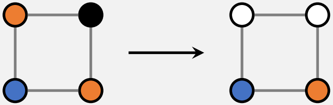
Example: block dynamics



- N: erase k u.r. verts
- P: recolor k u.r. verts

u.a.r. from valid colorings

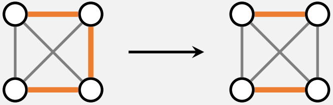
Example: block dynamics



- N: erase k u.r. verts
- P: recolor k u.r. verts

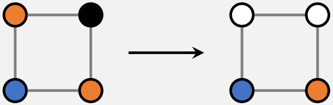
u.a.r. from valid colorings

Example: spanning trees (I)



- N: drop one edge u.a.r.
- P: then add edge u.a.r. from cut

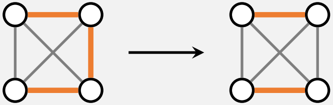
Example: block dynamics



- N: erase k u.r. verts
- P: recolor k u.r. verts

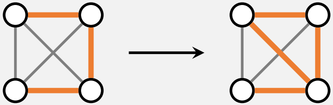
u.a.r. from valid colorings

Example: spanning trees (I)



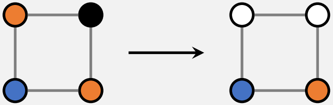
- N: drop one edge u.a.r.
- P: then add edge u.a.r. from cut

Example: spanning trees (II)



- N: add one edge u.a.r.
- P: drop edge u.a.r. from cycle

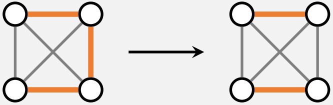
Example: block dynamics



- N: erase k u.r. verts
- P: recolor k u.r. verts

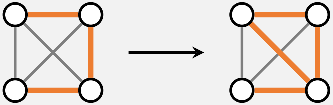
u.a.r. from valid colorings

Example: spanning trees (I)



- N: drop one edge u.a.r.
- P: then add edge u.a.r. from cut

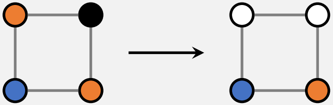
Example: spanning trees (II)



- N: add one edge u.a.r.
- P: drop edge u.a.r. from cycle

- Trivial example: let $\Omega' = \{\emptyset\}$ and N map everything to $\emptyset$.

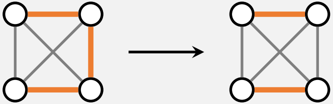
Example: block dynamics



- N: erase k u.r. verts
- P: recolor k u.r. verts

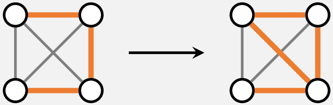
u.a.r. from valid colorings

Example: spanning trees (I)



- N: drop one edge u.a.r.
- P: then add edge u.a.r. from cut

Example: spanning trees (II)



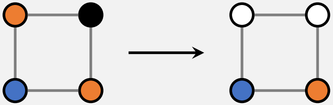
- N: add one edge u.a.r.
- P: drop edge u.a.r. from cycle

- Trivial example: let $\Omega' = \{\emptyset\}$ and N map everything to $\emptyset$.
- We get ideal Markov chain:

mixes in one step

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

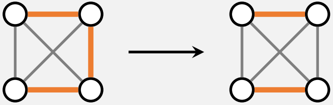
Example: block dynamics



- N: erase k u.r. verts
- P: recolor k u.r. verts

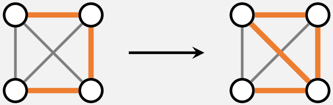
u.a.r. from valid colorings

Example: spanning trees (I)



- N: drop one edge u.a.r.
- P: then add edge u.a.r. from cut

Example: spanning trees (II)



- N: add one edge u.a.r.
- P: drop edge u.a.r. from cycle

- Trivial example: let $\Omega' = \{\emptyset\}$ and N map everything to $\emptyset$.
- We get ideal Markov chain:

mixes in one step

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

- Algorithmically useless!

Algorithmic implementation:

for $t = 0, 1, \ldots$ **do**

  sample $y_t \sim N(x_t, \cdot)$

  for z *with* $N(z, y_t) > 0$ **do**

    $p_z \leftarrow \mu(z) N(z, y_t)$

  sample z with prob $\propto p_z$

  $x_{t+1} \leftarrow$ sample

Algorithmic implementation:

for $t = 0, 1, \ldots$ **do**
- sample $y_t \sim N(x_t, \cdot)$
- **for** z *with* $N(z, y_t) > 0$ **do**
  - $p_z \leftarrow \mu(z) N(z, y_t)$
- sample z with prob $\propto p_z$
- $x_{t+1} \leftarrow$ sample

▷ Want sparse columns for N.

Algorithmic implementation:

for $t = 0, 1, \ldots$ **do**

  sample $y_t \sim N(x_t, \cdot)$

  for z *with* $N(z, y_t) > 0$ **do**

    $p_z \leftarrow \mu(z) N(z, y_t)$

  sample z with prob $\propto p_z$

  $x_{t+1} \leftarrow$ sample

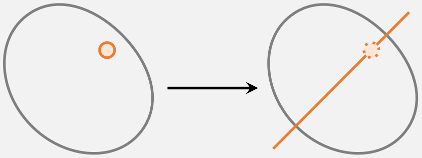
- Want sparse columns for N.
- Ideally we can simulate N, and its columns are not just sparse but efficiently enumerable.

Algorithmic implementation:

for $t = 0, 1, \ldots$ **do**
- sample $y_t \sim N(x_t, \cdot)$
- **for** z *with* $N(z, y_t) > 0$ **do**
  - $p_z \leftarrow \mu(z) N(z, y_t)$
- sample z with prob $\propto p_z$
- $x_{t+1} \leftarrow$ sample

- Want sparse columns for N.
- Ideally we can simulate N, and its columns are not just sparse but efficiently enumerable.

Example: hit-and-run ← infinite space



Algorithmic implementation:

for $t = 0, 1, \ldots$ **do**

  sample $y_t \sim N(x_t, \cdot)$

  for z *with* $N(z, y_t) > 0$ **do**

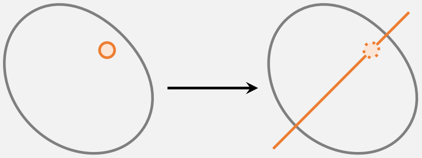
    $p_z \leftarrow \mu(z) N(z, y_t)$

  sample z with prob $\propto p_z$

  $x_{t+1} \leftarrow$ sample

- Want sparse columns for N.
- Ideally we can simulate N, and its columns are not just sparse but efficiently enumerable.

Example: hit-and-run ← infinite space



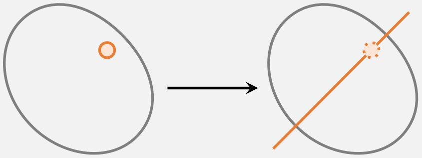
- μ is uniform on subset S of $\mathbb{R}^d$

Algorithmic implementation:

for $t = 0, 1, \ldots$ **do**
  sample $y_t \sim N(x_t, \cdot)$
  for z *with* $N(z, y_t) > 0$ **do**
    $p_z \leftarrow \mu(z) N(z, y_t)$
  sample z with prob $\propto p_z$
  $x_{t+1} \leftarrow$ sample

- Want sparse columns for N.
- Ideally we can simulate N, and its columns are not just sparse but efficiently enumerable.

Example: hit-and-run ← infinite space



- μ is uniform on subset S of $\mathbb{R}^d$
- N: $x \mapsto$ u.r. line ℓ through x

Algorithmic implementation:

for $t = 0, 1, \ldots$ **do**

  sample $y_t \sim N(x_t, \cdot)$

  for z *with* $N(z, y_t) > 0$ **do**

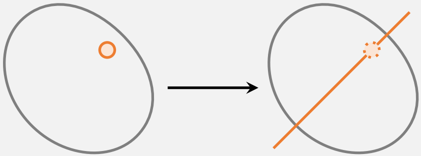
    $p_z \leftarrow \mu(z) N(z, y_t)$

  sample z with prob $\propto p_z$

  $x_{t+1} \leftarrow$ sample

- Want sparse columns for N.
- Ideally we can simulate N, and its columns are not just sparse but efficiently enumerable.

Example: hit-and-run ← infinite space



- μ is uniform on subset S of $\mathbb{R}^d$
- N: $x \mapsto$ u.r. line ℓ through x
- P: then choose u.a.r. from $\ell \cap S$

Algorithmic implementation:

for $t = 0, 1, \ldots$ **do**

  sample $y_t \sim N(x_t, \cdot)$

  for z *with* $N(z, y_t) > 0$ **do**

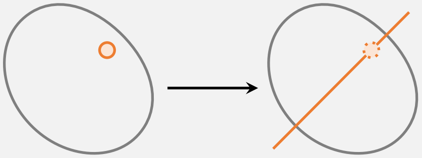
    $p_z \leftarrow \mu(z) N(z, y_t)$

  sample z with prob $\propto p_z$

  $x_{t+1} \leftarrow$ sample

- Want sparse columns for N.
- Ideally we can simulate N, and its columns are not just sparse but efficiently enumerable.

Example: hit-and-run ← infinite space



- μ is uniform on subset S of $\mathbb{R}^d$
- N: $x \mapsto$ u.r. line ℓ through x
- P: then choose u.a.r. from $\ell \cap S$

Example: restricted Gaussian

- μ: dist on $\mathbb{R}^d$
- N: $x \mapsto y = x + g$ for $g \sim \mathcal{N}(0, cI)$
- P: then sample z w.p. $\propto$

restricted Gaussian $\longrightarrow \mu(z) e^{-\|z-y\|^2/2c}$

Summary

Design P time-reversible w.r.t. μ:

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

Summary

Design P time-reversible w.r.t. μ:

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

1. Metropolis filter
   - Have some initial P
   - Modify it to

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

Summary

Design P time-reversible w.r.t. μ:

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

1. Metropolis filter
   - Have some initial P
   - Modify it to

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

2. Combination with time-reversal
   - Have some Markov kernel N
   - Form NN°

Summary

Design P time-reversible w.r.t. μ:

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

1. Metropolis filter
   - Have some initial P
   - Modify it to

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

2. Combination with time-reversal
   - Have some Markov kernel N
   - Form NN°

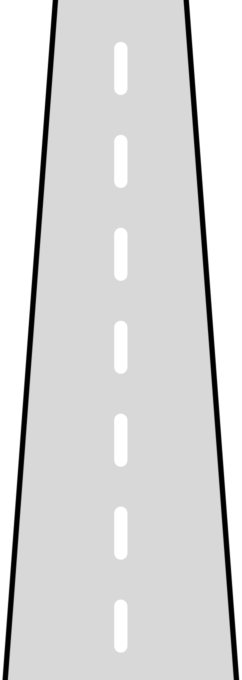
Question: do these guarantee irreducible/aperiodic?

Designing Markov Chains

- Markov kernels
- Combination with time-reversal

Mixing via Transport

- Wasserstein distance
- Path coupling

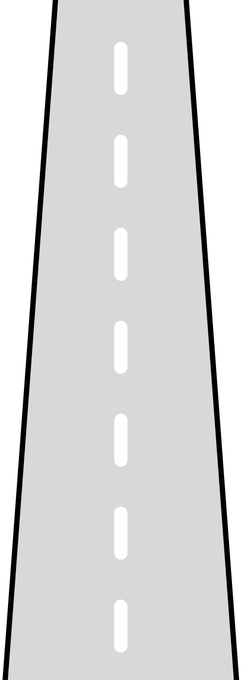


Designing Markov Chains

- Markov kernels
- Combination with time-reversal

Mixing via Transport

- Wasserstein distance
- Path coupling

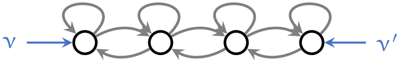


Transport distance

- Prevalent strategy for analyzing mixing time: contraction

Transport distance

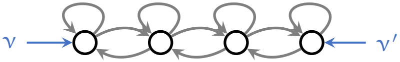
- Prevalent strategy for analyzing mixing time: contraction
- d_{TV} is too crude; doesn't contract every step



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

Transport distance

- Prevalent strategy for analyzing mixing time: contraction
- d_{TV} is too crude; doesn't contract every step



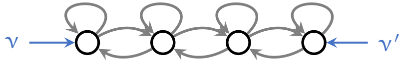
$$d_{TV}(\nu P, \nu' P) = d_{TV}(\nu, \nu')$$

- Fix: use a proxy for d_{TV}
  - Transport/Wasserstein/earth-mover distance ← today
  - f-divergences, variance, entropy

functional analysis, later

Transport distance

- Prevalent strategy for analyzing mixing time: contraction
- d_{TV} is too crude; doesn't contract every step

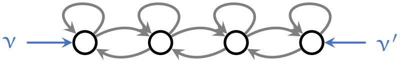


$$d_{TV}(\nu P, \nu' P) = d_{TV}(\nu, \nu')$$

- Fix: use a proxy for d_{TV}
  - Transport/Wasserstein/earth-mover distance ← today
  - f-divergences, variance, entropy
    ↑
    functional analysis, later
- Suppose Ω is equipped with metric $d : \Omega \times \Omega \to \mathbb{R}_{\geqslant 0}$.

Transport distance

- Prevalent strategy for analyzing mixing time: contraction
- d_{TV} is too crude; doesn't contract every step



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

- Fix: use a proxy for d_{TV}
  - Transport/Wasserstein/earth-mover distance ← today
  - f-divergences, variance, entropy

functional analysis, later

- Suppose Ω is equipped with metric $d : \Omega \times \Omega \to \mathbb{R}_{\geqslant 0}$.

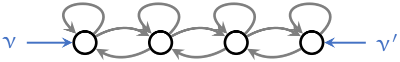
Wasserstein distance

We define the Wasserstein distance w.r.t. d as $\mathcal{W}(\mu, \nu) =$

$$\min\left\{\mathbb{E}_{(X,Y)\sim\pi}[d(X,Y)] \;\middle|\; \pi \text{ coupling}\right\}$$

Transport distance

- Prevalent strategy for analyzing mixing time: contraction
- d_{TV} is too crude; doesn't contract every step



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

- Fix: use a proxy for d_{TV}
  - Transport/Wasserstein/earth-mover distance ← today
  - f-divergences, variance, entropy

    ↑ functional analysis, later
- Suppose Ω is equipped with metric $d : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$.

Wasserstein distance

We define the Wasserstein distance w.r.t. d as $\mathcal{W}(\mu, \nu) =$

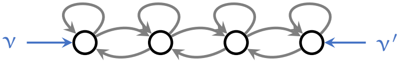
$$\min\{\mathbb{E}_{(X,Y)\sim\pi}[d(X, Y)] \mid \pi \text{ coupling}\}$$

Example: total variation

If we use $d(x, y) = \mathbb{1}[x \neq y]$: $\mathcal{W} = d_{TV}$

Transport distance

- Prevalent strategy for analyzing mixing time: contraction
- d_{TV} is too crude; doesn't contract every step



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

- Fix: use a proxy for d_{TV}
  - Transport/Wasserstein/earth-mover distance ⟵ today
  - f-divergences, variance, entropy

    ↑ functional analysis, later
- Suppose Ω is equipped with metric $d : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$.

Wasserstein distance

We define the Wasserstein distance w.r.t. d as $\mathcal{W}(\mu, \nu) =$

$$\min\left\{\mathbb{E}_{(X,Y)\sim\pi}[d(X, Y)] \;\middle|\; \pi \text{ coupling}\right\}$$

Example: total variation

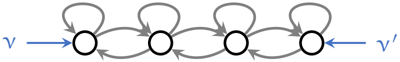
If we use $d(x, y) = \mathbb{1}[x \neq y]$: $\mathcal{W} = d_{\mathrm{TV}}$

Example: Hamming

$$\Omega = [q]^n \qquad d(x, y) = |\{i \mid x_i \neq y_i\}|$$

Transport distance

- Prevalent strategy for analyzing mixing time: contraction
- d_{TV} is too crude; doesn't contract every step



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

- Fix: use a proxy for d_{TV}
  - Transport/Wasserstein/earth-mover distance ⟵ today
  - f-divergences, variance, entropy

  ↑ functional analysis, later

- Suppose Ω is equipped with metric $d : \Omega \times \Omega \to \mathbb{R}_{\geq 0}$.

Wasserstein distance

We define the Wasserstein distance w.r.t. d as $\mathcal{W}(\mu, \nu) =$

$$\min\{\mathbb{E}_{(X,Y)\sim\pi}[d(X, Y)] \mid \pi \text{ coupling}\}$$

Example: total variation

If we use $d(x, y) = \mathbb{1}[x \neq y]$: $\mathcal{W} = d_{TV}$

Example: Hamming

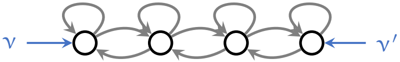
$$\Omega = [q]^n \quad d(x, y) = |\{i \mid x_i \neq y_i\}|$$

$$\mu = \text{unif on } \{(\bullet, \bullet, \bullet), (\bullet, \bullet, \bullet)\}$$

$$\nu = \text{unif on } \{(\bullet, \bullet, \bullet), (\bullet, \bullet, \bullet), (\bullet, \bullet, \bullet)\}$$

Transport distance

- Prevalent strategy for analyzing mixing time: contraction
- d_{TV} is too crude; doesn't contract every step



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

- Fix: use a proxy for d_{TV}
  - Transport/Wasserstein/earth-mover distance ← today
  - f-divergences, variance, entropy

    ↑ functional analysis, later
- Suppose Ω is equipped with metric $d : \Omega \times \Omega \to \mathbb{R}_{\geqslant 0}$.

Wasserstein distance

We define the Wasserstein distance w.r.t. d as $\mathcal{W}(\mu, \nu) =$

$$\min\left\{\mathbb{E}_{(X,Y)\sim\pi}[d(X, Y)] \;\middle|\; \pi \text{ coupling}\right\}$$

Example: total variation

If we use $d(x, y) = \mathbb{1}[x \neq y]$: $\mathcal{W} = d_{TV}$

Example: Hamming

$$\Omega = [q]^n \qquad d(x, y) = |\{i \mid x_i \neq y_i\}|$$

$\mu =$ unif on $\{({\color{orange}\bullet}, {\color{orange}\bullet}, {\color{orange}\bullet}), (\bullet, {\color{orange}\bullet}, {\color{blue}\bullet})\}$

$\nu =$ unif on $\{({\color{orange}\bullet}, {\color{orange}\bullet}, {\color{orange}\bullet}), ({\color{blue}\bullet}, {\color{blue}\bullet}, {\color{blue}\bullet}), (\bullet, \bullet, \bullet)\}$

$$\mathcal{W}(\mu, \nu) = \frac{1}{3} \cdot 0 + \frac{1}{6} \cdot 3 + \frac{1}{2} \cdot 2 = 1.5$$

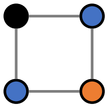
Coloring

- Input: graph G and $q \in \mathbb{N}$

Coloring

- Input: graph G and $q \in \mathbb{N}$
- Goal: sample proper colorings

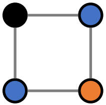
adjacent verts colored differently



Coloring

- Input: graph G and $q \in \mathbb{N}$
- Goal: sample proper colorings

adjacent verts colored differently

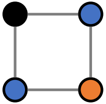


- NP-hard to even find one!

Coloring

- Input: graph G and $q \in \mathbb{N}$
- Goal: sample proper colorings

adjacent verts colored differently



- NP-hard to even find one!
- Easy when $q \geqslant \Delta + 1$

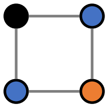
maximum degree

for *each vertex* v **do**
 pick a color from
 $[q] - \{$neighbors' colors$\}$

Coloring

- Input: graph G and $q \in \mathbb{N}$
- Goal: sample proper colorings

adjacent verts colored differently



- NP-hard to even find one!
- Easy when $q \geqslant \Delta + 1$

maximum degree

for *each vertex* v **do**
 pick a color from
 $[q] - \{$neighbors' colors$\}$

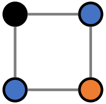
- Open: approx sample/count when

$$q \geqslant \Delta + 1$$

Coloring

- Input: graph G and $q \in \mathbb{N}$
- Goal: sample proper colorings

adjacent verts colored differently



- NP-hard to even find one!
- Easy when $q \geqslant \Delta + 1$

maximum degree

for *each vertex* v **do**
  pick a color from
  $[q] - \{$neighbors' colors$\}$

- Open: approx sample/count when

$$q \geqslant \Delta + 1$$

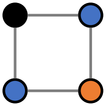
- Open: Metropolis/Glauber when:

$$q \geqslant \Delta + 2$$

Coloring

- Input: graph G and $q \in \mathbb{N}$
- Goal: sample proper colorings

adjacent verts colored differently



- NP-hard to even find one!
- Easy when $q \geqslant \Delta + 1$

maximum degree

for *each vertex* v **do**
  pick a color from
  $[q] - \{$neighbors' colors$\}$

- Open: approx sample/count when

$$q \geqslant \Delta + 1$$

- Open: Metropolis/Glauber when:

$$q \geqslant \Delta + 2$$

- Best-known [Chen-Delcourt-Moitra-Perarnau-Postle'18]:

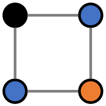
$$q \geqslant \left(\tfrac{11}{6} - \epsilon\right)\Delta$$

some tiny constant

Coloring

- Input: graph G and $q \in \mathbb{N}$
- Goal: sample proper colorings

adjacent verts colored differently



- NP-hard to even find one!
- Easy when $q \geqslant \Delta + 1$

maximum degree

for *each vertex* v **do**
  pick a color from
  $[q] - \{$neighbors' colors$\}$

- Open: approx sample/count when

$$q \geqslant \Delta + 1$$

- Open: Metropolis/Glauber when:

$$q \geqslant \Delta + 2$$

- Best-known [Chen-Delcourt-Moitra-Perarnau-Postle'18]:

$$q \geqslant (\tfrac{11}{6} - \epsilon)\Delta$$

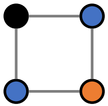
some tiny constant

- We will show $q \geqslant 4\Delta + 1$ works.

Coloring

- Input: graph G and $q \in \mathbb{N}$
- Goal: sample proper colorings

adjacent verts colored differently



- NP-hard to even find one!
- Easy when $q \geqslant \Delta + 1$

maximum degree

for *each vertex* v **do**
  pick a color from
  $[q] - \{$neighbors' colors$\}$

- Open: approx sample/count when
$$q \geqslant \Delta + 1$$
- Open: Metropolis/Glauber when:
$$q \geqslant \Delta + 2$$
- Best-known [Chen-Delcourt-Moitra-Perarnau-Postle'18]:
$$q \geqslant (\tfrac{11}{6} - \epsilon)\Delta$$

some tiny constant

- We will show $q \geqslant 4\Delta + 1$ works.
- Then we will improve to $q \geqslant 2\Delta + 1$.

- Strategy: show $\mathcal{W}$ contracts.

- Strategy: show $\mathcal{W}$ contracts.

Lemma

When $q \geqslant 4\Delta + 1$, for Metropolis P:

$$\mathcal{W}(\nu P, \nu' P) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathcal{W}(\nu, \nu')$$

- Strategy: show $\mathcal{W}$ contracts.

Lemma

When $q \geqslant 4\Delta + 1$, for Metropolis P:

$$\mathcal{W}(\nu P, \nu' P) \leqslant \left(1 - \frac{1}{\text{poly}(n)}\right) \mathcal{W}(\nu, \nu')$$

- This is enough because

$$\underbrace{\mathcal{W}(\nu P^t, \mu)}_{\text{upper bounds } d_{TV}} \leqslant \left(1 - \frac{1}{\text{poly}}\right)^t \underbrace{\mathcal{W}(\nu, \mu)}_{\text{at most } n}$$

- Strategy: show $\mathcal{W}$ contracts.

Lemma

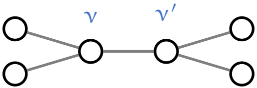
When $q \geqslant 4\Delta + 1$, for Metropolis P:

$$\mathcal{W}(\nu P, \nu' P) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathcal{W}(\nu, \nu')$$

- This is enough because

$$\underbrace{\mathcal{W}(\nu P^t, \mu)}_{\text{upper bounds } d_{TV}} \leqslant \left(1 - \frac{1}{\mathrm{poly}}\right)^t \underbrace{\mathcal{W}(\nu, \mu)}_{\text{at most } n}$$

- Note: unlike d_{TV}, weak contraction is NOT guaranteed.



- Strategy: show $\mathcal{W}$ contracts.

Lemma

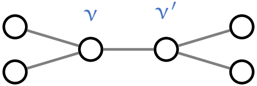
When $q \geqslant 4\Delta + 1$, for Metropolis P:

$$\mathcal{W}(\nu P, \nu' P) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathcal{W}(\nu, \nu')$$

- This is enough because

$$\underbrace{\mathcal{W}(\nu P^t, \mu)}_{\text{upper bounds } d_{TV}} \leqslant \left(1 - \frac{1}{\mathrm{poly}}\right)^t \underbrace{\mathcal{W}(\nu, \mu)}_{\text{at most } n}$$

- Note: unlike d_{TV}, weak contraction is NOT guaranteed.



- Use coupling:
  - Sample X_0, X_0' from optimal coupling of ν, ν'.
  - Evolve to get X_1, X_1' while minimizing $\mathbb{E}[d(X_1, X_1')]$

- Strategy: show $\mathcal{W}$ contracts.

Lemma

When $q \geqslant 4\Delta + 1$, for Metropolis P:

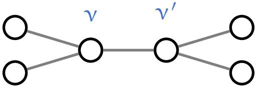
$$\mathcal{W}(\nu P, \nu' P) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathcal{W}(\nu, \nu')$$

- This is enough because

$$\mathcal{W}(\nu P^t, \mu) \leqslant \left(1 - \frac{1}{\mathrm{poly}}\right)^t \mathcal{W}(\nu, \mu)$$

upper bounds d_{TV} at most n

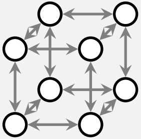
- Note: unlike d_{TV}, weak contraction is NOT guaranteed.



- Use coupling:
  - Sample X_0, X_0' from optimal coupling of ν, ν'.
  - Evolve to get X_1, X_1' while minimizing $\mathbb{E}[d(X_1, X_1')]$

Warmup example: hypercube

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



- Strategy: show $\mathcal{W}$ contracts.

Lemma

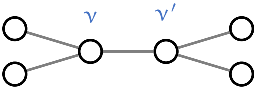
When $q \geqslant 4\Delta + 1$, for Metropolis P:

$$\mathcal{W}(\nu P, \nu' P) \leqslant \left(1 - \frac{1}{\text{poly}(n)}\right) \mathcal{W}(\nu, \nu')$$

- This is enough because

$$\underbrace{\mathcal{W}(\nu P^t, \mu)}_{\text{upper bounds } d_{TV}} \leqslant \left(1 - \frac{1}{\text{poly}}\right)^t \underbrace{\mathcal{W}(\nu, \mu)}_{\text{at most } n}$$

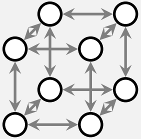
- Note: unlike d_{TV}, weak contraction is NOT guaranteed.



- Use coupling:
  - Sample X_0, X_0' from optimal coupling of ν, ν'.
  - Evolve to get X_1, X_1' while minimizing $\mathbb{E}[d(X_1, X_1')]$

Warmup example: hypercube

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



- Pick same i and same Ber$(\frac{1}{2})$

- Strategy: show $\mathcal{W}$ contracts.

Lemma

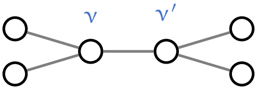
When $q \geqslant 4\Delta + 1$, for Metropolis P:

$$\mathcal{W}(\nu P, \nu' P) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathcal{W}(\nu, \nu')$$

- This is enough because

$$\underbrace{\mathcal{W}(\nu P^t, \mu)}_{\text{upper bounds } d_{TV}} \leqslant \left(1 - \frac{1}{\mathrm{poly}}\right)^t \underbrace{\mathcal{W}(\nu, \mu)}_{\text{at most } n}$$

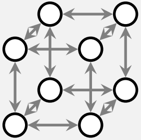
- Note: unlike d_{TV}, weak contraction is NOT guaranteed.



- Use coupling:
  - Sample X_0, X_0' from optimal coupling of ν, ν'.
  - Evolve to get X_1, X_1' while minimizing $\mathbb{E}[d(X_1, X_1')]$

Warmup example: hypercube

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



- Pick same i and same $\mathrm{Ber}(\frac{1}{2})$
- If $d(X_0, X_0') = k$, then

$$\mathbb{E}[d(X_1, X_1') \mid X_0, X_0'] = k - \frac{k}{n}$$

- Strategy: show $\mathcal{W}$ contracts.

Lemma

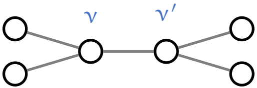
When $q \geqslant 4\Delta + 1$, for Metropolis P:

$$\mathcal{W}(\nu P, \nu' P) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathcal{W}(\nu, \nu')$$

- This is enough because

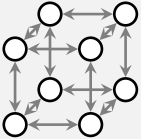
$$\underbrace{\mathcal{W}(\nu P^t, \mu)}_{\text{upper bounds } d_{TV}} \leqslant \left(1 - \frac{1}{\mathrm{poly}}\right)^t \underbrace{\mathcal{W}(\nu, \mu)}_{\text{at most } n}$$

- Note: unlike d_{TV}, weak contraction is NOT guaranteed.



- Use coupling:
  - Sample X_0, X_0' from optimal coupling of ν, ν'.
  - Evolve to get X_1, X_1' while minimizing $\mathbb{E}[d(X_1, X_1')]$

Warmup example: hypercube



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

- Pick same i and same $\mathrm{Ber}(\frac{1}{2})$
- If $d(X_0, X_0') = k$, then
$$\mathbb{E}[d(X_1, X_1') \mid X_0, X_0'] = k - \frac{k}{n}$$
- $\mathcal{W}(\nu P, \nu P') \leqslant (1 - 1/n)\, \mathcal{W}(\nu, \nu')$

- Strategy: show $\mathcal{W}$ contracts.

Lemma

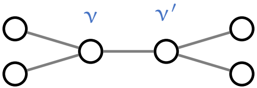
When $q \geqslant 4\Delta + 1$, for Metropolis P:

$$\mathcal{W}(\nu P, \nu' P) \leqslant \left(1 - \frac{1}{\text{poly}(n)}\right) \mathcal{W}(\nu, \nu')$$

- This is enough because

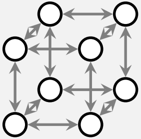
$$\underbrace{\mathcal{W}(\nu P^t, \mu)}_{\text{upper bounds } d_{TV}} \leqslant \left(1 - \frac{1}{\text{poly}}\right)^t \underbrace{\mathcal{W}(\nu, \mu)}_{\text{at most } n}$$

- Note: unlike d_{TV}, weak contraction is NOT guaranteed.



- Use coupling:
  - Sample X_0, X_0' from optimal coupling of ν, ν'.
  - Evolve to get X_1, X_1' while minimizing $\mathbb{E}[d(X_1, X_1')]$

Warmup example: hypercube



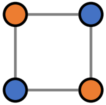
- $\Omega = \{0, 1\}^n$
- Pick u.r. $i \in [n]$
- Replace coord i with Ber$(\frac{1}{2})$
- Pick same i and same Ber$(\frac{1}{2})$
- If $d(X_0, X_0') = k$, then

$$\mathbb{E}[d(X_1, X_1') \mid X_0, X_0'] = k - \frac{k}{n}$$

- $\mathcal{W}(\nu P, \nu P') \leqslant (1 - 1/n)\, \mathcal{W}(\nu, \nu')$
- $t_{mix}(\epsilon) \leqslant n \log n + n \log(1/\epsilon)$

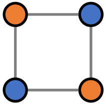
Take the Metropolis chain for colorings:

- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid



Take the Metropolis chain for colorings:

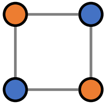
- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid



Coupling:

Take the Metropolis chain for colorings:

- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid

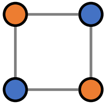


Coupling:

- Pick same v and same c

Take the Metropolis chain for colorings:

- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid

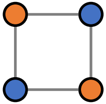


Coupling:

- Pick same v and same c
- If $d(X_0, X_0') = k$, then $d(X_1, X_1')$ is:
    - $k - 1$ (lucky)
    - $k + 1$ (unlucky)
    - k (neutral)

Take the Metropolis chain for colorings:

- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid



Coupling:

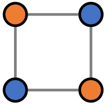
- Pick same v and same c
- If $d(X_0, X_0') = k$, then $d(X_1, X_1')$ is:
  - $k - 1$ (lucky)
  - $k + 1$ (unlucky)
  - k (neutral)
- $\mathbb{P}[\text{lucky}] \geqslant (k/n) \cdot (q - 2\Delta)/q$

pick differing v

c available to both

Take the Metropolis chain for colorings:

- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid



Coupling:

- Pick same v and same c
- If $d(X_0, X_0') = k$, then $d(X_1, X_1')$ is:
  - $k-1$ (lucky)
  - $k+1$ (unlucky)
  - k (neutral)
- $\mathbb{P}[\text{lucky}] \geqslant (k/n) \cdot (q-2\Delta)/q$

pick differing v

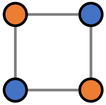
c available to both

- $\mathbb{P}[\text{unlucky}] \leqslant 2k\Delta/qn$

c color of differing neighbor in X_0 or X_0'

Take the Metropolis chain for colorings:

- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid



Coupling:

- Pick same v and same c
- If $d(X_0, X_0') = k$, then $d(X_1, X_1')$ is:
  - $k-1$ (lucky)
  - $k+1$ (unlucky)
  - k (neutral)
- $\mathbb{P}[\text{lucky}] \geqslant (k/n) \cdot (q-2\Delta)/q$

pick differing v c available to both

- $\mathbb{P}[\text{unlucky}] \leqslant 2k\Delta/qn$

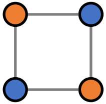
c color of differing neighbor in X_0 or X_0'

- We get $\mathbb{E}[d(X_1, X_1') \mid X_0, X_0'] \leqslant$

$$k - \frac{k(q-2\Delta)}{qn} + \frac{2k\Delta}{qn} = k \cdot \left(1 - \frac{q-4\Delta}{qn}\right)$$

Take the Metropolis chain for colorings:

- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid



Coupling:

- Pick same v and same c
- If $d(X_0, X_0') = k$, then $d(X_1, X_1')$ is:
  - $k - 1$ (lucky)
  - $k + 1$ (unlucky)
  - k (neutral)
- $\mathbb{P}[\text{lucky}] \geqslant (k/n) \cdot (q - 2\Delta)/q$

pick differing v c available to both

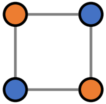
- $\mathbb{P}[\text{unlucky}] \leqslant 2k\Delta/qn$

c color of differing neighbor in X_0 or X_0'

- We get $\mathbb{E}[d(X_1, X_1') \mid X_0, X_0'] \leqslant$
$k - \frac{k(q-2\Delta)}{qn} + \frac{2k\Delta}{qn} = k \cdot \left(1 - \frac{q-4\Delta}{qn}\right)$
- As long as $q \geqslant 4\Delta + 1$, we have contraction. ☺

Take the Metropolis chain for colorings:

- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid



Coupling:

- Pick same v and same c
- If $d(X_0, X_0') = k$, then $d(X_1, X_1')$ is:
  - $k - 1$ (lucky)
  - $k + 1$ (unlucky)
  - k (neutral)
- $\mathbb{P}[\text{lucky}] \geqslant (k/n) \cdot (q - 2\Delta)/q$

pick differing v c available to both

- $\mathbb{P}[\text{unlucky}] \leqslant 2k\Delta/qn$

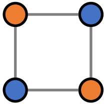
c color of differing neighbor in X_0 or X_0'

- We get $\mathbb{E}[d(X_1, X_1') \mid X_0, X_0'] \leqslant$
  $k - \frac{k(q-2\Delta)}{qn} + \frac{2k\Delta}{qn} = k \cdot \left(1 - \frac{q-4\Delta}{qn}\right)$
- As long as $q \geqslant 4\Delta + 1$, we have contraction.
- We get

$$t_{mix}(\epsilon) = O\left(\frac{q}{q-4\Delta} \cdot n \log(n/\epsilon)\right)$$

Take the Metropolis chain for colorings:

- Pick u.r. vertex v
- Pick u.r. color c
- Color v with c if valid



Coupling:

- Pick same v and same c
- If $d(X_0, X_0') = k$, then $d(X_1, X_1')$ is:
  - $k - 1$ (lucky)
  - $k + 1$ (unlucky)
  - k (neutral)
- $\mathbb{P}[\text{lucky}] \geqslant (k/n) \cdot (q - 2\Delta)/q$

  pick differing v c available to both

- $\mathbb{P}[\text{unlucky}] \leqslant 2k\Delta/qn$

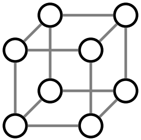
  c color of differing neighbor in X_0 or X_0'
- We get $\mathbb{E}[d(X_1, X_1') \mid X_0, X_0'] \leqslant$
  $k - \frac{k(q-2\Delta)}{qn} + \frac{2k\Delta}{qn} = k \cdot \left(1 - \frac{q-4\Delta}{qn}\right)$
- As long as $q \geqslant 4\Delta + 1$, we have contraction. ☺
- We get

$$t_{\text{mix}}(\epsilon) = O\left(\frac{q}{q-4\Delta} \cdot n \log(n/\epsilon)\right)$$

- Exercise: analyze Glauber this way.

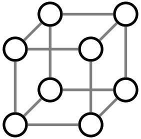
- Hamming distance is special.

- Hamming distance is special.
- There is a sparse graph s.t. $d(x, y)$ is shortest path from x to y.



$x \sim y$ when $x_i \neq y_i$ for one i

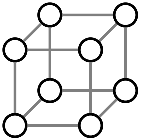
- Hamming distance is special.
- There is a sparse graph s.t. $d(x, y)$ is shortest path from x to y.



$x \sim y$ when $x_i \neq y_i$ for one i

- In general, if d is shortest path metric derived from a (possibly weighted) graph, we can use path coupling [Bubley-Dyer].

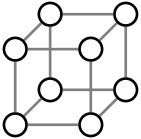
- Hamming distance is special.
- There is a sparse graph s.t. $d(x,y)$ is shortest path from x to y.



$x \sim y$ when $x_i \neq y_i$ for one i

- In general, if d is shortest path metric derived from a (possibly weighted) graph, we can use path coupling [Bubley-Dyer].
- Idea: only couple starting states X_0, X_0' that are adjacent.

- Hamming distance is special.
- There is a sparse graph s.t. $d(x, y)$ is shortest path from x to y.



$x \sim y$ when $x_i \neq y_i$ for one i

- In general, if d is shortest path metric derived from a (possibly weighted) graph, we can use path coupling [Bubley-Dyer].
- Idea: only couple starting states X_0, X_0' that are adjacent.

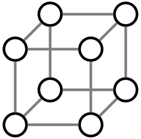
Path coupling lemma

Suppose for all adjacent $X_0 \sim X_0'$ we can couple X_1, X_1' s.t.

$$\mathbb{E}[d(X_1, X_1')] \leqslant (1-c)d(X_0, X_0').$$

Then $\mathcal{W}(\nu P, \nu' P) \leqslant (1-c)\,\mathcal{W}(\nu, \nu')$.

- Hamming distance is special.
- There is a sparse graph s.t. $d(x, y)$ is shortest path from x to y.



  $x \sim y$ when $x_i \neq y_i$ for one i
- In general, if d is shortest path metric derived from a (possibly weighted) graph, we can use path coupling [Bubley-Dyer].
- Idea: only couple starting states X_0, X_0' that are adjacent.

Path coupling lemma

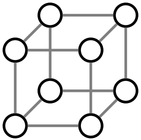
Suppose for all adjacent $X_0 \sim X_0'$ we can couple X_1, X_1' s.t.

$$\mathbb{E}[d(X_1, X_1')] \leqslant (1-c)d(X_0, X_0').$$

Then $\mathcal{W}(\nu P, \nu' P) \leqslant (1-c)\,\mathcal{W}(\nu, \nu')$.

Proof:

- Hamming distance is special.
- There is a sparse graph s.t. $d(x, y)$ is shortest path from x to y.



$x \sim y$ when $x_i \neq y_i$ for one i

- In general, if d is shortest path metric derived from a (possibly weighted) graph, we can use path coupling [Bubley-Dyer].
- Idea: only couple starting states X_0, X_0' that are adjacent.

Path coupling lemma

Suppose for all adjacent $X_0 \sim X_0'$ we can couple X_1, X_1' s.t.

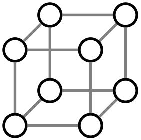
$$\mathbb{E}[d(X_1, X_1')] \leqslant (1-c) d(X_0, X_0').$$

Then $\mathcal{W}(\nu P, \nu' P) \leqslant (1-c)\, \mathcal{W}(\nu, \nu')$.

Proof:

- Take arbitrary X_0, X_0'.

- Hamming distance is special.
- There is a sparse graph s.t. $d(x, y)$ is shortest path from x to y.



$x \sim y$ when $x_i \neq y_i$ for one i

- In general, if d is shortest path metric derived from a (possibly weighted) graph, we can use path coupling [Bubley-Dyer].
- Idea: only couple starting states X_0, X_0' that are adjacent.

Path coupling lemma

Suppose for all adjacent $X_0 \sim X_0'$ we can couple X_1, X_1' s.t.

$$\mathbb{E}[d(X_1, X_1')] \leqslant (1-c)d(X_0, X_0').$$

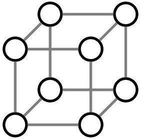
Then $\mathcal{W}(\nu P, \nu' P) \leqslant (1-c)\,\mathcal{W}(\nu, \nu')$.

Proof:

- Take arbitrary X_0, X_0'.
- Let shortest path be

$$X_0 = \nu_0 \to \nu_1 \to \cdots \to \nu_k = X_0'$$

- Hamming distance is special.
- There is a sparse graph s.t. $d(x, y)$ is shortest path from x to y.



$x \sim y$ when $x_i \neq y_i$ for one i

- In general, if d is shortest path metric derived from a (possibly weighted) graph, we can use path coupling [Bubley-Dyer].
- Idea: only couple starting states X_0, X_0' that are adjacent.

Path coupling lemma

Suppose for all adjacent $X_0 \sim X_0'$ we can couple X_1, X_1' s.t.

$$\mathbb{E}[d(X_1, X_1')] \leqslant (1-c)d(X_0, X_0').$$

Then $\mathcal{W}(\nu P, \nu' P) \leqslant (1-c)\,\mathcal{W}(\nu, \nu')$.

Proof:

- Take arbitrary X_0, X_0'.
- Let shortest path be

$$X_0 = \nu_0 \to \nu_1 \to \cdots \to \nu_k = X_0'$$

- By triangle ineq $\mathcal{W}(\mathbb{1}_{X_0} P, \mathbb{1}_{X_0'} P) \leqslant$

$$\textstyle\sum_i \mathcal{W}(\mathbb{1}_{\nu_i} P, \mathbb{1}_{\nu_{i+1}} P) \leqslant$$
$$(1-c)\textstyle\sum_i d(\nu_i, \nu_{i+1}) = (1-c)d(X_0, X_0')$$

Triangle inequality holds because couplings can be stitched together!

$$\nu_0 \xrightarrow{\pi_{0,1}} \nu_1 \xrightarrow{\pi_{1,2}} \cdots \xrightarrow{\pi_{k-1,k}} \nu_k$$

Exercise: there is joint dist with marginals $\pi_{i,i+1}$!