CS 263: Counting and Sampling

Nima Anari

Stanford University

slides for

Spectral Analysis

Review

- Influence: X, X' differing in coord j:

  $d_{\mathsf{TV}}\big(\mathsf{dist}(X_i \mid X_{-i}), \mathsf{dist}(X'_i \mid X'_{-i})\big)$

Review

- Influence: X, X' differing in coord j:

  $d_{\mathsf{TV}}\big(\mathsf{dist}(X_i \mid X_{-i}), \mathsf{dist}(X'_i \mid X'_{-i})\big)$
- Call maximum value $\mathcal{I}[j \to i]$.

Review

- Influence: X, X' differing in coord j:

$$d_{TV}\big(\mathrm{dist}(X_i \mid X_{-i}), \mathrm{dist}(X'_i \mid X'_{-i})\big)$$

- Call maximum value $\mathcal{I}[j \to i]$.

Dobrushin's condition

If columns of $\mathcal{I}$ sum to $\leqslant 1 - \delta$, then

$$\mathcal{W}(\nu P, \nu' P) \leqslant (1 - \delta/n)\, \mathcal{W}(\nu, \nu')$$

Review

- Influence: X, X' differing in coord j:

$$d_{\mathsf{TV}}\big(\mathsf{dist}(X_i \mid X_{-i}), \mathsf{dist}(X'_i \mid X'_{-i})\big)$$

- Call maximum value $\mathcal{I}[j \to i]$.

Dobrushin's condition

If columns of $\mathcal{I}$ sum to $\leqslant 1 - \delta$, then

$$\mathcal{W}(\nu P, \nu' P) \leqslant (1 - \delta/n)\, \mathcal{W}(\nu, \nu')$$

$$t_{\mathsf{mix}}(\epsilon) = O\big(\tfrac{1}{\delta} n \log(n/\epsilon)\big)$$

Review

- Influence: X, X' differing in coord j:

$$d_{TV}\big(\text{dist}(X_i \mid X_{-i}), \text{dist}(X'_i \mid X'_{-i})\big)$$

- Call maximum value $\mathcal{I}[j \to i]$.

Dobrushin's condition

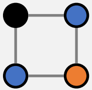
If columns of $\mathcal{I}$ sum to $\leqslant 1 - \delta$, then

$$\mathcal{W}(\nu P, \nu' P) \leqslant (1 - \delta/n)\, \mathcal{W}(\nu, \nu')$$

$$t_{\text{mix}}(\epsilon) = O\big(\tfrac{1}{\delta} n \log(n/\epsilon)\big)$$

Example: coloring

- $\Omega = [q]^n$
- $\mathcal{I} \leqslant \frac{1}{q-\Delta} \cdot \text{adj}$



Review

- Influence: X, X' differing in coord j:
$$d_{TV}\left(\text{dist}(X_i \mid X_{-i}), \text{dist}(X'_i \mid X'_{-i})\right)$$
- Call maximum value $\mathcal{J}[j \to i]$.

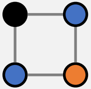
Dobrushin's condition

If columns of $\mathcal{J}$ sum to $\leqslant 1 - \delta$, then
$$\mathcal{W}(\nu P, \nu' P) \leqslant (1 - \delta/n)\, \mathcal{W}(\nu, \nu')$$

$$t_{\mathsf{mix}}(\epsilon) = O\left(\tfrac{1}{\delta} n \log(n/\epsilon)\right)$$

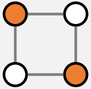
Example: coloring

- $\Omega = [q]^n$
- $\mathcal{J} \leqslant \frac{1}{q-\Delta} \cdot \text{adj}$



Example: hardcore

- $\Omega = \{0, 1\}^n$
- $\mathcal{J} \leqslant \frac{\lambda}{1+\lambda} \cdot \text{adj}$



Review

- Influence: X, X' differing in coord j:

$$d_{TV}\big(\mathrm{dist}(X_i \mid X_{-i}), \mathrm{dist}(X'_i \mid X'_{-i})\big)$$

- Call maximum value $\mathcal{J}[j \to i]$.

Dobrushin's condition

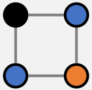
If columns of $\mathcal{J}$ sum to $\leqslant 1 - \delta$, then

$$\mathcal{W}(\nu P, \nu' P) \leqslant (1 - \delta/n)\, \mathcal{W}(\nu, \nu')$$

$$t_{\mathrm{mix}}(\epsilon) = O\big(\tfrac{1}{\delta} n \log(n/\epsilon)\big)$$

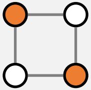
Example: coloring

- $\Omega = [q]^n$
- $\mathcal{J} \leqslant \frac{1}{q - \Delta} \cdot \mathrm{adj}$



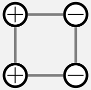
Example: hardcore

- $\Omega = \{0, 1\}^n$
- $\mathcal{J} \leqslant \frac{\lambda}{1+\lambda} \cdot \mathrm{adj}$



Example: Ising

- $\Omega = \{\pm 1\}^n$
- $\mathcal{J}[j \to i] \leqslant |\beta_{ij}|$



Review

- Influence: X, X' differing in coord j:

$$d_{\mathrm{TV}}\big(\mathrm{dist}(X_i \mid X_{-i}), \mathrm{dist}(X'_i \mid X'_{-i})\big)$$

- Call maximum value $\mathcal{J}[j \to i]$.

Dobrushin's condition

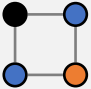
If columns of $\mathcal{J}$ sum to $\leqslant 1 - \delta$, then

$$\mathcal{W}(\nu P, \nu' P) \leqslant (1 - \delta/n)\, \mathcal{W}(\nu, \nu')$$

$$t_{\mathrm{mix}}(\epsilon) = O\big(\tfrac{1}{\delta} n \log(n/\epsilon)\big)$$

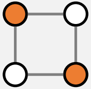
Example: coloring

- $\Omega = [q]^n$
- $\mathcal{J} \leqslant \frac{1}{q-\Delta} \cdot \mathrm{adj}$



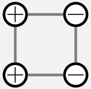
Example: hardcore

- $\Omega = \{0, 1\}^n$
- $\mathcal{J} \leqslant \frac{\lambda}{1+\lambda} \cdot \mathrm{adj}$



Example: Ising

- $\Omega = \{\pm 1\}^n$
- $\mathcal{J}[j \to i] \leqslant |\beta_{ij}|$



- Dobrushin++: if $c\,\mathcal{J} < (1 - \delta)c$

$$t_{\mathrm{mix}}(\epsilon) = O\left(\frac{n}{\delta} \log\left(\frac{n \cdot c_{\max}}{\epsilon \cdot c_{\min}}\right)\right)$$

Review

- Influence: X, X' differing in coord j:
$$d_{TV}\big(\text{dist}(X_i \mid X_{-i}), \text{dist}(X'_i \mid X'_{-i})\big)$$
- Call maximum value $\mathcal{J}[j \to i]$.

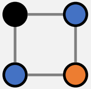
Dobrushin's condition

If columns of $\mathcal{J}$ sum to $\leqslant 1 - \delta$, then
$$\mathcal{W}(\nu P, \nu' P) \leqslant (1 - \delta/n)\, \mathcal{W}(\nu, \nu')$$

$$t_{\text{mix}}(\epsilon) = O\big(\tfrac{1}{\delta} n \log(n/\epsilon)\big)$$

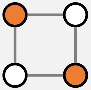
Example: coloring

- $\Omega = [q]^n$
- $\mathcal{J} \leqslant \frac{1}{q - \Delta} \cdot \text{adj}$



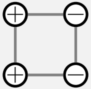
Example: hardcore

- $\Omega = \{0, 1\}^n$
- $\mathcal{J} \leqslant \frac{\lambda}{1+\lambda} \cdot \text{adj}$



Example: Ising

- $\Omega = \{\pm 1\}^n$
- $\mathcal{J}[j \to i] \leqslant |\beta_{ij}|$



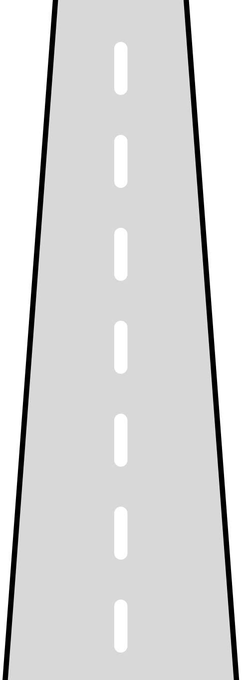
- Dobrushin++: if $c\,\mathcal{J} < (1 - \delta)c$
$$t_{\text{mix}}(\epsilon) = O\left(\frac{n}{\delta} \log\left(\frac{n \cdot c_{\max}}{\epsilon \cdot c_{\min}}\right)\right)$$
- Existence: $\lambda_{\max}(\mathcal{J}) < 1$

Functional Analysis

- Divergences
- Poincaré and modified log-Sobolev
- Data processing
- Spectral analysis

Fourier Analysis

- Abelian walks
- Characters

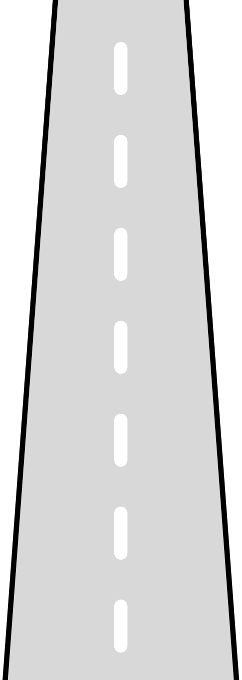


Functional Analysis

- Divergences
- Poincaré and modified log-Sobolev
- Data processing
- Spectral analysis

Fourier Analysis

- Abelian walks
- Characters



Divergences

ϕ-entropy

For function ϕ and $f : \Omega \to \mathbb{R}$ define

$$\mathrm{Ent}_\mu^\phi[f] = \mathbb{E}_\mu[\phi \circ f] - \phi(\mathbb{E}_\mu[f]).$$

Divergences

ϕ-entropy

For function ϕ and $f : \Omega \to \mathbb{R}$ define

$$\mathsf{Ent}_{\mu}^{\phi}[f] = \mathbb{E}_{\mu}[\phi \circ f] - \phi(\mathbb{E}_{\mu}[f]).$$

- When ϕ is convex, ϕ-entropy is $\geqslant 0$ (Jensen's inequality).

Divergences

ϕ-entropy

For function ϕ and $f : \Omega \to \mathbb{R}$ define

$$\mathrm{Ent}_{\mu}^{\phi}[f] = \mathbb{E}_{\mu}[\phi \circ f] - \phi(\mathbb{E}_{\mu}[f]).$$

- When ϕ is convex, ϕ-entropy is $\geqslant 0$ (Jensen's inequality).
- Equal to 0 when f is constant.

Divergences

ϕ-entropy

For function ϕ and $f : \Omega \to \mathbb{R}$ define

$$\mathrm{Ent}_{\mu}^{\phi}[f] = \mathbb{E}_{\mu}[\phi \circ f] - \phi(\mathbb{E}_{\mu}[f]).$$

- When ϕ is convex, ϕ-entropy is $\geqslant 0$ (Jensen's inequality).
- Equal to 0 when f is constant.

usually f in the literature

ϕ-divergence

For measure ν and dist μ define

$$\mathcal{D}_{\phi}(\nu \,\|\, \mu) = \mathrm{Ent}_{\mu}^{\phi}\left[\frac{\nu}{\mu}\right]$$

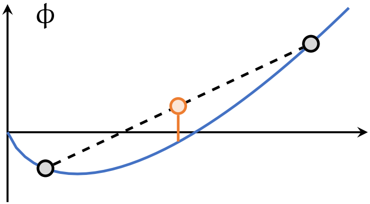
Divergences

ϕ-entropy

For function ϕ and $f : \Omega \to \mathbb{R}$ define

$$\mathrm{Ent}_\mu^\phi[f] = \mathbb{E}_\mu[\phi \circ f] - \phi(\mathbb{E}_\mu[f]).$$

- When ϕ is convex, ϕ-entropy is $\geqslant 0$ (Jensen's inequality).
- Equal to 0 when f is constant.



usually f in the literature

ϕ-divergence

For measure ν and dist μ define

$$\mathcal{D}_\phi(\nu \parallel \mu) = \mathrm{Ent}_\mu^\phi\left[\frac{\nu}{\mu}\right]$$

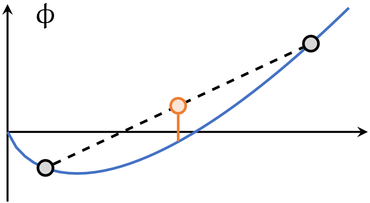
Divergences

ϕ-entropy

For function ϕ and $f : \Omega \to \mathbb{R}$ define

$$\mathrm{Ent}_\mu^\phi[f] = \mathbb{E}_\mu[\phi \circ f] - \phi(\mathbb{E}_\mu[f]).$$

- When ϕ is convex, ϕ-entropy is $\geqslant 0$ (Jensen's inequality).
- Equal to 0 when f is constant.



- How far from $\frac{\nu}{\mu} \equiv$ const?

usually f in the literature

ϕ-divergence

For measure ν and dist μ define

$$\mathcal{D}_\phi(\nu \parallel \mu) = \mathrm{Ent}_\mu^\phi\left[\frac{\nu}{\mu}\right]$$

Divergences

ϕ-entropy

For function ϕ and $f : \Omega \to \mathbb{R}$ define

$$\mathrm{Ent}^{\phi}_{\mu}[f] = \mathbb{E}_{\mu}[\phi \circ f] - \phi(\mathbb{E}_{\mu}[f]).$$

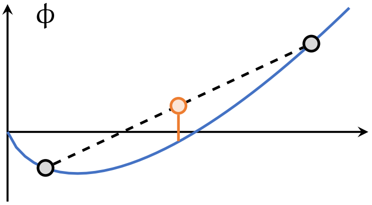
- When ϕ is convex, ϕ-entropy is $\geqslant 0$ (Jensen's inequality).
- Equal to 0 when f is constant.

usually f in the literature

ϕ-divergence

For measure ν and dist μ define

$$\mathcal{D}_{\phi}(\nu \parallel \mu) = \mathrm{Ent}^{\phi}_{\mu}\left[\frac{\nu}{\mu}\right]$$



- How far from $\frac{\nu}{\mu} \equiv \text{const}$?

Example: total variation

If $\phi(x) = \frac{1}{2}|x-1|$, then

$$\mathcal{D}_{\phi}(\nu \parallel \mu) = d_{TV}(\nu, \mu)$$

Divergences

ϕ-entropy

For function ϕ and $f : \Omega \to \mathbb{R}$ define

$$\mathrm{Ent}_{\mu}^{\phi}[f] = \mathbb{E}_{\mu}[\phi \circ f] - \phi(\mathbb{E}_{\mu}[f]).$$

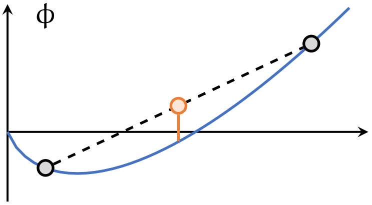
- When ϕ is convex, ϕ-entropy is $\geqslant 0$ (Jensen's inequality).
- Equal to 0 when f is constant.

usually f in the literature

ϕ-divergence

For measure ν and dist μ define

$$\mathcal{D}_{\phi}(\nu \,\|\, \mu) = \mathrm{Ent}_{\mu}^{\phi}\left[\frac{\nu}{\mu}\right]$$



- How far from $\frac{\nu}{\mu} \equiv$ const?

Example: total variation

If $\phi(x) = \frac{1}{2}|x - 1|$, then

$$\mathcal{D}_{\phi}(\nu \,\|\, \mu) = d_{TV}(\nu, \mu)$$

- Note: in general $\mathcal{D}_{\phi}$ is asymmetric and doesn't satisfy triangle ineq.

Proxy for d_{TV}

Contraction: $\mathcal{D}_\phi(\nu P \parallel \mu) \leqslant (1-\rho)\, \mathcal{D}_\phi(\nu \parallel \mu)$ for stationary μ.

Proxy for d_{TV}

Contraction: $\mathcal{D}_\phi(\nu P \parallel \mu) \leqslant (1-\rho)\,\mathcal{D}_\phi(\nu \parallel \mu)$ for stationary μ.

Variance

$$\phi(x) := x^2$$

- $\mathrm{Ent}^{\phi}_{\mu}[f] = \mathrm{Var}_{\mu}[f]$
- $\mathcal{D}_\phi(\nu \parallel \mu) = \chi^2(\nu \parallel \mu)$
- It is a proxy by Cauchy-Schwarz:

$$d_{TV}(\nu,\mu) \leqslant O\left(\sqrt{\chi^2(\nu \parallel \mu)}\right)$$

- Contraction related to eigs of P.

called Poincaré inequality

Entropy

$$\phi(x) := x \log x$$

- $\mathrm{Ent}^{\phi}_{\mu}[f] = \mathrm{Ent}_{\mu}[f]$
- $\mathcal{D}_\phi(\nu \parallel \mu) = \mathcal{D}_{KL}(\nu \parallel \mu)$
- It is a proxy by Pinsker:

$$d_{TV}(\nu,\mu) \leqslant O\left(\sqrt{\mathcal{D}_{KL}(\nu \parallel \mu)}\right)$$

- Contraction: very hard!

called modified log-Sobolev inequality

- Why care about entropy?

- Why care about entropy?
- Suppose $\nu = \mathbb{1}_x$. Then

$$\chi^2(\nu \parallel \mu) = \frac{1}{\mu(x)} - 1 \leftarrow \text{ignore}$$

$$\mathcal{D}_{\mathsf{KL}}(\nu \parallel \mu) = \log\left(\frac{1}{\mu(x)}\right)$$

- Why care about entropy?
- Suppose $\nu = \mathbb{1}_x$. Then

$$\chi^2(\nu \parallel \mu) = \frac{1}{\mu(x)} - 1 \leftarrow \text{ignore}$$

$$\mathcal{D}_{\mathsf{KL}}(\nu \parallel \mu) = \log\left(\frac{1}{\mu(x)}\right)$$

- Contraction by $1 - \rho$ implies

$$t_{\mathsf{mix}} \leqslant \frac{\log(1/\mu(x))}{\rho}$$

$$t_{\mathsf{mix}} \leqslant \frac{\log\log(1/\mu(x))}{\rho}$$

- Why care about entropy?
- Suppose $\nu = \mathbb{1}_x$. Then

$$\chi^2(\nu \parallel \mu) = \frac{1}{\mu(x)} - 1 \leftarrow \text{ignore}$$

$$\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu) = \log\left(\frac{1}{\mu(x)}\right)$$

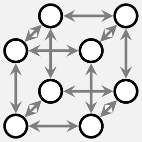
- Contraction by $1 - \rho$ implies

$$t_{\mathrm{mix}} \leqslant \frac{\log(1/\mu(x))}{\rho}$$

$$t_{\mathrm{mix}} \leqslant \frac{\log\log(1/\mu(x))}{\rho}$$

Example: hypercube

- $\rho = \Theta(1/n)$
- $1/\mu(x) = 2^n$
- $t_{\mathrm{mix}} = O(n^2)$ vs.
  $t_{\mathrm{mix}} = O(n \log n)$



- Why care about entropy?
- Suppose $\nu = \mathbb{1}_x$. Then

$$\chi^2(\nu \parallel \mu) = \frac{1}{\mu(x)} - 1 \leftarrow \text{ignore}$$

$$\mathcal{D}_{KL}(\nu \parallel \mu) = \log\left(\frac{1}{\mu(x)}\right)$$

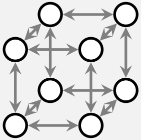
- Contraction by $1-\rho$ implies

$$t_{mix} \leqslant \frac{\log(1/\mu(x))}{\rho}$$

$$t_{mix} \leqslant \frac{\log\log(1/\mu(x))}{\rho}$$

Example: hypercube

- $\rho = \Theta(1/n)$
- $1/\mu(x) = 2^n$
- $t_{mix} = O(n^2)$ vs.
  $t_{mix} = O(n \log n)$



- However, entropy contraction is much harder to prove. ☹ We will focus mostly on variance for now.

- Why care about entropy?
- Suppose $\nu = \mathbb{1}_x$. Then

$$\chi^2(\nu \parallel \mu) = \frac{1}{\mu(x)} - 1 \leftarrow \text{ignore}$$

$$\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu) = \log\left(\frac{1}{\mu(x)}\right)$$

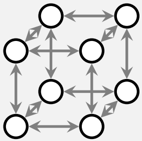
- Contraction by $1 - \rho$ implies

$$t_{\mathrm{mix}} \leqslant \frac{\log(1/\mu(x))}{\rho}$$

$$t_{\mathrm{mix}} \leqslant \frac{\log\log(1/\mu(x))}{\rho}$$

Example: hypercube

- $\rho = \Theta(1/n)$
- $1/\mu(x) = 2^n$
- $t_{\mathrm{mix}} = O(n^2)$ vs.
  $t_{\mathrm{mix}} = O(n \log n)$



- However, entropy contraction is much harder to prove. ☹ We will focus mostly on variance for now.
- Divergences have one major benefit: weak contraction. ☺

- Why care about entropy?
- Suppose $\nu = \mathbb{1}_x$. Then

$$\chi^2(\nu \parallel \mu) = \frac{1}{\mu(x)} - 1 \leftarrow \text{ignore}$$

$$\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu) = \log\left(\frac{1}{\mu(x)}\right)$$

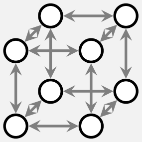
- Contraction by $1 - \rho$ implies

$$t_{\mathrm{mix}} \leqslant \frac{\log(1/\mu(x))}{\rho}$$

$$t_{\mathrm{mix}} \leqslant \frac{\log\log(1/\mu(x))}{\rho}$$

Example: hypercube

- $\rho = \Theta(1/n)$
- $1/\mu(x) = 2^n$
- $t_{\mathrm{mix}} = O(n^2)$ vs. $t_{\mathrm{mix}} = O(n \log n)$



- However, entropy contraction is much harder to prove. ☹ We will focus mostly on variance for now.
- Divergences have one major benefit: weak contraction. ☺

Lemma: data processing

Suppose N is Markov kernel. Then

$$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant \mathcal{D}_\phi(\nu \parallel \mu)$$

- Why care about entropy?
- Suppose $\nu = \mathbb{1}_x$. Then

$$\chi^2(\nu \,\|\, \mu) = \frac{1}{\mu(x)} - 1 \leftarrow \text{ignore}$$

$$\mathcal{D}_{\mathrm{KL}}(\nu \,\|\, \mu) = \log\left(\frac{1}{\mu(x)}\right)$$

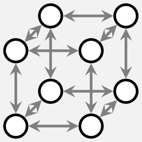
- Contraction by $1 - \rho$ implies

$$t_{\mathrm{mix}} \leqslant \frac{\log(1/\mu(x))}{\rho}$$

$$t_{\mathrm{mix}} \leqslant \frac{\log\log(1/\mu(x))}{\rho}$$

Example: hypercube

- $\rho = \Theta(1/n)$
- $1/\mu(x) = 2^n$
- $t_{\mathrm{mix}} = O(n^2)$ vs.
  $t_{\mathrm{mix}} = O(n \log n)$



- However, entropy contraction is much harder to prove. ☹ We will focus mostly on variance for now.
- Divergences have one major benefit: weak contraction. ☺

Lemma: data processing

Suppose N is Markov kernel. Then

$$\mathcal{D}_\phi(\nu N \,\|\, \mu N) \leqslant \mathcal{D}_\phi(\nu \,\|\, \mu)$$

- Markov chain P with stationary μ:

$$\mathcal{D}_\phi(\nu P \,\|\, \mu) \leqslant \mathcal{D}_\phi(\nu \,\|\, \mu)$$

- Why care about entropy?
- Suppose $\nu = \mathbb{1}_x$. Then

$$\chi^2(\nu \parallel \mu) = \frac{1}{\mu(x)} - 1 \leftarrow \text{ignore}$$

$$\mathcal{D}_{\mathrm{KL}}(\nu \parallel \mu) = \log\left(\frac{1}{\mu(x)}\right)$$

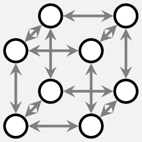
- Contraction by $1 - \rho$ implies

$$t_{\mathrm{mix}} \leqslant \frac{\log(1/\mu(x))}{\rho}$$

$$t_{\mathrm{mix}} \leqslant \frac{\log\log(1/\mu(x))}{\rho}$$

Example: hypercube

- $\rho = \Theta(1/n)$
- $1/\mu(x) = 2^n$
- $t_{\mathrm{mix}} = O(n^2)$ vs. $t_{\mathrm{mix}} = O(n \log n)$



- However, entropy contraction is much harder to prove. ☹ We will focus mostly on variance for now.
- Divergences have one major benefit: weak contraction. ☺

Lemma: data processing

Suppose N is Markov kernel. Then

$$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant \mathcal{D}_\phi(\nu \parallel \mu)$$

- Markov chain P with stationary μ:

$$\mathcal{D}_\phi(\nu P \parallel \mu) \leqslant \mathcal{D}_\phi(\nu \parallel \mu)$$

- Useful for $P = NN^\circ$. Only need to show strong contraction for N (or possibly N°). ☺

Proof:

- N°: time-reversal of N w.r.t. μ.

Proof:

- N°: time-reversal of N w.r.t. μ.
- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$.

Proof:

- N°: time-reversal of N w.r.t. μ.
- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$.
- We have

$$g(y) = \frac{\sum_x f(x)\mu(x)N(x,y)}{\sum_x \mu(x)N(x,y)}$$

Proof:

- N°: time-reversal of N w.r.t. μ.
- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$.
- We have

$$g(y) = \frac{\sum_x f(x)\mu(x)N(x,y)}{\sum_x \mu(x)N(x,y)}$$

- This means $g = N^\circ f$.

column vector

column vector

Proof:

- N°: time-reversal of N w.r.t. μ.
- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$.
- We have

$$g(y) = \frac{\sum_x f(x)\mu(x)N(x,y)}{\sum_x \mu(x)N(x,y)}$$

- This means $g = N^\circ f$.

column vector

column vector

- So we have $\mathbb{E}_\mu[\phi \circ f] - \mathbb{E}_{\mu N}[\phi \circ g] =$

$$\mathbb{E}_{y \sim \mu N}\left[\mathrm{Ent}^{\phi}_{N^\circ(y,\cdot)}[f]\right] \geqslant 0.$$

Proof:

- N°: time-reversal of N w.r.t. μ.
- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$.
- We have

$$g(y) = \frac{\sum_x f(x)\mu(x)N(x,y)}{\sum_x \mu(x)N(x,y)}$$

- This means $g = N^\circ f$.

column vector   column vector

- So we have $\mathbb{E}_\mu[\phi \circ f] - \mathbb{E}_{\mu N}[\phi \circ g] =$

$$\mathbb{E}_{y \sim \mu N}\left[\mathrm{Ent}^{\phi}_{N^\circ(y,\cdot)}[f]\right] \geqslant 0.$$

- On the other hand, $\mathbb{E}_\mu[f] = \mathbb{E}_{\mu N}[g]$, so

$$\phi(\mathbb{E}_\mu[f]) = \phi(\mathbb{E}_{\mu N}[g]).$$

Proof:

- N°: time-reversal of N w.r.t. μ.
- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$.
- We have

$$g(y) = \frac{\sum_x f(x)\mu(x)N(x,y)}{\sum_x \mu(x)N(x,y)}$$

- This means $g = N^\circ f$.

column vector   column vector

- So we have $\mathbb{E}_\mu[\phi \circ f] - \mathbb{E}_{\mu N}[\phi \circ g] =$

$$\mathbb{E}_{y \sim \mu N}\left[\mathrm{Ent}^{\phi}_{N^\circ(y,\cdot)}[f]\right] \geqslant 0.$$

- On the other hand, $\mathbb{E}_\mu[f] = \mathbb{E}_{\mu N}[g]$, so

$$\phi(\mathbb{E}_\mu[f]) = \phi(\mathbb{E}_{\mu N}[g]).$$

- Therefore

$$\mathrm{Ent}^{\phi}_{\mu}[f] \geqslant \mathrm{Ent}^{\phi}_{\mu N}[g].$$

Proof:

- N°: time-reversal of N w.r.t. μ.
- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$.
- We have

$$g(y) = \frac{\sum_x f(x)\mu(x)N(x,y)}{\sum_x \mu(x)N(x,y)}$$

- This means $g = N^\circ f$.

column vector   column vector

- So we have $\mathbb{E}_\mu[\phi \circ f] - \mathbb{E}_{\mu N}[\phi \circ g] =$

$$\mathbb{E}_{y\sim\mu N}\left[\mathrm{Ent}^{\phi}_{N^\circ(y,\cdot)}[f]\right] \geqslant 0.$$

- On the other hand, $\mathbb{E}_\mu[f] = \mathbb{E}_{\mu N}[g]$, so

$$\phi(\mathbb{E}_\mu[f]) = \phi(\mathbb{E}_{\mu N}[g]).$$

- Therefore

$$\mathrm{Ent}^{\phi}_{\mu}[f] \geqslant \mathrm{Ent}^{\phi}_{\mu N}[g].$$

Lemma: data processing

Suppose N is Markov kernel and ϕ convex. Then

$$\mathcal{D}_\phi(\nu N \parallel \mu N) \leqslant \mathcal{D}_\phi(\nu \parallel \mu)$$

Spectral analysis

Contraction of χ^2 is determined by eigenvalues:

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(N N^\circ)$$

Spectral analysis

Contraction of χ^2 is determined by eigenvalues:

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(NN^\circ)$$

- When P is time-reversible w.r.t. μ:

$$\mathrm{diag}(\mu)P = \underset{\text{symmetric matrix}}{\underset{\uparrow}{Q}}$$

Spectral analysis

Contraction of χ^2 is determined by eigenvalues:

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N\|\mu N)}{\chi^2(\nu\|\mu)}\right\} = \lambda_2(NN^\circ)$$

- When P is time-reversible w.r.t. μ:

$$\mathrm{diag}(\mu)P = \underset{\substack{\uparrow\\ \text{symmetric matrix}}}{Q}$$

- So we have

$$\underbrace{\mathrm{diag}(\mu)^{1/2}\cdot P\cdot \mathrm{diag}(\mu)^{-1/2} = \mathrm{diag}(\mu)^{-1/2}\cdot Q\cdot \mathrm{diag}(\mu)^{-1/2}}_{\text{still symmetric}}$$

Spectral analysis

Contraction of χ^2 is determined by eigenvalues:

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N\|\mu N)}{\chi^2(\nu\|\mu)}\right\} = \lambda_2(NN^\circ)$$

- When P is time-reversible w.r.t. μ:

$$\mathrm{diag}(\mu)P = \underset{\uparrow \atop \text{symmetric matrix}}{Q}$$

- So we have

$$\mathrm{diag}(\mu)^{1/2}\cdot P\cdot \mathrm{diag}(\mu)^{-1/2} = \underbrace{\mathrm{diag}(\mu)^{-1/2}\cdot Q\cdot \mathrm{diag}(\mu)^{-1/2}}_{\text{still symmetric}}$$

- This means eigs are real!

so λ_2 has meaning

Spectral analysis

Contraction of χ^2 is determined by eigenvalues:

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(NN^\circ)$$

- When P is time-reversible w.r.t. μ:

$$\mathrm{diag}(\mu)P = \underset{\uparrow \atop \text{symmetric matrix}}{Q}$$

- So we have

$$\underbrace{\mathrm{diag}(\mu)^{1/2} \cdot P \cdot \mathrm{diag}(\mu)^{-1/2} = \mathrm{diag}(\mu)^{-1/2} \cdot Q \cdot \mathrm{diag}(\mu)^{-1/2}}_{\text{still symmetric}}$$

- This means eigs are real!

so λ_2 has meaning

- We will show later that eigs are $\in [-1, 1]$ for time-reversible P.

Spectral analysis

Contraction of χ^2 is determined by eigenvalues:

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N\|\mu N)}{\chi^2(\nu\|\mu)}\right\} = \lambda_2(NN^\circ)$$

- When P is time-reversible w.r.t. μ:

$$\mathrm{diag}(\mu)P = \underset{\uparrow \text{ symmetric matrix}}{Q}$$

- So we have

$$\mathrm{diag}(\mu)^{1/2}\cdot P\cdot \mathrm{diag}(\mu)^{-1/2} = \underbrace{\mathrm{diag}(\mu)^{-1/2}\cdot Q\cdot \mathrm{diag}(\mu)^{-1/2}}_{\text{still symmetric}}$$

- This means eigs are real!

  ↑ so λ_2 has meaning

- We will show later that eigs are $\in[-1,1]$ for time-reversible P.
- For $P = NN^\circ$, they are $\geqslant 0$!

Spectral analysis

Contraction of χ^2 is determined by eigenvalues:

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(NN^\circ)$$

- When P is time-reversible w.r.t. μ:

$$\mathrm{diag}(\mu)P = \underset{\substack{\uparrow \\ \text{symmetric matrix}}}{Q}$$

- So we have

$$\mathrm{diag}(\mu)^{1/2} \cdot P \cdot \mathrm{diag}(\mu)^{-1/2} = \underbrace{\mathrm{diag}(\mu)^{-1/2} \cdot Q \cdot \mathrm{diag}(\mu)^{-1/2}}_{\text{still symmetric}}$$

- This means eigs are real!

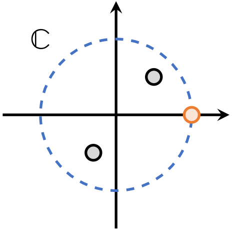
  so λ_2 has meaning

- We will show later that eigs are $\in [-1, 1]$ for time-reversible P.
- For $P = NN^\circ$, they are $\geqslant 0$!
- When N is time-reversible chain:

$$\lambda_2(NN^\circ) = \max\{\lambda_2(N), |\lambda_{\min}(N)|\}^2$$

Eigenvalues

[Perron-Frobenius] for Markov chains:

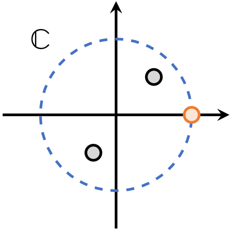


Eigenvalues

[Perron-Frobenius] for Markov chains:

- 1 is special eig:

  $$\mu P = \mu, P\mathbb{1} = \mathbb{1}$$



Eigenvalues

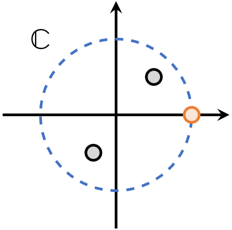
[Perron-Frobenius] for Markov chains:

- 1 is special eig:

  $$\mu P = \mu, P\mathbb{1} = \mathbb{1}$$

- Other eigs have $|\cdot| \leqslant 1$.

  strict if ergodic



Eigenvalues

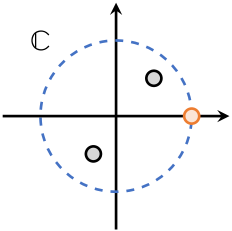
[Perron-Frobenius] for Markov chains:

- 1 is special eig:

  $$\mu P = \mu, P\mathbb{1} = \mathbb{1}$$

- Other eigs have $|\cdot| \leqslant 1$.

  strict if ergodic



Proof: $(Pv)_i$ is an average of v_js, so

$$|(Pv)_i| \leqslant \max\{|v_j|\}.$$

So if $Pv = \lambda v$, we must have $|\lambda| \leqslant 1$.

Eigenvalues

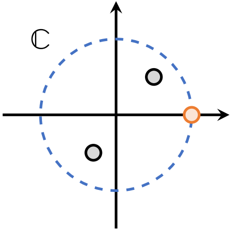
[Perron-Frobenius] for Markov chains:

- 1 is special eig:

$$\mu P = \mu,\ P\mathbb{1} = \mathbb{1}$$

- Other eigs have $|\cdot| \leqslant 1$.

strict if ergodic

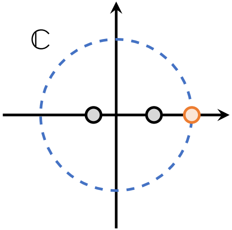


Proof: $(Pv)_i$ is an average of v_js, so

$$|(Pv)_i| \leqslant \max\{|v_j|\}.$$

So if $Pv = \lambda v$, we must have $|\lambda| \leqslant 1$.

- If P is time-reversible the picture is

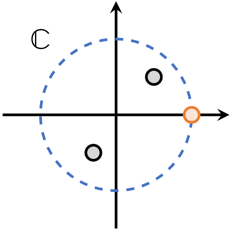


Eigenvalues

[Perron-Frobenius] for Markov chains:

- 1 is special eig:

$$\mu P = \mu, P\mathbb{1} = \mathbb{1}$$



- Other eigs have $|\cdot| \leqslant 1$.

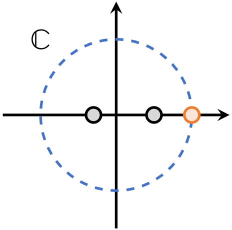
strict if ergodic

Proof: $(Pv)_i$ is an average of v_js, so

$$|(Pv)_i| \leqslant \max\{|v_j|\}.$$

So if $Pv = \lambda v$, we must have $|\lambda| \leqslant 1$.

- If P is time-reversible the picture is



- Use convention

$$1 = \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant -1$$

Eigenvalues

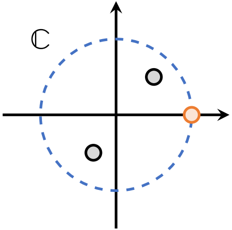
[Perron-Frobenius] for Markov chains:

- 1 is special eig:

$$\mu P = \mu, P\mathbb{1} = \mathbb{1}$$

- Other eigs have $|\cdot| \leqslant 1$.

strict if ergodic

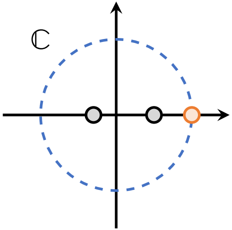


Proof: $(Pv)_i$ is an average of v_js, so

$$|(Pv)_i| \leqslant \max\{|v_j|\}.$$

So if $Pv = \lambda v$, we must have $|\lambda| \leqslant 1$.

- If P is time-reversible the picture is



- Use convention

$$1 = \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant -1$$

- Spectral gap: usually $1 - \lambda_2$, in some places $1 - \max\{\lambda_2, |\lambda_n|\}$.

Eigenvalues

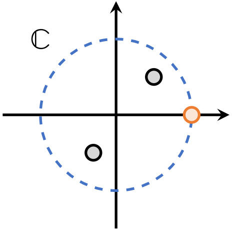
[Perron-Frobenius] for Markov chains:

- 1 is special eig:

  $$\mu P = \mu, P\mathbb{1} = \mathbb{1}$$

- Other eigs have $|\cdot| \leqslant 1$.

  strict if ergodic

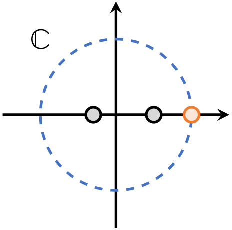


Proof: $(Pv)_i$ is an average of v_js, so

$$|(Pv)_i| \leqslant \max\{|v_j|\}.$$

So if $Pv = \lambda v$, we must have $|\lambda| \leqslant 1$.

- If P is time-reversible the picture is



- Use convention

  $$1 = \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant -1$$

- Spectral gap: usually $1 - \lambda_2$, in some places $1 - \max\{\lambda_2, |\lambda_n|\}$.
- If $P = NN^\circ$, we will show all $\lambda \geqslant 0$.

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(N N^\circ)$$

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.
- Additive shift doesn't change Var:

$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[f + c],$$

  because

$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 = \mathbb{E}_\mu[(f - \mathbb{E}_\mu[f])^2]$$

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(N N^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.
- Additive shift doesn't change Var:
$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[f + c],$$
because
$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 = \mathbb{E}_\mu[(f - \mathbb{E}_\mu[f])^2]$$
- Can assume $\mathbb{E}_\mu[f] = 0$, which means $\mathbb{E}_{\mu^\circ}[g] = 0$.

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.
- Additive shift doesn't change Var:
$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[f + c],$$
because
$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 = \mathbb{E}_\mu[(f - \mathbb{E}_\mu[f])^2]$$
- Can assume $\mathbb{E}_\mu[f] = 0$, which means $\mathbb{E}_{\mu^\circ}[g] = 0$.
- Then $\mathrm{Var}_\mu[f] = f^\intercal \mathrm{diag}(\mu) f$, and $\mathrm{Var}_{\mu^\circ}[g] = g^\intercal \mathrm{diag}(\mu^\circ) g$.

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N\|\mu N)}{\chi^2(\nu\|\mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.
- Additive shift doesn't change Var:
$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[f + c],$$
because
$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 = \mathbb{E}_\mu[(f - \mathbb{E}_\mu[f])^2]$$
- Can assume $\mathbb{E}_\mu[f] = 0$, which means $\mathbb{E}_{\mu^\circ}[g] = 0$.
- Then $\mathrm{Var}_\mu[f] = f^\intercal \mathrm{diag}(\mu) f$, and $\mathrm{Var}_{\mu^\circ}[g] = g^\intercal \mathrm{diag}(\mu^\circ) g$.

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.
- Additive shift doesn't change Var:
$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[f + c],$$
because
$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 = \mathbb{E}_\mu[(f - \mathbb{E}_\mu[f])^2]$$
- Can assume $\mathbb{E}_\mu[f] = 0$, which means $\mathbb{E}_{\mu^\circ}[g] = 0$.
- Then $\mathrm{Var}_\mu[f] = f^\intercal \mathrm{diag}(\mu) f$, and $\mathrm{Var}_{\mu^\circ}[g] = g^\intercal \mathrm{diag}(\mu^\circ) g$.
- So if $u = \mathrm{diag}(\mu)^{1/2} f$, then we are after $u^\intercal M u / \|u\|^2$ for $M =$
$$\mathrm{diag}(\mu)^{-1/2} (N^\circ)^\intercal \mathrm{diag}(\mu^\circ) N^\circ \mathrm{diag}(\mu)^{-1/2}$$

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu \| \mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.
- Additive shift doesn't change Var:
$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[f + c],$$
because
$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 = \mathbb{E}_\mu[(f - \mathbb{E}_\mu[f])^2]$$
- Can assume $\mathbb{E}_\mu[f] = 0$, which means $\mathbb{E}_{\mu^\circ}[g] = 0$.
- Then $\mathrm{Var}_\mu[f] = f^\intercal \mathrm{diag}(\mu) f$, and $\mathrm{Var}_{\mu^\circ}[g] = g^\intercal \mathrm{diag}(\mu^\circ) g$.
- So if $u = \mathrm{diag}(\mu)^{1/2} f$, then we are after $u^\intercal M u / \|u\|^2$ for $M =$
$$\mathrm{diag}(\mu)^{-1/2}(N^\circ)^\intercal \mathrm{diag}(\mu^\circ) N^\circ \mathrm{diag}(\mu)^{-1/2}$$
- Note that $M = AA^\intercal$, so $\geqslant 0$ eigs.

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N\|\mu N)}{\chi^2(\nu\|\mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.
- Additive shift doesn't change Var:
$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[f+c],$$
because
$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 = \mathbb{E}_\mu[(f-\mathbb{E}_\mu[f])^2]$$
- Can assume $\mathbb{E}_\mu[f] = 0$, which means $\mathbb{E}_{\mu^\circ}[g] = 0$.
- Then $\mathrm{Var}_\mu[f] = f^\intercal \mathrm{diag}(\mu) f$, and $\mathrm{Var}_{\mu^\circ}[g] = g^\intercal \mathrm{diag}(\mu^\circ) g$.
- So if $u = \mathrm{diag}(\mu)^{1/2} f$, then we are after $u^\intercal M u / \|u\|^2$ for $M =$
$$\mathrm{diag}(\mu)^{-1/2}(N^\circ)^\intercal \mathrm{diag}(\mu^\circ) N^\circ \mathrm{diag}(\mu)^{-1/2}$$
- Note that $M = AA^\intercal$, so $\geqslant 0$ eigs.
- By detailed balance $\mathrm{diag}(\mu) N = (\mathrm{diag}(\mu^\circ) N^\circ)^\intercal$, so
$$M = \mathrm{diag}(\mu)^{1/2} N N^\circ \mathrm{diag}(\mu)^{-1/2}$$

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N\|\mu N)}{\chi^2(\nu\|\mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.
- Additive shift doesn't change Var:
$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[f+c],$$
because
$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 = \mathbb{E}_\mu[(f-\mathbb{E}_\mu[f])^2]$$
- Can assume $\mathbb{E}_\mu[f] = 0$, which means $\mathbb{E}_{\mu^\circ}[g] = 0$.
- Then $\mathrm{Var}_\mu[f] = f^\intercal \mathrm{diag}(\mu) f$, and $\mathrm{Var}_{\mu^\circ}[g] = g^\intercal \mathrm{diag}(\mu^\circ) g$.
- So if $u = \mathrm{diag}(\mu)^{1/2} f$, then we are after $u^\intercal M u / \|u\|^2$ for $M =$
$$\mathrm{diag}(\mu)^{-1/2}(N^\circ)^\intercal \mathrm{diag}(\mu^\circ) N^\circ \mathrm{diag}(\mu)^{-1/2}$$
- Note that $M = AA^\intercal$, so $\geqslant 0$ eigs.
- By detailed balance $\mathrm{diag}(\mu) N = (\mathrm{diag}(\mu^\circ) N^\circ)^\intercal$, so
$$M = \mathrm{diag}(\mu)^{1/2} N N^\circ \mathrm{diag}(\mu)^{-1/2}$$
- Similar to NN°, so same eigs.

Lemma

Suppose N is Markov kernel and N° is time-reversal w.r.t. μ. Then

$$\max\left\{\frac{\chi^2(\nu N \| \mu N)}{\chi^2(\nu\|\mu)}\right\} = \lambda_2(NN^\circ)$$

Proof:

- Let $f = \nu/\mu$ and $g = (\nu N)/(\mu N)$, and $\mu^\circ = \mu N$.
- We can equivalently consider $\mathrm{Var}_\mu[f]$ vs. $\mathrm{Var}_{\mu^\circ}[g]$.
- Additive shift doesn't change Var:
$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[f+c],$$
because
$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[f^2] - \mathbb{E}_\mu[f]^2 = \mathbb{E}_\mu[(f - \mathbb{E}_\mu[f])^2]$$
- Can assume $\mathbb{E}_\mu[f] = 0$, which means $\mathbb{E}_{\mu^\circ}[g] = 0$.
- Then $\mathrm{Var}_\mu[f] = f^\intercal \mathrm{diag}(\mu) f$, and $\mathrm{Var}_{\mu^\circ}[g] = g^\intercal \mathrm{diag}(\mu^\circ) g$.
- So if $u = \mathrm{diag}(\mu)^{1/2} f$, then we are after $u^\intercal M u / \|u\|^2$ for $M =$
$$\mathrm{diag}(\mu)^{-1/2}(N^\circ)^\intercal \mathrm{diag}(\mu^\circ) N^\circ \mathrm{diag}(\mu)^{-1/2}$$
- Note that $M = AA^\intercal$, so $\geqslant 0$ eigs.
- By detailed balance $\mathrm{diag}(\mu) N = (\mathrm{diag}(\mu^\circ) N^\circ)^\intercal$, so
$$M = \mathrm{diag}(\mu)^{1/2} N N^\circ \mathrm{diag}(\mu)^{-1/2}$$
- Similar to NN°, so same eigs.
- Top eigenvec of M: $\mathrm{diag}(\mu)^{1/2}\mathbb{1}$. We want u orthogonal. So we get
$$\lambda_2(M) = \lambda_2(NN^\circ).$$

- As a corollary, for chain P with stationary μ:

$$\chi^2(\nu P \parallel \mu) \leqslant \lambda_2(PP^\circ)\,\chi^2(\nu \parallel \mu).$$

- As a corollary, for chain P with stationary μ:

$$\chi^2(\nu P \parallel \mu) \leqslant \lambda_2(PP^\circ)\,\chi^2(\nu \parallel \mu).$$

- To get mixing we need one more ingredient:

- As a corollary, for chain P with stationary μ:

$$\chi^2(\nu P \parallel \mu) \leqslant \lambda_2(PP^\circ)\,\chi^2(\nu \parallel \mu).$$

- To get mixing we need one more ingredient:

Lemma: χ^2 proxy for d_{TV}

$$d_{TV}(\nu, \mu) \leqslant O\left(\sqrt{\chi^2(\nu \parallel \mu)}\right)$$

- As a corollary, for chain P with stationary μ:

$$\chi^2(\nu P \parallel \mu) \leqslant \lambda_2(PP^\circ)\,\chi^2(\nu \parallel \mu).$$

- To get mixing we need one more ingredient:

Lemma: χ^2 proxy for d_{TV}

$$d_{TV}(\nu, \mu) \leqslant O\left(\sqrt{\chi^2(\nu \parallel \mu)}\right)$$

Proof: we have $d_{TV}(\nu, \mu) =$

$$\frac{1}{2}\mathbb{E}_\mu\left[\left|\frac{\nu}{\mu} - 1\right|\right] \leqslant \frac{1}{2}\sqrt{\mathbb{E}_\mu\left[\left(\frac{\nu}{\mu} - 1\right)^2\right]} = O\left(\sqrt{\chi^2(\nu \parallel \mu)}\right)$$

- As a corollary, for chain P with stationary μ:

$$\chi^2(\nu P \parallel \mu) \leqslant \lambda_2(PP^\circ)\,\chi^2(\nu \parallel \mu).$$

- To get mixing we need one more ingredient:

Lemma: χ^2 proxy for d_{TV}

$$d_{TV}(\nu, \mu) \leqslant O\left(\sqrt{\chi^2(\nu \parallel \mu)}\right)$$

Proof: we have $d_{TV}(\nu, \mu) =$

$$\frac{1}{2}\mathbb{E}_\mu\left[\left|\frac{\nu}{\mu} - 1\right|\right] \leqslant \frac{1}{2}\sqrt{\mathbb{E}_\mu\left[\left(\frac{\nu}{\mu} - 1\right)^2\right]} = O\left(\sqrt{\chi^2(\nu \parallel \mu)}\right)$$

Corollary: mixing

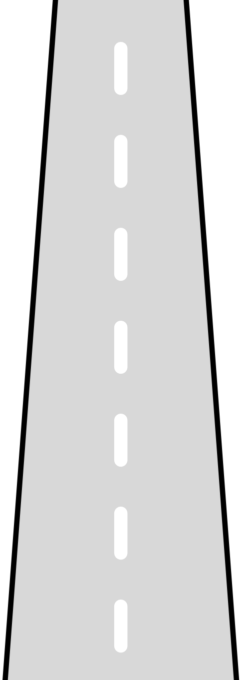
$$t_{mix}(\epsilon) = O\left(\frac{1}{1-\lambda_2(PP^\circ)}\log\left(\frac{\chi^2(\nu_0 \| \mu)}{\epsilon}\right)\right)$$

Functional Analysis

- Divergences
- Poincaré and modified log-Sobolev
- Data processing
- Spectral analysis

Fourier Analysis

- Abelian walks
- Characters

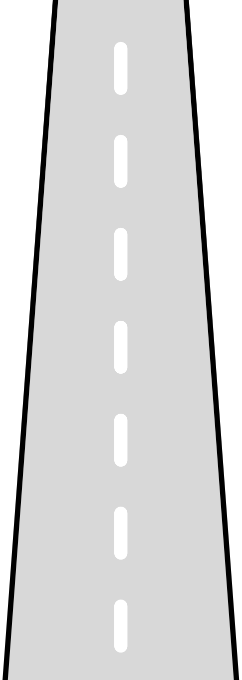


Functional Analysis

- Divergences
- Poincaré and modified log-Sobolev
- Data processing
- Spectral analysis

Fourier Analysis

- Abelian walks
- Characters



Abelian walks

- Finite Abelian group (with $+$):

$$G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$$

Abelian walks

- Finite Abelian group (with $+$):

$$G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$$

- Take dist π over G.

sparse support

Abelian walks

- Finite Abelian group (with $+$):

$$G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$$

- Take dist π over G.

  sparse support

- We get Markov chain P:

$$X_t \mapsto X_{t+1} = X_t + Z_t$$

where Z_t are i.i.d. samples from π.

Abelian walks

- Finite Abelian group (with $+$):

$$G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$$

- Take dist π over G.

sparse support

- We get Markov chain P:

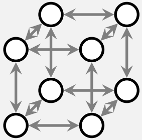
$$X_t \mapsto X_{t+1} = X_t + Z_t$$

where Z_t are i.i.d. samples from π.

Example: hypercube

Distribution π:

- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



Abelian walks

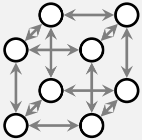
- Finite Abelian group (with $+$):
$$G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$$
- Take dist π over G.

sparse support

- We get Markov chain P:
$$X_t \mapsto X_{t+1} = X_t + Z_t$$
where Z_t are i.i.d. samples from π.

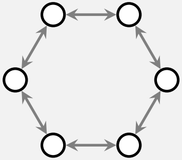
Example: hypercube

Distribution π:
- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



Example: cycle

Distribution π:
- $+1$ w.p. $1/2$
- -1 w.p. $1/2$



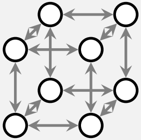
Abelian walks

- Finite Abelian group (with $+$):
$$G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$$
- Take dist π over G.

  sparse support
- We get Markov chain P:
$$X_t \mapsto X_{t+1} = X_t + Z_t$$
where Z_t are i.i.d. samples from π.

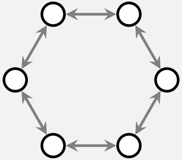
Example: hypercube

Distribution π:
- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



Example: cycle

Distribution π:
- $+1$ w.p. $1/2$
- -1 w.p. $1/2$



- Fact: $\mu =$ uniform is always stationary

Abelian walks

- Finite Abelian group (with $+$):

$$G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$$

- Take dist π over G.

  sparse support

- We get Markov chain P:

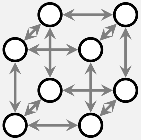
$$X_t \mapsto X_{t+1} = X_t + Z_t$$

  where Z_t are i.i.d. samples from π.

Example: hypercube

Distribution π:

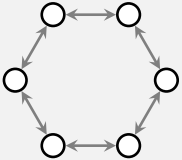
- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



Example: cycle

Distribution π:

- $+1$ w.p. $1/2$
- -1 w.p. $1/2$



- Fact: $\mu =$ uniform is always stationary
- Fact: P time-reversible iff π is symmetric, i.e.,

$$\pi(x) = \pi(-x)$$

Abelian walks

- Finite Abelian group (with $+$):

$$G = \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_k}$$

- Take dist π over G.

sparse support

- We get Markov chain P:

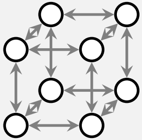
$$X_t \mapsto X_{t+1} = X_t + Z_t$$

where Z_t are i.i.d. samples from π.

Example: hypercube

Distribution π:

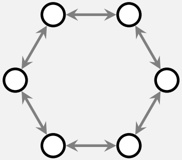
- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



Example: cycle

Distribution π:

- $+1$ w.p. $1/2$
- -1 w.p. $1/2$



- Fact: $\mu =$ uniform is always stationary
- Fact: P time-reversible iff π is symmetric, i.e.,

$$\pi(x) = \pi(-x)$$

- Fact: P irreducible iff $\mathbf{supp}(\pi)$ generates G.

Characters

- Abelian walks are extremely easy for spectral analysis. 🙂

Characters

- Abelian walks are extremely easy for spectral analysis. ☺
- Eigvecs are always the characters.

Characters

- Abelian walks are extremely easy for spectral analysis. ☺
- Eigvecs are always the characters.

Character

A function $\chi : G \to \mathbb{C} - \{0\}$ where

$$\chi(x + y) = \chi(x)\chi(y)$$

Characters

- Abelian walks are extremely easy for spectral analysis. ☺
- Eigvecs are always the characters.

Character

A function $\chi : G \to \mathbb{C} - \{0\}$ where

$$\chi(x+y) = \chi(x)\chi(y)$$

Proof: we have $(P\chi)(x) =$

$$\sum_y \pi(y-x)\chi(y) = \chi(x) \sum_y \chi(y-x)\pi(y-x),$$

so $P\chi = \lambda\chi$, where

$$\lambda = \mathbb{E}_{z\sim\pi}[\chi(z)].$$

Characters

- Abelian walks are extremely easy for spectral analysis. 🙂
- Eigvecs are always the characters.

Character

A function $\chi : G \to \mathbb{C} - \{0\}$ where

$$\chi(x+y) = \chi(x)\chi(y)$$

Proof: we have $(P\chi)(x) =$

$$\sum_y \pi(y-x)\chi(y) = \chi(x) \sum_y \chi(y-x)\pi(y-x),$$

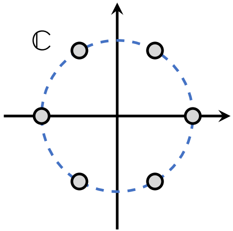
so $P\chi = \lambda\chi$, where

$$\lambda = \mathbb{E}_{z\sim\pi}[\chi(z)].$$

- We know characters of $\mathbb{Z}_n$:

$$\chi(x) = \exp(2\pi i \cdot kx/n)$$

for $k = 0, \ldots, n-1$.



Characters

- Abelian walks are extremely easy for spectral analysis. ☺
- Eigvecs are always the characters.

Character

A function $\chi : G \to \mathbb{C} - \{0\}$ where

$$\chi(x + y) = \chi(x)\chi(y)$$

Proof: we have $(P\chi)(x) =$

$$\sum_y \pi(y-x)\chi(y) = \chi(x) \sum_y \chi(y-x)\pi(y-x),$$

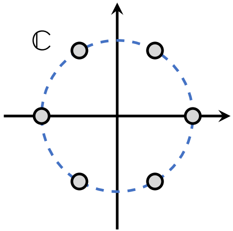
so $P\chi = \lambda\chi$, where

$$\lambda = \mathbb{E}_{z\sim\pi}[\chi(z)].$$

- We know characters of $\mathbb{Z}_n$:

$$\chi(x) = \exp(2\pi i \cdot kx/n)$$

for $k = 0, \ldots, n-1$.



- There are exactly n of them! ☺

Characters

- Abelian walks are extremely easy for spectral analysis. 🙂
- Eigvecs are always the characters.

Character

A function $\chi : G \to \mathbb{C} - \{0\}$ where

$$\chi(x+y) = \chi(x)\chi(y)$$

Proof: we have $(P\chi)(x) =$

$$\sum_y \pi(y-x)\chi(y) = \chi(x) \sum_y \chi(y-x)\pi(y-x),$$

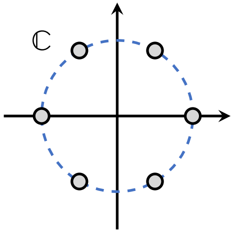
so $P\chi = \lambda\chi$, where

$$\lambda = \mathbb{E}_{z\sim\pi}[\chi(z)].$$

- We know characters of $\mathbb{Z}_n$:

$$\chi(x) = \exp(2\pi i \cdot kx/n)$$

for $k = 0, \ldots, n-1$.



- There are exactly n of them! 🙂
- Characters of $G_1 \times G_2$:

$$\chi(x, y) = \chi_1(x)\chi_2(y).$$

Characters

- Abelian walks are extremely easy for spectral analysis. ☺
- Eigvecs are always the characters.

Character

A function $\chi : G \to \mathbb{C} - \{0\}$ where

$$\chi(x+y) = \chi(x)\chi(y)$$

Proof: we have $(P\chi)(x) =$

$$\sum_y \pi(y-x)\chi(y) = \chi(x) \sum_y \chi(y-x)\pi(y-x),$$

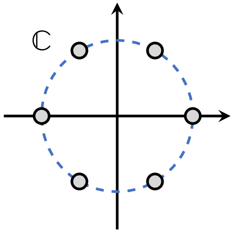
so $P\chi = \lambda\chi$, where

$$\lambda = \mathbb{E}_{z\sim\pi}[\chi(z)].$$

- We know characters of $\mathbb{Z}_n$:

$$\chi(x) = \exp(2\pi i \cdot kx/n)$$

for $k = 0, \ldots, n-1$.



- There are exactly n of them! ☺
- Characters of $G_1 \times G_2$:

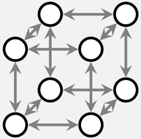
$$\chi(x, y) = \chi_1(x)\chi_2(y).$$

- For G, we get $|G|$ characters. ☺

Example: hypercube

Distribution π:

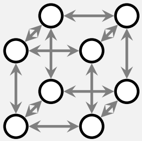
- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



Example: hypercube

Distribution π:

- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$

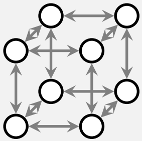


- There are 2^n characters.

Example: hypercube

Distribution π:

- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



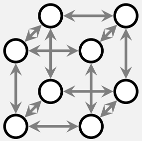
- There are 2^n characters.
- $\binom{n}{k}$ of them have eigenval

$$k/n$$

Example: hypercube

Distribution π:

- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$

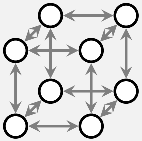


- There are 2^n characters.
- $\binom{n}{k}$ of them have eigenval
$$k/n$$
- Spectral gap:
$$1-(n-1)/n=1/n$$

Example: hypercube

Distribution π:

- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



- There are 2^n characters.
- $\binom{n}{k}$ of them have eigenval

$$k/n$$

- Spectral gap:

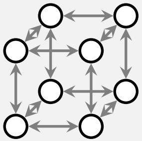
$$1 - (n-1)/n = 1/n$$

- $t_{\text{mix}} \leqslant O(n^2)$

Example: hypercube

Distribution π:

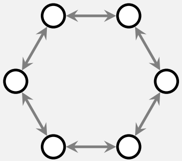
- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



Example: cycle

Distribution π:

- $+1$ w.p. $1/2$
- -1 w.p. $1/2$



- There are 2^n characters.
- $\binom{n}{k}$ of them have eigenval

$$k/n$$

- Spectral gap:

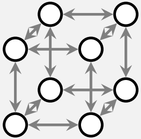
$$1 - (n-1)/n = 1/n$$

- $t_{mix} \leqslant O(n^2)$

Example: hypercube

Distribution π:

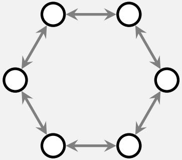
- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$



Example: cycle

Distribution π:

- $+1$ w.p. $1/2$
- -1 w.p. $1/2$



- There are 2^n characters.
- $\binom{n}{k}$ of them have eigenval

$$k/n$$

- Spectral gap:

$$1-(n-1)/n = 1/n$$

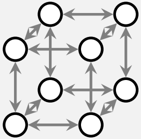
- $t_{\text{mix}} \leqslant O(n^2)$

- There are n characters.

Example: hypercube

Distribution π:

- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$

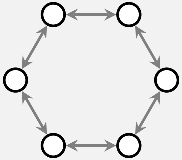


- There are 2^n characters.
- $\binom{n}{k}$ of them have eigenval
$$k/n$$
- Spectral gap:
$$1-(n-1)/n = 1/n$$
- $t_{\text{mix}} \leqslant O(n^2)$

Example: cycle

Distribution π:

- $+1$ w.p. $1/2$
- -1 w.p. $1/2$

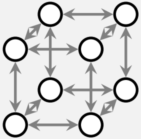


- There are n characters.
- Each has eigenval
$$\cos(2\pi k/n)$$

Example: hypercube

Distribution π:

- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$

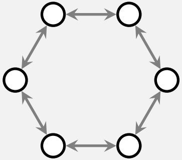


- There are 2^n characters.
- $\binom{n}{k}$ of them have eigenval
$$k/n$$
- Spectral gap:
$$1 - (n-1)/n = 1/n$$
- $t_{\text{mix}} \leqslant O(n^2)$

Example: cycle

Distribution π:

- $+1$ w.p. $1/2$
- -1 w.p. $1/2$

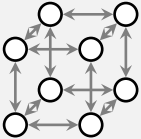


- There are n characters.
- Each has eigenval
$$\cos(2\pi k/n)$$
- Spectral gap:
$$1 - \cos(2\pi/n) \simeq \Theta(1/n^2)$$

Example: hypercube

Distribution π:

- 0 w.p. $1/2$
- $\mathbb{1}_i$ w.p. $1/2n$

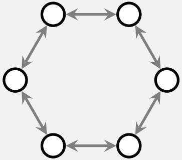


- There are 2^n characters.
- $\binom{n}{k}$ of them have eigenval
$$k/n$$
- Spectral gap:
$$1 - (n-1)/n = 1/n$$
- $t_{\mathrm{mix}} \leqslant O(n^2)$

Example: cycle

Distribution π:

- $+1$ w.p. $1/2$
- -1 w.p. $1/2$



- There are n characters.
- Each has eigenval
$$\cos(2\pi k/n)$$
- Spectral gap:
$$1 - \cos(2\pi/n) \simeq \Theta(1/n^2)$$
- $t_{\mathrm{mix}} \leqslant O(n^2 \log n)$?