CPSC 661: Sampling Algorithms in ML

Andre Wibisono

February 8, 2021

Last time

- Markov chain $P = (P_x : x \in \mathcal{X})$
- Reversibility (detailed balance) $\Rightarrow$ Stationary

$$\nu(x) \cdot P_x(y) = \nu(y) \cdot P_y(x) \Rightarrow \nu(y) = \sum_{x \in \mathcal{X}} \nu(x) \cdot P_x(y)$$

$\Leftrightarrow$ if $X_0 \sim \nu$, $X_1 | X_0 \sim P_{X_0}$ } $(X_0, X_1) \overset{d}{=} (X_1, X_0)$

$X_1 \sim \nu$

Today: Spectral theory (eigenvalue analysis) of reversible MC

Motivation from discrete space

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

$\mathcal{P}(\mathcal{X}) = \Delta_n = \{ p = (p_1, \dots, p_n) : p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \}$

Markov chain P = stochastic matrix

eg. RW on graph $G = (V, E)$

adjacency matrix A

diagonal degree matrix D

Random walk matrix $P = D^{-1}A$

Laplacian $L = I - P$

Markov chain: $X_k \sim \rho_k$

$X_{k+1} | X_k \sim P_{X_k}$

$\Rightarrow X_{k+1} \sim \rho_{k+1} = \rho_k \cdot P$

$= P$

Then $\boxed{\delta_k = \delta_0 \cdot P^k}$ ← Power method for computing eigenvector with max eigenvalue

1) Assume P is symmetric : $P_x(y) = P_y(x)$

(eg. G is a regular graph)

↳ linear algebra : Spectral theorem:

P has real eigenvalues $1 = \lambda_1 \geq \dots \geq \lambda_n$

with orthonormal eigenvectors $v = v_1, \dots, v_n \in \mathbb{R}^n$

convergence of power method characterized by

gap $\lambda_1 - \lambda_2 = 1 - \lambda_2$

2) What about P general? (P not sym $\Rightarrow$ complex eigenvalues)

Assume reversibility :

$$\tilde{P}(x,y) = \sqrt{v(x)} \cdot \frac{P(x,y)}{\sqrt{v(y)}} = \tilde{P}(y,x)$$

↑ symmetric

$\Rightarrow$ spectral theorem

(eg. $\tilde{P} = D^{-\frac{1}{2}} A D^{\frac{1}{2}}$

$\tilde{L} = I - \tilde{P} = I - D^{-\frac{1}{2}} A D^{\frac{1}{2}}$ normalized Laplacian)

this is equivalent to working $L^2(v)$

$L^2(\nu)$

Let ν be a probability distribution on $\mathcal{X}$

Define $L^2(\nu) = \{ f : \mathcal{X} \to \mathbb{R} \mid \int_{\mathcal{X}} f(x)^2 \, d\nu(x) < \infty \}$

This is a Hilbert space (complete inner product space)

inner product : $\langle f, g \rangle_\nu := \int_{\mathcal{X}} f(x) g(x) \, d\nu(x) = \int fg \, d\nu = \mathbb{E}_\nu[fg]$

squared norm : $\|f\|_\nu^2 = \langle f, f \rangle_\nu = \int_{\mathcal{X}} f(x)^2 \, d\nu(x) = \mathbb{E}_\nu[f^2]$

note: * $\mathbb{E}_\nu[f] = \int_{\mathcal{X}} f(x) \cdot 1 \, d\nu(x) = \langle f, \mathbb{1} \rangle_\nu$

* if $\mathbb{E}_\nu[f] = 0$, then $\mathrm{Var}_\nu(f) = \mathbb{E}_\nu[f^2] = \|f\|_\nu^2$

χ^2-divergence

How to measure distance between $\mathbb{P}$ and ν?

$\Rightarrow$ measure squared norm distance in $L^2(\nu)$

Density $h = \frac{d\mathbb{P}}{d\nu} : \mathcal{X} \to \mathbb{R} \quad (h \in L^2(\nu))$

$\mathbb{1} = \frac{d\nu}{d\nu} : \mathcal{X} \to \mathbb{R}$

$\left(\mathbb{E}_\nu[h] = \int_{\mathcal{X}} h \, d\nu = \int_{\mathcal{X}} d\mathbb{P} = \mathbb{P}(\mathcal{X}) = 1\right)$

Define: χ^2-divergence $\quad \chi^2_\nu(\mathbb{P}) = \| h - \mathbb{1} \|_\nu^2$

$= \mathbb{E}_\nu[(h(X) - 1)^2]$

$= \mathrm{Var}_\nu(h)$

$= \mathrm{Var}_\nu\left(\frac{d\mathbb{P}}{d\nu}\right)$

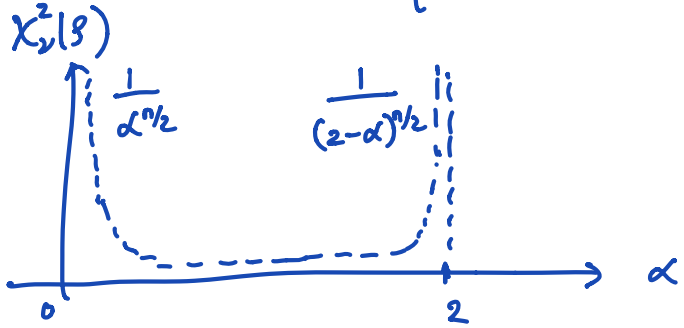
note: * not symmetric: $\chi^2_\nu(\mu) \neq \chi^2_\mu(\nu)$

* range of χ^2 can be large (exponential in n) $\leftarrow \chi = \mathbb{R}^n$

eg. $\mu = \mathcal{N}(0, \alpha I)$, $\alpha > 0$ on $\chi = \mathbb{R}^n$

$\nu = \mathcal{N}(0, I)$

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



Markov chain as linear operator

Let P be reversible wrt ν on X.

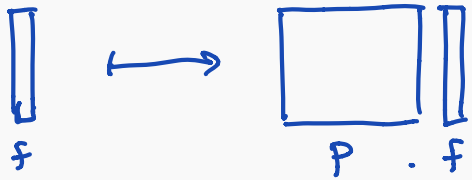
P defines a linear operator:

$$P : L^2(\nu) \longrightarrow L^2(\nu)$$

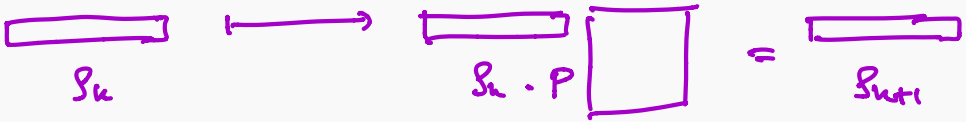
$$f \longmapsto Pf$$

given by $(Pf)(x) = \int_X f(y)\, dP_x(y) = \mathbb{E}[f(X_1) \mid X_0 = x]$

notes: * in discrete space, this is matrix mult. on the left:



* recall for distributions, matrix mult. on the right:



Reversibility ⇒ Symmetry

Lemma
Assume P is reversible with respect to ν. Then P is self-adjoint: $: L^2(\nu) \to L^2(\nu)$

$$\forall f, g \in L^2(\nu): \qquad \langle f, Pg \rangle_\nu = \langle Pf, g \rangle_\nu$$

Proof:
$$\langle f, Pg \rangle_\nu = \int f(x)\,(Pg)(x)\,d\nu(x)$$
$$= \iint f(x)\,g(y)\,\underbrace{dP_x(y)\,d\nu(x)}$$
$$= dP_y(x)\,d\nu(y) \text{ by reversibility}$$
$$= \iint f(x)\,g(y)\,dP_y(x)\,d\nu(y)$$
$$= \int g(y)\,(Pf)(y)\,d\nu(y)$$
$$= \langle Pf, g \rangle_\nu$$

□

Spectral theorem

Theorem
A self-adjoint operator P on $L^2(\nu)$ has real eigenvalues $\lambda_i \in \mathbb{R}$:

$$P\phi_i = \lambda_i \phi_i$$

for $i = 1, 2, \ldots, n$ ($n = \infty$) with an orthonormal basis of eigenfunctions $\phi_i \in L^2(\nu)$:

$$\langle \phi_i, \phi_j \rangle_\nu = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

- We assume $L^2(\nu)$ is separable (has countable basis); see [BGL, Appendix A.4][1] for general/continuous spectrum
- e.g. $\mathcal{X} = \mathbb{R}^n$, $\nu = \mathcal{N}(0, I)$: Eigenfunctions are Hermite polynomials

Courant-Fischer characterization of eigenvalues

Theorem
Assume P is self-adjoint with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ ($n = \infty$) and corresponding eigenfunctions $\phi_1, \ldots, \phi_n \in L^2(\nu)$. Then:

$$\lambda_1 = \max_{0 \neq f \in L^2(\nu)} \frac{\langle f, Pf \rangle_\nu}{\langle f, f \rangle_\nu}$$

$$\lambda_2 = \max_{\substack{0 \neq f \in L^2(\nu) \\ f \perp \phi_1 \\ (\langle f, \phi_1 \rangle_\nu = 0)}} \frac{\langle f, Pf \rangle_\nu}{\langle f, f \rangle_\nu}$$

$$\lambda_n = \min_{0 \neq f \in L^2(\nu)} \frac{\langle f, Pf \rangle_\nu}{\langle f, f \rangle_\nu}$$

Bounds on eigenvalues

Lemma
Assume P is reversible with respect to ν. For all $f \in L^2(\nu)$:

✓ (a) $\langle f, Pf\rangle_\nu \le \langle f, f\rangle_\nu$

✓ (b) $\langle f, Pf\rangle_\nu \ge -\langle f, f\rangle_\nu$

write $-I \preceq P \preceq I$

($\Leftrightarrow \forall f: -\langle f,f\rangle_\nu \le \langle f, Pf\rangle_\nu \le \langle f,f\rangle_\nu$)

Proof: Write

$$\langle f,f\rangle_\nu = \frac{1}{2}\int_\chi f(x)^2 \, d\nu(x) + \frac{1}{2}\int_\chi f(y)^2 \, d\nu(y)$$

$$= \frac{1}{2}\int_\chi\int_\chi f(x)^2 \, dP_x(y) \, d\nu(x) + \frac{1}{2}\iint_{\chi\times\chi} f(y)^2 \, dP_y(x) \, d\nu(y)$$

(equal by rev.)

$$= \frac{1}{2}\int_\chi\int_\chi \left(f(x)^2 + f(y)^2\right) dP_x(y) \, d\nu(x)$$

Also, $\langle f, Pf\rangle_\nu = \int_\chi\int_\chi f(x) f(y) \, dP_x(y) \, d\nu(x)$

a) $\langle f, f\rangle_\nu - \langle f, Pf\rangle_\nu = \frac{1}{2}\int_X\int_X \left(f(x)^2 + f(y)^2 - 2f(x)f(y)\right) dP_x(y)\, d\nu(x)$

$= \langle f, (I-P)f\rangle_\nu \quad = \frac{1}{2}\int_X\int_X (f(x)-f(y))^2\, dP_x(y)\, d\nu(x)$

$= \langle f, Lf\rangle_\nu \quad \geq 0$

b) $\langle f, f\rangle_\nu + \langle f, Pf\rangle_\nu = \frac{1}{2}\int_X\int_X (f(x)+f(y))^2\, dP_x(y)\, d\nu(x)$

≥ 0

□

So all eigenvalues of P satisfy:

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

$\lambda_2 < 1$ if P is irreducible

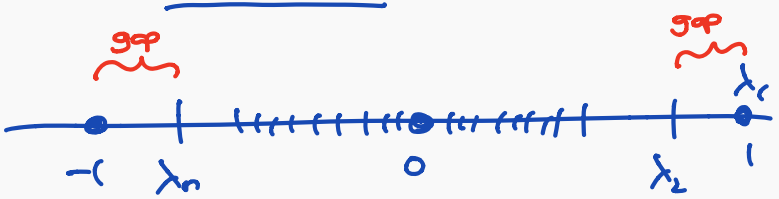
$\lambda_n = -1$ iff bipartite / periodic

Spectral Gap

P has eigenvalues

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > -1$$

Def: (Absolute) Spectral gap: $\gamma^* = \min\{1 - \lambda_2, 1 + \lambda_n\}$



so all $|\lambda_i| \leq 1 - \gamma^*$

Note: Can shift all $\lambda_i \geq 0$ by lazifying P:

$$\hat{P} = \frac{1}{2} I + \frac{1}{2} P$$

(at each step:
- w.p. $\frac{1}{2}$, stay
- w.p. $\frac{1}{2}$, jump following P)

then $\hat{P} \geq 0$

(since $\langle f, \hat{P} f \rangle_\nu = \frac{1}{2}\langle f, f \rangle_\nu + \frac{1}{2}\langle f, Pf \rangle_\nu$

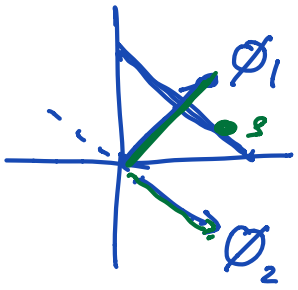
≥ 0

Def: Spectral Gap: $\gamma = 1 - \lambda_2$

($\gamma = \gamma^*$ if P is lazy)

$$\pi_1 = \pi_0 \cdot P$$

$$\pi_t(y) = \sum_x \underbrace{\pi_0(x)}_{\geq 0} \cdot \underbrace{P_x(y)}_{\geq 0} \geq 0$$



P has max eigenvalue $\lambda_1 = 1$

with eigenvector $\phi_1 = \mathbb{1}$

Laplacian

Def: Laplacian $L = I - P : L^2(\nu) \to L^2(\nu)$

P reversible wrt. ν $\Rightarrow$ L is symmetric, positive semidefinite:

Dirichlet form $\mathcal{E}(f,f) := \langle f, Lf \rangle_\nu$

$$= \frac{1}{2} \mathbb{E}_{(\nu,P)} \left[(f(X_1) - f(X_0))^2 \right] \geq 0$$

$X_0 \sim \nu$

$X_1 | X_0 \sim P_{X_0}$

$$L \geq 0$$

eigenvalues $0 = \lambda_1 < \lambda_2 \leq \dots \leq \lambda_n \ (< 2)$

eigenfunction $\mathbb{1} = \phi_1, \phi_2, \dots, \phi_n$

Spectral Gap: $\gamma = \lambda_2(L) = \min_{\substack{f \in L^2(\nu) \\ \langle f, \mathbb{1} \rangle_\nu = 0}} \frac{\langle f, Lf \rangle_\nu}{\langle f, f \rangle_\nu} = \min_{f \in L^2(\nu)} \frac{\mathcal{E}(f,f)}{\mathrm{Var}_\nu(f)}$

Mixing time in χ^2-divergence

Theorem
Let P be a Markov chain reversible with respect to ν with spectral gap γ $(= \gamma^*)$. For any $X_0 \sim \rho_0$, along the Markov chain $X_k \sim \rho_k$:

$$\chi^2_\nu(\rho_k) \leq (1-\gamma)^{2k} \chi^2_\nu(\rho_0)$$

$\Rightarrow$ Corollary: To reach $\chi^2_\nu(\rho_k) \leq \varepsilon$, enough to take

mixing time in χ^2_ν :

$$\tau(\varepsilon, \rho_0) = \frac{1}{2\gamma} \log \frac{\chi^2_\nu(\rho_0)}{\varepsilon}$$

want ↓

because $\chi^2_\nu(\rho_k) \leq (1-\gamma)^{2k} \chi^2_\nu(\rho_0) \leq e^{-2\gamma k} \chi^2_\nu(\rho_0) \leq \varepsilon$

$$\Leftrightarrow \quad k \geq \frac{1}{2\gamma} \log \frac{\chi^2_\nu(\rho_0)}{\varepsilon}$$

eg, ρ_0, ν Gaussin, $\chi^2_\nu(\rho_0) \sim \frac{1}{\alpha^n}$, so $\log \chi^2_\nu(\rho_0) = O(n)$

Proof

$X_n \sim \rho_n, \qquad X_{n+1} \sim \rho_{n+1}$

density $h_n = \frac{d\rho_n}{d\nu}$, $h_{n+1} = \frac{d\rho_{n+1}}{d\nu} \in L^2(\nu)$

Lemma: P reversible wrt $\nu \Rightarrow \boxed{h_{n+1} = P h_n}$

$\Rightarrow$ then $h_n = P^k h_0$

do eigenvalue analysis of $P \Rightarrow \chi^2_\nu(\rho_n) = \mathrm{Var}_\nu(h_n) = \| h_n - \mathbb{1} \|^2_\nu$