CPSC 661: Sampling Algorithms in ML

Andre Wibisono

March 1, 2021

Yale University

Last time

- Reversible Markov chain $\nu(x) \cdot P_x(y) = \nu(y) \cdot P_y(x)$
- s-Conductance $\Rightarrow$ Mixing time in TV distance

Φ_s $\qquad \tau(\varepsilon) = \frac{2}{\Phi_s^2} \log \frac{2M}{\varepsilon}, \quad s = \frac{\varepsilon}{2M}, \quad M = M_\nu^\infty(\xi_0)$

- Metropolis-Hastings : accept - reject step

from x, go to $x' = \begin{cases} y \sim P_x & \text{with prob } \min\left\{1, \frac{\nu(y) \cdot P_y(x)}{\nu(x) \cdot P_x(y)}\right\} = a_x(y) \\ x & \text{with prob } 1 - a_x(y) \end{cases}$

- Isoperimetry of ν

$$\psi = \inf_{S \subset X} \frac{\nu(\partial S)}{\min\{\nu(S), \nu(S^c)\}}$$

Today:

- From isoperimetry to conductance
- Metropolis Random Walk (MRW)

References

- Vempala, *Geometric Random Walk: A Survey*, Combinatorial and Computational Geometry, 2005

- Dwivedi, Chen, Wainwright, and Yu, *Log-Concave Sampling: Metropolis-Hastings Algorithms are Fast*, Journal of Machine Learning Research, 2019

Strongly Log-Concave Distribution

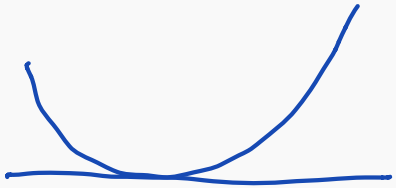
Recall $\nu \propto e^{-f}$ on $\mathbb{R}^n$ is α- strongly log-concave (α-SLC)

if $f = -\log \nu : \mathbb{R}^n \to \mathbb{R}$ is α-strongly convex

$\Leftrightarrow$ $\boxed{\nabla^2 f(x) \succeq \alpha I}$ $\forall\ x \in \mathbb{R}^n$

(Recall this means all eigenvalues λ of $\nabla^2 f(x)$ are lower bounded:

$\lambda \geq \alpha$)



- example: $\nu = N(\mu, \Sigma)$, $\mu \in \mathbb{R}^n$, $\Sigma \succ 0$

$$\nu(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}}\, e^{-\frac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}}$$

$$f(x) = \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) + \frac{1}{2}\log\det(2\pi\Sigma)$$

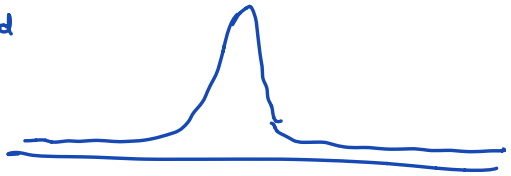
$$\nabla^2 f(x) = \Sigma^{-1} \succeq \frac{1}{\lambda_{max}(\Sigma)} I \quad \leftarrow \quad \alpha = \frac{1}{\lambda_{max}(\Sigma)}$$

large Σ $\Rightarrow$ ν is flat

$\Rightarrow$ less SLC

small Σ $\Rightarrow$ ν is peaked

$\Rightarrow$ more SLC



1. Isoperimetry

Theorem
If ν is α-strongly log-concave on $\mathbb{R}^n$, then it satisfies isoperimetry with constant:

$$\psi \geq (\log 2)\sqrt{\alpha} \geq \frac{\sqrt{\alpha}}{4}$$

- Cousins & Vempala, *Gaussian Cooling and $O^*(n^3)$ Algorithms for Volume and Gaussian Volume*, SIAM Journal of Computing, 2018 (Theorem 5.4)
- Proof by localization: Reduce problem from n dimension ⟶ 1 dimension
- Nice: dimension free
  (cf. KLS conjecture for weakly log-concave)

2. Effective Diameter

Prob distribution $\nu \propto e^{-f}$ on $\mathbb{R}^n$

- Mode: $x^* = \arg\max_{x \in \mathbb{R}^n} \nu(x)$

$= \arg\min_{x \in \mathbb{R}^n} f(x)$

[Convex optimization: if f is α-strongly convex, L-smooth

$$\alpha I \preceq \nabla^2 f(x) \preceq L I$$

Define condition number

$$\kappa = \frac{L}{\alpha} \geq 1$$

- Gradient descent can find x^* in time $\tilde{O}(\kappa)$
- Accelerated gradient descent can find x^* in time $\tilde{O}(\sqrt{\kappa})$ (this is optimal)]

- **mean**: $\bar{x} = \mathbb{E}_\nu[X] = \dfrac{\int_{\mathbb{R}^n} x\, e^{-f(x)}\, dx}{\int_{\mathbb{R}^n} e^{-f(x)}\, dx}$

Lemma: Suppose ν is α-SLC on $\mathbb{R}^n$. Then:

1) $\mathrm{Var}_\nu(X) = \mathbb{E}_\nu[\|X - \bar{x}\|_2^2] \le \frac{n}{\alpha}$
2) $\mathbb{E}_\nu[\|X - x^*\|_2^2] \le \frac{n}{\alpha}$
3) $\|\bar{x} - x^*\|_2^2 \le \frac{n}{\alpha}$

$\therefore$ effective diameter $\sim \sqrt{\frac{n}{\alpha}}$

- **example**: $\nu = \mathcal{N}(\mu, \Sigma)$, α-SLC with $\alpha = \dfrac{1}{\lambda_{\max}(\Sigma)}$

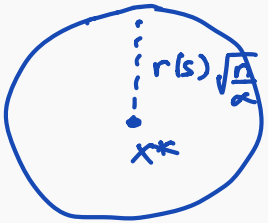
$$\mathrm{Var}_\nu(X) = \mathrm{Tr}(\Sigma) = \sum_{i=1}^{n} \lambda_i(\Sigma) \le n \cdot \lambda_{\max}(\Sigma) = \frac{n}{\alpha}$$

3. Concentration

For $s > 0$, let $r(s) = 2 + 2\max\left\{\left(\frac{1}{n}\log\frac{1}{s}\right)^{\frac{1}{4}}, \left(\frac{1}{n}\log\frac{1}{s}\right)^{\frac{1}{2}}\right\}$

(note:
- if $s \geq e^{-n}$ then $\frac{1}{n}\log\frac{1}{s} \leq 1 \Rightarrow r(s) \leq 4$
- if $s < e^{-n}$ then $r(s) \sim \left(\frac{1}{n}\log\frac{1}{s}\right)^{\frac{1}{2}}$
- if $s = e^{-cn}$ then $r(s) \sim \sqrt{c}$)

Let $\mathcal{R}_s = \mathbb{B}\left(x^*, r(s)\sqrt{\frac{n}{\alpha}}\right)$



Lemma: If ν is α-SLC on $\mathbb{R}^n$

then $\nu(\mathcal{R}_s) \geq 1 - s$

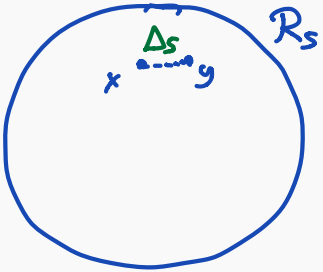
[DCWY '19, Lemma 1]

From isoperimetry to conductance

Lemma (DCWY'19 Lemma 2)
Suppose ν is α-strongly log-concave on $\mathbb{R}^n$. Fix $0 < s \le \frac{1}{2}$. Let P be a Markov chain with stationary distribution ν and satisfying the following **one-step overlap** *property:*

$$TV(P_x, P_y) \le \frac{3}{4} \qquad (\text{or } \le 1-c \text{ for some } c>0)$$

for all $x, y \in \mathcal{R}_s$ with $\|x - y\|_2 \le \Delta_s$. Then P has s-conductance:



$$\phi_s \ge \min\left\{\frac{1}{16}, \frac{\sqrt{\alpha}\,\Delta_s}{128}\right\} = \Omega\left(\sqrt{\alpha}\cdot\Delta_s\right)$$

Corollary: Mixing time in TV: $\tau(\varepsilon) = \tilde{O}\left(\frac{1}{\phi_s^2}\right) = \tilde{O}\left(\frac{1}{\alpha\cdot\Delta_s^2}\right)$

Proof: Let $A \subset \mathbb{R}^n$ with $s < \nu(A) < 1-s$.

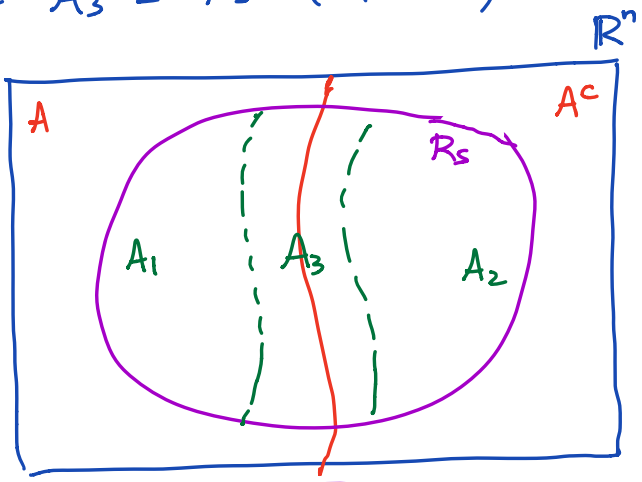
We will bound $C(A) = \dfrac{\Phi(A)}{\min\{\nu(A)-s,\ \nu(A^c)-s\}}$

(1) Consider points which are "deep inside" and unlikely to cross over

$$A_1 = \{ x \in A \cap R_s \mid P_x(A^c) < \tfrac{1}{8} \}$$

$$A_2 = \{ y \in A^c \cap R_s \mid P_y(A) < \tfrac{1}{8} \}$$

and define $A_3 = R_s \setminus (A_1 \cup A_2)$



Claim: $\boxed{d(A_1, A_2) \geq \Delta_s}$ ... (1)

Proof of claim: $\forall\ x \in A_1,\ y \in A_2$:

$$TV(P_x, P_y) = \sup_{B \subset \mathbb{R}^n} |P_x(B) - P_y(B)|$$

$$\geq P_x(A) - P_y(A)$$

$$= 1 - P_x(A^c) - P_y(A)$$

$$\geq 1 - \tfrac{1}{8} - \tfrac{1}{8} = \tfrac{3}{4}$$

So by (the contrapositive of) the one-step overlap:

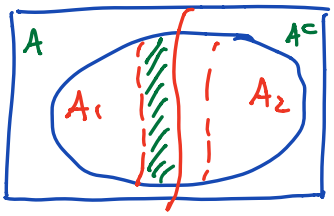
$$\|x-y\|_2 > \Delta_s$$

Then: $d(A_1, A_2) = \inf_{\substack{x\in A_1 \\ y\in A_2}} \|x-y\|_2 \geq \Delta_s.$ □

---

② Claim: if $\nu(A_1) \leq \frac{\nu(A\cap R_s)}{2}$, then $C(A) \geq \frac{1}{16}$

Proof:



$$\Phi(A) = \int_A P_x(A^c)\, d\nu(x)$$

$$\geq \int_{(A\cap R_s)\setminus A_1} \underbrace{P_x(A^c)}_{\geq \frac{1}{8}}\, d\nu(x)$$

$$\geq \frac{1}{8}\,\nu((A\cap R_s)\setminus A_1)$$

$$\geq \frac{1}{16}\cdot \nu(A\cap R_s) \qquad \text{since } \nu(A_1) \leq \frac{\nu(A\cap R_s)}{2}$$

$$= \frac{1}{16}\left(\nu(A) - \nu(A\cap R_s^c)\right)$$

$$\geq \frac{1}{16}\left(\nu(A) - \nu(R_s^c)\right)$$

$$\geq \frac{1}{16}\left(\nu(A) - s\right) \qquad \text{since } \nu(R_s) \geq 1-s$$

$$\geq \frac{1}{16}\min\left\{\nu(A) - s,\ \nu(A^c) - s\right\}$$

$$\Rightarrow C(A) \geq \frac{1}{16}.$$

□

Similarly, if $\nu(A_2) \leq \dfrac{\nu(A^c \cap R_s)}{2}$, then $C(A) \geq \frac{1}{16}$.

③ Now assume:

$$\nu(A_1) \geq \frac{\nu(A \cap R_s)}{2} \geq \frac{\nu(A) - s}{2}$$

$$\nu(A_2) \geq \frac{\nu(A^c \cap R_s)}{2} \geq \frac{\nu(A^c) - s}{2} \quad \ldots(2)$$

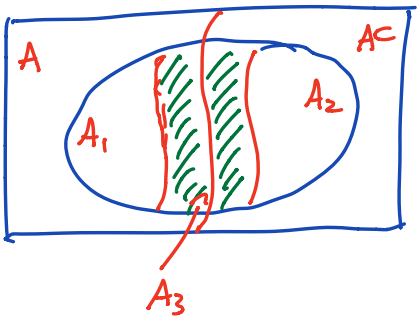
Now compute:

$$\Phi(A) = \frac{1}{2}\left(\Phi(A) + \Phi(A^c)\right)$$

$$= \frac{1}{2}\left(\int_A P_x(A^c)\, d\nu(x) + \int_{A^c} P_x(A)\, d\nu(x)\right)$$

$$\geq \frac{1}{2}\left(\int_{(A \cap R_s) \setminus A_1} \underbrace{P_x(A^c)}_{\geq \frac{1}{8}}\, d\nu(x) + \int_{(A^c \cap R_s) \setminus A_2} \underbrace{P_x(A)}_{\frac{1}{8}}\, d\nu(x)\right)$$

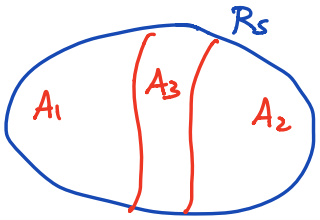
$$\geq \frac{1}{16}\,\nu(A_3) \quad \ldots [3)$$



④ Let $\nu_s = \nu|_{R_s}$ be restriction of ν on R_s

$$\nu_s(B) = \frac{\nu(B \cap R_s)}{\nu(R_s)}$$

Nice: ν is α-SLC, R_S is convex set $\Big\}$ ν_S is also α-SLC



By isoperimetry for ν_S:

for partition $R_S = A_1 \cup A_2 \cup A_3$

$$\nu_S(A_3) \geq \frac{\sqrt{\alpha}}{4} \cdot d(A_1, A_2) \cdot \min\{\nu_S(A_1), \nu_S(A_2)\}$$

$$\Leftrightarrow \nu(A_3) \geq \frac{\sqrt{\alpha}}{4} \cdot d(A_1, A_2) \cdot \min\{\nu(A_1), \nu(A_2)\}$$

$$\overset{(1)}{\geq} \frac{\sqrt{\alpha}}{4} \cdot \Delta_S \cdot \min\{\nu(A_1), \nu(A_2)\}$$

$$\overset{(2)}{\geq} \frac{\sqrt{\alpha} \cdot \Delta_S}{8} \cdot \min\{\nu(A) - s, \nu(A^c) - s\}$$

Plugging in to (3):

$$\overline{\Phi}(A) \geq \frac{1}{16} \cdot \nu(A_3)$$

$$\geq \frac{\sqrt{\alpha} \cdot \Delta_S}{128} \cdot \min\{\nu(A) - s, \nu(A^c) - s\}$$

$$\Rightarrow C(A) \geq \frac{\sqrt{\alpha} \cdot \Delta_S}{128}.$$

□

Recap

To sample from ν on $\mathbb{R}^n$:

1. Start from any Markov chain P
2. Apply Metropolis-Hastings filter to get $\tilde{P}$ reversible wrt ν
3. Assume ν is α-SLC, so isoperimetric with $\psi = \Omega(\sqrt{\alpha})$
4. Show $\tilde{P}$ satisfies one-step overlap property:

$$x, y \in \mathcal{R}_s, \|x - y\|_2 \leq \Delta_s \quad \Rightarrow \quad \mathrm{TV}(\tilde{P}_x, \tilde{P}_y) \leq \frac{3}{4}$$

$\Rightarrow$ $\tilde{P}$ has s-conductance $\phi_s = \Omega(\sqrt{\alpha}\,\Delta_s)$

$\Rightarrow$ mixing time in TV distance: $\tau(\epsilon) = \tilde{O}\left(\frac{1}{\phi_s^2}\right) = \tilde{O}\left(\frac{1}{\alpha\,\Delta_s^2}\right)$

What random walk?

1. P = Brownian motion (Gaussian walk)

   $\Rightarrow \tilde{P}$ = Metropolis Random Walk (MRW)

   (Today)

2. P = Unadjusted Langevin Algorithm (ULA)

   $\Rightarrow \tilde{P}$ = Metropolis-Adjusted Langevin Algorithm (MALA)

   (Next time)

Kullback-Leibler (KL) Divergence

Let ρ and ν be probability distributions on $\mathcal{X}$.

The **Kullback-Leibler (KL) divergence** of ρ with respect to ν is

$$H_\nu(\rho) = \int_{\mathcal{X}} \rho(x) \log \frac{\rho(x)}{\nu(x)} \, dx$$

Notes:

- Also called Relative entropy
- Relative form of Shannon entropy $H(\delta) = -\int_{\mathcal{X}} \delta(x) \log \delta(x) \, dx$
- Non-negative: $H_\nu(\delta) \geq 0 \quad \forall \, \delta, \nu$

  $H_\nu(\delta) = 0 \iff \delta = \nu$
- Not a metric: not symmetric

  does not satisfy triangle-inequality

- If $\mathcal{S}$ has density $h = \frac{d\mathcal{S}}{d\nu} : \mathcal{X} \to \mathbb{R}$ wrt ν

  then $H_\nu(\mathcal{S}) = \mathbb{E}_\nu[h(x) \log h(x)]$

- Recall: $TV(\mathcal{S}, \nu) = \frac{1}{2}\mathbb{E}_\nu[|h(x) - 1|]$

- Pinsker's Inequality: $\boxed{TV(\mathcal{S}, \nu) \leq \sqrt{\frac{1}{2} H_\nu(\mathcal{S})}}$