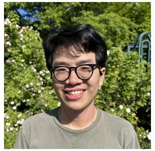
Fit Like You Sample: Sample-Efficient Generalized Score Matching from Fast Mixing Diffusions



Yilong Qin

Carnegie Mellon University
ML MACHINE LEARNING DEPARTMENT

Andrej Risteski

Carnegie Mellon University
ML MACHINE LEARNING DEPARTMENT

What is it about?

Our Question:

What makes Annealed Score Matching (diffusion model) so successful?

Specifically, why is Annealed Score Matching so much better than Vanilla Score Matching?

Our Approach:

We tackle this question from the perspective of asymptotic sample complexity.

We show that for multimodal distributions, the sample complexity of annealed score matching is polynomial in problem parameters.

The previous known bound is exponential for vanilla score matching (*KHR 2022*).

Score Matching

[*H2005*] Score Matching objective

$$D_{SM}\left(p,q\right)=\frac{1}{2}\mathbb{E}_p\left\|\nabla_x\log p-\nabla_x\log q\right\|_2^2$$
$$=\frac{1}{2}\mathbb{E}_p\left\|\frac{\nabla_x p}{p}-\frac{\nabla_x q}{q}\right\|_2^2$$

[*L2012*] Generalized Score Matching objective replaces the gradient with an arbitrary linear operator

$$D_{GSM}\left(p,q\right)=\frac{1}{2}\mathbb{E}_p\left\|\frac{\mathcal{O}p}{p}-\frac{\mathcal{O}q}{q}\right\|_2^2$$

[*SE2019*] Annealed Score Matching objective

$$\mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p^\beta}\|\nabla_x\log p(x|\beta)-\nabla_x\log p_\theta(x|\beta)\|^2$$

is a weighted sum of score matching objectives at different temperatures beta, with each distribution tempered with different amount of noise.

Annealing process (Image Credit: Yang Song’s blog)

Asymptotic Sample Complexity

[*V2000*] Under suitable conditions, the empirical estimator converges to the following asymptotic normal distribution by an extension to the Central Limit Theorem

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, (\nabla_\theta^2 L(\theta^*))^{-1} Cov(\nabla_\theta \ell(x;\theta^*))(\nabla_\theta^2 L(\theta^*))^{-1}\right)$$

Interpretation:

The “larger” the asymptotic covariance, the more sample is required to ensure the estimator is sufficiently close to the true parameter.

We will use the operator norm of the asymptotic covariance to quantify the sample complexity.

Multimodality via Markov Chain

Intuition:

The multimodality of a distribution could be characterized by the mixing time of a Markov Chain (e.g. Langevin Dynamics) that converges to the distribution as its stationary distribution. For multimodal distributions, the chain suffers from slow mixing.

[*Definition 3*] Markov Semigroup of a Markov Process

$$\mathbf{E}[f(X_{t+s})|\{X_r\}_{r\leq t}] = (P_s f)(X_t).$$

[*Definition 4*] Infinitesimal Generator

$$\mathcal{L}g = \lim_{t\to 0}\frac{P_t g - g}{t}.$$

[*Definition 5*] Poincaré constant is defined as the rate of convergence (i.e. mixing time) in chi-square divergence

$$\chi^2(p_t, p) \leq e^{-2t/C_P}\chi^2(p_0, p).$$

Equivalently, it is the smallest C that satisfies the Poincaré Inequality for all g

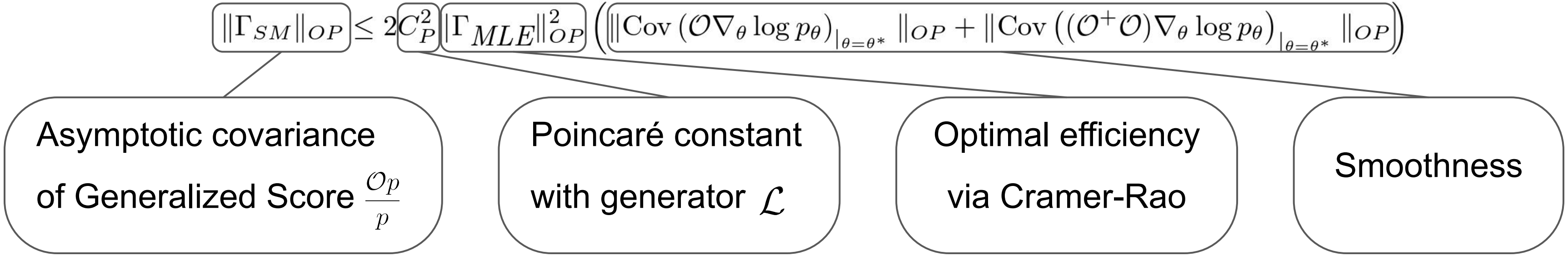
$$\mathcal{E}(g) \geq \frac{1}{C}\mathrm{Var}_p(g).$$

where the Dirichlet form is given by

$$\mathcal{E}(g) = -\mathbb{E}_p\langle g, \mathcal{L}g\rangle.$$

Framework for Statistical Gap

We extend the results from *KHR (2022)* for exponential family: Under conditions (asymptotic normality, realizability) the operator norm of the asymptotic covariance of the Generalized Score Matching objective can be upper bounded by the following bound (Informal Theorem 3)



$$\boxed{\|\Gamma_{SM}\|_{OP}} \leq 2\boxed{C_P^2}\boxed{\|\Gamma_{MLE}\|_{OP}^2}\left(\boxed{\|\mathrm{Cov}\left(\mathcal{O}\nabla_\theta \log p_\theta\right)_{|\theta=\theta^*}\|_{OP} + \|\mathrm{Cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\right)_{|\theta=\theta^*}\|_{OP}}\right)$$

Asymptotic covariance of Generalized Score $\frac{\mathcal{O}p}{p}$

Poincaré constant with generator $\mathcal{L}$

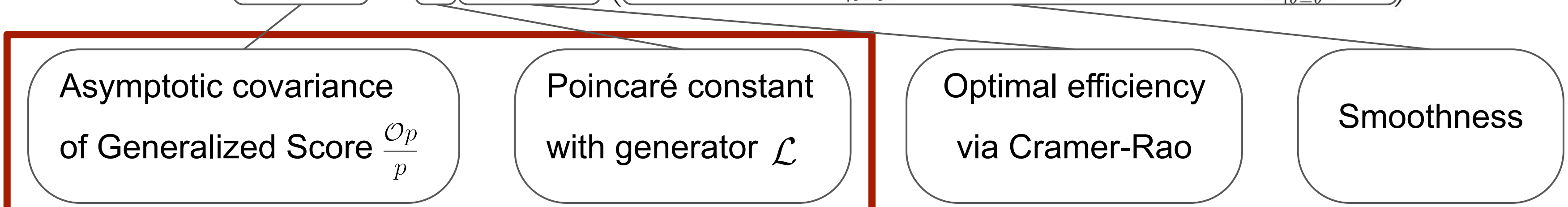
Optimal efficiency via Cramer-Rao

Smoothness

Framework for Statistical Gap

We extend the results from *KHR (2022)* for exponential family: Under conditions (asymptotic normality, realizability) the operator norm of the asymptotic covariance of the Generalized Score Matching objective can be upper bounded by the following bound (Informal Theorem 3)

$$\boxed{\|\Gamma_{SM}\|_{OP}} \leq 2\boxed{C_P^2}\boxed{\|\Gamma_{MLE}\|_{OP}^2}\left(\boxed{\left\|\mathrm{Cov}\left(\mathcal{O}\nabla_\theta \log p_\theta\right)_{|\theta=\theta*}\right\|_{OP} + \left\|\mathrm{Cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\right)_{|\theta=\theta*}\right\|_{OP}}\right)$$



Asymptotic covariance of Generalized Score $\frac{\mathcal{O}p}{p}$

Poincaré constant with generator $\mathcal{L}$

Optimal efficiency via Cramer-Rao

Smoothness

For a Markov Chain with some generator, its mixing time determines the sample complexity of a corresponding SM objective.

Generalized Score and Markov Chains

For a Markov Chain with some generator, its mixing time determines the sample complexity of a corresponding SM objective.

[*Theorem 3*] For **every** continuous time Markov Chain, the corresponding SM objective is in fact a pre-conditioned score loss:

$$\frac{1}{2}\mathbb{E}_p \left\| \sqrt{D(x)}\,(\nabla_x \log p - \nabla_x \log q) \right\|_2^2$$

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \boxed{(\nabla_\theta^2 L(\theta^*))^{-1}} Cov(\nabla_\theta \ell(x;\theta^*)) \boxed{(\nabla_\theta^2 L(\theta^*))^{-1}}\right)$$

Key ingredients:

[*MCF 2015*] For every continuous time Markov Chain with stationary distribution, it admits the following Ito's diffusion representation:

$$dX_t = (-(D(X_t) + Q(X_t))\nabla f(X_t) + \Gamma(X_t))\,dt + \sqrt{2D(X_t)}dB_t$$

where D is PSD, Q is skew-symmetric, and

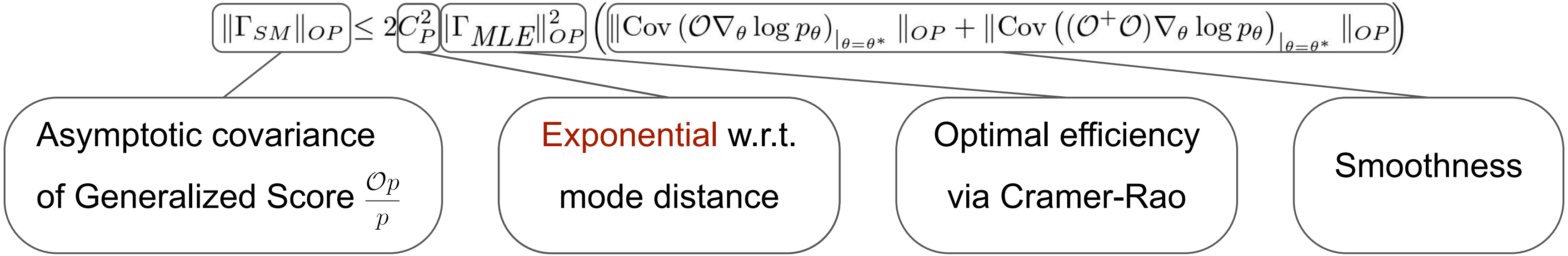
$$\Gamma_i(x) := \sum_j \partial_j (D_{ij}(x) + Q_{ij}(x)).$$

[*Lemma 3*] The above Markov Chain has the following Dirichlet form:

$$\mathcal{E}(g) = \mathbb{E}_p \| \sqrt{D(x)} \nabla g(x) \|_2^2$$

which is closely related to the Hessian of the loss.

KHR (2022): Vanilla Score Matching - **slow** mixing



$$\boxed{\|\Gamma_{SM}\|_{OP}} \le 2\boxed{C_P^2}\boxed{\|\Gamma_{MLE}\|_{OP}^2}\left(\boxed{\left\|\mathrm{Cov}\left(\mathcal{O}\nabla_\theta \log p_\theta\right)_{|\theta=\theta^*}\right\|_{OP} + \left\|\mathrm{Cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\right)_{|\theta=\theta^*}\right\|_{OP}}\right)$$

Asymptotic covariance of Generalized Score $\frac{\mathcal{O}p}{p}$

Exponential w.r.t. mode distance

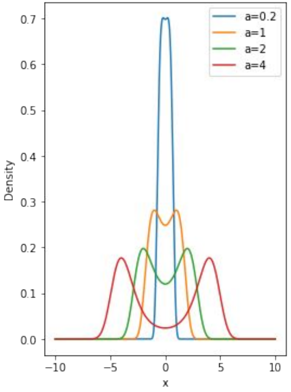
Optimal efficiency via Cramer-Rao

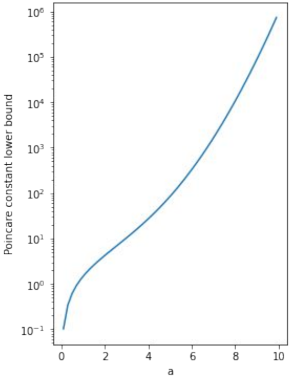
Smoothness

[*Example 2 from KHR (2022).*] For an exponential family distribution with sufficient statistic

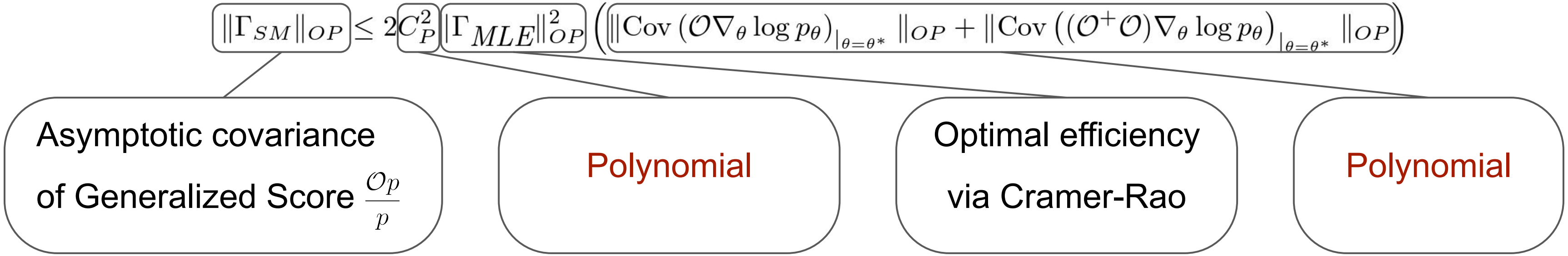
$$F_1(x) = -\frac{1}{8a^2}(x-a)^2(x+a)^2$$

The multiplicative gap is exponential w.r.t. a.





Ours: Annealed Score Matching - **fast** mixing



$$\boxed{\|\Gamma_{SM}\|_{OP}} \leq 2\boxed{C_P^2}\boxed{\|\Gamma_{MLE}\|_{OP}^2}\left(\boxed{\|\mathrm{Cov}\left(\mathcal{O}\nabla_\theta \log p_\theta\right)_{|\theta=\theta^*}\|_{OP} + \|\mathrm{Cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\right)_{|\theta=\theta^*}\|_{OP}}\right)$$

Asymptotic covariance of Generalized Score $\frac{\mathcal{O}p}{p}$

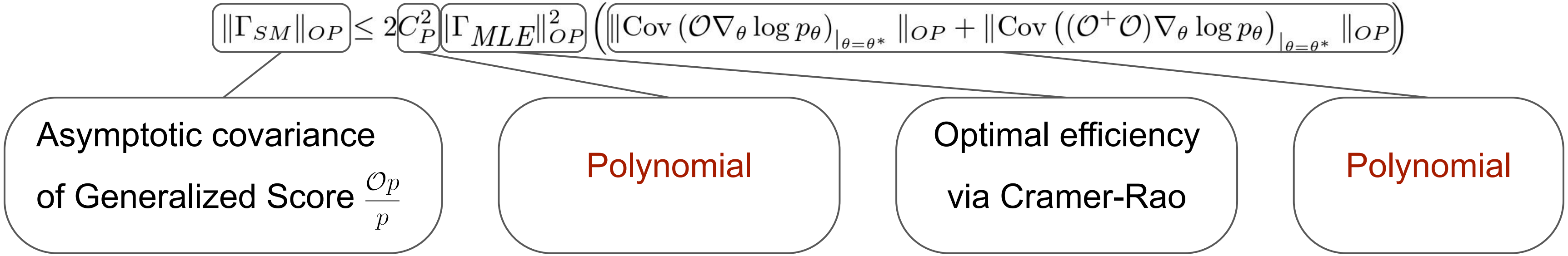
Polynomial

Optimal efficiency via Cramer-Rao

Polynomial

1. Annealed Score Matching (especially higher order) objective is very similar to Score Matching on an “lifted” state space augmented with temperature. We analyze Score Matching on this “lifted” space.
2. Under a Gaussian Mixture distribution (with shared covariance), the Score Matching loss (that arises from tempering dynamics) enjoys a poly(dimension, diameter of means, eigenvalues of covariance) bound under its natural parameterization.

Proof Outline



$$\boxed{\|\Gamma_{SM}\|_{OP}} \leq 2\boxed{C_P^2}\boxed{\|\Gamma_{MLE}\|_{OP}^2}\left(\boxed{\left\|\mathrm{Cov}\left(\mathcal{O}\nabla_\theta \log p_\theta\right)_{|\theta=\theta^*}\right\|_{OP} + \left\|\mathrm{Cov}\left((\mathcal{O}^+\mathcal{O})\nabla_\theta \log p_\theta\right)_{|\theta=\theta^*}\right\|_{OP}}\right)$$

Asymptotic covariance of Generalized Score $\frac{\mathcal{O}p}{p}$

Polynomial

Optimal efficiency via Cramer-Rao

Polynomial

Polynomial mixing time bound:

- Langevin Dynamics on the “lifted” (x, beta) distribution corresponds to an Annealed Score Matching loss.
- Mixing time analysis uses Markov Chain Decomposition Theorem *GLR (2018)*.

Polynomial Smoothness bound:

- Relate score of mixture with score of component via the convexity of perspective map.
- Single Gaussian derivative is given by the Hermite polynomial.

Continuously Tempered Langevin Dynamics

[*Definition 7*] We define Continuous Tempered Langevin Dynamics as the Langevin Dynamics over the joint (x, beta) state space, with reflection at the boundary of support of beta.

$$\begin{cases} dX_t = \nabla_x \log p^\beta(X_t)dt + \sqrt{2}dB_t \\ d\beta_t = \nabla_\beta \log r(\beta_t)dt + \nabla_\beta \log p^\beta(X_t)dt \\ \qquad + \nu_t L(dt) + \sqrt{2}dB_t \end{cases}$$

where

$$r(\beta) \propto \exp\left(-\frac{7D^2}{\lambda_{\min}(1+\beta)}\right) \textit{ and } \beta_{\max} = \frac{14D^2}{\lambda_{\min}} - 1.$$

and

$$p^\beta := p * \mathcal{N}(0, \beta\lambda_{\min} I_d)$$

[*Proposition 4*] The Score Matching loss on the joint state space satisfies

$$\left[\nabla_\theta^2 D_{GSM}(p, p_{\theta^*})\right]^{-1} \preceq C_P \Gamma_{MLE}$$

Moreover,

$$D_{GSM}(p, p_\theta)$$

SM on lifted dist.

$$= \mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p^\beta}(\|\nabla_x \log p(x,\beta) - \nabla_x \log p_\theta(x,\beta)\|^2$$
$$+ \|\nabla_\beta \log p(x,\beta) - \nabla_\beta \log p_\theta(x,\beta)\|^2)$$

Annealed SM

$$= \mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p^\beta}\|\nabla_x \log p(x|\beta) - \nabla_x \log p_\theta(x|\beta)\|^2$$

Higher order SM

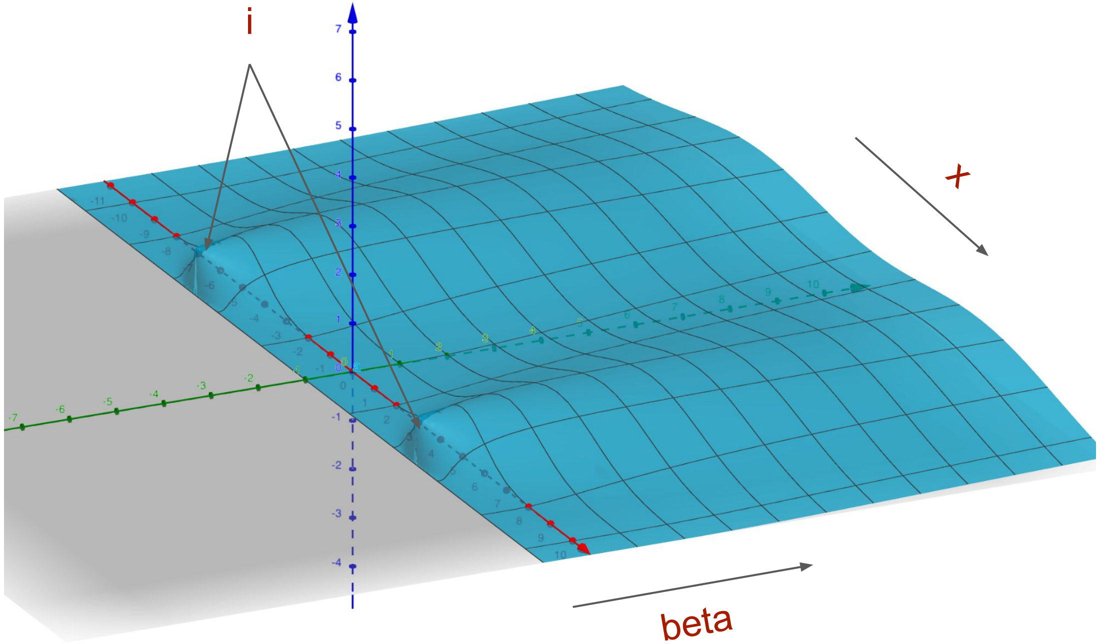
$$+ \lambda_{\min}\mathbb{E}_{\beta\sim r(\beta)}\mathbb{E}_{x\sim p^\beta}$$
$$\left(\left(\operatorname{Tr} \nabla_x^2 \log p(x|\beta) - \operatorname{Tr} \nabla_x^2 \log p_\theta(x|\beta)\right)\right.$$
$$\left. + \left(\|\nabla_x \log p(x|\beta)\|_2^2 - \|\nabla_x \log p_\theta(x|\beta)\|_2^2\right)^2\right.$$

Lifted distribution $p(x, \beta, i)$

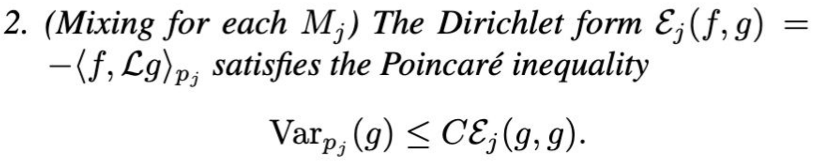
x: original state space

beta: temperature

i: mode index



Fast Mixing within mode $p(x, \beta|i)$ across temperatures

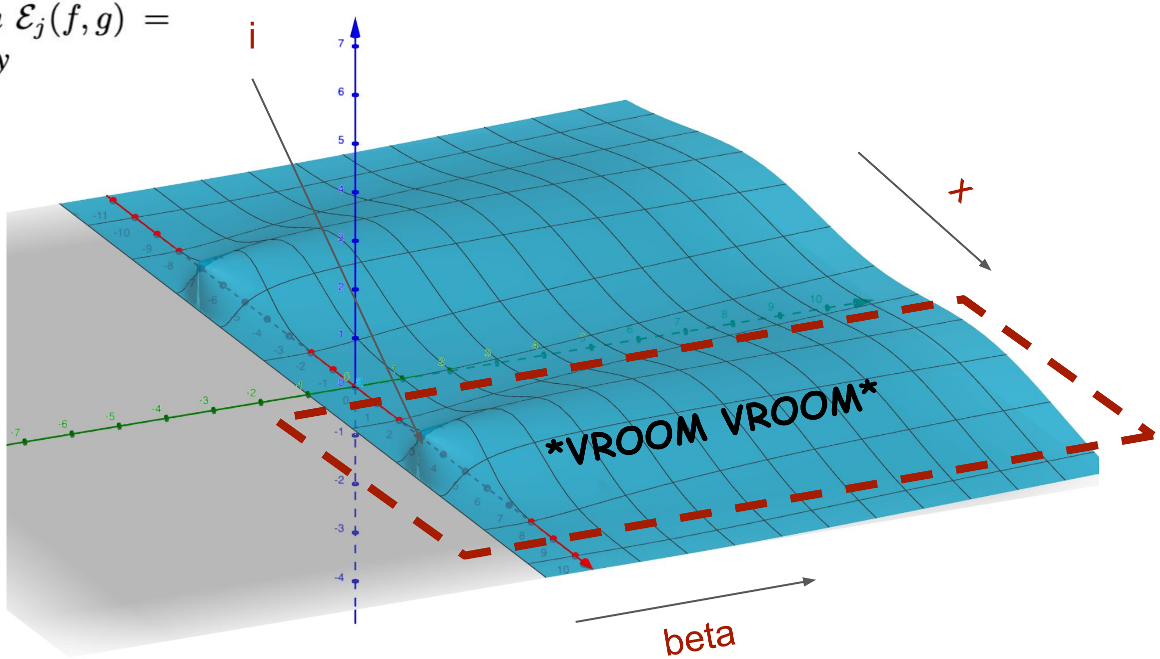


2. *(Mixing for each M_j) The Dirichlet form $\mathcal{E}_j(f, g) = -\langle f, \mathcal{L}g\rangle_{p_j}$ satisfies the Poincaré inequality*

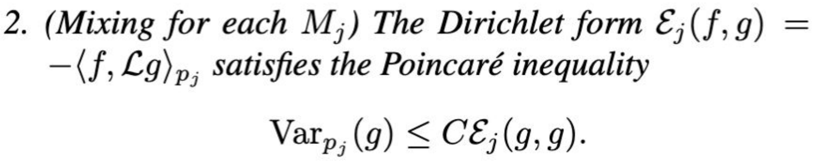
$$\mathrm{Var}_{p_j}(g) \leq C\mathcal{E}_j(g, g).$$

[*Lemma 7*] In our mixture analysis, this is polynomial

$$C_{x,\beta|i} \lesssim D^{20} d^2 \lambda_{\max}^9 \lambda_{\min}^{-1}$$



Fast Mixing within mode $p(x, \beta|i)$ across temperatures

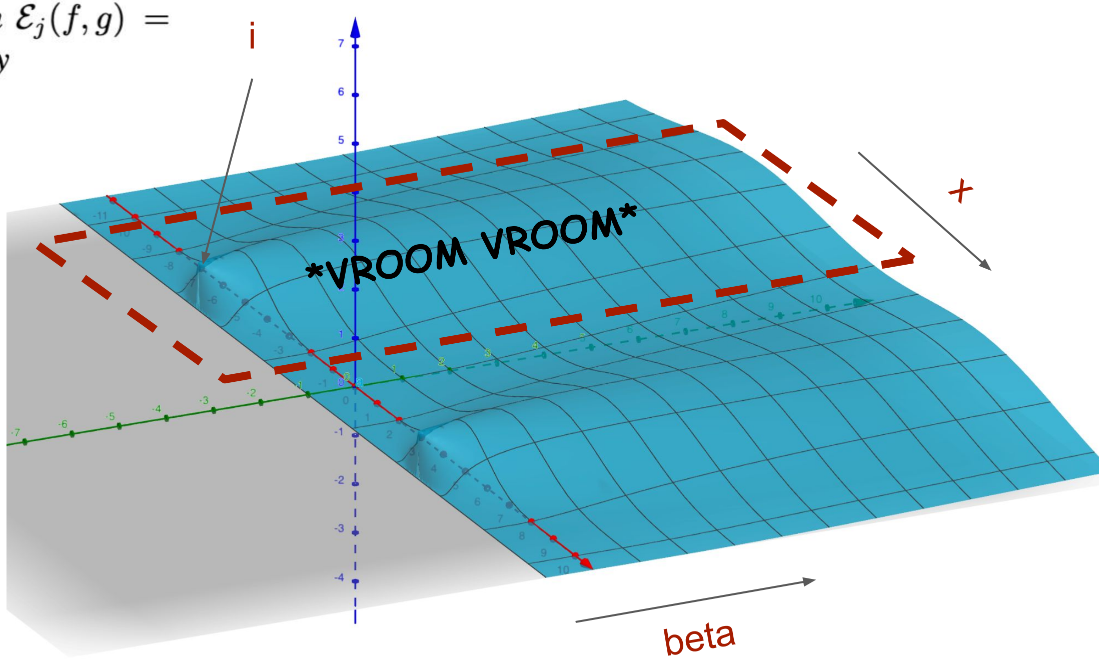


2. *(Mixing for each M_j) The Dirichlet form $\mathcal{E}_j(f, g) = -\langle f, \mathcal{L}g \rangle_{p_j}$ satisfies the Poincaré inequality*

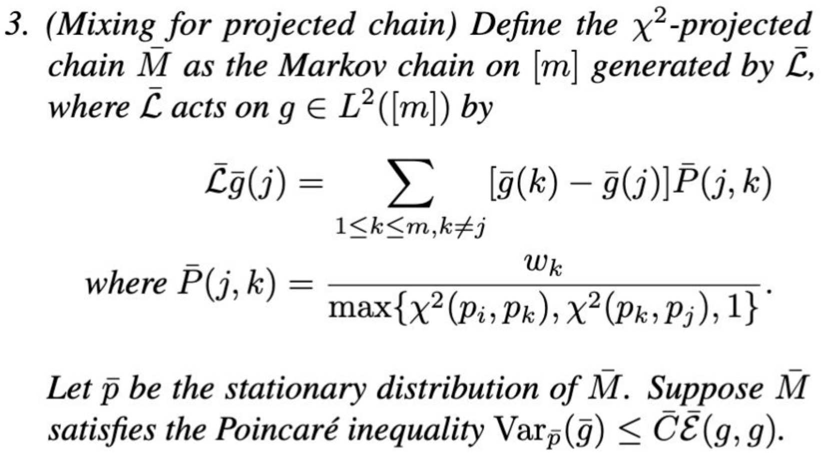
$$\mathrm{Var}_{p_j}(g) \leq C\mathcal{E}_j(g, g).$$

[*Lemma 7*] In our mixture analysis, this is polynomial

$$C_{x,\beta|i} \lesssim D^{20} d^2 \lambda_{\max}^{9} \lambda_{\min}^{-1}$$



Fast Mixing across modes $p(i)$ at high temperature

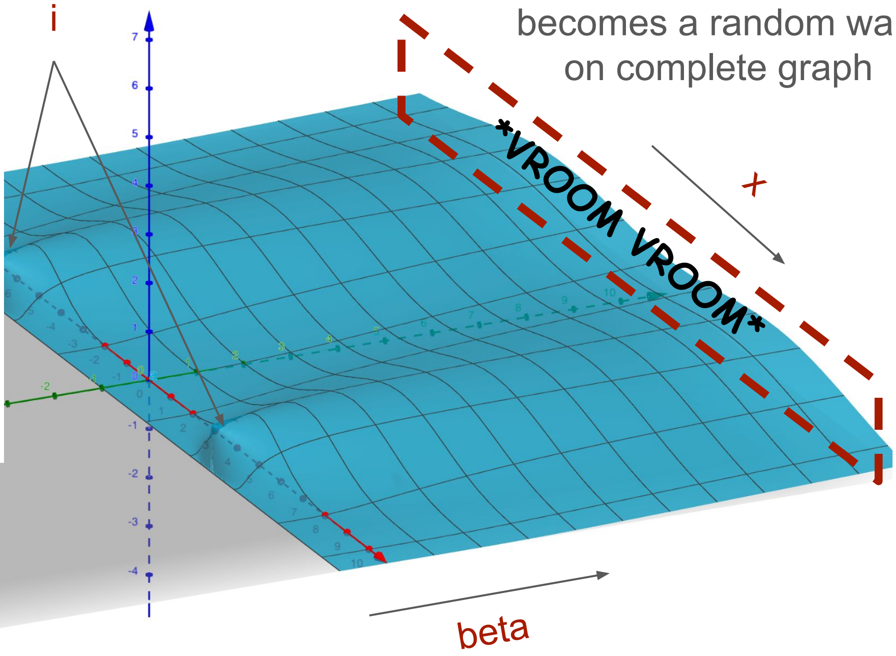


3. *(Mixing for projected chain) Define the χ^2-projected chain $\bar{M}$ as the Markov chain on $[m]$ generated by $\bar{\mathcal{L}}$, where $\bar{\mathcal{L}}$ acts on $g \in L^2([m])$ by*

$$\bar{\mathcal{L}}\bar{g}(j) = \sum_{1 \le k \le m, k \neq j} [\bar{g}(k) - \bar{g}(j)]\bar{P}(j,k)$$

$$\text{where } \bar{P}(j,k) = \frac{w_k}{\max\{\chi^2(p_i, p_k), \chi^2(p_k, p_j), 1\}}.$$

Let $\bar{p}$ be the stationary distribution of $\bar{M}$. Suppose $\bar{M}$ satisfies the Poincaré inequality $\mathrm{Var}_{\bar{p}}(\bar{g}) \le \bar{C}\bar{\mathcal{E}}(g,g)$.



The projected chain on i becomes a random walk on complete graph

[*Lemma 8*] In our mixture analysis, this is polynomial

$$\bar{C} \lesssim D^2 \lambda_{\min}^{-1}.$$

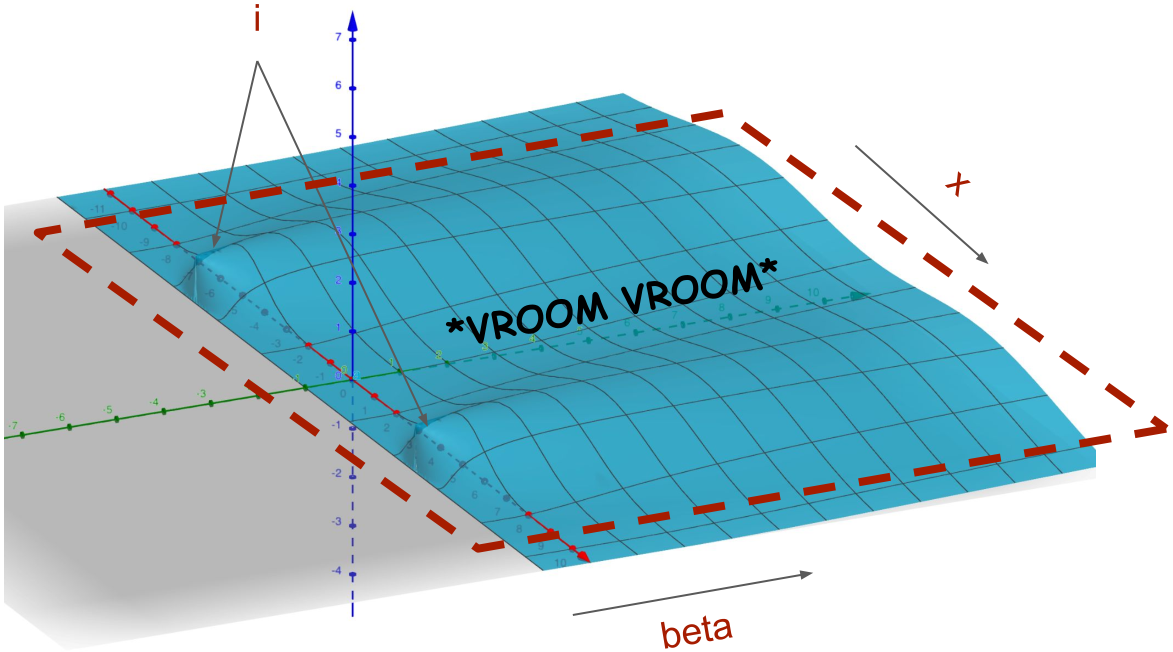
Putting it all together

The overall Poincaré constant is bounded by the two Poincaré constants.

$$\mathrm{Var}_p(g) \le C\left(1 + \frac{\bar{C}}{2}\right)\mathcal{E}(g,g).$$

[*Theorem 4*] In our mixture analysis, this is polynomial

$$C_P \lesssim D^{22} d^2 \lambda_{\max}^{9} \lambda_{\min}^{-2}$$



Takeaways

1. Sample complexity of score matching is governed by mixing time of Markov Chains. Given a (continuous time) Markov Chain, we can design score matching loss with different sample complexities.
2. Annealed score matching loss corresponds to Langevin Dynamics over the lifted (x, beta) distribution, which mixes much faster than Langevin on the original distribution over x.

Thank you!