A Proximal Algorithm for Sampling

Yongxin Chen

Georgia Institute of Technology

Oct 31, 2023

Georgia State University

Georgia Tech

Collaborators



Sinho Chewi

(IAS)



Adil Salim

(Microsoft Research)



Andre Wibisono

(Yale)



Jiaming Liang

(Rochester)



Bo Yuan

(Georgia Tech)



Jiaojiao Fan

(Georgia Tech)

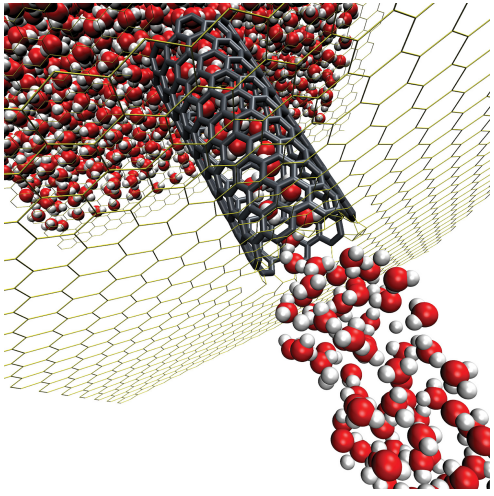
Georgia Tech

Papers

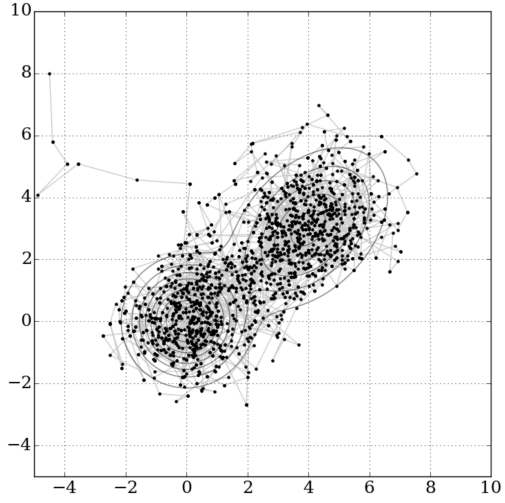
1. Chen, Chewi, Salim, Wibisono. *Improved Analysis for A Proximal Algorithm for Sampling*, 2022
2. Liang, Chen. *A Proximal Algorithm for Sampling from Non-convex Potentials*, 2022
3. Fan, Yuan, Chen. *Improved Dimension Dependence of A Proximal Algorithm for Sampling*, 2023
4. Yuan, Fan, Liang, Wibisono, Chen. *On A Class of Gibbs Sampling over Networks*, 2023

Georgia Tech

Markov chain Monte Carlo sampling



- Uncertainty quantification
- Inference, filtering
- Reliability analysis
- Design optimization
- Molecular dynamics



P. Diaconis (2009), “The Markov chain Monte Carlo revolution":

...asking about applications of Markov chain Monte Carlo (MCMC) is a little like asking about applications of the quadratic formula... you can take any area of science, from hard to social, and find a burgeoning MCMC literature specifically tailored to that area.

Georgia Tech

Sampling and optimization

Given a target function $f : \mathbb{R}^d \to \mathbb{R}$

Optimization:
Output an (approximate) minimizer of f

Sampling:
Output an (approximate) sample from the target density $\pi \propto \exp(-f)$

Sampling is an optimization over the manifold of probability distributions $\mathcal{P}(\mathbb{R}^d)$

$$\min_{\mu \in \mathcal{P}} H_\pi(\mu) = \int \mathrm{d}\mu \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi}$$

these connections furnish new algorithms and theory for sampling

Georgia Tech

Langevin diffusion

Basic approach to sampling: discretize Langevin diffusion

$$\mathrm{d}X_t = \underbrace{-\nabla f(X_t)\,\mathrm{d}t}_{\text{gradient flow}} + \underbrace{\sqrt{2}\,\mathrm{d}W_t}_{\text{Brownian motion}} \quad , \quad X_0 \sim \mu_0$$

which has π as stationary distribution

$$X_t \sim \mu_t \to \pi \propto \exp(-f)$$

Langevin diffusion is the gradient flow

of the KL divergence $\mu \mapsto H_\pi(\mu)$

over the Wasserstein space $(\mathcal{P}(\mathbb{R}^d), W_2)$

Jordan, Kinderlehrer, Otto 98

Fast convergence rates under mild assumptions

Georgia Tech

Discretization of Langevin diffusion

Langevin Monte Carlo (Euler-Maruyama discretization)

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \sqrt{2\eta}\, \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, I_d)$$

asymptotic bias: $x_k \sim \mu_k \to \mu_\infty \neq \pi$, low accuracy

Is there any better discretization scheme for Langevin diffusion?

Proximal point method in optimization:

$$x_{k+1} = \mathrm{prox}_{\eta f}(x_k)$$

proximal operator

$$\mathrm{prox}_{\eta f}(y) := \mathrm{argmin}_{x \in \mathbb{R}^d} \left\{ f(x) + \frac{1}{2\eta} \|x - y\|^2 \right\}$$

Georgia Tech

Previous attempts

- Proximal Langevin algorithm [Pereyra 16; Bernton 18; Wibisono 19]

$$x_{k+1} = \mathrm{prox}_{\eta f}(x_k + \sqrt{2\eta}\,\epsilon_k)\,, \qquad \epsilon_k \sim \mathcal{N}(0, I_d)$$

  . . . asymptotic bias

- JKO scheme [Jordan, Kinderlehrer, Otto '98]

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \left\{ H_\pi(\mu) + \frac{1}{2\eta}\, W_2^2(\mu, \mu_k) \right\}$$

  proximal point method in the Wasserstein space

  . . . not implementable

Georgia Tech

Proximal sampler

Augment the target density by

$$\pi^{XY}(x,y) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x-y\|^2\right)$$

Lee, Shen, Tian 21

Algorithm (Gibbs sampling):

1. Draw $y_k \sim \pi^{Y|X=x_k} = \mathcal{N}(x_k, \eta I_d)$
2. Draw $x_{k+1} \sim \pi^{X|Y=y_k}$

unbiased algorithm

Restricted Gaussian oracle (RGO) for a fixed y (η: stepsize)

$$\pi^{X|Y=y}(x) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|y-x\|^2\right)$$

Proximal operator for sampling

Georgia Tech

Proximal sampler

Questions to answer:

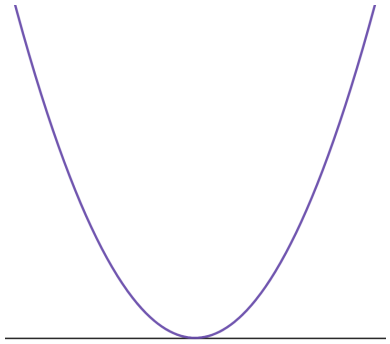
1. How fast does the proximal sampler converge?
2. How to implement the RGO efficiently?

Georgia Tech

Convergence Rate of Proximal Sampler

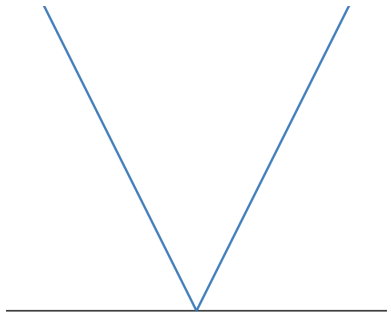
Georgia Tech

Assumptions on the target $\pi \propto \exp(-f)$



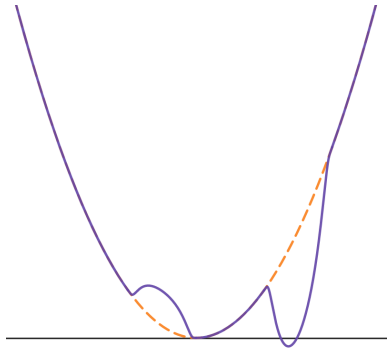
strong convexity of f

interpretation: strong convexity of H_π



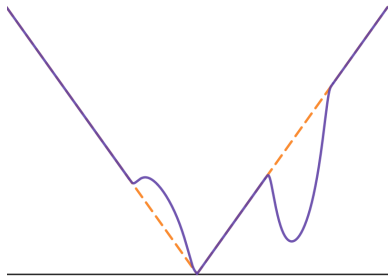
convexity of f

interpretation: convexity of H_π



log-Sobolev inequality

interpretation: PL inequality for H_π



Poincaré inequality

interpretation: spectral gap

Georgia Tech

Convergence of the Langevin diffusion

1. α-strong convexity of f $\Longrightarrow$

$$W_2^2(\mu_t, \pi) \leq \exp(-2\alpha t)\, W_2^2(\mu_0, \pi)$$

2. convexity of f $\Longrightarrow$

$$H_\pi(\mu_t) \leq \frac{1}{2t}\, W_2^2(\mu_0, \pi)$$

3. α-LSI $\Longrightarrow$

$$H_\pi(\mu_t) \leq \exp(-2\alpha t)\, H_\pi(\mu_0)$$

4. α-PI $\Longrightarrow$

$$\chi_\pi^2(\mu_t) \leq \exp(-2\alpha t)\, \chi_\pi^2(\mu_0)$$

Georgia Tech

Convergence of the proximal sampler

Proximal sampler

1. Draw $y_k \sim \pi^{Y|X=x_k} = \mathcal{N}(x_k, \eta I_d)$
2. Draw $x_{k+1} \sim \pi^{X|Y=y_k}$

Let ρ_k^X denote the law of the iterates, i.e., $x_k \sim \rho_k^X$

1. [Lee, Shen, Tian 21] α-strong convexity of f $\Longrightarrow$

$$W_2^2(\rho_k^X, \pi) \leq \frac{1}{(1+\alpha\eta)^{2k}} W_2^2(\rho_0^X, \pi)$$

Compare with optimization: if f is α-strongly convex

$$\|\mathrm{prox}_{\eta f}(x) - x^\star\|^2 \leq \frac{1}{(1+\alpha\eta)^2} \|x - x^\star\|^2$$

Georgia Tech

Convergence of the proximal sampler

Theorem (C., Chewi, Salim, Wibisono 22):

2. convexity of f $\Longrightarrow$

$$H_\pi(\rho_k^X) \le \frac{1}{k\eta} W_2^2(\rho_0^X, \pi)$$

3. α-LSI $\Longrightarrow$

$$H_\pi(\rho_k^X) \le \frac{1}{(1+\alpha\eta)^{2k}} H_\pi(\rho_0^X)$$

4. α-PI $\Longrightarrow$

$$\chi^2_\pi(\rho_k^X) \le \frac{1}{(1+\alpha\eta)^{2k}} \chi^2_\pi(\rho_0^X)$$

Georgia Tech

Proof: strong data processing inequality

Proximal sampler as forward/backward diffusions ($\pi_t = \pi * \mathcal{N}(0, tI_d)$)

$$y_k \sim \pi^{Y|X=x_k} : \quad \mathrm{d}Z_t = \mathrm{d}W_t,\ Z_0 = x_k,\ y_k = Z_\eta$$

$$x_{k+1} \sim \pi^{X|Y=y_k} : \quad \mathrm{d}Z_t = -\nabla \log \pi_t(Z_t)\mathrm{d}t + \mathrm{d}W_t,\ Z_\eta = y_k,\ x_{k+1} = Z_0$$

Theorem (C., Chewi, Salim, Wibisono 22): Given an arbitrary diffusion

$$\mathrm{d}Z_t = b_t(Z_t)\mathrm{d}t + \sigma \mathrm{d}W_t$$

let μ_t, ν_t be the distribution of Z_t initialized with $Z_0 \sim \mu_0$ and $Z_0 \sim \nu_0$ respectively. Assume ν_t satisfies α_t-LSI, then for any $T \geq 0$

$$H_{\nu_T}(\mu_T) \leq H_{\nu_0}(\mu_0) \exp\left(-\int_0^T \sigma^2 \alpha_t \mathrm{d}t\right)$$

Georgia Tech

Implementations of Restricted Gaussian Oracle

Georgia Tech

RGO for smooth potentials

Suppose f is L-smooth: $\|\nabla^2 f\| \leq L$, then RGO

$$\pi^{X|Y=y}(x) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|y-x\|^2\right)$$

is strongly log-concave and log-smooth for small η

$\Longrightarrow$ can be implemented via other samplers

Lemma: If we take step size $\eta \asymp \frac{1}{Ld}$, then rejection sampling implements RGO using $\mathcal{O}(1)$ queries to f in expectation

Georgia Tech

Rejection sampling for RGO

RGO: given y, sample from

$$\exp(-f_\eta^y(x)) := \exp\left(-f(x) - \frac{1}{2\eta}\|y-x\|^2\right)$$

Algorithm 1 RGO Rejection Sampling

Input: Proposal $\exp(-h(x))$ satisfies $h(x) \le f_\eta^y(x)$

1. Generate sample $X \sim \exp(-h(x))$
2. Generate sample $U \sim \mathcal{U}[0,1]$
3. If

$$U \le \frac{\exp(-f_\eta^y(X))}{\exp(-h(X))}$$

then accept/return X; otherwise, reject X and go to step 1

Georgia Tech

Sample complexity for smooth potentials

Theorem (C., Chewi, Salim, Wibisono 22): Suppose that π satisfies α-PI and f is L-smooth. Then, the proximal sampler outputs ρ_k^X with $\chi^2_\pi(\rho_k^X) \le \varepsilon$ and the expected number of queries to f is

$$\tilde{\mathcal{O}}\Big(\frac{Ld^2}{\alpha}\Big)$$

Existing results: [Chewi et al 21] $\quad \tilde{\mathcal{O}}\Big(\frac{L^2 d^3}{\alpha^2 \epsilon}\Big)$

Dependence on d is not good enough

Georgia Tech

Approximate Rejection Sampling for RGO

RGO: given y, sample from

$$\exp(-f_\eta^y(x)) = \exp\left(-f(x) - \frac{1}{2\eta}\|y-x\|^2\right)$$

larger η: faster convergence, smaller η: $\exp(-f_\eta^y(x))$ closer to Gaussian

---

Algorithm 2 Approximate Rejection Sampling for RGO

---

1. Solve $x_y = \operatorname{argmin}[f(x) + \frac{1}{2\eta}\|y-x\|^2]$
2. Define $\hat{f}(x) = f(x) - \langle \nabla f(x_y), x\rangle$
3. Generate sample $X, Z \sim \mathcal{N}(x_y, \eta I_d)$
4. Generate sample $U \sim \mathcal{U}[0,1]$
5. If

$$U \le \frac{1}{2}\exp(\hat{f}(Z) - \hat{f}(X))$$

then accept/return X; otherwise, reject X and go to step 3

Georgia Tech

RGO with $\mathcal{O}(1)$ complexity

Assumption 1: $f(x)$ is L_γ-semi-smooth of order γ

$$\|\nabla f(u) - \nabla f(v)\| \le L_\gamma \|u - v\|^\gamma, \quad \forall u, v \in \mathbb{R}^d$$

1) $\gamma = 1$, smooth, 2) $\gamma = 0$, nonsmooth
3) $0 < \gamma < 1$, weakly smooth

Theorem (Fan, Yuan, C. 23): Under Assumption 1, suppose

$$\eta \le \frac{C}{L_\gamma^{\frac{2}{1+\gamma}} d^{\frac{\gamma}{1+\gamma}} \log(1/\delta)}$$

for some small constant C and accuracy δ, then Algorithm 2 returns a sample that has δ total variation distance to the distribution $\exp\left(-f(x) - \frac{1}{2\eta}\|y - x\|^2\right)$, and it accesses only $\mathcal{O}(1)$ queries of f and its gradient

Georgia Tech

Concentration inequality for semi-smooth functions

Lemma (Fan, Yuan, C. 23): Let $X \sim \mathcal{N}(0, \eta I_d)$ and g be a L_γ-semi-smooth of order $\gamma \in [0, 1]$ satisfying $\nabla g(0) = 0$. Then for all $t \geq 0$,

$$\Pr[g(X) - \mathbb{E}[g(X)] \geq t] \leq C_1 \exp\left(-\frac{C_2\, t^{2/(1+\gamma)}}{\eta L_\gamma^{\frac{2}{1+\gamma}} d^{\frac{\gamma}{1+\gamma}}}\right)$$

for some universal constants C_1, C_2

It generalizes

i) Gaussian concentration for Lipschitz functions

ii) Hanson-Wright inequality for quadratic functions

Georgia Tech

Concentration inequalities: special cases

1. Lipschitz g: $\gamma = 0$

Gaussian concentration (Ledoux 99): Let $X \sim \mathcal{N}(0, \eta I_d)$ and g be a L_0-Lipschitz continuous function. Then, for all $t \geq 0$,

$$\Pr[g(X) - \mathbb{E}[g(X)] \geq t] \leq \exp\left(-\frac{t^2}{2\eta L_0^2}\right)$$

2. quadratic $g(x) = x^\top Ax$: $\gamma = 1$

Hanson-Wright (Rudelson, Vershynin 13): Let $X \sim \mathcal{N}(0, \eta I_d)$ and A be an $d \times d$ matrix. Then, for all $t \geq 0$,

$$\Pr(X^\top AX - \mathbb{E}(X^\top AX) > t) \leq \exp\left(-C \min\left(\frac{4t^2}{\eta^2 \|A\|_{\mathrm{F}}^2}, \frac{2t}{\eta \|A\|}\right)\right)$$

Georgia Tech

Proximal sampler complexity

Theorem (Fan, Yuan, C. 23): Suppose f is L_1-smooth. With $\eta \asymp 1/(L_1 d^{1/2})$, the proximal sampler with RGO by Algorithm 2 has complexity bound

$$\tilde{\mathcal{O}}\left(\frac{L_1 d^{1/2}}{\alpha}\right)$$

to achieve ϵ error to $\pi \propto \exp(-f)$ in total variation, if either f is α-strongly convex or π satisfies α-LSI. Each iteration needs $\mathcal{O}(1)$ queries of f and its gradient

Existing results:

Strongly log-concave: [Wu, Schmidler, Chen 22] $\tilde{\mathcal{O}}\left(\frac{L_1 d^{1/2}}{\alpha}\right)$ with warm start

α-LSI: [Liang, Chen 22] $\tilde{\mathcal{O}}\left(\frac{L_1 d}{\alpha}\right)$

Concurrent results: [Altschuler, Chewi 23]

Georgia Tech

Gibbs sampling over networks

Composite distribution

$$\pi(x,y) \propto \exp(-f_1(x) - \frac{1}{2\eta}\|x-y\|^2 - f_2(y))$$

Joint distribution over a graph $G = (V, E)$

$$\pi(x_1, x_2, \ldots, x_n) \propto \exp\left(-\sum_{i=1}^{n} f_i(x_i) - \sum_{(i,j)\in E} \frac{w_{ij}}{2\eta}\|x_i - x_j\|^2\right)$$

Georgia Tech

Takeaway

The proximal sampler is the sampling analogue of the proximal point method from optimization. Unlike Langevin Monte Carlo, it does not induce bias (discretization error)

1. new convergence results under weaker assumptions
2. efficient realizations of RGO even for non-convex and semi-smooth potentials
3. new state-of-the-art results for sampling

Georgia Tech

Papers

1. Chen, Chewi, Salim, Wibisono. *Improved Analysis for A Proximal Algorithm for Sampling*, 2022
2. Liang, Chen. *A Proximal Algorithm for Sampling from Non-convex Potentials*, 2022
3. Fan, Yuan, Chen. *Improved Dimension Dependence of A Proximal Algorithm for Sampling*, 2023
4. Yuan, Fan, Liang, Wibisono, Chen. *On A Class of Gibbs Sampling over Networks*, 2023

Georgia Tech

Acknowledgment








National Science Foundation

SIMONS INSTITUTE for the Theory of Computing

Thank you for your attention!

Georgia Tech