A Proximal Sampling Algorithm

Jiaming Liang

Department of Computer Science
Yale University

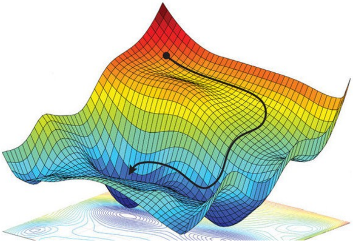
November 2, 2022

Management Science & Information Systems, Rutgers Business School

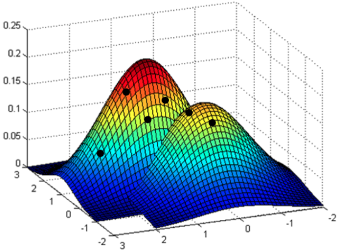
Joint works with Yongxin Chen (Georgia Tech)

Introduction

Design and analysis of fast algorithms for sampling problems by leveraging tools from optimization.



(a) Optimization, $\min f(x)$



(b) Sampling, samp $\exp(-f(x))$

Overview

- A proximal sampling algorithm for nonconvex, semi-smooth and composite potentials

- Improved complexity to sample from a distribution ε-close to the target distribution in KL, χ^2 and Rényi divergences

- Close interplay between sampling and optimization
  Proximal point framework

Assumptions

Problem: sample from $\nu(x) \propto \exp(-f(x))$

(A1) f is semi-smooth, i.e., there exist $\alpha_i \in [0,1]$ and $L_i > 0$, $i = 1, \ldots, n$, s.t.

$$\|f'(u) - f'(v)\| \le \sum_{i=1}^{n} L_{\alpha_i} \|u - v\|^{\alpha_i}, \quad \forall u, v \in \mathbb{R}^d$$

where $f'(x)$ is in the Frechet subdiffernetial $\tilde{\partial}\phi(x)$;

Examples: $n = 1$

1) $\alpha_1 = 1$, smooth, 2) $\alpha_1 = 0$, nonsmooth, 3) $0 < \alpha_1 < 1$, weakly smooth

(A2) ν satisfies log-Sobolev inequality (LSI) or Poincaré inequality (PI).

LSI: $H_\nu(\rho) \le \frac{C_{LSI}}{2} J_\rho(\nu)$, PI: $\mathbb{E}_\nu[(\psi - \mathbb{E}_\nu[\psi])^2] \le C_{PI} \mathbb{E}_\nu[\|\nabla\psi\|^2]$

Observations: ν is not necessarily log-concave, f is not necessarily convex.

Comparison

Source	Complexity	Assumption	Metric
Chewi et al.	$\tilde{\mathcal{O}}\left(\frac{C_{\mathrm{PI}}^{1+1/\alpha} L_\alpha^{2/\alpha} d^{2+1/\alpha}}{\varepsilon^{1/\alpha}}\right)$	weakly smooth $\alpha > 0$, PI	Rényi
This work	$\tilde{\mathcal{O}}\left(C_{\mathrm{PI}} L_\alpha^{2/(1+\alpha)} d^2\right)$	semi-smooth, PI	Rényi

Table: Complexity bounds for sampling from non-convex semi-smooth potentials.

Source	Complexity	Assumption	Metric
Nguyen et al.	$\tilde{\mathcal{O}}\left(C_{\mathrm{LSI}}^{1+\max\{\frac{1}{\alpha_i}\}}\left[\frac{n\max\{L_{\alpha_i}^2\}d}{\varepsilon}\right]^{\max\{\frac{1}{\alpha_i}\}}\right)$	weakly smooth $\alpha_i > 0$, LSI	KL
This work	$\tilde{\mathcal{O}}\left(C_{\mathrm{LSI}} \sum_{i=1}^n L_{\alpha_i}^{2/(\alpha_i+1)} d\right)$	semi-smooth, LSI	KL
This work	$\tilde{\mathcal{O}}\left(C_{\mathrm{PI}} \sum_{i=1}^n L_{\alpha_i}^{2/(\alpha_i+1)} d\right)$	semi-smooth, PI	Rényi

Table: Complexity bounds for sampling from non-convex composite potentials.

Alternating Sampling Framework (ASF)

Joint distribution $\pi(x,y) \propto \exp[-f(x) - \frac{1}{2\eta}\|x-y\|^2]$

Algorithm 1 ASF (Shen, Tian and Lee 2021)

1. Sample $y_k \sim \pi^{Y|X}(y \mid x_k) \propto \exp[-\frac{1}{2\eta}\|x_k - y\|^2]$
2. Sample $x_{k+1} \sim \pi^{X|Y}(x \mid y_k) \propto \exp[-f(x) - \frac{1}{2\eta}\|x - y_k\|^2]$

Restricted Gaussian Oracle (RGO)
Given y, sample from

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

Alternating Sampling Framework (ASF)

Joint distribution $\pi(x, y) \propto \exp[-f(x) - \frac{1}{2\eta}\|x - y\|^2]$

Algorithm 2 ASF (Shen, Tian and Lee 2021)

1. Sample $y_k \sim \pi^{Y|X}(y \mid x_k) \propto \exp[-\frac{1}{2\eta}\|x_k - y\|^2]$
2. Sample $x_{k+1} \sim \pi^{X|Y}(x \mid y_k) \propto \exp[-f(x) - \frac{1}{2\eta}\|x - y_k\|^2]$

Restricted Gaussian Oracle (RGO)
Given y, sample from

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

Without an implementable and provable RGO, ASF is only conceptual.

Nontrivial

Proximal Point Framework (PPF)

Proximal point framework: constructs a sequence of proximal problems

$$x_{k+1} \leftarrow \mathrm{prox}_{\eta f}(x_k) = \operatorname*{argmin}_x \left\{ f(x) + \frac{1}{2\eta}\|x - x_k\|^2 \right\} \qquad (*) \tag{1}$$

E.g., Chambolle-Pock for saddle point, ADMM for distributed optimization

Algorithm 3 PPF

1. $y_k \leftarrow \operatorname*{argmin}_x \frac{1}{2\eta}\|x - x_k\|^2 = x_k$
2. $x_{k+1} \leftarrow \operatorname*{argmin}_x \left\{ f^\eta_{y_k}(x) := f(x) + \frac{1}{2\eta}\|x - y_k\|^2 \right\}$

ASF for sampling $\longleftrightarrow$ PPF for optimization

RGO in sampling $\longleftrightarrow$ proximal mapping in optimization

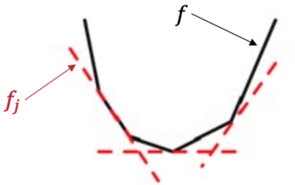
Relaxed Proximal Bundle Method (L. and Monteiro 2021)

f is convex and Lipschitz continuous (nonsmooth, $\alpha_1 = 0$). Subgradient method.

Approximately solve (1) by the cutting-plane method (implementable)

$$z_j \leftarrow \mathsf{prox}_{\eta f_j}(x_0) = \min_z \left\{ f_j(z) + \frac{1}{2\eta}\|z - z_0\|^2 \right\}, \quad z_0 = x_k$$

where $f_j(z) = \max\{f(w) + \langle f'(w), z - w\rangle : w \in \{z_0, z_1, \ldots, z_{j-1}\}\}$



Complexities: PPF $\mathcal{O}(\varepsilon^{-1})\times$ cutting-plane $\mathcal{O}(\varepsilon^{-1}) \implies$ total $\mathcal{O}(\varepsilon^{-2})$ optimal

implementable and provable

RGO Implementation

RGO: given y, sample from $\exp(-f_y^\eta(x))$

Algorithm 4 RGO Rejection Sampling

1. Compute an approximate stationary point w of f_y^η
2. Generate sample $X \sim \exp(-h_1(x))$
3. Generate sample $U \sim \mathcal{U}[0,1]$
4. If
$$U \leq \frac{\exp(-f_y^\eta(X))}{\exp(-h_1(X))},$$
then accept/return X; otherwise, reject X and go to step 2.

Proposal: $\exp(-h_1(x))$ where $h_1(x) \leq f_y^\eta(x)$

Rejection Sampling

$X \sim \pi^{X|Y}(\cdot|y)$ and

$$\begin{aligned}\mathbb{P}(X \text{ is accepted}) =& \mathbb{P}\left(U \leq \frac{\exp(-f_y^\eta(X))}{\exp(-h_1(X))}\right) \\ =& \frac{\int \exp(-f_y^\eta(x))dx}{\int \exp(-h_1(x))dx} \geq \frac{\int \exp(-h_2(x))dx}{\int \exp(-h_1(x))dx}\end{aligned} \tag{2}$$

Want to find h_1 and h_2 such that

i) sampling $\exp(-h_1(x))$ is easy,

ii) $h_1(x) \leq f_y^\eta(x) \leq h_2(x)$,

iii) (2) is bounded from below.

A Key Lemma

Consider $n = 1$, $\alpha \in [0, 1]$ and $L_\alpha > 0$

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

Lemma

Assume f is L_α-semi-smooth, then for $\delta > 0$ and every $u, v \in \mathbb{R}^d$, we have

$$|f(u) - f(v) - \langle f'(v), u - v\rangle| \le \frac{M}{2}\|u - v\|^2 + \frac{(1-\alpha)\delta}{2}, \quad M = \frac{L_\alpha^{\frac{2}{\alpha+1}}}{[(\alpha+1)\delta]^{\frac{1-\alpha}{\alpha+1}}}.$$

Proof:

$$|f(u) - f(v) - \langle f'(v), u - v\rangle| \le \frac{L_\alpha}{\alpha+1}\|u - v\|^{\alpha+1}$$

Young's inequality $ab \le \frac{a^p}{p} + \frac{b^q}{q}$, $\frac{1}{p} + \frac{1}{q} = 1$ with

$$a = \frac{L_\alpha}{(\alpha+1)\delta^{\frac{1-\alpha}{2}}}\|u - v\|^{\alpha+1}, \quad b = \delta^{\frac{1-\alpha}{2}}, \quad p = \frac{2}{\alpha+1}, \quad q = \frac{2}{1-\alpha}.$$

Construction

Definition

A stationary point w^* of f_y^η is such that $f'(w^*) + \frac{1}{\eta}(w^* - y) = 0$.

Definition

An approximate stationary point w of f_y^η is s.t. $\|f'(w) + \frac{1}{\eta}(w - y)\| \le \sqrt{Md}$.

$$h_1(x) := f(w) + \langle f'(w), x - w\rangle - \frac{M}{2}\|x - w\|^2 + \frac{1}{2\eta}\|x - y\|^2 - \frac{(1-\alpha)\delta}{2},$$
$$h_2(x) := f(w^*) + \langle f'(w^*), x - w^*\rangle + \frac{M}{2}\|x - w^*\|^2 + \frac{1}{2\eta}\|x - y\|^2 + \frac{(1-\alpha)\delta}{2}.$$

Answers: i) sampling $\exp(-h_1(x))$ is easy;
ii) verify $h_1(x) \le f_y^\eta(x) \le h_2(x)$ by the key lemma.

Remaining Questions

Q1. Rejection sampling complexity

$$[\mathbb{P}(X \text{ is accepted})]^{-1} \leq \frac{\int \exp(-h_1(x))dx}{\int \exp(-h_2(x))dx}$$

Q2. Optimization complexity to find an approx. stat. pt. w s.t.

$$\left\| f'(w) + \frac{1}{\eta}(w - y) \right\| \leq \sqrt{Md}$$

Answer to Q1 – RGO complexity

Proposition

Assume

$$\eta \leq \frac{1}{Md} = \frac{[(\alpha+1)\delta]^{\frac{1-\alpha}{\alpha+1}}}{L_\alpha^{\frac{2}{\alpha+1}} d},$$

then the expected number of rejection steps in Algorithm 4 is at most $\exp\left(\frac{3(1-\alpha)\delta}{2} + 3\right)$.

Intuition: if η is small enough, h_1 and h_2 are convex quadratic functions, so

$$\frac{\int \exp(-h_1(x))dx}{\int \exp(-h_2(x))dx} \approx \left(\frac{1+\eta M}{1-\eta M}\right)^{d/2} \leq (1+4\eta M)^{d/2} \leq \left(1+\frac{4}{d}\right)^{d/2} \leq e^2.$$

Answer to Q2 – Optimization complexity

Lemma

Let $f_y^\eta := f + \frac{1}{2\eta}\|\cdot - y\|^2$ *and* $(f_y^\eta)' := f' + \frac{1}{\eta}(\cdot - y)$, *then for every* $u, v \in \mathbb{R}^d$,

$$\frac{1}{2}\left(\frac{1}{\eta} - M\right)\|u - v\|^2 - \frac{(1-\alpha)\delta}{2} \le f_y^\eta(u) - f_y^\eta(v) - \langle (f_y^\eta)'(v), u - v\rangle$$
$$\le \frac{1}{2}\left(\frac{1}{\eta} + M\right)\|u - v\|^2 + \frac{(1-\alpha)\delta}{2}.$$

f_y^η is nearly $(\eta^{-1} - M)$-strongly convex and $(\eta^{-1} + M)$-smooth

Proposition

Assume $\eta \le \frac{1}{Md}$, *then the iteration-complexity to find the approx. stat. pt.* w *s.t.* $\left\| f'(w) + \frac{1}{\eta}(w - y)\right\| \le \sqrt{Md}$ *by Nesterov acceleration is* $\tilde{\mathcal{O}}(1)$.

$$\mu = \frac{1}{\eta} - M \approx M(d-1), \quad L = \frac{1}{\eta} + M \approx M(d+1), \quad \sqrt{L/\mu} \approx 1$$

RGO and ASF Complexities

Putting previous results together, we can implement RGO with $\tilde{\mathcal{O}}(1)$ subgradients of f and $\mathcal{O}(1)$ samples from Gaussian distribution in expectation.

Other ingredients for total complexity: **Convergence rate analysis of ASF**

Theorem (Chen, Chewi, Salim and Wibisono 2022)

If $\nu \propto \exp(-f)$ satisfies LSI with $C_{\mathsf{LSI}} > 0$, then x_k of ASF $\sim \rho_k$, which satisfies

$$H_\nu(\rho_k) \le \frac{H_\nu(\rho_0)}{\left(1 + \frac{\eta}{C_{\mathsf{LSI}}}\right)^{2k}}.$$

Theorem (Chen, Chewi, Salim and Wibisono 2022)

If $\nu \propto \exp(-f)$ satisfies PI with $C_{\mathrm{PI}} > 0$, then x_k of ASF $\sim \rho_k$, which satisfies

$$\chi^2_\nu(\rho_k) \le \frac{\chi^2_\nu(\rho_0)}{\left(1 + \frac{\eta}{C_{\mathrm{PI}}}\right)^{2k}}.$$

Main Result

Theorem

Suppose f is L_α-semi-smooth and ν satisfies PI. With $\eta \asymp 1/(L_\alpha^{\frac{2}{\alpha+1}} d)$, then ASF with RGO by rejection has complexity bound

$$\tilde{\mathcal{O}}\left(C_{\mathrm{PI}} L_\alpha^{\frac{2}{\alpha+1}} d\right)$$

to achieve ε error to ν in terms of χ^2 divergence. Each iteration queries $\tilde{\mathcal{O}}(1)$ subgradients of f and generates $\mathcal{O}(1)$ samples in expectation from Gaussian distribution.

General Results – LSI

$$\|f'(u) - f'(v)\| \le \sum_{i=1}^{n} L_{\alpha_i} \|u - v\|^{\alpha_i}, \quad \forall u, v \in \mathbb{R}^d;$$

Theorem

Suppose f is semi-smooth and ν satisfies LSI. With $\eta \asymp \left[\sum_{i=1}^{n} L_{\alpha_i}^{\frac{2}{\alpha_i+1}} d\right]^{-1}$, then ASF with RGO by rejection has complexity bound

$$\tilde{\mathcal{O}}\left(C_{\mathrm{LSI}} \sum_{i=1}^{n} L_{\alpha_i}^{\frac{2}{\alpha_i+1}} d\right)$$

to achieve ε error to ν in terms of χ^2 divergence. Each iteration queries $\tilde{\mathcal{O}}(1)$ subgradients of f and generates $\mathcal{O}(1)$ samples in expectation from Gaussian distribution.

General Results – PI

Theorem

Suppose f is semi-smooth and ν satisfies PI. With $\eta \asymp \left[\sum_{i=1}^{n} L_{\alpha_i}^{\frac{2}{\alpha_i+1}} d\right]^{-1}$, then ASF with RGO by rejection has complexity bound

$$\tilde{\mathcal{O}}\left(C_{\mathrm{PI}} \sum_{i=1}^{n} L_{\alpha_i}^{\frac{2}{\alpha_i+1}} d\right)$$

to achieve ε error to ν in terms of χ^2 divergence. Each iteration queries $\tilde{\mathcal{O}}(1)$ subgradients of f and generates $\mathcal{O}(1)$ samples in expectation from Gaussian distribution.

Interpretation of Unadjusted Langevin Algorithm (ULA)

Algorithm 5 ASF

1. Sample $y_k \sim \pi^{Y|X}(y \mid x_k) \propto \exp[-\frac{1}{2\eta}\|x_k - y\|^2]$
2. Sample $x_{k+1} \sim \pi^{X|Y}(x \mid y_k) \propto \exp[-f(x) - \frac{1}{2\eta}\|x - y_k\|^2]$

Algorithm 6 ULA

1. Sample $y_k \sim \pi^{Y|X}(y \mid x_k) \propto \exp[-\frac{1}{2\eta}\|x_k - y\|^2]$
2. Sample $x_{k+1} \sim e^{-\langle \nabla f(y_k), x - y_k\rangle - \frac{1}{2\eta}\|x - y_k\|^2} \propto e^{-\frac{1}{2\eta}\|x - (y_k - \eta\nabla f(y_k))\|^2}$

$$x_{k+1} = y_k - \eta\nabla f(y_k) + \sqrt{\eta} z_k, \quad z_k \sim N(0, I),$$
$$y_{k+1} = x_{k+1} + \sqrt{\eta} z_k', \quad z_k' \sim N(0, I).$$

$$\implies y_{k+1} = y_k - \eta\nabla f(y_k) + \sqrt{\eta}(z_k + z_k') = y_k - \eta\nabla f(y_k) + \sqrt{2\eta} z, \quad z \sim N(0, I)$$

ULA can be viewed as ASF with RGO implemented without rejection

$$h_1(x) = f(y_k) + \langle \nabla f(y_k), x - y_k\rangle + \frac{1}{2\eta}\|x - y_k\|^2 \le f(x) + \frac{1}{2\eta}\|x - y_k\|^2 = f^{\eta}_{y_k}(x)$$

Conclusions

- A proximal sampling algorithm for $\nu \propto \exp(-f)$.

  f nonconvex, semi-smooth, composite. ν satisfies either LSI or PI.

- Total complexity $\tilde{\mathcal{O}}\left(C \sum_{i=1}^{n} L_{\alpha_i}^{\frac{2}{\alpha_i+1}} d\right)$ where $C = C_{\mathrm{LSI}}$ or $C = C_{\mathrm{PI}}$.

  Each iteration takes $\tilde{\mathcal{O}}(1)$ subgradients of f and $\mathcal{O}(1)$ samples from Gaussian.

- Inspired by proximal point framework and proximal mapping.

  Leverage tools from optimization to design and analyze sampling algorithms.

  E.g., acceleration in sampling for weakly smooth potentials.

References

- Jiaming Liang and Yongxin Chen. A Proximal Algorithm for Sampling from Non-convex Potentials. 2022
- Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved Analysis for a Proximal Algorithm for Sampling. COLT 2022
- Ruoqi Shen, Kevin Tian, and Yin Tat Lee. Structured Logconcave Sampling with a Restricted Gaussian Oracle. COLT 2021
- Dao Nguyen, Xin Dang, and Yixin Chen Unadjusted Langevin Algorithm for Non-convex Weakly Smooth Potentials. 2021
- Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Shunshi Zhang. Analysis of Langevin Monte Carlo from Poincare to Log-Sobolev. COLT 2022
- Jiaming Liang and Renato Monteiro. A Proximal Bundle Variant with Optimal Iteration-complexity for A Large Range of Prox Stepsizes. SIAM Journal of Optimization 2021

Thank you!

Definition (Frechet ε-subdiffernetial)

Let $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a proper closed function, then the Frechet ε subdiffernetial is defined as

$$\tilde{\partial}_\varepsilon \phi(x) = \left\{ v \in \mathbb{R}^n : \liminf_{y \to x} \frac{\phi(y) - \phi(x) - \langle v, y - x \rangle}{\|y - x\|} \geq -\varepsilon \right\}$$

When $\varepsilon = 0$, we denote $\tilde{\partial}_0 \phi(x)$ simply by $\tilde{\partial} \phi(x)$.

Lemma

If $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is an m-weakly convex function, then for any $x, c \in \mathbb{R}^n$, we have

$$\tilde{\partial} \phi(x) = \partial \left(\phi_m(\cdot; c) \right)(x) - m(x - c) \tag{3}$$

where

$$\phi_m(\cdot; c) := \phi(\cdot) + \frac{m}{2} \| \cdot - c \|^2.$$