A Universal Proximal Framework for Optimization and Sampling

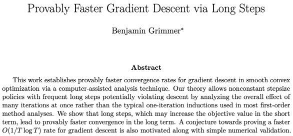
Jiaming Liang

Department of Computer Science
Goergen Institute for Data Science

September 19, 2023

Motivation – Gradient Method with Long Steps

(a) Gradient descent



Provably Faster Gradient Descent via Long Steps

Benjamin Grimmer*

Abstract

This work establishes provably faster convergence rates for gradient descent in smooth convex optimization via a computer-assisted analysis technique. Our theory allows nonconstant stepsize policies with frequent long steps potentially violating descent by analyzing the overall effect of many iterations at once rather than the typical one-iteration inductions used in most first-order method analyses. We show that long steps, which may increase the objective value in the short term, lead to provably faster convergence in the long term. A conjecture towards proving a faster $O(1/T \log T)$ rate for gradient descent is also motivated along with simple numerical validation.

(b) The paper

Gradient descent $x_{k+1} = x_k - \frac{h_k}{L} \nabla f\left(x_k\right)$

Classical result: $h_k \leq 1$,

$$f\left(x_k\right) - f\left(x_\star\right) \leq \frac{LD^2}{2k}$$

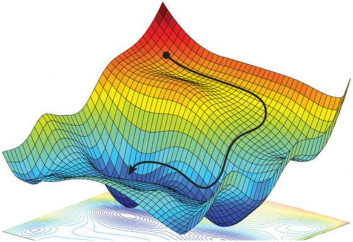
New result: $h_k = (2.9, 1.5, 2.9, 1.5, \ldots)$,

$$f\left(x_k\right) - f\left(x_\star\right) \leq \frac{LD^2}{2.2k} + O\left(\frac{1}{k^2}\right)$$

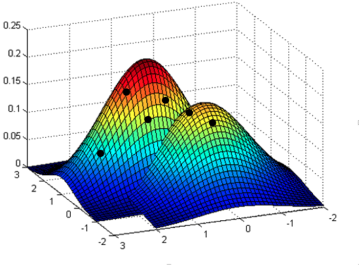
Proximal bundle method and restarted Nesterov's accelerated gradient.

Optimization and Sampling

Fast algorithm design for solving fundamental optimization and sampling problems using the proximal point framework.



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



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

Algorithms for Optimization and Sampling

- Stochastic gradient descent, $\min_x \mathbb{E}_\xi[F(x,\xi)]$

$$x_{k+1} = x_k - \lambda_k s(x_k, \xi_k), \quad s(x_k,\xi_k) \in \partial F(x_k, \xi_k)$$

- Accelerated gradient descent, $\min_x f(x)$

$$\tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_{k+1}}, \quad y_{k+1} = \tilde{x}_k - \lambda_k \nabla f(\tilde{x}_k), \quad x_{k+1} = \frac{A_{k+1}}{a_k} y_{k+1} - \frac{A_k}{a_k} y_k$$

- Unadjusted Langevin algorithm, sample from $\nu(x) \propto \exp(-f(x))$

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k) + \sqrt{2\lambda_k} z, \quad z \sim \mathcal{N}(0, I)$$

Algorithms for Optimization and Sampling

- Stochastic gradient descent, $\min_x \mathbb{E}_\xi[F(x,\xi)]$

$$x_{k+1} = x_k - \lambda_k s(x_k, \xi_k), \quad s(x_k,\xi_k) \in \partial F(x_k, \xi_k)$$

- Accelerated gradient descent, $\min_x f(x)$

$$\tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_{k+1}}, \quad y_{k+1} = \tilde{x}_k - \lambda_k \nabla f(\tilde{x}_k), \quad x_{k+1} = \frac{A_{k+1}}{a_k} y_{k+1} - \frac{A_k}{a_k} y_k$$

- Unadjusted Langevin algorithm, sample from $\nu(x) \propto \exp(-f(x))$

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k) + \sqrt{2\lambda_k} z, \quad z \sim \mathcal{N}(0, I)$$

Algorithms for Optimization and Sampling

- Stochastic gradient descent, $\min_x \mathbb{E}_\xi[F(x,\xi)]$

$$x_{k+1} = x_k - \lambda_k s(x_k, \xi_k), \quad s(x_k, \xi_k) \in \partial F(x_k, \xi_k)$$

- Accelerated gradient descent, $\min_x f(x)$

$$\tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_{k+1}}, \quad y_{k+1} = \tilde{x}_k - \lambda_k \nabla f(\tilde{x}_k), \quad x_{k+1} = \frac{A_{k+1}}{a_k} y_{k+1} - \frac{A_k}{a_k} y_k$$

- Unadjusted Langevin algorithm, sample from $\nu(x) \propto \exp(-f(x))$

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k) + \sqrt{2\lambda_k} z, \quad z \sim \mathcal{N}(0, I)$$

A Universal Proximal Framework

Optimization

Algorithm Proximal Point Framework

1. $y_k \leftarrow \underset{x}{\operatorname{argmin}} \frac{1}{2\lambda}\|x - x_k\|^2 = x_k$
2. $x_{k+1} \leftarrow \underset{x}{\operatorname{argmin}} \left\{ f(x) + \frac{1}{2\lambda}\|x - y_k\|^2 \right\}$

E.g., GD, SGD, AGD, Newton, Chambolle-Pock, ADMM, proximal bundle ...

Sampling

Algorithm Alternating Sampling Framework

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

E.g., ULA, proximal Langevin algorithm, symmetric Langevin algorithm ...

A Universal Proximal Framework

Optimization

Algorithm Proximal Point Framework

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

E.g., GD, SGD, AGD, Newton, Chambolle-Pock, ADMM, proximal bundle ...

Sampling

Algorithm Alternating Sampling Framework

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

E.g., ULA, proximal Langevin algorithm, symmetric Langevin algorithm ...

Outline

Outline

1. Nonsmooth Optimization
2. Stochastic Optimization
3. High-dimensional Sampling

Assumptions

Convex nonsmooth composite problem

$$\phi_* := \min \{\phi(x) := f(x) + h(x) : x \in \mathbb{R}^n\}$$

(A1) bounded subgradient

$$\|f'(x)\| \leq M;$$

(A2) h is μ-strongly convex ($\mu \geq 0$).

Motivation - Proximal Bundle Method

Goal: find $\hat{x}$ such that $\phi(\hat{x}) - \phi_* \leq \varepsilon$

- Subgradient, Mirror descent, Bundle-level, and Prox Level method are optimal.
- Proximal bundle method $\mathcal{O}(\varepsilon^{-3})$ $\leftarrow$ previously best, improvable?
- Lower complexity bound $\Omega(\varepsilon^{-2})$

Proximal bundle method is not optimal in general

We close the gap by showing the tight upper bound $\mathcal{O}(\varepsilon^{-2})$
through a new proximal bundle method and a refined analysis

Motivation - Proximal Bundle Method

Goal: find $\hat{x}$ such that $\phi(\hat{x}) - \phi_* \leq \varepsilon$

- Subgradient, Mirror descent, Bundle-level, and Prox Level method are optimal.
- Proximal bundle method $\mathcal{O}(\varepsilon^{-3})$ $\leftarrow$ previously best, improvable?
- Lower complexity bound $\Omega(\varepsilon^{-2})$

Proximal bundle method is not optimal in general

We close the gap by showing the tight upper bound $\mathcal{O}(\varepsilon^{-2})$
through a new proximal bundle method and a refined analysis

Motivation - Proximal Bundle Method

Goal: find $\hat{x}$ such that $\phi(\hat{x}) - \phi_* \leq \varepsilon$

- Subgradient, Mirror descent, Bundle-level, and Prox Level method are optimal.
- Proximal bundle method $\mathcal{O}(\varepsilon^{-3})$ $\leftarrow$ previously best, improvable?
- Lower complexity bound $\Omega(\varepsilon^{-2})$

Proximal bundle method is not optimal in general

We close the gap by showing the tight upper bound $\mathcal{O}(\varepsilon^{-2})$
through a new proximal bundle method and a refined analysis

Motivation - Proximal Bundle Method

Goal: find $\hat{x}$ such that $\phi(\hat{x}) - \phi_* \leq \varepsilon$

- Subgradient, Mirror descent, Bundle-level, and Prox Level method are optimal.
- Proximal bundle method $\mathcal{O}(\varepsilon^{-3})$ $\leftarrow$ previously best, improvable?
- Lower complexity bound $\Omega(\varepsilon^{-2})$

Proximal bundle method is not optimal in general

We close the gap by showing the tight upper bound $\mathcal{O}(\varepsilon^{-2})$
through a new proximal bundle method and a refined analysis

Motivation - Proximal Bundle Method

Goal: find $\hat{x}$ such that $\phi(\hat{x}) - \phi_* \leq \varepsilon$

- Subgradient, Mirror descent, Bundle-level, and Prox Level method are optimal.
- Proximal bundle method $\mathcal{O}(\varepsilon^{-3})$ $\leftarrow$ previously best, improvable?
- Lower complexity bound $\Omega(\varepsilon^{-2})$

Proximal bundle method is not optimal in general

We close the gap by showing the tight upper bound $\mathcal{O}(\varepsilon^{-2})$
through a new proximal bundle method and a refined analysis

Review of the Proximal Bundle Method

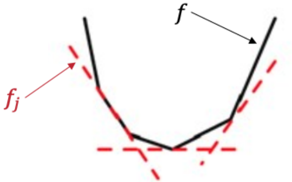
Proximal point framework: constructs a sequence of proximal problems.

Approximately solve the proximal problem by an iterative process

$$x^+ \leftarrow \min_{z\in\mathbb{R}^n} \left\{ f(z) + h(z) + \frac{1}{2\lambda}\|z - x^c\|^2 \right\}.$$

Recursively build up a cutting-plane model

$$f_j(z) = \max\{f(z_i) + \langle f'(z_i), z - z_i\rangle : 0 \le i \le j-1\}$$



Relaxed Proximal Bundle Method (L. and Monteiro, 2021)

Consider a proximal problem

$$\min_{u \in \mathbb{R}^n} \left\{ f(u) + h(u) + \frac{1}{2\lambda}\|u - x^c\|^2 \right\}$$

Algorithm RPB (one stage)

If find an $(\varepsilon/2)$-solution to the current proximal problem, **then** change the prox-center; $\leftarrow$ serious

Otherwise, keep the prox-center, update the cutting-plane model and solve the prox subproblem based on the current model, i.e., $\leftarrow$ null

$$x_j = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ f_j(u) + h(u) + \frac{1}{2\lambda}\|u - x^c\|^2 \right\}.$$

Main Results (L. and Monteiro, 2021)

We establish **improved** upper bounds and **matching** lower bounds.

Table: Upper and lower complexity bounds

	Convex	Strongly convex
Upper bound	$\mathcal{O}\left(\frac{M^2 d_0^2}{\varepsilon^2}\right)$	$\mathcal{O}\left(\frac{M^2}{\mu\varepsilon}\log\frac{\mu d_0^2}{\varepsilon}\right)$
Lower bound	$\Omega\left(\frac{M^2 d_0^2}{\varepsilon^2}\right)$	$\Omega\left(\frac{M^2}{\mu\varepsilon}\right)$

Optimal for convex and nearly optimal for strongly convex

Outline

Motivation

Main problem

$$\phi_* := \min_{x \in \mathbb{R}^n} \{\phi(x) := f(x) + h(x)\}, \quad f(x) = \mathbb{E}_\xi[F(x,\xi)]$$

Applications: Two-stage SP, Statistical learning, Statistical inference

$$\min_{P_\theta \in \mathcal{P}_2(\mathbb{R}^n)} KL(P_{\theta_0}||P_\theta) = \min_{P_\theta \in \mathcal{P}_2(\mathbb{R}^n)} \int \log \frac{P_{\theta_0}}{P_\theta} P_{\theta_0}(x)dz$$
$$= \int \log P_{\theta_0} P_{\theta_0}(z)dz - \max_{\theta \in \Theta} \mathbb{E}_{z \sim P_{\theta_0}} [\log P_\theta(z)].$$

Maximum likelihood estimation (MLE) is a sample average approximation (SAA)

$$\max_{\theta \in \Theta} \left\{ \ell(\theta|Z) := \frac{1}{N} \sum_{i=1}^{N} \log P_\theta(Z_i) \right\} \quad \leftarrow \text{offline}$$

Goal: stochastic approximation (SA) based on proximal bundle ← online

Motivation

Main problem

$$\phi_* := \min_{x \in \mathbb{R}^n} \{\phi(x) := f(x) + h(x)\}, \quad f(x) = \mathbb{E}_\xi[F(x, \xi)]$$

Applications: Two-stage SP, Statistical learning, Statistical inference

$$\begin{aligned}\min_{P_\theta \in \mathcal{P}_2(\mathbb{R}^n)} KL(P_{\theta_0} || P_\theta) &= \min_{P_\theta \in \mathcal{P}_2(\mathbb{R}^n)} \int \log \frac{P_{\theta_0}}{P_\theta} P_{\theta_0}(x) dz \\ &= \int \log P_{\theta_0} P_{\theta_0}(z) dz - \max_{\theta \in \Theta} \mathbb{E}_{z \sim P_{\theta_0}} [\log P_\theta(z)].\end{aligned}$$

Maximum likelihood estimation (MLE) is a sample average approximation (SAA)

$$\max_{\theta \in \Theta} \left\{ \ell(\theta | Z) := \frac{1}{N} \sum_{i=1}^{N} \log P_\theta(Z_i) \right\} \quad \leftarrow \text{offline}$$

Goal: stochastic approximation (SA) based on proximal bundle $\leftarrow$ online

Motivation

Main problem

$$\phi_* := \min_{x\in\mathbb{R}^n} \left\{\phi(x) := f(x) + h(x)\right\}, \quad f(x) = \mathbb{E}_\xi[F(x,\xi)]$$

Applications: Two-stage SP, Statistical learning, Statistical inference

$$\begin{aligned}\min_{P_\theta\in\mathcal{P}_2(\mathbb{R}^n)} KL(P_{\theta_0}||P_\theta) &= \min_{P_\theta\in\mathcal{P}_2(\mathbb{R}^n)} \int \log \frac{P_{\theta_0}}{P_\theta} P_{\theta_0}(x)dz \\ &= \int \log P_{\theta_0} P_{\theta_0}(z)dz - \max_{\theta\in\Theta} \mathbb{E}_{z\sim P_{\theta_0}}[\log P_\theta(z)].\end{aligned}$$

Maximum likelihood estimation (MLE) is a sample average approximation (SAA)

$$\max_{\theta\in\Theta}\left\{\ell(\theta|Z) := \frac{1}{N}\sum_{i=1}^{N} \log P_\theta(Z_i)\right\} \quad \leftarrow \text{offline}$$

Goal: stochastic approximation (SA) based on proximal bundle $\leftarrow$ online

Assumptions

Stochastic convex composite optimization

$$\phi_* := \min\left\{\phi(x) := f(x) + h(x) : x \in \mathbb{R}^n\right\}, \quad f(x) = \mathbb{E}_\xi[F(x,\xi)]$$

(A1) unbiased estimators

$$\mathbb{E}[F(x,\xi)] = f(x), \quad \mathbb{E}[s(x,\xi)] = f'(x) \in \partial f(x);$$

(A2) bounded variance

$$\mathbb{E}[\|s(x,\xi)\|^2] \le M^2.$$

A Motivating Question

- Stochastic gradient descent, $\min_x \mathbb{E}_\xi[F(x,\xi)]$

$$x_{k+1} = x_k - \lambda_k s(x_k,\xi_k), \quad s(x_k,\xi_k) \in \partial F(x_k,\xi_k)$$

Approximation by a single cut: $\mathbb{E}[f(y) + \langle s(y;\xi), x - y\rangle] \leq f(x)$

A Motivating Question

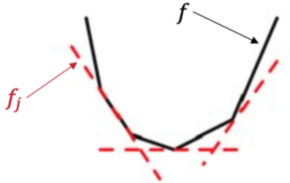
- Stochastic gradient descent, $\min_x \mathbb{E}_\xi[F(x,\xi)]$

$$x_{k+1} = x_k - \lambda_k s(x_k,\xi_k), \quad s(x_k,\xi_k) \in \partial F(x_k,\xi_k)$$

Approximation by a single cut: $\mathbb{E}[f(y) + \langle s(y;\xi), x-y\rangle] \le f(x)$

- Cutting-plane model: approximation by multiple cuts

$$f_j(x) = \max\{f(x_i) + \langle f'(x_i), x - x_i\rangle : 0 \le i \le j-1\} \le f(x)$$



- In the stochastic setting, is it still true?

$$\mathbb{E}[f_j(x)] \le f(x)?$$

A Motivating Question

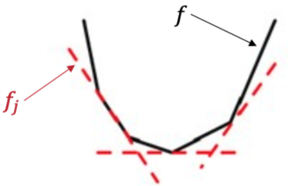
- Stochastic gradient descent, $\min_x \mathbb{E}_\xi[F(x,\xi)]$

$$x_{k+1} = x_k - \lambda_k s(x_k,\xi_k), \quad s(x_k,\xi_k) \in \partial F(x_k,\xi_k)$$

  Approximation by a single cut: $\mathbb{E}[f(y) + \langle s(y;\xi), x-y\rangle] \leq f(x)$

- Cutting-plane model: approximation by multiple cuts

$$f_j(x) = \max\{f(x_i) + \langle f'(x_i), x - x_i\rangle : 0 \leq i \leq j-1\} \leq f(x)$$



- In the stochastic setting, is it still true?

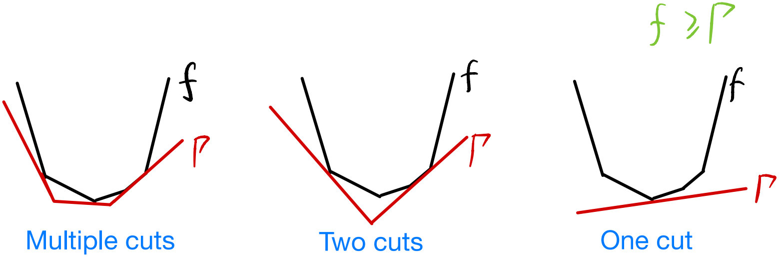
$$\mathbb{E}[f_j(x)] \leq f(x)?$$

Other bundle models

(E1) **single cut update**[1]**:** $\Gamma^+ = \Gamma^+_\tau := \tau\Gamma + (1-\tau)\ell_f(\cdot;x)$.

(E2) **two cuts update:** $\Gamma^+ = \max\{A_f^+, \ell_f(\cdot;x)\}$ where
$A_f^+ = \theta A_f + (1-\theta)\ell_f(\cdot;x^-)$.

Bundle of past information $\{(x_i, f(x_i), f'(x_i))\}$



[1]Liang and Monteiro, 2021. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems.

Convergence of SCPB

Let pair (λ, K) and constant $m \geq 1$ be given

- Number of iterations within $\mathcal{C}_k$, or number of null steps

$$|\mathcal{C}_k| \leq \left\lceil (m+1) \ln\left(\frac{\lambda k}{C} + 1\right) \right\rceil + 1.$$

- Convergence of SCPB

$$\mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{2D^2}{\lambda K} + \frac{2\lambda M^2}{m}.$$

- Its expected overall iteration complexity is $\tilde{\mathcal{O}}(mK)$.

Comparison with Robust Stochastic Approximation [2]

RSA is basically SGD with constant stepsize λ

$$\text{RSA: } \mathbb{E}[\phi(x_K^a)] - \phi_* \leq \frac{2D^2}{\lambda K} + 2\lambda M^2$$

$$\text{SCPB: } \mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \leq \frac{2D^2}{\lambda K} + \frac{2\lambda M^2}{m}$$

Taking the optimal stepsize for SCPB $\lambda = \frac{\sqrt{m}D}{M\sqrt{K}}$

- RSA has iteration complexity $\mathcal{O}\left(\frac{mM^2D^2}{\varepsilon^2}\right)$;
- SCPB has iteration complexity $\tilde{\mathcal{O}}\left(\frac{M^2D^2}{\varepsilon^2}\right)$.

[2] Nemirovski, Juditsky, Lan and Shapiro, 2009. Robust stochastic approximation approach to stochastic programming.

Comparison with Robust Stochastic Approximation [2]

RSA is basically SGD with constant stepsize λ

$$\text{RSA: } \mathbb{E}[\phi(x_K^a)] - \phi_* \le \frac{2D^2}{\lambda K} + 2\lambda M^2$$

$$\text{SCPB: } \mathbb{E}[\phi(\hat{y}_K^a)] - \phi_* \le \frac{2D^2}{\lambda K} + \frac{2\lambda M^2}{m}$$

Taking the optimal stepsize for SCPB $\lambda = \frac{\sqrt{m}D}{M\sqrt{K}}$

- RSA has iteration complexity $\mathcal{O}\left(\frac{mM^2D^2}{\varepsilon^2}\right)$;
- SCPB has iteration complexity $\tilde{\mathcal{O}}\left(\frac{M^2D^2}{\varepsilon^2}\right)$.

[2] Nemirovski, Juditsky, Lan and Shapiro, 2009. Robust stochastic approximation approach to stochastic programming.

Two-stage Stochastic Program

$$\begin{cases} \min \; c^T x_1 + \mathbb{E}[Q(x_1,\xi)] \\ x_1 \in \mathbb{R}^n : x_1 \geq 0, \sum_{i=1}^n x_1(i) = 1 \end{cases}$$

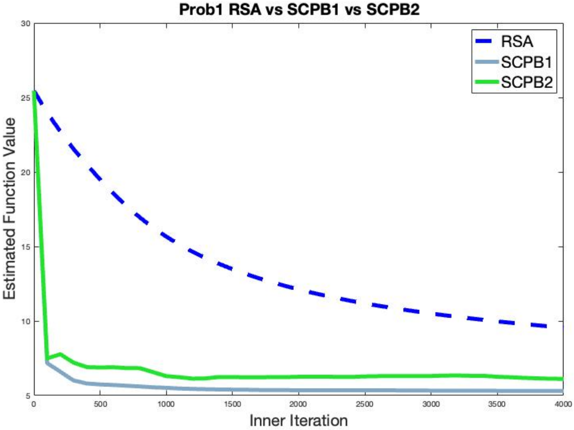
where the second stage recourse function is given by

$$Q(x_1,\xi) = \begin{cases} \min\limits_{x_2 \in \mathbb{R}^n} \; \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \left(\xi\xi^T + \lambda_0 I_{2n}\right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \xi^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ x_2 \geq 0, \sum\limits_{i=1}^n x_2(i) = 1. \end{cases}$$

Table: $n = 50, \; N = 4000$

Statistics	RSA	SCPB1	SCPB2
λ	7.4×10^{-7}	10^{-3}	10^{-3}
Min Inner	1	9	2
Max Inner	1	52	43
Avg Inner	1	43	5

Two-stage Stochastic Program



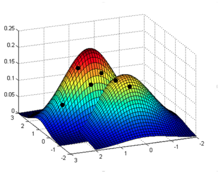
Take-away Message

- Optimal complexity for large stepsizes
- Non-trivial variance reduction by PPF

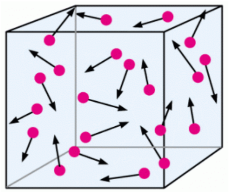
Outline

Sampling - Generation from Data

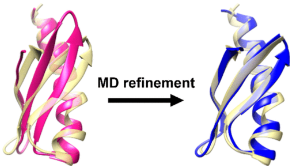
Sample from a probability distribution $\propto \exp(-f(x))$ where f has certain properties, such as convexity and smoothness



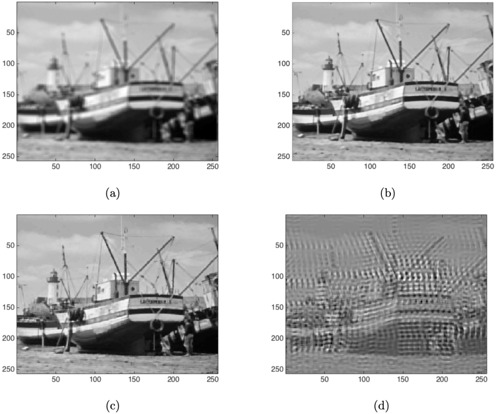
Extensively used in Bayesian inference and scientific computing



(e) Statistical Mechanics



(f) Molecular Dynamics



$$p(\mathcal{M}_1|y) = 0.964, \quad p(\mathcal{M}_2|y) = 0.036, \quad p(\mathcal{M}_3|y) < 0.001$$

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 $L_{\alpha_i} > 0$, $i = 1, \ldots, n$, s.t.

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

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

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

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 $L_{\alpha_i} > 0$, $i = 1, \ldots, n$, s.t.

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

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

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

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 $L_{\alpha_i} > 0$, $i = 1, \ldots, n$, s.t.

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

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.

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 $L_{\alpha_i} > 0$, $i = 1, \ldots, n$, s.t.

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

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

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

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

Algorithm 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

Alternating Sampling Framework

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

Algorithm 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

RGO Implementation

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

Algorithm 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 $h_1(x) \leq f_y^\eta(x)$, construct the proposal as a Gaussian

Rejection Sampling Efficiency (L. and Chen, 2022)

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 RGO Rejection Sampling is at most $\exp\left(\frac{3(1-\alpha)\delta}{2} + 3\right)$.

Proposition

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

Rejection Sampling Efficiency (L. and Chen, 2022)

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 RGO Rejection Sampling is at most $\exp\left(\frac{3(1-\alpha)\delta}{2} + 3\right)$.

Proposition

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

ASF Complexity

Another ingredient for total complexity: **Convergence rate analysis of ASF**

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 (L. and Chen, 2022)

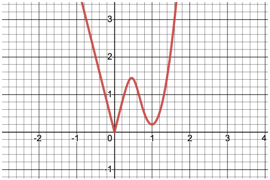
Theorem

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

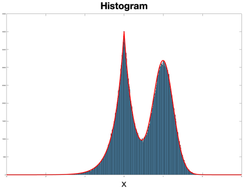
$$\tilde{\mathcal{O}}\left(C_{\mathrm{PI}} L_\alpha^{\frac{2}{\alpha+1}} d\right)$$

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.

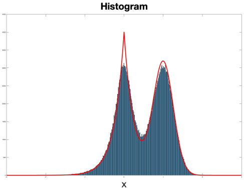
$$\nu(x) = 0.5(2\pi)^{-d/2}\sqrt{\det Q}\,\exp\left(-\frac{1}{2}(x-\mathbf{1})^\top Q(x-\mathbf{1})\right) + 0.5(2^d)\exp(-\|4x\|_1)$$



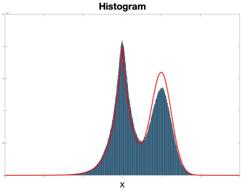
(g) $f(x) = -\ln \nu(x)$



(h) Histogram ASF



(i) Histogram ULA



(j) Histogram ULA with small η

Conclusion

- A universal proximal framework
  - Nonsmooth optimization
  - Stochastic optimization
  - High-dimensional sampling
  - Beyond gradient descent
  - Restarted Nesterov's acceleratedgradient
- Optimization and sampling + X

  statistical signal processing, medical imaging, biostatistics, ...

References

- Chen, Chewi, Salim, and Wibisono. Improved Analysis for a Proximal Algorithm for Sampling. COLT 2022
- Chewi, Erdogdu, Li, Shen, and Zhang. Analysis of Langevin Monte Carlo from Poincare to Log-Sobolev. COLT 2022
- Lee, Shen, and Tian. Structured Logconcave Sampling with a Restricted Gaussian Oracle. COLT 2021
- L. and Monteiro. A Proximal Bundle Variant with Optimal Iteration-complexity for A Large Range of Prox Stepsizes. SIOPT 2021
- L. and Monteiro. A Unified Analysis of A Class of Proximal Bundle Methods for Solving Hybrid Convex Composite Optimization Problems. MOR 2023
- L., Guigues, and Monteiro. A Single Cut Proximal Bundle Method for Stochastic Convex Composite Optimization. 2022
- L. and Chen. A Proximal Algorithm for Sampling. TMLR 2022
- Nemirovski, Juditsky, Lan, and Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. SIOPT 2009
- Nguyen, Dang, and Chen. Unadjusted Langevin Algorithm for Non-convex Weakly Smooth Potentials. 2021

Thank you!