Proximal Oracles for Optimization and Sampling

Jiaming Liang

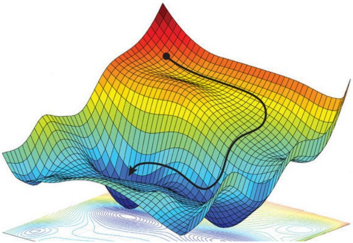
Department of Computer Science & Goergen Institute for Data Science
Univerity of Rochester

March 24, 2024

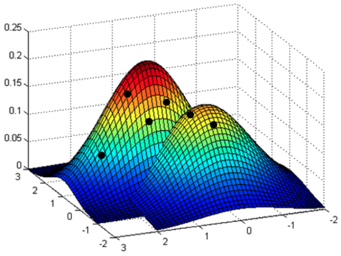
Joint work with Yongxin Chen (Georgia Tech)

IOS Conference 2024, Houston, Texas

Algorithm design for optimization and sampling using proximal oracles.



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



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

Algorithms for Optimization and Sampling

- Gradient descent

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

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

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

  equivalent to sampling $x_{k+1} \sim p(y|x_k)$ where

$$p(y|x_k) \propto \exp\left(-\frac{1}{2\eta}\|x - (x_k - \eta \nabla f(x_k))\|^2\right)$$

Sampling as Optimization

Langevin dynamics

$$\mathrm{d}X_t = -\nabla f(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

Fokker-Planck equation (continuity equation)

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu}\right) = -\mathsf{grad}_{\rho_t} H_\nu(\rho_t)$$

Jordan, Kinderlehrer, and Otto 1998: Langevin dynamics in space corresponds to the gradient flow of the relative entropy in the space of measures with the Wasserstein metric

$$\min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ H_\nu(\rho) = \int \rho \log \frac{\rho}{\nu} \right\}$$

Metropolis-Hastings

ULA is biased: $\rho_k \to \bar{\rho}$ as $k \to \infty$, but $H_\nu(\rho_\infty) > 0$.

Metropolis-adjusted Langevin algorithm: take ULA as a proposal density $p(\cdot|x_k)$, draw $y_k \sim p(y_k|x_k)$, and accept y_k with probability

$$\min\left\{1, \frac{\nu(y_k)p(x_k|y_k)}{\nu(x_k)p(y_k|x_k)}\right\}$$

MH filter makes the Markov chain reversible and hence ν is the stationary distribution.

Gibbs Sampling

Joint distribution

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

Gibbs sampling:

- given x_k, sample $y_k \sim \pi^{Y|X}(\cdot|x_k)$
- given y_k, sample $x_{k+1} \sim \pi^{X|Y}(\cdot|y_{k+1})$

It is known from Gibbs sampling that $(x_k, y_k)_{k\geq 1}$ form a reversible MC with stationary distribution $\pi(x,y)$, whose x-marginal is $\nu(x) \propto \exp(-f(x))$.

Proximal Frameworks

Optimization

Algorithm Proximal Point Framework

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

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

Sampling

Algorithm Alternating Sampling Framework (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 $x_{k+1} \sim \pi^{X|Y}(x \mid y_k) \propto \exp[-f(x) - \frac{1}{2\eta}\|x - y_k\|^2]$

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

Assumptions

(A1) f is semi-smooth, i.e., there exist $\alpha \in [0,1]$ and $L_\alpha > 0$, s.t.

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

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

(A2) f is composite, i.e., there exist $\alpha_i \in [0,1]$ and $L_{\alpha_i} > 0$, $i = 1, \ldots, n$, s.t.

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

Outline

1. Regularized Cutting-Plane Method
2. Adaptive Proximal Bundle Method
3. Proximal Sampling Algorithm

Outline

1. Regularized Cutting-Plane Method
2. Adaptive Proximal Bundle Method
3. Proximal Sampling Algorithm

Regularized Cutting-plane Method

Proximal subproblem

$$f_y^\eta(x^*) = \min_{x \in \mathbb{R}^d} \left\{ f_y^\eta(x) = f(x) + \frac{1}{2\eta}\|x - y\|^2 \right\}$$

Algorithm Regularized Cutting-Plane Method (RCPM)

1. Let $y \in \mathbb{R}^d$, $\eta > 0$, and $\delta > 0$ be given, and set $x_0 = \tilde{x}_0 = y$, and $j = 1$.
2. Update $f_j(x) = \max_{0 \le i \le j-1} \{f(x_i) + \langle f'(x_i), x - x_i \rangle\}$.
3. Compute

$$x_j = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f_j^\eta(x) := f_j(x) + \frac{1}{2\eta}\|x - y\|^2 \right\},$$

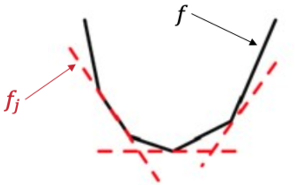
$$\tilde{x}_j = \operatorname{argmin} \left\{ f_y^\eta(x) : x \in \{x_j, \tilde{x}_{j-1}\} \right\}.$$

4. If $f_y^\eta(\tilde{x}_j) - f_j^\eta(x_j) \le \delta$, then **stop** and **return** $J = j, x_J, \tilde{x}_J$; else, go to step 5.
5. Set $j \leftarrow j + 1$ and go to step 2.

Cutting-Plane Model

Recursively build up a cutting-plane model

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



Convergence Analysis

Define $\delta_j := f_y^\eta(\tilde{x}_j) - f_j^\eta(x_j)$. Note that $\delta_j \geq f_y^\eta(\tilde{x}_j) - f_j^\eta(x^*)$.

Recall that we want to find $\delta_J \leq \delta$. If $\delta_j > \delta$, then $(1+\beta)\delta_j \leq \delta_{j-1}$ where

$$\beta = \frac{1}{2\eta}\left(\frac{\alpha+1}{L_\alpha}\right)^{\frac{2}{\alpha+1}} \delta^{\frac{1-\alpha}{\alpha+1}}.$$

The complexity is $\tilde{\mathcal{O}}(\beta^{-1}+1)$

Theorem

If f is semi-smooth, RCPM takes $\tilde{\mathcal{O}}\left(\eta L_\alpha^{\frac{2}{\alpha+1}}\left(\frac{1}{\delta}\right)^{\frac{1-\alpha}{\alpha+1}}+1\right)$ iterations to terminate.

If f is composite, RCPM takes $\tilde{\mathcal{O}}\left(\eta M+1\right)$ iterations to terminate, where

$$M = \sum_{i=1}^{n} \frac{L_{\alpha_i}^{\frac{2}{\alpha_i+1}}}{\left[(\alpha_i+1)\delta\right]^{\frac{1-\alpha_i}{\alpha_i+1}}}.$$

Outline

Intuition

$$\text{Goal: } \min_{x \in \mathbb{R}^d} f(x) \text{ where } f \text{ is semi-smooth}$$

proximal bundle method $\approx$ proximal point framework + RCPM

Inner complexity is $\tilde{\mathcal{O}}(\beta^{-1}+1)$. In practice, it is desirable to have a relatively small number, say 10. Prescribe this number by choosing $\beta_0 \in (0,1])$ and check

$$(1+\beta_0)\delta_j \leq \delta_{j-1}.$$

If always true, we have complexity $\tilde{\mathcal{O}}(\beta_0^{-1}+1)$. Otherwise, reduce η in the next cycle. This is because $(1+\beta)\delta_j \leq \delta_{j-1}$ where

$$\beta = \frac{1}{2\eta}\left(\frac{\alpha+1}{L_\alpha}\right)^{\frac{2}{\alpha+1}} \delta^{\frac{1-\alpha}{\alpha+1}}.$$

This approach is **adaptive and parameter-free**.

Adaptive Proximal Bundle Method

Inequality to check

$$(1+\beta_0)\delta_j \le \delta_{j-1}. \tag{*}$$

Algorithm Adaptive Proximal Bundle Method (APBM)

1. Let $y_0 \in \mathbb{R}^d$, $\eta_0 > 0$, $\beta_0 \in (0,1]$, and $\varepsilon > 0$ be given, and set $k = 1$
2. Call RCPM with $(y, \eta, \delta) = (y_{k-1}, \eta_{k-1}, \varepsilon/2)$ and output $(y_k, \tilde{y}_k) = (x_J, \tilde{x}_J)$
3. In the execution of RCPM, if (*) is always true, then set $\eta_k = \eta_{k-1}$; otherwise, set $\eta_k = \eta_{k-1}/2$
4. Set $k \leftarrow k+1$ and go to step 2.

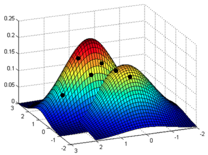
Theorem

The complexity of APBM to find an ε-solution is $\tilde{\mathcal{O}}\left(\frac{L_\alpha^{\frac{2}{\alpha+1}}\|y_0 - x_*\|^2}{\varepsilon^{\frac{2}{\alpha+1}}} + 1\right)$.

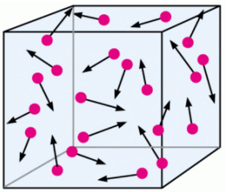
Outline

Sampling - Generation from Data

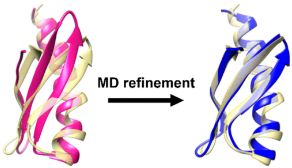
Sample from a probability distribution $\nu \propto \exp\left(-f(x)\right)$ where f has certain properties, such as convexity and smoothness



Extensively used in Bayesian inference and scientific computing



(c) Statistical Mechanics



(d) Molecular Dynamics

Proximal Sampling

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

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.

Proximal Sampling

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

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.

RGO Implementation

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

Algorithm RGO Rejection Sampling

1. Run RCPM to compute x_J and $\tilde{x}_J$
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.

Rejection Sampling

Define

$$h_1 := \frac{1}{2\eta}\| \cdot - x_J\|^2 + f_y^\eta(\tilde{x}_J) - \delta,$$
$$h_2 := \frac{1}{2\eta}\| \cdot - x^*\|^2 + \frac{L_\alpha}{\alpha+1}\| \cdot - x^*\|^{\alpha+1} + f_y^\eta(x^*).$$

We have $h_1(x) \leq f_y^\eta(x) \leq h_2(x)$.

The intuition is to build a proposal as a Gaussian close to $\exp(-f_y^\eta(x))$. Similar to the Laplace approximation of a density.

Known for RJ: X is an unbiased sample from $\exp(-f_y^\eta(x))$ and the probability that X is accepted is

$$\mathbb{P}\left(U \leq \frac{\exp(-f_y^\eta(X))}{\exp(-h_1(X))}\right) = \frac{\int \exp(-f_y^\eta(x))\mathrm{d}x}{\int \exp(-h_1(x))\mathrm{d}x} \geq \frac{\int \exp(-h_2(x))\mathrm{d}x}{\int \exp(-h_1(x))\mathrm{d}x}.$$

Rejection Sampling Efficiency

Lemma

Let $\alpha \in [0,1]$, $\eta > 0$, $a \geq 0$ *and* $d \geq 1$. *If*

$$2a(\eta d)^{(\alpha+1)/2} \leq 1,$$

then

$$\int \exp\left(-\frac{1}{2\eta}\|x\|^2 - a\|x\|^{\alpha+1}\right) \mathrm{d}x \geq \frac{(2\pi\eta)^{d/2}}{2}.$$

Proposition

Assume f *is convex and* L_α*-semi-smooth. If*

$$\eta \leq \frac{(\alpha+1)^{\frac{2}{\alpha+1}}}{(2L_\alpha)^{\frac{2}{\alpha+1}} d},$$

then the expected number of iterations in the rejection sampling of RGO is at most $2\exp(\delta)$.

ASF Complexity

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

Theorem (Chen, Chewi, Salim and Wibisono 2022)

If $\nu \propto \exp(-f)$ *is log-concave, then* x_k *of ASF* $\sim \rho_k$*, which satisfies*

$$H_\nu(\rho_k) \leq \frac{W_2^2(\rho_0, \nu)}{k\eta}.$$

If $\nu \propto \exp(-f)$ *satisfies log-Sobolev inequality with* $C_{LSI} > 0$*, then*

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

If $\nu \propto \exp(-f)$ *satisfies Poincaré inequality with* $C_{\mathrm{PI}} > 0$*, then*

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

Total Complexity

Combining complexities of **ASF**, **RGO**, and **RCPM**

Theorem

Assume f is convex and L_α-semi-smooth, then ASF using the RGO implementation, initialized with ρ_0 and stepsize $\eta \asymp 1/(L_\alpha^{\frac{2}{\alpha+1}} d)$, has the iteration-complexity bound

$$\mathcal{O}\left(\frac{L_\alpha^{\frac{2}{\alpha+1}} d W_2^2(\rho_0, \nu)}{\varepsilon}\right) \tag{1}$$

to achieve ε error to the target $\nu \propto \exp(-f)$ in terms of KL divergence. Each RGO requires $\tilde{\mathcal{O}}\left(\frac{1}{d}\left(\frac{1}{\delta}\right)^{\frac{1-\alpha}{\alpha+1}} + 1\right)$ subgradient evaluations of f and $2\exp(\delta)$ rejection steps in expectation.

Generalize to LSI, PI, composite.

Approximation

Algorithm 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 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

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

Conclusion

Interplay between optimization and sampling

- Proximal frameworks
  - Proximal point framework
  - Alternating sampling framework
- Proximal oracles
  - Proximal map
  - Restricted Gaussian oracle
- Applications
  - Adaptive proximal bundle method
  - Proximal sampling algorithm
- Simplifications
  - Subgradient method
  - Unadjusted Langevin algorithm

Future directions: Parameter-free sampling? Acceleration in sampling?

References

- Chen, Chewi, Salim, and Wibisono. Improved Analysis for a Proximal Algorithm for Sampling. Conference on Learning Theory 2022
- Jordan, Kinderlehrer, and Otto. The variational formulation of the Fokker–Planck equation. SIAM Journal on Mathematical Analysis 1998
- Lee, Shen, and Tian. Structured Logconcave Sampling with a Restricted Gaussian Oracle. Conference on Learning Theory 2021

Thank you!