A Proximal Algorithm for Sampling from Non-smooth Potentials

Jiaming Liang

Department of Computer Science
Yale University

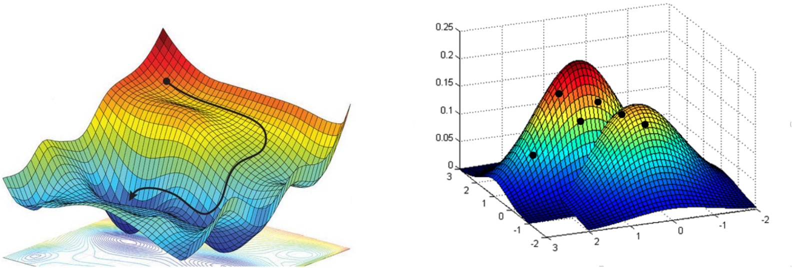
December 13, 2022

2022 Winter Simulation Conference, Singapore

Joint work 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 convex and nonsmooth potentials

- Improved complexity to sample from a distribution ε-close to the target distribution in total variation

- Close interplay between sampling and optimization

Story of the Smooth Setting

Sampling from $\nu(x) \propto \exp(-f(x))$ where f is convex and L-smooth, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d.$$

Starting from $x_0 \sim \rho_0$, unadjusted Langevin algorithm (ULA) iterates as

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

Suppose $x_k \sim \rho_k$ and define $\bar{\rho}_k = \frac{1}{k}\sum_{i=1}^{k} \rho_i$, then the iteration-complexity for ULA to obtain $KL(\bar{\rho}_k|\nu) \leq \varepsilon$ is

$$\mathcal{O}\left(\frac{W_2^2(\rho_0, \nu)Ld}{\varepsilon^2}\right),$$

where $KL(\rho|\nu) = \int \rho(x) \log \frac{\rho(x)}{\nu(x)} dx$.

Assumptions

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

- f is convex and M-Lipschitz continuous, i.e.,

  $$\|f(u) - f(v)\| \leq M\|u - v\|, \quad \forall u, v \in \mathbb{R}^d,$$

  or equivalently,

  $$\|f'(u)\| \leq M, \quad \forall u \in \mathbb{R}^d.$$

Typical approaches of dealing with nonsmoothness:

- Moreau envelop
- Gaussian smoothing
- Assuming existence of proximal mapping

Comparison

Source	Complexity	Convergence
Chatterji et al. 2020	$\tilde{\mathcal{O}}(M^6 d^5 \mathcal{M}_4^{3/2} \varepsilon^{-10})$	last iterate
Durmus et al. 2019	$\mathcal{O}(M^2 W_2^2(\rho_0, \nu) \varepsilon^{-4})$	average iterate
Lee and Vempala 2017	$\mathcal{O}(d^{5/2} \log(\beta \varepsilon^{-1}))$	last iterate
This work	$\tilde{\mathcal{O}}(M^2 d \mathcal{M}_4^{1/2} \varepsilon^{-1})$	last iterate

Table: Complexity bounds for sampling from convex nonsmooth potentials.

$\mathcal{M}_4$: 4th moment, $\mathcal{M}_4 \approx d^2$ in the isotropic case

β: warmness, $\log \beta \approx d$ if the initial distribution is not warm started

$W_2^2(\rho_0, \nu) \approx d$, $M \approx \sqrt{d}$ in typical problems

- Chatterji et al. 2020: $\tilde{\mathcal{O}}(M^6 d^8 \varepsilon^{-10})$
- Durmus et al. 2019: $\mathcal{O}(M^2 d \varepsilon^{-4})$
- Lee and Vempala 2017: $\tilde{\mathcal{O}}(d^{7/2})$
- This work: $\tilde{\mathcal{O}}(M^2 d^2 \varepsilon^{-1})$

Alternating Sampling Framework (ASF)

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

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

Theorem (Lee, Shen, and Tian 2021)

Let $\pi \propto \exp(-f)$ be a distribution on $\mathbb{R}^d$ and suppose f is μ-strongly convex. Let $\eta \in (0, 1/\mu]$ and $\varepsilon > 0$ be given. ASF, initialized at the minimizer of f, requires

$$\Theta\left(\frac{1}{\eta\mu}\log\frac{d}{\eta\mu\varepsilon}\right)$$

iterations to obtain a sample whose distribution is within ε total variation distance to π.

Alternating Sampling Framework (ASF)

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

Algorithm 2 ASF (Lee, Shen, and Tian 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

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

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

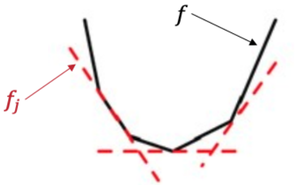
Relaxed Proximal Bundle Method (L. and Monteiro 2021)

f is convex and M-Lipschitz continuous.

Approximately solve (1) by the cutting-plane method

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

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



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

Sampling: ASF $\tilde{\mathcal{O}}((\eta\mu)^{-1})\times$ RGO **?**

RGO Implementation – with an Oracle

Goal: sample from $\exp(-g(x))$ where $g(x) := f(x) + \frac{\mu}{2}\|x - x^0\|^2$

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

$$x^* = \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ g_y^\eta(x) := g(x) + \frac{1}{2\eta}\|x - y\|^2 \right\}$$

Algorithm 4 RGO Rejection Sampling

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

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

Rejection Sampling

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

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

Want to find functions h_1 and h_2 such that

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

ii) $h_1(x) \leq g_y^\eta(x) \leq h_2(x) \quad \forall x \in \mathbb{R}^d$,

iii) (2) is bounded from below.

Construction

$$h_1(x) := \frac{1}{2\eta_\mu}\|x - x^*\|^2 + g_y^\eta(x^*),$$
$$h_2(x) := \frac{1}{2\eta_\mu}\|x - x^*\|^2 + 2M\|x - x^*\| + g_y^\eta(x^*).$$

Observations:
i) sampling from $\exp(-h_1(x))$ is easy;
ii) It follows from

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

and the fact that $g_y^\eta(x)$ is η_μ-strongly convex that

$$g_y^\eta(x) \geq g_y^\eta(x^*) + \frac{1}{2\eta_\mu}\|x - x^*\|^2 = h_1(x)$$

Proof Sketch of $g_y^\eta(x) \le h_2(x)$

It follows from the optimality condition of x^* that

$$0 \in \partial f(x^*) + \mu(x^* - x_0) + \frac{x^* - y}{\eta}, \quad -\mu(x^* - x_0) - \frac{x^* - y}{\eta} \in \partial f(x^*),$$

and hence that

$$\left\| \mu(x^* - x_0) + \frac{x^* - y}{\eta} \right\| \le M.$$

We have

$$\begin{aligned} g_y^\eta(x) - g_y^\eta(x^*) &\le f(x) - f(x^*) + \|x - x^*\| \left\| \mu(x^* - x_0) + \frac{x^* - y}{\eta} \right\| + \frac{1}{2\eta_\mu}\|x - x^*\|^2 \\ &\le 2M\|x - x^*\| + \frac{1}{2\eta_\mu}\|x - x^*\|^2, \end{aligned}$$

and hence

$$g_y^\eta(x) \le g_y^\eta(x^*) + 2M\|x - x^*\| + \frac{1}{2\eta_\mu}\|x - x^*\|^2 = h_2(x).$$

Remaining Question

Rejection sampling complexity

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

$$\int \exp(-h_1(x))dx = \int \exp\left(-\frac{1}{2\eta_\mu}\|x - x^*\|^2 - g_y^\eta(x^*)\right) dx$$
$$= \exp\left(-g_y^\eta(x^*)\right) (2\pi\eta_\mu)^{d/2}$$
$$\int \exp(-h_2(x))dx = \exp(-g_y^\eta(x^*)) \int \exp\left(-\frac{1}{2\eta_\mu}\|x - x^*\|^2 - 2M\|x - x^*\|\right) dx$$

Proposition (nontrivial)

For $\lambda > 0$, $a \geq 0$ and $d \geq 1$, if $\lambda \leq \frac{1}{4a^2 d}$, then

$$\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2\lambda}\|x\|^2 - a\|x\|\right) dx \geq \frac{(2\pi\lambda)^{d/2}}{2}.$$

RGO complexity

Proposition

Assume f is convex and M-Lipschitz continuous. If $\eta_\mu \le \frac{1}{16M^2 d}$, then the expected number of rejection steps in Algorithm 4 is at most 2.

Proof sketch

$$
\begin{aligned}
&\frac{\int \exp(-h_1(x))dx}{\int \exp(-h_2(x))dx} \\
=&\frac{\exp\left(-g_y^\eta(x^*)\right)(2\pi\eta_\mu)^{d/2}}{\exp(-g_y^\eta(x^*))\int \exp\left(-\frac{1}{2\eta_\mu}\|x-x^*\|^2 - 2M\|x-x^*\|\right)dx} \\
\le&\frac{(2\pi\eta_\mu)^{d/2}}{(2\pi\eta_\mu)^{d/2}/2} = 2.
\end{aligned}
$$

RGO Implementation – without an Oracle

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

$$x_J, \tilde{x}_J \approx \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ g_y^\eta(x) := g(x) + \frac{1}{2\eta} \|x - y\|^2 \right\}$$

Algorithm 5 RGO Rejection Sampling

1. Compute x_J and $\tilde{x}_J$ as in Algorithm 6;
2. Generate $X \sim \exp(-h_1(x))$;
3. Generate $U \sim \mathcal{U}[0,1]$;
4. If

$$U \leq \frac{\exp(-g_y^\eta(X))}{\exp(-h_1(X))},$$

then accept/return X; otherwise, reject X and go to step 2.

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

Cutting-plane method

Algorithm 6 Proximal Bundle Method Subroutine

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

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

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

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

$\tilde{x}_j$ is a δ-solution to $g_y^\eta(x)$

$$g_y^\eta(\tilde{x}_j) - g_y^\eta(x^*) \le g_y^\eta(\tilde{x}_j) - g_j^\eta(x_j) \le \delta$$

Construction

$$h_1 := \frac{1}{2\eta_\mu}\| \cdot - x_J\|^2 + g_y^\eta(\tilde{x}_J) - \delta,$$

$$h_2 := \frac{1}{2\eta_\mu}\| \cdot - \tilde{x}_J\|^2 + \left(2M + \frac{\sqrt{2\delta}}{\sqrt{\eta_\mu}}\right)\| \cdot - \tilde{x}_J\| + g_y^\eta(\tilde{x}_J).$$

Observations:
i) Sampling from $\exp(-h_1(x))$ is easy;
ii) It holds that $h_1(x) \le g_y^\eta(x) \le h_2(x) \quad \forall x \in \mathbb{R}^d$;
iii) RGO complexity is bounded from above.

Proposition

Assume f is convex and M-Lipschitz continuous. If

$$\eta_\mu \le \frac{1}{64M^2 d}, \quad \delta \le \frac{1}{32d},$$

then the expected number of rejection steps in Algorithm 5 is at most 2.

Remaining Question

Optimization complexity to find approximate solutions $x_J, \tilde{x}_J$ s.t.

$$g_y^\eta(\tilde{x}_J) - g_J^\eta(x_J) \le \delta.$$

Proposition

Algorithm 6 takes $\mathcal{O}(\eta_\mu M^2/\delta + 1)$ iterations to terminate, and each iteration takes one subgradient of f and solves an affinely constrained convex quadratic programming.

In particular, taking

$$\eta_\mu = \frac{1}{64M^2 d}, \quad \delta = \frac{1}{64d},$$

we have

$$\mathcal{O}\left(\frac{\eta_\mu M^2}{\delta} + 1\right) = \mathcal{O}(1).$$

Each RGO needs $\mathcal{O}(1)$ subgradients of f and $\mathcal{O}(1)$ samples from Gaussian distribution in expectation.

Main Results – Strongly Convex

Theorem

Let $x^0 \in \mathbb{R}^d$, $\varepsilon > 0$, $M > 0$, and $\mu > 0$ be given. Assume f is convex and M-Lipschitz continuous and let $g(x) = f(x) + \frac{\mu}{2}\|x - x^0\|^2$. Set

$$\delta = \frac{1}{64d}, \quad \eta = \frac{1}{64M^2 d}.$$

Then the ASF with Algorithm 5 as an RGO achieves ε error in terms of total variation with respect to the target distribution $\pi \propto \exp(-g)$ in $\tilde{\mathcal{O}}\left(\frac{M^2 d}{\mu}\right)$ iterations, and each iteration queries $\mathcal{O}(1)$ subgradient oracles of f and $\mathcal{O}(1)$ Gaussian distribution sampling oracles.

Main Results – Convex

Theorem

Let $\nu(x) \propto \exp(-f(x))$ *where* f *is convex and* M*-Lipschitz continuous on* $\mathbb{R}^d$*. Let* $x^0 \in \mathbb{R}^d$ *and* $\varepsilon > 0$ *be given and*

$$\mu = \frac{\varepsilon}{\sqrt{2}\left(\sqrt{\mathcal{M}_4} + \|x^0 - x_{min}\|^2\right)}$$

where $\mathcal{M}_4 = \int_{x \in \mathbb{R}^d} \|x - x_{min}\|^4 d\nu(x)$ *and* $x_{min} = \operatorname{argmin}\{f(x) : x \in \mathbb{R}^d\}$*. Choose*

$$\delta = \frac{1}{64d}, \quad \eta = \frac{1}{64M^2 d}.$$

and consider ASF using Algorithm 5 as an RGO for step 1, applied to $g(x) = f(x) + \frac{\mu}{2}\|x - x^0\|^2$*. Then, the iteration-complexity bound to achieve* ε *error to* ν *in terms of total variation is*

$$\tilde{\mathcal{O}}\left(\frac{M^2 d\left(\sqrt{\mathcal{M}_4} + \|x^0 - x_{min}\|^2\right)}{\varepsilon}\right).$$

Interpretation of Unadjusted Langevin Algorithm (ULA)

Algorithm 7 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 8 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

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

Conclusions

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

  f is convex and M-Lipschitz continuous

- Total complexity $\tilde{\mathcal{O}}\left(\frac{M^2 d\left(\sqrt{\mathcal{M}_4}+\|x^0-x_{\mathsf{min}}\|^2\right)}{\varepsilon}\right)$

  Each iteration takes $\mathcal{O}(1)$ 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.

References

- Liang and Chen. A Proximal Algorithm for Sampling from Non-smooth Potentials. WSC 2022
- Lee, Shen, and Tian. Structured Logconcave Sampling with a Restricted Gaussian Oracle. COLT 2021
- Durmus, Majewski, and Miasojedow. Analysis of Langevin Monte Carlo via Convex Optimization. JMLR 2019
- Chatterji, Diakonikolas, Jordan, and Bartlett. Langevin Monte Carlo without Smoothness. AISTATS 2020
- Lee and Vempala. Eldan's Stochastic Localization And the KLS Hyperplane Conjecture: An Improved Lower Bound for Expansion. FOCS 2017
- Liang and Monteiro. A Proximal Bundle Variant with Optimal Iteration-complexity for A Large Range of Prox Stepsizes. SIOPT 2021

Thank you!

Extensions – Improved ASF Complexities

Theorem (Chen, Chewi, Salim and Wibisono 2022)

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

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

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

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

Extensions – Nonconvex and Semi-smooth Potentials

Sampling from $\nu(x) \propto \exp(-f(x))$ where

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

Theorem (Liang and Chen 2022)

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 KL divergence. Each iteration queries $\tilde{\mathcal{O}}(1)$ subgradients of f and generates $\mathcal{O}(1)$ samples in expectation from Gaussian distribution.