

Zeroth-Order Methods for Nonconvex Stochastic Problems with Decision-Dependent Distributions

Yuya Hikima^{*1} and Akiko Takeda^{1,2}

¹Graduate School of Information Science and Technology, University of Tokyo, Tokyo, Japan

²Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan

Abstract

In this study, we consider an optimization problem with uncertainty dependent on decision variables, which has recently attracted attention due to its importance in machine learning and pricing applications. In this problem, the gradient of the objective function cannot be obtained explicitly because the decision-dependent distribution is unknown. Therefore, several zeroth-order methods have been proposed, which obtain noisy objective values by sampling and update the iterates. Although these existing methods have theoretical convergence for optimization problems with decision-dependent uncertainty, they require strong assumptions about the function and distribution or exhibit large variances in their gradient estimators. To overcome these issues, we propose two zeroth-order methods under mild assumptions. First, we develop a zeroth-order method with a new one-point gradient estimator including a variance reduction parameter. The proposed method updates the decision variables while adjusting the variance reduction parameter. Second, we develop a zeroth-order method with a two-point gradient estimator. There are situations where only one-point estimators can be used, but if both one-point and two-point estimators are available, it is more practical to use the two-point estimator. As theoretical results, we show the convergence of our methods to stationary points and provide the worst-case iteration and sample complexity analysis. Our simulation experiments with real data on a retail service application show that our methods output solutions with lower objective values than the conventional zeroth-order methods.

1. Introduction

In this study, we consider the following problem:

$$(P) \quad \min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x})} [f(\mathbf{x}, \boldsymbol{\xi})],$$

where $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is generally non-convex. The main feature of this problem is that the probability distribution $D(\mathbf{x})$ depends on the decision variable \mathbf{x} . This feature appears in a wide range of applications such as performative prediction [Perdomo et al., 2020] and price optimization [Ray et al., 2022, Hikima and Takeda, 2023]. Because of its importance for applications, various studies have addressed this problem [Perdomo et al., 2020, Mendler-Dünner et al., 2020, Ray et al., 2022, Chen et al., 2023].¹

In general, problem (P) is a non-convex problem: even if $f(\mathbf{x}, \boldsymbol{\xi})$ is convex with respect to \mathbf{x} , it is generally non-convex since the probability distribution $D(\mathbf{x})$ depends on \mathbf{x} . Moreover, there exists a difficulty: we cannot access the gradient of the objective function because the probability distribution $D(\mathbf{x})$ is unknown in practical applications.

^{*}Corresponding author. E-mail: yuya-hikima@g.ecc.u-tokyo.ac.jp

¹Codes of our experiments can be found in <https://github.com/Yuya-Hikima/AAAI25-Zeroth-Order-Methods-for-Nonconvex-Stochastic-Problems-with-Decision-Dependent-Distributions>.

Recently, several studies [Ray et al., 2022, Liu et al., 2024, Chen et al., 2023] have proposed zeroth-order methods for (P). Concretely, Ray et al. [2022] proposed a condition under which the objective function of (P) is strongly convex, and employed a zeroth-order method based on the following one-point gradient estimator:

$$\mathbf{g}_k := \frac{d}{\mu} f(\mathbf{x}_k + \mu \mathbf{v}_k, \boldsymbol{\xi}_k) \mathbf{v}_k,$$

where \mathbf{v}_k is a unit vector sampled uniformly from the unit sphere in dimension d , $\boldsymbol{\xi}_k \sim D(\mathbf{x}_k + \mu \mathbf{v}_k)$, and $\mu \in \mathbb{R}_{\geq 0}$. Liu et al. [2024] considered non-convex problem (P) with a state-dependent setting where the distribution evolves according to an underlying controlled Markov chain. They proposed a zeroth-order method based on the above one-point gradient estimator and showed its convergence to a stationary point. Chen et al. [2023] supposed that their optimization problem can be reduced to convex ones, and they employed a zeroth-order method based on the following gradient estimator:

$$\mathbf{g}_k := \frac{d}{2\mu} (f(\mathbf{x}_k + \mu \mathbf{v}_k, \boldsymbol{\xi}_k^1) - f(\mathbf{x}_k - \mu \mathbf{v}_k, \boldsymbol{\xi}_k^2)) \mathbf{v}_k,$$

where \mathbf{v}_k is a unit vector sampled uniformly from the unit sphere in dimension d , $\boldsymbol{\xi}_k^1 \sim D(\mathbf{x}_k + \mu \mathbf{v}_k)$, $\boldsymbol{\xi}_k^2 \sim D(\mathbf{x}_k - \mu \mathbf{v}_k)$, and $\mu \in \mathbb{R}_{\geq 0}$.

Although these existing studies have successfully proposed zeroth-order methods with theoretical convergence for problem (P), there exist limitations. First, zeroth-order methods with one-point gradient estimators [Ray et al., 2022, Liu et al., 2024] exhibit large variances of their gradient estimators, which may increase the number of iterations and sampling. Specifically, although the variances of their gradient estimators are bounded by using $G := \sup_{\mathbf{x}, \boldsymbol{\xi}} |f(\mathbf{x}, \boldsymbol{\xi})|$, G is generally large or unbounded. Second, the existing zeroth-order method [Chen et al., 2023] with a two-point gradient estimator supposes that the distribution $D(\mathbf{x})$ is included in a special exponential family and the function f is required to satisfy a certain (intuitively difficult to understand) inequality [Chen et al., 2023, Section 3, Appendix F].

In this paper, we propose two zeroth-order methods under mild assumptions, which do not restrict f and $D(\mathbf{x})$ to any particular one. First, we develop a new zeroth-order method with an improved one-point gradient estimator. Specifically, we propose the following gradient estimator including variance reduction parameter c_k :

$$\mathbf{g}_k := \frac{1}{\mu_k} (f(\mathbf{x}_k + \mu_k \mathbf{u}_k, \boldsymbol{\xi}_k) - c_k) \mathbf{u}_k, \quad (1)$$

where \mathbf{u}_k is sampled from the standard normal distribution $\mathcal{N}(0, I_d)$ in dimension d , $\boldsymbol{\xi}_k \sim D(\mathbf{x}_k + \mu_k \mathbf{v}_k)$, $\mu_k \in \mathbb{R}_{>0}$, and $c_k \in \mathbb{R}$. Then, we theoretically show that if c_k is close to $\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_k)} [f(\mathbf{x}_k, \boldsymbol{\xi})]$, the variance of the above gradient estimator is reduced. Based on this fact, we propose a zeroth-order method that updates the iterates \mathbf{x}_k using the proposed gradient estimator (1) while adjusting c_k . Second, we develop a zeroth-order method with the following two-point gradient estimator:

$$\mathbf{g}_k := \frac{1}{2\mu_k} (f(\mathbf{x}_k + \mu_k \mathbf{u}_k, \boldsymbol{\xi}_k^1) - f(\mathbf{x}_k - \mu_k \mathbf{u}_k, \boldsymbol{\xi}_k^2)) \mathbf{u}_k, \quad (2)$$

where \mathbf{u}_k is sampled from the standard normal distribution $\mathcal{N}(0, I_d)$ in dimension d , $\boldsymbol{\xi}_k^1 \sim D(\mathbf{x}_k + \mu_k \mathbf{v}_k)$, $\boldsymbol{\xi}_k^2 \sim D(\mathbf{x}_k - \mu_k \mathbf{v}_k)$, and $\mu_k \in \mathbb{R}_{>0}$. Although our method is similar to the zeroth-order method employed by [Chen et al., 2023], it does not assume strong assumptions for the distribution $D(\mathbf{x})$ and the function f . Moreover, we also include the Gaussian homotopy technique [Iwakiri et al., 2022] in both proposed methods to find potentially better stationary points.

As theoretical results, we show the convergence of our methods to stationary points and provide the worst-case iteration and sample complexity analysis. In particular, we show that the worst-case sample

complexity of our methods is $O(d^{\frac{9}{2}}\varepsilon^{-6})$. Since the sample complexity of the existing zeroth-order method [Liu et al., 2024] for the non-convex problem (P) is $O(G^6 d^2 \varepsilon^{-6})$, the proposed method has an advantage when $G = \sup_{\mathbf{x}, \boldsymbol{\xi}} |f(\mathbf{x}, \boldsymbol{\xi})|$ is large or unbounded.

We conducted simulation experiments with real-data on a retail service application. The results show that our methods output solutions with lower objective values than conventional zeroth-order methods.

Our contributions are as follows.

- We propose a new zeroth-order method with an improved one-point gradient estimator. When $G := \sup_{\mathbf{x}, \boldsymbol{\xi}} |f(\mathbf{x}, \boldsymbol{\xi})|$ is large or unbounded, the sample complexity of the proposed method has an advantage over existing zeroth-order methods [Ray et al., 2022, Liu et al., 2024], which also utilize one-point gradient estimators.
- We develop a zeroth-order method with a two-point gradient estimator under milder assumptions than the existing ones [Chen et al., 2023], which utilize a two-point gradient estimator. Although (P) is generally non-convex under our loose assumptions, we show the convergence of the method to stationary points and provide its sample complexity.

Notation. Bold lowercase symbols (e.g., \mathbf{x}, \mathbf{y}) denote vectors, and $\|\mathbf{x}\|$ denotes the Euclidean norm of a vector \mathbf{x} . Let $\mathbb{R}_{>0}$ ($\mathbb{R}_{\geq 0}$) be the set of positive (non-negative) real numbers. The gradient for a real-valued function $F(\mathbf{x})$ is denoted by $\nabla F(\mathbf{x})$. A binomial coefficient of a pair of integers m and n is written as $\binom{m}{n}$. Let $[N]$ be the set of $\{1, 2, \dots, N\}$. The standard normal distribution in dimension d is written as $\mathcal{N}(0, I_d)$.

2. Related Work

2.1 Zeroth-order Methods

Zeroth-order methods are a class of powerful optimization tools to solve many complex problems, whose explicit gradients are infeasible to access. Various types of zeroth-order methods have been proposed and their convergence has been analyzed [Ghadimi and Lan, 2013, Nesterov and Spokoiny, 2017]. Since then, multiple techniques for zeroth-order methods have been proposed, such as techniques to reduce the variance of the gradient estimator [Liu et al., 2018, Ji et al., 2019], to make the methods scalable through variable selection [Wang et al., 2018], and to deal with regularized problems [Cai et al., 2022].

Recently, zeroth-order methods have also been proposed for problems with decision-dependent uncertainty [Ray et al., 2022, Liu et al., 2024, Chen et al., 2023]. Ray et al. [2022] and Liu et al. [2024] proposed zeroth-order methods using one-point gradient estimators for (P) where the distribution changes dynamically with time. Chen et al. [2023] showed that (P) can be reduced to a convex optimization problem for particular combinations of the distribution $D(\mathbf{x})$ and the function f . They then employed a zeroth-order method using a two-point gradient estimator. In this study, we propose a new zeroth-order method that extends the conventional one-point gradient estimator used in [Ray et al., 2022, Liu et al., 2024] by including a variance reduction parameter. Moreover, we also develop a zeroth-order method with a two-point gradient estimator without strong assumptions on the distribution and function.

2.2 Other Methods for Stochastic Problems with Decision-dependent Uncertainty

We describe different techniques, other than zeroth-order methods, for solving (P).²

²Another formulation dealing with decision-dependent uncertainties is *decision-dependent distributionally robust optimization* [Luo and Mehrotra, 2020, Basciftci et al., 2021], which aims to find an optimal solution in the worst case when the probability distribution has ambiguity. Although such formulations are effective when the ambiguity set of the probability distribution is known, we consider situations where the probability distribution is completely unknown.

Retraining methods [Perdomo et al., 2020, Mendler-Dünner et al., 2020]. Retraining methods update the current iterate by fixing the distribution at each iteration. Specifically, Perdomo et al. [2020] proposed *repeated gradient descent*: $\mathbf{x}_{k+1} := \text{proj}_{\mathcal{C}}(\mathbf{x}_k - \eta_k \mathbb{E}_{\xi \sim D(\mathbf{x}_k)}[\nabla_{\mathbf{x}} f(\mathbf{x}_k, \xi)])$, where \mathcal{C} is the feasible region and $\text{proj}_{\mathcal{C}}$ is the Euclidean projection operator onto \mathcal{C} . They showed that the repeated gradient descent method converges to a *performatively stable point* defined as $\mathbf{x}_{\text{PS}} := \arg \min_{\mathbf{x}} \mathbb{E}_{\xi \sim D(\mathbf{x}_{\text{PS}})}[f(\mathbf{x}, \xi)]$. However, these methods assume the strong convexity of $f(\mathbf{x}, \xi)$ w.r.t. \mathbf{x} and are not applicable to our problem.

Stochastic gradient descent methods [Hikima and Takeda, 2023, Sutton and Barto, 2018]. These methods update the current iterate by the stochastic gradient such as $\mathbf{x}_{k+1} := \mathbf{x}_k - \eta_k \mathbf{g}_k^{\dagger}$, where \mathbf{g}_k^{\dagger} is an unbiased stochastic gradient for the objective function. Although these methods converge to stationary points, they require access to $\Pr(\xi | \mathbf{x})$ and $\nabla_{\mathbf{x}} \Pr(\xi | \mathbf{x})$ to find an unbiased stochastic gradient \mathbf{g}_k^{\dagger} . Since we assume that the probability density function of $D(\mathbf{x})$ is unknown, these methods are not applicable to our problem.

Two-stage approach [Miller et al., 2021]. This approach estimates a model of the distribution map $D(\cdot)$ in the first stage and optimizes the proxy function of the objective function by using the estimated distribution in the second stage. Since this approach assumes that the distribution map is included in location-scale families [Miller et al., 2021, Eq. (2)], it is not applicable to our problem.

Bayesian optimization [Brochu et al., 2010, Frazier, 2018]. This method is a process of learning for global optimization of black-box functions. When this method is applied to our problem, it is necessary to obtain a large number of samples $\xi \sim D(\mathbf{x})$ in order to widely search the entire space of decision variables. This is not desirable from a practical point of view, since it is necessary to deploy decisions (\mathbf{x}) in the real-world to obtain $\xi \sim D(\mathbf{x})$.

3. Preliminaries

3.1 Problem Definition

We restate the problem under consideration:

$$(P) \quad \min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) := \mathbb{E}_{\xi \sim D(\mathbf{x})}[f(\mathbf{x}, \xi)],$$

where F is generally non-convex. Note that problem (P) is a broad problem class that also includes the following optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \mathbb{E}_{\xi \sim D}[f(\mathbf{x}, \xi)].$$

Our goal is to obtain better solutions with low sample complexity for $D(\mathbf{x})$. This is because, to obtain a sample $\xi \sim D(\mathbf{x})$, a decision (\mathbf{x}) must be deployed in the real-world (e.g., a decision maker must sell some products at price \mathbf{x} to obtain a sample of demand $\xi \sim D(\mathbf{x})$), so the fewer samples the better.

3.2 Assumptions

We make the following assumptions.

Assumption 1. For any $\mathbf{x} \in \mathbb{R}^d$, there exists a constant $\sigma \in \mathbb{R}_{\geq 0}$ satisfying

$$\mathbb{E}_{\xi \sim D(\mathbf{x})}[(F(\mathbf{x}) - f(\mathbf{x}, \xi))^2] \leq \sigma^2. \quad (3)$$

Assumption 2. For any $\mathbf{x} \in \mathbb{R}^d$, there exists a constant $\alpha \in \mathbb{R}_{\geq 0}$ satisfying

$$W(D(\mathbf{x}), D(\mathbf{x}')) \leq \alpha \|\mathbf{x} - \mathbf{x}'\|,$$

where W represents the Wasserstein-1 distance.

Assumption 3. $f(\mathbf{x}, \boldsymbol{\xi})$ is $L_{\boldsymbol{\xi}}$ -Lipschitz continuous in $\boldsymbol{\xi}$. Moreover, $f(\mathbf{x}, \boldsymbol{\xi})$ is $L_{\mathbf{x}}$ -Lipschitz continuous in \mathbf{x} .

Assumption 4. $F(\mathbf{x})$ is H_F -smooth.

For Assumption 4, the following lemma has been given as a sufficient condition.

Lemma 1. [Ray et al., 2022, Lemma 1] Suppose that there exist matrix \mathbf{A} and distribution D' such that

$$\boldsymbol{\xi} \sim D(\mathbf{x}) \iff \boldsymbol{\xi} = \boldsymbol{\nu} + \mathbf{A}\mathbf{x},$$

where $\boldsymbol{\nu}$ has mean $\bar{\boldsymbol{\nu}} := \mathbb{E}_{\boldsymbol{\nu} \sim D'}[\boldsymbol{\nu}]$ and co-variance $\mathbb{E}_{\boldsymbol{\nu} \sim D'}[(\boldsymbol{\nu} - \bar{\boldsymbol{\nu}})(\boldsymbol{\nu} - \bar{\boldsymbol{\nu}})^\top]$. Moreover, suppose that $f(\mathbf{x}, \boldsymbol{\xi})$ is ρ -smooth with respect to both \mathbf{x} and $\boldsymbol{\xi}$. Then, Assumption 4 holds with

$$H_F := \sqrt{\rho^2(1 + \|\mathbf{A}\|_{\text{op}}^2) \max(1, \|\mathbf{A}\|_{\text{op}}^4)},$$

where $\|\mathbf{A}\|_{\text{op}}$ is the operator norm of \mathbf{A} .

Our assumptions are looser than those of an existing study [Ray et al., 2022] that tackles a problem similar to ours. Assumption 5 in [Ray et al., 2022] implies that $E_{\boldsymbol{\xi} \sim D(\mathbf{x})}[(F(\mathbf{x}) - f(\mathbf{x}, \boldsymbol{\xi}))^2] \leq (2G)^2$ for $G := \sup_{\mathbf{x}, \boldsymbol{\xi}} |f(\mathbf{x}, \boldsymbol{\xi})|$, which yields our Assumption 1. Moreover, Assumptions 1 and 3 in [Ray et al., 2022] yields our Assumptions 2–4. Conversely, we do not require Assumption 1(c) and Assumption 2 in [Ray et al., 2022]. The price of our loose assumptions is that our problem becomes a non-convex problem. Therefore, we develop zeroth-order methods that converge to stationary points, rather than optimal solutions.

3.3 Gaussian Smoothed Function

To propose our method, let us define the Gaussian smoothed function for F .

Definition 1. We call the following function the Gaussian smoothed function of F .

$$F_\mu(\mathbf{x}) := \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)}[F(\mathbf{x} + \mu\mathbf{u})].$$

Function F_μ is an approximation of F . Here, the smoothing parameter μ controls the level of approximation: when μ is small, F_μ is close to F .

In this paper, we propose two unbiased gradient estimators for F_μ . Our methods update \mathbf{x}_k using the proposed gradient estimators while reducing the smoothing parameter μ .

4. Proposed Method with One-point Gradient Estimator

4.1 One-point Gradient Estimator

We propose the following one-point gradient estimator:³

$$\mathbf{g}_1(\mathbf{x}, \mu, c, \mathbf{u}, \boldsymbol{\xi}) := \frac{f(\mathbf{x} + \mu\mathbf{u}, \boldsymbol{\xi}) - c}{\mu} \mathbf{u}, \quad (4)$$

where $\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)$ and $\boldsymbol{\xi} \sim D(\mathbf{x} + \mu\mathbf{u})$. Here, $\mathbf{g}_1(\mathbf{x}, \mu, c, \mathbf{u}, \boldsymbol{\xi})$ is an unbiased gradient estimator of the function F_μ as stated in the following lemma.

Lemma 2. For any $\mathbf{x} \in \mathbb{R}^d$, $\mu \in \mathbb{R}_{>0}$, and $c \in \mathbb{R}$, the following holds.

$$\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x} + \mu\mathbf{u})} [\mathbf{g}_1(\mathbf{x}, \mu, c, \mathbf{u}, \boldsymbol{\xi})]] = \nabla F_\mu(\mathbf{x}).$$

The above gradient estimator is a generalization of the one-point gradient estimator in [Liu et al., 2024]: when $c = 0$, our estimator is the same as that of [Liu et al., 2024, Eq. (4)]. While c does not affect the unbiasedness of the gradient estimator, it affects its variance. This fact is shown in the following lemma.

Lemma 3. Suppose that Assumptions 1 and 4 hold. Then, the following holds for any $\mathbf{x} \in \mathbb{R}^d$, $c \in \mathbb{R}$, and $m \in \mathbb{N}$.

$$\begin{aligned} & \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^j\}_{j=1}^m \sim D(\mathbf{x} + \mu\mathbf{u})} \left[\left\| \frac{1}{m} \sum_{j=1}^m \mathbf{g}_1(\mathbf{x}, \mu, c, \mathbf{u}, \boldsymbol{\xi}^j) \right\|^2 \right] \right] \\ & \leq \frac{3}{2} \mu^2 H_F^2 (d+6)^3 + 6(d+4) \|\nabla F(\mathbf{x})\|^2 + \frac{3\sigma^2 d}{\mu^2 m} + \frac{3d(F(\mathbf{x}) - c)^2}{\mu^2}. \end{aligned}$$

From Lemma 3, when c is close to $F(\mathbf{x})$, the variance of our gradient estimator is reduced. It is widely known in the field of stochastic optimization that reducing the variance of the gradient estimator has a positive impact on the convergence of the method [Johnson and Zhang, 2013]. Therefore, in the next section, we give a method for setting c_k at each iteration k to reduce the variance of the proposed gradient estimator.

Discussion. In the existing studies [Ray et al., 2022, Liu et al., 2024], the variance of their gradient estimators ([Ray et al., 2022, (3)] and [Liu et al., 2024, (4)]) are bounded by using $G := \max_{\mathbf{x}, \boldsymbol{\xi}} f(\mathbf{x}, \boldsymbol{\xi})$. However, since G is generally large, their upper bounds are also generally large. In contrast to existing studies, Lemma 3 provides an upper bound independent of G .

4.2 Setting of Variance Reduction Parameter c

To reduce the variance in Lemma 3, the parameter c should be close to $F(\mathbf{x})$. Therefore, during the iterations of the algorithm, we update c_k to bring it closer to the target value $F(\mathbf{x}_k)$. A naive method is to obtain samples $\{\boldsymbol{\xi}_k^j\}_{j=1}^{j_{\max}} \sim D(\mathbf{x}_k)$ for some $j_{\max} \in \mathbb{N}$, and set c_k as follows:

$$c_k := \frac{1}{j_{\max}} \sum_{j=1}^{j_{\max}} f(\mathbf{x}_k, \boldsymbol{\xi}_k^j).$$

³The term ‘‘one-point gradient estimator’’ in this paper refers to the gradient estimator obtained by queries of the objective function at a single point $(\mathbf{x} + \mu\mathbf{u})$. Note that ‘‘one-point gradient estimator’’ is also used to refer to a gradient estimator obtained by one query of the objective function, but in our case, Algorithm 1 proposed later requires additional calculations of $f(\mathbf{x}, \boldsymbol{\xi})$ to set parameter c . Moreover, if we use mini-batch gradients, Algorithm 1 requires multiple computations of $f(\mathbf{x}, \boldsymbol{\xi})$ for different $\boldsymbol{\xi}$.

However, this method requires new samples to estimate c_k , increasing the sample complexity. We propose an approach, inspired by [Jagadeesan et al., 2022], to approximate $F(\mathbf{x}_k)$ from samples obtained in past iterations. Specifically, using the sample set $\{\xi_i^j\}_{j=1}^{m_i} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)$ at each past $i \in \{k-s, \dots, k-1\}$ iteration, which is obtained to compute the gradient estimator, we compute c_k as follows.

$$c_k := \sum_{i=k-s}^{k-1} \frac{a_i}{m_i} \sum_{j=1}^{m_i} f(\mathbf{x}_k, \xi_i^j), \quad (5)$$

where $a_i \in [0, 1]$ and $\sum_{i=k-s}^{k-1} a_i = 1$. Here, a_i represents the importance of the samples of the i -th iteration. For example, if \mathbf{x}_k and $\mathbf{x}_i + \mu_i \mathbf{u}_i$ are close, then $D(\mathbf{x}_k)$ and $D(\mathbf{x}_i + \mu_i \mathbf{u}_i)$ have similar distributions from Assumption 2. Then, the samples of the i -th iteration are important, and it is better to make a_i larger.

To set a_i for each $i \in \{k-s, \dots, k-1\}$ appropriately, we first show the following lemma.

Lemma 4. Suppose that Assumptions 1–3 hold. Let c_k be defined by (5). Then, the following holds:

$$(F(\mathbf{x}_k) - c_k)^2 \leq 2 \sum_{i=k-s}^{k-1} a_i^2 \left(L_\xi^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i} \right). \quad (6)$$

From Lemma 4, a_i that minimizes the right-hand side of (6) is desirable to bring c_k close to $F(\mathbf{x}_k)$. Such a_i can be obtained in closed form from the following lemma.

Lemma 5. Consider the following optimization problem:

$$\begin{aligned} \min_{\mathbf{a}} \quad & \sum_{i=k-s}^{k-1} a_i^2 \left(L_\xi^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i} \right) \\ \text{s.t.} \quad & a_i \geq 0, \forall i \in \{k-s, \dots, k-1\}, \\ & \sum_{i=k-s}^{k-1} a_i = 1. \end{aligned}$$

Then, the optimal solution is \mathbf{a}^* such that $a_i^* := \frac{1}{b_i \sum_{j=k-s}^{k-1} \frac{1}{b_j}}$, where $b_i := L_\xi^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i}$.

Using the results of Lemmas 4 and 5, we can show the theoretical convergence of our method proposed in the next section.

4.3 Proposed Method and Theoretical Results

We propose Algorithm 1. Lines 2–4 of Algorithm 1 update the iterate based on the gradient estimator (4). Line 5 adjusts the smoothing parameter μ_k : Algorithm 1 starts with a sufficiently large μ_0 and gradually reduces μ_k so that the smoothed function F_{μ_k} approaches the original objective function F . Lines 6–8 calculate the variance reduction parameter c_{k+1} from the past samples. We define s_{\max} as the maximum window size, which indicates how far back in the past the samples are considered for calculating c_{k+1} . Note that lines 6–8, although requiring oracle computation of the function f , do not increase the sample complexity since those do not need new samples.

Remark 1. The proposed method incorporates the Gaussian Homotopy technique [Iwakiri et al., 2022] to obtain better stationary points. Specifically, the method starts from solving an almost convex smoothed function $F_{\mu_0}(\mathbf{x})$ with sufficiently large $\mu_0 \geq 0$ and gradually changes the optimization problem $F_{\mu_k}(\mathbf{x})$ to the original one $F(\mathbf{x})$. It is known that a better local solution can be potentially obtained by the technique [Mobahi and Fisher, 2015, Mobahi and Fisher III, 2015, Hazan et al., 2016, Iwakiri et al., 2022].

To demonstrate the convergence of Algorithm 1, we first present the following lemma.

Algorithm 1 Zeroth-order method with the improved one-point gradient estimator

Input: $\mathbf{x}_0, \mu_0, \mu_{\min}, c_0, s_{\max}, \beta, \gamma, \{m_k\}_{k=0}^T, M, T$.

- 1: **for** $k = 0, 1, \dots, T$:
 - 2: Sample \mathbf{u}_k from $\mathcal{N}(0, \mathbf{I}_d)$ and $\{\xi_k^j\}_{j=1}^{m_k}$ from $D(\mathbf{x}_k + \mu_k \mathbf{u}_k)$
 - 3: $\mathbf{g}_k \leftarrow \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{g}_1(\mathbf{x}_k, \mu_k, c_k, \mathbf{u}_k, \xi_k^j)$
 - 4: $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \beta \mathbf{g}_k$
 - 5: $\mu_{k+1} \leftarrow \max(\gamma \mu_k, \mu_{\min})$
 - 6: $s \leftarrow \min(s_{\max}, k + 1)$
 - 7: $a_i \leftarrow \frac{1}{b_i \sum_{j=k-s+1}^k \frac{1}{b_j}}$ for all $i \in [k - s + 1, k]$, where $b_i := M \|\mathbf{x}_{k+1} - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{1}{m_i}$
 - 8: $c_{k+1} \leftarrow \sum_{i=k-s+1}^k \frac{a_i}{m_i} \sum_{j=1}^{m_i} f(\mathbf{x}_{k+1}, \xi_i^j)$
-

Lemma 6. Suppose that Assumptions 1–4 hold. Let $\{\mathbf{x}_k\}$ and $\{c_k\}$ be the sequence generated by Algorithm 1 with $m_k \geq m_{\min}$ for $k \in \{0, 1, \dots, T\}$, $\beta \leq \frac{\mu_{\min}}{2L\xi\alpha\sqrt{6d}}$, and $M := \frac{L\xi^2\alpha^2}{\sigma^2}$. Then, for any setting parameter $\mathbf{x}_0 \in \mathbb{R}^d$, $\mu_0 \in \mathbb{R}_{>0}$, $\mu_{\min} \leq \mu_0$, $c_0 \in \mathbb{R}$, $s_{\max} \in \mathbb{N}$, and $\gamma \in (0, 1)$, the following holds:

$$\mathbb{E}_{\zeta_{[0,k-1]}}[\delta_k^2] \leq \left(\frac{1}{2}\right)^k \delta_0^2 + \frac{\mu_{\min}^2 \mu_0^2 H_F^2 (d+6)^3}{2d} + 2\mu_{\min}^2 \frac{d+4}{d} L_F^2 + 8L\xi^2 \alpha^2 \mu_0^2 d + \frac{5}{m_{\min}} \sigma^2,$$

where $\delta_k := F(\mathbf{x}_k) - c_k$, $k \in \{1, \dots, T\}$, and $\zeta_{[0,k-1]} := \{\mathbf{u}_0, \xi_0, \dots, \mathbf{u}_{k-1}, \xi_{k-1}\}$.

The above lemma shows that the estimation error of c_k (i.e., δ_k) is bounded by a constant at each iteration k . Intuitively, by keeping the step size below a certain threshold ($\mu_{\min}/2L\xi\alpha\sqrt{6d}$), we ensure that the past iterates are not too far from the current iterate. This reduces the difference between the distribution at past iterates and that at the current iterate, which in turn reduces the estimation error of c_k .

From Lemma 6, we show the convergence of Algorithm 1.

Theorem 1. Suppose that Assumptions 1–4 hold. Let $\{\mathbf{x}_k\}$ be the sequence generated by Algorithm 1 with $m_k = \Theta(\varepsilon^{-2}d^2)$ for all $k \in \{0, \dots, T\}$, $M := \frac{L\xi^2\alpha^2}{\sigma^2}$, $\mu_{\min} = \Theta(\varepsilon d^{-\frac{3}{2}})$, $\mu_0 = \Theta(\varepsilon d^{-\frac{3}{2}})$ such that $\mu_0 \geq \mu_{\min}$, and $\beta := \min\left(\frac{1}{12(d+4)H_F}, \frac{1}{(T+1)^{\frac{1}{2}}d^{\frac{3}{4}}}, \frac{\mu_{\min}}{2L\xi\alpha\sqrt{6d}}\right)$. Let $\hat{\mathbf{x}} := \mathbf{x}_{k'}$, where k' is chosen from a uniform distribution over $\{0, \dots, T\}$. Then, for any $\mathbf{x}_0 \in \mathbb{R}^d$, $c_0 \in \mathbb{R}$, $s_{\max} \in \{1, \dots, T\}$, and $\gamma \in (0, 1)$, the iteration complexity required to obtain $\mathbb{E}[\|\nabla F(\hat{\mathbf{x}})\|^2] \leq \varepsilon^2$ is $O(d^{\frac{5}{2}}\varepsilon^{-4})$. Moreover, the sample complexity is $O(d^{\frac{9}{2}}\varepsilon^{-6})$.

Remark 2. The iteration complexity of our method is $O(d^{\frac{1}{2}})$ larger than that of the Gaussian homotopy method [Iwakiri et al., 2022, Theorem C.2], which considers the case where the oracle of the objective function contains noise, but the random variables are independent of the decision variables. This increase in iteration complexity is reasonable, given that we are dealing with a complex situation where the random variables depend on the decision variables. Moreover, since the sample complexity of the existing zeroth-order method [Liu et al., 2024] for non-convex problem (P) is $O(G^6 d^2 \varepsilon^{-6})$,⁴ where $G := \sup_{\mathbf{x}, \xi} |f(\mathbf{x}, \xi)|$, the proposed method has an advantage when G is large or unbounded.

⁴[Liu et al., 2024] claim that the sample complexity of their method is $O(\varepsilon^{-3})$, but in our definition of the stationary point ($\mathbb{E}[\|\nabla F(\hat{\mathbf{x}})\|^2] \leq \varepsilon^2$), it becomes $O(\varepsilon^{-6})$.

Remark 3. In order to set the parameters of Theorem 1, it is necessary to know in advance σ of Assumption 1 and α of Assumption 2. While the existing study [Ray et al., 2022] also needs such information (e.g., γ in their paper), it may not be known practically. In such cases, it is possible to start with a sufficiently large value and estimate it from the information obtained during the iterations.

5. Proposed Method with Two-point Gradient Estimator

5.1 Two-point Gradient Estimator

We consider the following gradient estimator:

$$g_2(x, \mu, u, \xi^1, \xi^2) := \frac{f(x + \mu u, \xi^1) - f(x - \mu u, \xi^2)}{2\mu} u, \quad (7)$$

where $u \sim \mathcal{N}(0, I_d)$, $\xi^1 \sim D(x + \mu u)$, $\xi^2 \sim D(x - \mu u)$, and $\mu \in \mathbb{R}_{>0}$.⁵

Then, the following lemmas hold for the two-point gradient estimator.

Lemma 7. For any $x \in \mathbb{R}^d$ and $\mu \in \mathbb{R}_{>0}$, the following holds.

$$\mathbb{E}_{u \sim \mathcal{N}(0, I_d)} [\mathbb{E}_{\xi^1 \sim D(x + \mu u), \xi^2 \sim D(x - \mu u)} [g_2(x, \mu, u, \xi^1, \xi^2)]] = \nabla F_\mu(x).$$

Lemma 8. Suppose that Assumptions 1 and 4 hold. For any $x \in \mathbb{R}^d$, the following holds.

$$\begin{aligned} & \mathbb{E}_{u \sim \mathcal{N}(0, I_d)} \left[\mathbb{E}_{\{\xi^{1,j}\}_{j=1}^m, \{\xi^{2,j}\}_{j=1}^m} \left[\left\| \frac{1}{m} \sum_{j=1}^m g_2(x, \mu, u, \xi^{1,j}, \xi^{2,j}) \right\|^2 \right] \right] \\ & \leq \frac{3}{2} \mu^2 H_F^2 (d+6)^3 + 6(d+4) \|\nabla F(x)\|^2 + \frac{3\sigma^2 d}{2\mu^2 m}, \end{aligned}$$

where $\{\xi^{1,j}\}_{j=1}^m \sim D(x + \mu u)$ and $\{\xi^{2,j}\}_{j=1}^m \sim D(x - \mu u)$.

5.2 Proposed Method and Theoretical Results

We propose Algorithm 2. It changes the gradient estimator of Algorithm 1 to the two-point gradient estimator (7). From Lemmas 7 and 8, we show the convergence of Algorithm 2.

Theorem 2. Suppose that Assumptions 1–4 hold. Let $\{x_k\}$ be the sequence generated by Algorithm 2 with $m_k = \Theta(\varepsilon^{-2} d^2)$ for all $k \in \{0, \dots, T\}$, $\mu_{\min} = \Theta(\varepsilon d^{-\frac{3}{2}})$, $\mu_0 = \Theta(\varepsilon d^{-\frac{3}{2}})$ such that $\mu_0 \geq \mu_{\min}$, and $\beta := \min \left(\frac{1}{12(d+4)H_F}, \frac{1}{(T+1)^{\frac{1}{2}} d^{\frac{3}{4}}} \right)$. Let $\hat{x} := x_{k'}$, where k' is chosen from a uniform distribution over $\{0, \dots, T\}$. Then, for any $x_0 \in \mathbb{R}^d$ and $\gamma \in (0, 1)$, the iteration complexity required to obtain $\mathbb{E}[\|\nabla F(\hat{x})\|^2] \leq \varepsilon^2$ is $O(d^{\frac{5}{2}} \varepsilon^{-4})$. Moreover, the sample complexity is $O(d^{\frac{9}{2}} \varepsilon^{-6})$.

Remark 4. Algorithm 2 has the same sample complexity as Algorithm 1. However, in practice, if we can use both the one-point and two-point gradient estimators, the two-point gradient estimator often achieves better performance (See the experimental results in Section 6.3). On the other hand, there are problems where only a one-point gradient estimator can be used. For example, in the case where the distribution changes with time, sampling in the same environment is not possible [Ray et al., 2022, Liu et al., 2024].

⁵Although the two-point estimator is also discussed in [Liu et al., 2024] as $g_{2\text{pt-II}}$, they do not prove the convergence of zeroth-order methods using the estimator without $G = \sup_{x, \xi} |f(x, \xi)|$.

Algorithm 2 Zeroth-order method with the two-point gradient estimator

Input: $\mathbf{x}_0, \mu_0, \mu_{\min}, \beta, \gamma, \{m_k\}_{k=0}^T, T$.

- 1: **for** $k = 0, 1, \dots, T$:
 - 2: Sample \mathbf{u}_k from $\mathcal{N}(0, \mathbf{I}_d)$, $\{\boldsymbol{\xi}_k^{1,j}\}_{j=1}^{m_k}$ from $D(\mathbf{x}_k + \mu_k \mathbf{u}_k)$, and $\{\boldsymbol{\xi}_k^{2,j}\}_{j=1}^{m_k}$ from $D(\mathbf{x}_k - \mu_k \mathbf{u}_k)$.
 - 3: $\mathbf{g}_k \leftarrow \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{g}_2(\mathbf{x}_k, \mu_k, \mathbf{u}_k, \boldsymbol{\xi}_k^{1,j}, \boldsymbol{\xi}_k^{2,j})$
 - 4: $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \beta \mathbf{g}_k$
 - 5: $\mu_{k+1} \leftarrow \max(\gamma \mu_k, \mu_{\min})$
-

6. Experiments

We conducted experiments on an application of *multiproduct pricing* to show that Algorithms 1 and 2 output solutions with lower objective values compared with the existing methods. We performed simulation experiments with real retail data from a supermarket service provider in Japan.⁶ The details of our experiments are in the supplementary material.

6.1 Problem Setup

We consider a variant of [Gallego and Wang, 2014] where a seller determines the prices of multiple ($n = 10$) products for multiple ($m = 40$) buyers. Each buyer can purchase at most one copy of any product. Let $\mathbf{x} := (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ be the price vector for the products. Let $\boldsymbol{\xi} \in \{0, 1, \dots, m\}^{n+1}$ denote a random vector, where ξ_0 represents the number of buyers not purchasing any product and ξ_i for $i = 1, \dots, n$ represents the number of sales of each product. Let $s(\mathbf{x}, \boldsymbol{\xi})$ and $c(\boldsymbol{\xi})$ be real-valued functions representing the sales and costs of products, respectively, and defined as:

$$s(\mathbf{x}, \boldsymbol{\xi}) := \sum_{i=1}^n x_i \xi_i, \quad c(\boldsymbol{\xi}) := \sum_{i=1}^n c_i(\xi_i),$$

where

$$c_i(\xi_i) := \begin{cases} 2w_i \xi_i, & \xi_i \leq l_i, \\ w_i(\xi_i - l_i) + 2w_i l_i, & l_i < \xi_i \leq u_i, \\ 3w_i(\xi_i - u_i) + w_i(u_i - l_i) + 2w_i l_i, & \xi_i > u_i. \end{cases}$$

Here, $l_i := \frac{0.5m}{n}$, $u_i := \frac{1.5m}{n}$, and $w_i := \rho_i \theta_i$, where ρ_i is the random variable generated from a uniform distribution of $[0.25, 0.5]$ and θ_i is the normalized recorded average selling price for each product i . The function c_i represents the case where the cost rate varies with the number of sold products.⁷ Then, the revenue-maximizing problem is as follows:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x})} [-s(\mathbf{x}, \boldsymbol{\xi}) + c(\boldsymbol{\xi})],$$

where $D(\mathbf{x})$ is the probability distribution that $\boldsymbol{\xi}$ follows.

⁶We used publicly available data, “New Product Sales Ranking”, provided by KSP-SP Co., Ltd, <http://www.ksp-sp.com>. Accessed August 15, 2024.

⁷Such piecewise linear cost functions are considered in existing studies [Shaw and Wagelmans, 1998, Tunc et al., 2016, Ou, 2017].

Settings of unknown parameters for $D(\mathbf{x})$. We assume that buyers choose one product stochastically. Each buyer chooses product $i \in I := \{1, \dots, n\}$ with probability $p_i(\mathbf{x}) = \frac{e^{\gamma_i(\theta_i - x_i)}}{a_0 + \sum_{j=1}^n e^{\gamma_j(\theta_j - x_j)}}$ or does not choose any product with probability $p_0(\mathbf{x}) = \frac{a_0}{a_0 + \sum_{j=1}^n e^{\gamma_j(\theta_j - x_j)}}$. Here, we let $\gamma_i := \frac{2\pi}{\sqrt{6}\alpha_i}$ and let $a_0 := 0.1n$. Then, $\Pr(\xi_i | \mathbf{x}) = \binom{m}{\xi_i} p_i(\mathbf{x})^{\xi_i}$. Note that the information of $D(\mathbf{x})$ is not used by each method.

6.2 Compared Methods and Settings

We implemented the following methods.

Proposed-1 (mini-batch). We implemented Algorithm 1 with $\mu_0 := 0.19$, $\mu_{\min} := 0.0001$, $c_0 := \sum_{j=1}^{20} f(\mathbf{x}_0, \xi^j(\mathbf{x}_0))$, $s_{\max} := 10$, $\beta := 0.001 \cdot 0.95^{k+1}$, $\gamma = 0.95$, $m_k = 30 + 2k$, and $M = 0.1$, where k is the current iteration number.

Proposed-1 (batch size 1). We implemented Algorithm 1 with $m_k = 1$ and other same parameters as Proposed-1 (mini-batch).

Proposed-2 (mini-batch). We implemented Algorithm 2 with $\mu_0 := 0.19$, $\mu_{\min} := 0.0001$, $\beta := 0.001 \cdot 0.95^{k+1}$, $m_k = 30 + 2k$, and $\gamma = 0.95$.

Proposed-2 (batch size 1). We implemented Algorithm 2 with $m_k = 1$ and other same parameters as Proposed-2 (mini-batch).

CZO-1 (mini-batch). This method is a Conventional Zeroth-Order (CZO) method with a one-point gradient estimator. It is consistent with Algorithm 1 where $c_k = 0$ for all $k \in \{0, \dots, T\}$, $\beta = 10^{-5}$, $\mu_k = 0.001$, and the other parameters are the same as Proposed-1 (mini-batch). This method is analogous to the zeroth-order method used in existing studies [Ray et al., 2022, Liu et al., 2024].

CZO-1 (batch size 1). This method is the same as the CZO-1 (mini-batch) with batch size $m_k = 1$ for all $k \in \{0, \dots, T\}$.

We performed our experiments under the following settings.

Initial points. For all methods, we set the initial points as $\mathbf{x}_0 := 0.5\mathbf{e}$, where $\mathbf{e} \in \mathbb{R}^n$ is a vector with all elements 1.

Metric. For the output $\hat{\mathbf{x}}$ of each method, we computed $obj := \frac{1}{10^3} \sum_{q=1}^{10^3} (-s(\hat{\mathbf{x}}, \xi^q(\hat{\mathbf{x}})) + c(\xi^q(\hat{\mathbf{x}})))$, where $\xi^q(\hat{\mathbf{x}}) \sim D(\hat{\mathbf{x}})$.

Termination criteria. We terminated each method when it had taken 5000 samples from $D(\mathbf{x})$ for some \mathbf{x} .

6.3 Experimental Results

Table 1 shows the results of the experiments using real data from different weeks. All proposed methods were superior to the baselines (CZO-1) for all weeks of data.

Figure 1 shows the change in the objective value (obj) for each method with respect to the number of samples. The figure implies that the proposed methods reduce the objective value more stably than CZO-1. In particular, the Proposed-2 method shows high performance. While the Proposed-2 method requires twice as many samples as the Proposed-1 method at each iteration, it still reduces the objective value with fewer samples. Although the Proposed-1 method is slightly inferior to the Proposed-2 method, it outperforms the CZO-1 in many cases. This fact shows the advantage of introducing the variance reduction parameter c .

7. Conclusion

We proposed two zeroth-order methods for non-convex problems with decision-dependent uncertainty: a method using a new one-point gradient estimator including a variance reduction parameter and the one using a two-point gradient estimator. As theoretical results, we showed their convergence to stationary points and

Table 1: Results of simulation experiments with real data for 20 randomly generated problem instances. The *obj* (*sd*) column represents the average (standard deviation) of the *obj*. The best value of the average *obj* for each experiment is in bold. In all experiments, the differences between the proposed method with the best value of the average *obj* and CZO-1 (both types) are significant (two-sided t-test: $p < 0.05$).

date	Proposed-1 (mini-batch)		Proposed-1 (batch size 1)		Proposed-2 (mini-batch)		Proposed-2 (batch size 1)		CZO-1 (mini-batch)		CZO-1 (batch size 1)	
	obj	sd	obj	sd	obj	sd	obj	sd	obj	sd	obj	sd
02/21–02/27	-7.39	1.92	-2.62	4.54	-7.26	1.68	-6.62	2.01	1.83	3.34	2.83	5.10
03/21–03/27	-7.61	1.80	-2.73	4.45	-7.78	1.48	-6.70	1.98	2.41	5.55	4.08	11.55
05/23–05/29	-5.28	1.74	-1.89	3.79	-6.38	1.54	-5.35	1.77	0.21	2.88	0.65	2.94
06/20–06/26	-5.58	1.69	1.81	4.62	-5.54	1.35	-5.61	1.87	7.04	9.73	9.83	11.51
07/18–07/24	-3.52	2.15	3.99	3.53	-4.10	2.02	-4.17	2.08	10.39	11.85	10.29	12.57
08/08–08/14	-6.40	1.68	-2.97	4.79	-7.01	1.29	-6.17	1.54	-1.05	1.35	-0.47	2.05
09/19–09/25	-3.21	2.26	3.32	4.94	-3.88	2.23	-3.83	2.44	10.38	14.92	12.15	12.09
12/05–12/11	-4.61	1.78	2.80	3.89	-4.88	2.10	-4.27	2.84	7.66	7.58	5.57	5.99

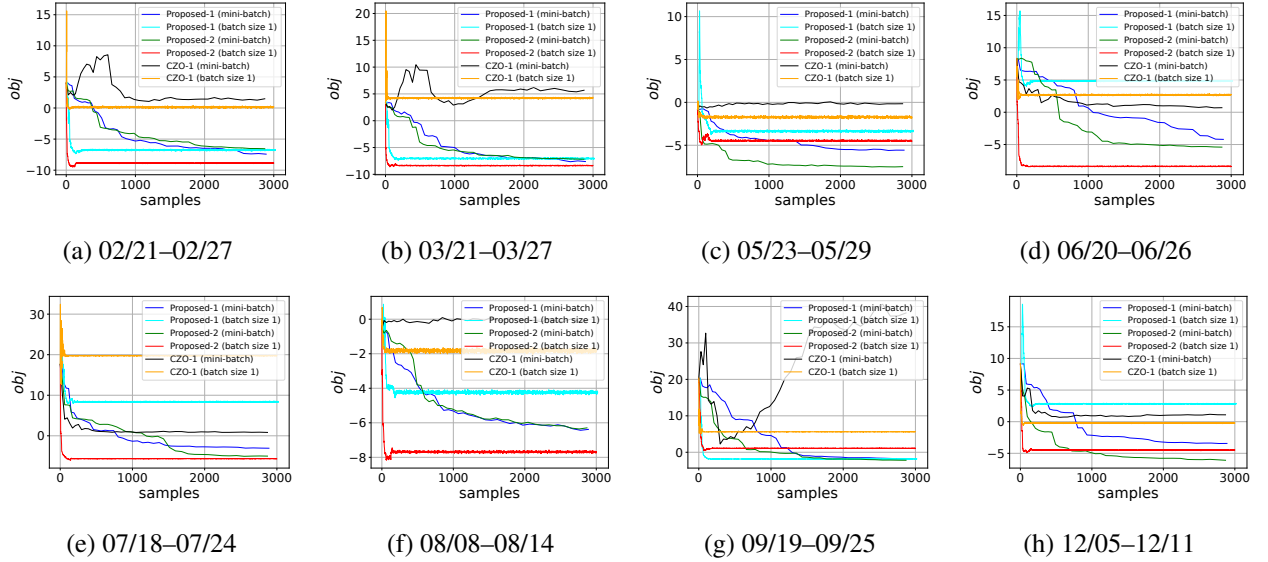


Figure 1: Change in *obj* in the first 3000 samples in the simulation experiment with real data. Each graph shows the result in one problem instance for each week. The horizontal axis indicates the number of samples, and the vertical axis indicates *obj*.

provided the worst-case iteration and sample complexity analysis. Our simulation experiments with real data on a retail service application showed good performance of our methods compared to the conventional zeroth-order methods.

In future work, we would like to conduct theoretical analyses of the proposed method in different situations. For example, one might apply our method to problem (P) where the distribution changes dynamically with time [Ray et al., 2022, Liu et al., 2024]. For such problems, since it is not possible to sample random variables more than once in the same environment, only a one-point gradient estimator can be used. Therefore, our improved one-point gradient estimator may be effective. Another future direction is to integrate existing techniques [Wang et al., 2018, Cai et al., 2022, Liu et al., 2018] for zeroth-order methods into our methods. For example, incorporating the variable selection technique [Wang et al., 2018] may enhance the scalability of our method. Additionally, applying the variance reduction technique [Liu et al., 2018] for

gradient estimation has the potential to speed up our methods.

References

- B. Basciftci, S. Ahmed, and S. Shen. Distributionally robust facility location problem under decision-dependent stochastic demand. *European Journal of Operational Research*, 292(2):548–561, 2021.
- E. Brochu, V. M. Cora, and N. De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.
- H. Cai, D. McKenzie, W. Yin, and Z. Zhang. Zeroth-order regularized optimization (zoro): Approximately sparse gradients and adaptive sampling. *SIAM Journal on Optimization*, 32(2):687–714, 2022.
- Y. Chen, W. Tang, C.-J. Ho, and Y. Liu. Performative prediction with bandit feedback: Learning through reparameterization. *arXiv preprint arXiv:2305.01094*, 2023.
- P. I. Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.
- J. E. Freund and R. E. Walpole. *Mathematical statistics*. Prentice-Hall, Inc., 1986.
- G. Gallego and R. Wang. Multiproduct price optimization and competition under the nested logit model with product-differentiated price sensitivities. *Operations Research*, 62(2):450–461, 2014.
- S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.
- E. Hazan, K. Y. Levy, and S. Shalev-Shwartz. On graduated optimization for stochastic non-convex problems. In *International Conference on Machine Learning*, pages 1833–1841, 2016.
- Y. Hikima and A. Takeda. Stochastic approach for price optimization problems with decision-dependent uncertainty. *arXiv preprint arXiv:2307.00492 (to appear in European Journal of Operational Research)*, 2023.
- S. Ito and R. Fujimaki. Large-scale price optimization via network flow. *Advances in Neural Information Processing Systems*, 29:3862–3870, 2016.
- H. Iwakiri, Y. Wang, S. Ito, and A. Takeda. Single loop gaussian homotopy method for non-convex optimization. *Advances in Neural Information Processing Systems*, 35:7065–7076, 2022.
- M. Jagadeesan, T. Zrnic, and C. Mendler-Dünner. Regret minimization with performative feedback. In *International Conference on Machine Learning*, pages 9760–9785, 2022.
- K. Ji, Z. Wang, Y. Zhou, and Y. Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, pages 3100–3109, 2019.
- R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26:315–323, 2013.
- H. Liu, Q. Li, and H. T. Wai. Two-timescale derivative free optimization for performative prediction with Markovian data. In *International Conference on Machine Learning*, pages 31425–31450, 2024.

- S. Liu, B. Kailkhura, P.-Y. Chen, P. Ting, S. Chang, and L. Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31:3731–3741, 2018.
- F. Luo and S. Mehrotra. Distributionally robust optimization with decision dependent ambiguity sets. *Optimization Letters*, 14:2565–2594, 2020.
- C. Mendler-Dünner, J. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.
- J. P. Miller, J. C. Perdomo, and T. Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720, 2021.
- H. Mobahi and J. W. Fisher. On the link between gaussian homotopy continuation and convex envelopes. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 43–56, 2015.
- H. Mobahi and J. Fisher III. A theoretical analysis of optimization by gaussian continuation. In *AAAI Conference on Artificial Intelligence*, pages 1205–1211, 2015.
- Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.
- J. Ou. Improved exact algorithms to economic lot-sizing with piecewise linear production costs. *European Journal of Operational Research*, 256(3):777–784, 2017.
- J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609, 2020.
- M. Ray, L. J. Ratliff, D. Drusvyatskiy, and M. Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *AAAI Conference on Artificial Intelligence*, pages 8081–8088, 2022.
- D. X. Shaw and A. P. Wagelmans. An algorithm for single-item capacitated economic lot sizing with piecewise linear production costs and general holding costs. *Management Science*, 44(6):831–838, 1998.
- R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- H. Tunc, O. A. Kilic, S. A. Tarim, and B. Eksioglu. The stochastic lot sizing problem with piecewise linear concave ordering costs. *Computers & Operations Research*, 65:104–110, 2016.
- Y. Wang, S. Du, S. Balakrishnan, and A. Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 1356–1365, 2018.

A. Details of Our Experiments

A.1 Common Settings

All experiments were conducted on a computer with an AMD EPYC 7413 24-Core Processor, 503.6 GiB of RAM, and Ubuntu 20.04.6 LTS. The program code was implemented in Python 3.8.3. Our used libraries are the following: numpy, scipy, math, GPyOpt, time, sys, pickle, matplotlib.

A.2 Details of Used Data

We used publicly available data, “New Product Sales Ranking”, provided by KSP-SP Co., Ltd, <http://www.ksp-sp.com>. The data are actual retail data from a middle-size supermarket located in Tokyo. The parameters of our simulation experiments (θ_i in our paper) were determined using the sales prices of confectionery products recorded in the data. We chose to use these data for the following two reasons. First, since the data are open, we can make the experimental code publicly available. Second, since the data have been used in an existing price optimization study [Ito and Fujimaki, 2016], we believe it is appropriate for our experiments on price optimization.

A.3 Hyper Parameter Settings

To determine the parameters of the implemented methods, we performed some preliminary experiments under the same setting as the experiments in our paper. As the parameters of the Proposed-1 (minibatch) method, we adopted the ones that resulted in low objective values. We tested the following parameters: $\mu_0 \in \{0.019, 0.19, 1.9\}$, $\mu_{\min} \in \{10^{-3}, 10^{-4}, 10^{-5}\}$, $c_0 \in \{\sum_{j=1}^{20} f(\mathbf{x}_0, \boldsymbol{\xi}^j(\mathbf{x}_0)), \sum_{j=1}^{100} f(\mathbf{x}_0, \boldsymbol{\xi}^j(\mathbf{x}_0))\}$, $s_{\max} \in \{1, 10, 100\}$, $\beta_k \in \{10^{-3} \times 0.95^{k+1}, 10^{-4} \times 0.95^{k+1}, 10^{-5} \times 0.95^{k+1}\}$, $\gamma \in \{0.9, 0.95, 0.99\}$, $m_k \in \{30, 30 + 2k, 30 + 5k\}$, and $M \in \{0.01, 0.1, 1\}$. Then, the parameters of the other methods were set based on the parameters of the Proposed-1 method: for the Proposed-2 (mini-batch) method, all parameters (i.e., μ_0 , μ_{\min} , β , γ , and m_k) were set the same as those of the Proposed-1 method; for the CZO (mini-batch) method, the parameters γ and m_k were set the same as those of the Proposed-1 method, and $\beta := 10^{-5}$ and $\mu_k := 0.001$ for all iteration k . Here, parameters β and μ_k of the CZO method were selected from $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ and $\{10^{-3}, 10^{-4}, 10^{-5}\}$, respectively, with low objective values through preliminary experiments as in the Proposed-1 method. The Proposed-1 (batch size 1), the Proposed-2 (batch size 1), and the CZO (batch size 1) are identical to the Proposed-1 (mini-batch), the Proposed-2 (mini-batch), and the CZO (mini-batch), respectively, except that $m_k = 1$ for all iteration k .

A.4 Random Seed Settings

We set the random seed in our experiments by “numpy.random.seed(2024)” in our code. Note that “2024” is the used seed.

B. Proofs

B.1 Technical lemmas

We provide some technical lemmas, which are used to prove the lemmas and theorems in our paper.

Lemma B.1. [Jagadeesan et al., 2022, Lemma 2.1] Suppose that Assumptions 2 and 3 hold. Then, $F(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x})}[f(\mathbf{x}, \boldsymbol{\xi})]$ is L_F -Lipschitz continuous with respect to \mathbf{x} , where $L_F := L_{\boldsymbol{\xi}} + \alpha L_{\mathbf{x}}$.

Lemma B.2. [Jagadeesan et al., 2022, Section 3.1] Suppose that Assumptions 2 and 3 hold. Then, for any $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{x}' \in \mathbb{R}^d$, the following holds:

$$\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}')} [f(\mathbf{x}, \boldsymbol{\xi})] - L_{\boldsymbol{\xi}} \alpha \|\mathbf{x} - \mathbf{x}'\| \leq F(\mathbf{x}) \leq \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x})} [f(\mathbf{x}, \boldsymbol{\xi})] + L_{\boldsymbol{\xi}} \alpha \|\mathbf{x} - \mathbf{x}'\|.$$

Lemma B.3. [Nesterov and Spokoiny, 2017, Eq. (14)]

$$\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\|\mathbf{u}\|^2] = d.$$

Lemma B.4. [Iwakiri et al., 2022, Lemma 3.2] When F is H_F -smooth, F_μ is H_F -smooth for any $\mu \in \mathbb{R}_{>0}$.

Lemma B.5. [Iwakiri et al., 2022, Lemma 3.3] When F is L_F -Lipschitz, F_μ is $L_F\sqrt{d}$ -Lipschitz with respect to μ , that is, for any $\mathbf{x} \in \mathbb{R}^d$, $\mu_1 \in \mathbb{R}_{>0}$, and $\mu_2 \in \mathbb{R}_{>0}$,

$$|F_{\mu_1}(\mathbf{x}) - F_{\mu_2}(\mathbf{x})| \leq L_F\sqrt{d}|\mu_1 - \mu_2|.$$

Lemma B.6. [Freund and Walpole, 1986, p183] For any $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbb{E}_{\{\xi^j\}_{j=1}^m \sim D(\mathbf{x})} \left[\left\| \frac{1}{m} \sum_{j=1}^m f(\mathbf{x}, \xi^j) - \mathbb{E}_{\xi \sim D(\mathbf{x})}[f(\mathbf{x}, \xi)] \right\|^2 \right] \leq \frac{1}{m} \mathbb{E}_{\xi' \sim D} \left[\|f(\mathbf{x}, \xi') - \mathbb{E}_{\xi \sim D}[f(\mathbf{x}, \xi)]\|^2 \right].$$

Lemma B.7. [Nesterov and Spokoiny, 2017, Lemma 4] Suppose that Assumption 4 holds. Then, for any $\mathbf{x} \in \mathbb{R}^d$ and $\mu \in \mathbb{R}_{>0}$,

$$\|\nabla F(\mathbf{x})\|^2 \leq 2\|\nabla F_\mu(\mathbf{x})\|^2 + \frac{\mu^2}{2} H_F^2(d+6)^3.$$

Lemma B.8. [Nesterov and Spokoiny, 2017, Theorem 4] Suppose that Assumption 4 holds. Then, for any $\mathbf{x} \in \mathbb{R}^d$ and $\mu \in \mathbb{R}_{>0}$, the following holds.

$$\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\left\| \frac{1}{\mu} (F(\mathbf{x} + \mu\mathbf{u}) - F(\mathbf{x}))\mathbf{u} \right\|^2 \right] \leq \frac{\mu^2}{2} H_F^2(d+6)^3 + 2(d+4)\|\nabla F(\mathbf{x})\|^2.$$

Remark. Lemma B.8 corresponds to the first inequality of (35) in [Nesterov and Spokoiny, 2017, Theorem 4]. Nesterov and Spokoiny [2017] show the inequality holds under the assumption that f (which corresponds to F in Lemma B.8) is convex. However, the inequality holds even if the function f is non-convex. This is because the convexity of f is not used in the proof. Therefore, Lemma B.8 does not require the convexity of F .

Lemma B.9. Suppose that Assumption 4 holds. Then, for any $\mathbf{x} \in \mathbb{R}^d$ and $\mu \in \mathbb{R}_{>0}$, the following holds.

$$\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\left\| \frac{1}{2\mu} (F(\mathbf{x} + \mu\mathbf{u}) - F(\mathbf{x} - \mu\mathbf{u}))\mathbf{u} \right\|^2 \right] \leq \frac{\mu^2}{2} H_F^2(d+6)^3 + 2(d+4)\|\nabla F(\mathbf{x})\|^2.$$

Proof. Since F is H_F -smooth,

$$\begin{aligned} F(\mathbf{x} + \mu\mathbf{u}) - F(\mathbf{x} - \mu\mathbf{u}) &= F(\mathbf{x} + \mu\mathbf{u}) - F(\mathbf{x}) + F(\mathbf{x}) - F(\mathbf{x} - \mu\mathbf{u}) \\ &\stackrel{(*)}{\leq} \mu\langle \nabla F(\mathbf{x}), \mathbf{u} \rangle + \frac{\mu^2}{2} H_F \|\mathbf{u}\|^2 + \mu\langle \nabla F(\mathbf{x}), \mathbf{u} \rangle + \frac{\mu^2}{2} H_F \|\mathbf{u}\|^2 \\ &= 2\mu\langle \nabla F(\mathbf{x}), \mathbf{u} \rangle + \mu^2 H_F \|\mathbf{u}\|^2, \end{aligned}$$

where $(*)$ comes from the fact that

$$\begin{aligned} F(\mathbf{x} + \mu\mathbf{u}) &\leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mu\mathbf{u} \rangle + \frac{1}{2} H_F \|\mu\mathbf{u}\|^2, \\ F(\mathbf{x} - \mu\mathbf{u}) &\geq F(\mathbf{x}) - \langle \nabla F(\mathbf{x}), \mu\mathbf{u} \rangle - \frac{1}{2} H_F \|\mu\mathbf{u}\|^2. \end{aligned}$$

Similarly, we have $F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x} - \mu \mathbf{u}) \geq 2\mu \langle \nabla F(\mathbf{x}), \mathbf{u} \rangle - \mu^2 H_F \|\mathbf{u}\|^2$. Therefore,

$$(F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x} - \mu \mathbf{u}))^2 \leq 8\mu^2 (\nabla F(\mathbf{x})^\top \mathbf{u})^2 + 2\mu^4 H_F^2 \|\mathbf{u}\|^4.$$

Then,

$$\begin{aligned} E_{\mathbf{u}} \left[\left\| \frac{1}{2\mu} (F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x} - \mu \mathbf{u})) \mathbf{u} \right\|^2 \right] &= \frac{1}{4\mu^2} E_{\mathbf{u}} [(F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x} - \mu \mathbf{u}))^2 \|\mathbf{u}\|^2] \\ &\leq \frac{1}{2\mu^2} (E_{\mathbf{u}} [\mu^4 H_F^2 \|\mathbf{u}\|^6] + E_{\mathbf{u}} [4\mu^2 \langle \nabla F(\mathbf{x}), \mathbf{u} \rangle^2 \|\mathbf{u}\|^2]) \\ &\stackrel{(*)}{\leq} \frac{\mu^2}{2} H_F^2 (d+6)^3 + 2(d+4) \|\nabla F(\mathbf{x})\|^2, \end{aligned}$$

where (*) comes from [Nesterov and Spokoiny, 2017, (17) and (32)]. \square

B.2 Proof of Lemma 2

Lemma 2. For any $\mathbf{x} \in \mathbb{R}^d$, $\mu \in \mathbb{R}_{>0}$, and $c \in \mathbb{R}$, the following holds.

$$\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x} + \mu \mathbf{u})} [\mathbf{g}_1(\mathbf{x}, \mu, c, \mathbf{u}, \boldsymbol{\xi})]] = \nabla F_\mu(\mathbf{x}).$$

Proof.

$$\begin{aligned} &\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x} + \mu \mathbf{u})} [\mathbf{g}_1(\mathbf{x}, \mu, c, \mathbf{u}, \boldsymbol{\xi})]] \\ &= \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x} + \mu \mathbf{u})} \left[\frac{f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}) - c}{\mu} \mathbf{u} \right] \right] \\ &= \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\frac{1}{\mu} \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x} + \mu \mathbf{u})} [f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}) \mathbf{u}] \right] - \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\frac{c}{\mu} \mathbf{u} \right] \\ &\stackrel{(*)}{=} \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\frac{1}{\mu} F(\mathbf{x} + \mu \mathbf{u}) \mathbf{u} \right] \\ &\stackrel{(**)}{=} \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\frac{1}{\mu} F(\mathbf{x} + \mu \mathbf{u}) \mathbf{u} \right] - \frac{1}{\mu} F(\mathbf{x}) \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\mathbf{u}] \\ &= \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\frac{1}{\mu} (F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x})) \mathbf{u} \right] \\ &\stackrel{(***)}{=} \nabla F_\mu(\mathbf{x}), \end{aligned}$$

where (*) and (**) follow from the fact that $\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\mathbf{u}] = 0$, and (***) comes from [Nesterov and Spokoiny, 2017, Eq. (21)]. \square

B.3 Proof of Lemma 3

Lemma 3. Suppose that Assumptions 1 and 4 hold. Then, the following holds for any $\mathbf{x} \in \mathbb{R}^d$, $c \in \mathbb{R}$, and $m \in \mathbb{N}$.

$$\begin{aligned} &\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^j\}_{j=1}^m \sim D(\mathbf{x} + \mu \mathbf{u})} \left[\left\| \frac{1}{m} \sum_{j=1}^m \mathbf{g}_1(\mathbf{x}, \mu, c, \mathbf{u}, \boldsymbol{\xi}^j) \right\|^2 \right] \right] \\ &\leq \frac{3}{2} \mu^2 H_F^2 (d+6)^3 + 6(d+4) \|\nabla F(\mathbf{x})\|^2 + \frac{3\sigma^2 d}{\mu^2 m} + \frac{3d(F(\mathbf{x}) - c)^2}{\mu^2}. \end{aligned}$$

Proof. From the inequality of arithmetic and geometric means, the following holds for any $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^d$, and $\mathbf{z} \in \mathbb{R}^d$:

$$\|\mathbf{x}\|\|\mathbf{y}\| \leq \frac{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}{2}, \quad \|\mathbf{y}\|\|\mathbf{z}\| \leq \frac{\|\mathbf{y}\|^2 + \|\mathbf{z}\|^2}{2}, \quad \|\mathbf{z}\|\|\mathbf{x}\| \leq \frac{\|\mathbf{z}\|^2 + \|\mathbf{x}\|^2}{2}.$$

Then,

$$\begin{aligned} \|\mathbf{x} + \mathbf{y} + \mathbf{z}\|^2 &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{z}\|^2 + 2\mathbf{x}^\top \mathbf{y} + 2\mathbf{y}^\top \mathbf{z} + 2\mathbf{x}^\top \mathbf{z} \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + \|\mathbf{z}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + 2\|\mathbf{y}\|\|\mathbf{z}\| + 2\|\mathbf{x}\|\|\mathbf{z}\| \\ &\leq 3\|\mathbf{x}\|^2 + 3\|\mathbf{y}\|^2 + 3\|\mathbf{z}\|^2. \end{aligned} \tag{8}$$

We then show the inequality of Lemma 3:

$$\begin{aligned} &\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^j\}_{j=1}^m \sim D(\mathbf{x} + \mu \mathbf{u})} \left[\left\| \frac{1}{m} \sum_{j=1}^m \mathbf{g}_1(\mathbf{x}, \mu, c, \mathbf{u}, \boldsymbol{\xi}^j) \right\|^2 \right] \right] \\ &= \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^j\}_{j=1}^m \sim D(\mathbf{x} + \mu \mathbf{u})} \left[\left\| \frac{\frac{1}{m} \sum_{j=1}^m f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}^j) - c}{\mu} \mathbf{u} \right\|^2 \right] \right] \\ &= \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^j\}_{j=1}^m \sim D(\mathbf{x} + \mu \mathbf{u})} \left[\left\| \frac{F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x}) - F(\mathbf{x} + \mu \mathbf{u}) + \frac{1}{m} \sum_{j=1}^m f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}^j) + F(\mathbf{x}) - c}{\mu} \mathbf{u} \right\|^2 \right] \right] \\ &\stackrel{(*)}{\leq} \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^j\}_{j=1}^m \sim D(\mathbf{x} + \mu \mathbf{u})} \left[3 \left\| \frac{F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x})}{\mu} \mathbf{u} \right\|^2 + 3 \left\| \frac{F(\mathbf{x} + \mu \mathbf{u}) - \frac{1}{m} \sum_{j=1}^m f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}^j)}{\mu} \mathbf{u} \right\|^2 + 3 \left\| \frac{F(\mathbf{x}) - c}{\mu} \mathbf{u} \right\|^2 \right] \right] \\ &= 3 \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\left\| \frac{(F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x}))}{\mu} \mathbf{u} \right\|^2 \right] + 3 \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^j\}_{j=1}^m \sim D(\mathbf{x} + \mu \mathbf{u})} \left[\frac{(F(\mathbf{x} + \mu \mathbf{u}) - \frac{1}{m} \sum_{j=1}^m f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}^j))^2}{\mu^2} \|\mathbf{u}\|^2 \right] \right] \\ &\quad + \frac{3(F(\mathbf{x}) - c)^2}{\mu^2} \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\|\mathbf{u}\|^2] \\ &\stackrel{(**)}{\leq} \frac{3}{2} \mu^2 H_F^2 (d+6)^3 + 6(d+4) \|\nabla F(\mathbf{x})\|^2 + \frac{3\sigma^2 d}{\mu^2 m} + \frac{3d(F(\mathbf{x}) - c)^2}{\mu^2}, \end{aligned}$$

where (*) follows from (8), and (**) follows from Assumption 1, Lemmas B.3, B.6, and B.8. \square

B.4 Proof of Lemma 4

Lemma 4. Let $c_k := \sum_{i=k-s}^{k-1} \frac{a_i}{m_i} \sum_{j=1}^{m_i} f(\mathbf{x}_k, \boldsymbol{\xi}_i^j)$, where $\{\boldsymbol{\xi}_i^j\}_{j=1}^{m_i} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)$ for each $i \in \{k-s, \dots, k-1\}$. Then, the following holds:

$$(F(\mathbf{x}_k) - c_k)^2 \leq 2 \sum_{i=k-s}^{k-1} a_i^2 \left(L_{\boldsymbol{\xi}}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i} \right).$$

Proof.

$$\begin{aligned}
\delta_k^2 &= \left(\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_k)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \sum_{i=k-s}^{k-1} \frac{a_i}{m_i} \sum_{j=1}^{m_i} f(\mathbf{x}_k, \boldsymbol{\xi}_i^j) \right)^2 \\
&= \left(\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_k)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \sum_{i=k-s}^{k-1} a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] \right. \\
&\quad \left. + \sum_{i=k-s}^{k-1} a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \sum_{i=k-s}^{k-1} \frac{a_i}{m_i} \sum_{j=1}^{m_i} f(\mathbf{x}_k, \boldsymbol{\xi}_i^j) \right)^2 \\
&\leq 2 \left(\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_k)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \sum_{i=k-s}^{k-1} a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] \right)^2 \\
&\quad + 2 \left(\sum_{i=k-s}^{k-1} a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \sum_{i=k-s}^{k-1} \frac{a_i}{m_i} \sum_{j=1}^{m_i} f(\mathbf{x}_k, \boldsymbol{\xi}_i^j) \right)^2 \\
&\stackrel{(*)}{=} 2 \left(\sum_{i=k-s}^{k-1} a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_k)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \sum_{i=k-s}^{k-1} a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] \right)^2 \\
&\quad + 2 \left(\sum_{i=k-s}^{k-1} a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \sum_{i=k-s}^{k-1} \frac{a_i}{m_i} \sum_{j=1}^{m_i} f(\mathbf{x}_k, \boldsymbol{\xi}_i^j) \right)^2 \\
&\stackrel{(**)}{\leq} 2 \sum_{i=k-s}^{k-1} \left(a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_k)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] \right)^2 \\
&\quad + 2 \sum_{i=k-s}^{k-1} \left(a_i \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \frac{a_i}{m_i} \sum_{j=1}^{m_i} f(\mathbf{x}_k, \boldsymbol{\xi}_i^j) \right)^2 \\
&= 2 \sum_{i=k-s}^{k-1} a_i^2 \left(\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_k)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] \right)^2 \\
&\quad + 2 \sum_{i=k-s}^{k-1} a_i^2 \left(\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_i + \mu_i \mathbf{u}_i)}[f(\mathbf{x}_k, \boldsymbol{\xi})] - \frac{1}{m_i} \sum_{j=1}^{m_i} f(\mathbf{x}_k, \boldsymbol{\xi}_i^j) \right)^2 \\
&\stackrel{(***)}{\leq} 2 \sum_{i=k-s}^{k-1} a_i^2 L_{\boldsymbol{\xi}}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + 2 \sum_{i=k-s}^{k-1} a_i^2 \frac{\sigma^2}{m_i} \\
&= 2 \sum_{i=k-s}^{k-1} a_i^2 \left(L_{\boldsymbol{\xi}}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i} \right),
\end{aligned}$$

where (*) follows from $\sum_{i=k-s}^{k-1} a_i = 1$, (**) is due to Jensen's inequality and the convexity of L2 norm, (***) comes from Assumption 1, Lemma B.2 and Lemma B.6. \square

B.5 Proof of Lemma 5

Lemma 5. Consider the following optimization problem:

$$\begin{aligned} \min_{\mathbf{a}} \quad & \sum_{i=k-s}^{k-1} b_i a_i^2 \\ \text{s.t.} \quad & a_i \geq 0, \forall i \in I := \{k-s, \dots, k-1\}, \\ & \sum_{i=k-s}^{k-1} a_i = 1, \end{aligned}$$

where $b_i := L_{\xi}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i}$. Then, the optimal solution is \mathbf{a}^* such that $a_i^* := \frac{1}{b_i \sum_{j=k-s}^{k-1} \frac{1}{b_j}}$.

Proof. Let the Lagrangian function be as follows:

$$\mathcal{L}(\mathbf{a}, \boldsymbol{\eta}, \lambda) = \sum_{i \in I} b_i a_i^2 + \sum_{i \in I} \eta_i a_i + \lambda \left(\sum_{i \in I} a_i - 1 \right).$$

The given optimization problem is a convex problem with a strongly convex objective function and satisfies Slater's condition. Therefore, if $(\mathbf{a}, \boldsymbol{\eta}, \lambda)$ satisfies the following KKT conditions, \mathbf{a} is the optimal solution for the given optimization problem.

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathbf{a}, \boldsymbol{\eta}, \lambda)}{\partial a_i} &= 2b_i a_i + \eta_i + \lambda = 0, \quad \forall i \in I, \\ \sum_{i \in I} a_i - 1 &= 0, \\ \eta_i a_i &= 0, \quad a_i \geq 0, \eta_i \leq 0, \quad \forall i \in I. \end{aligned}$$

Here, \mathbf{a}^* such that $a_i^* := \frac{1}{b_i \sum_{j \in I} \frac{1}{b_j}}$, $\lambda^* := -\frac{2}{\sum_{j \in I} \frac{1}{b_j}}$, and $\eta_i^* = 0$ for all $i \in I$ satisfy the above condition. \square

B.6 Proof of Lemma 6

Lemma 6. Suppose that Assumptions 1–4 hold. Let $\{\mathbf{x}_k\}$ and $\{c_k\}$ be the sequence generated by Algorithm 1 with $m_k \geq m_{\min}$ for $k \in \{0, 1, \dots, T\}$, $\beta \leq \frac{\mu_{\min}}{2L_{\xi}\alpha\sqrt{6d}}$, and $M := \frac{L_{\xi}^2 \alpha^2}{\sigma^2}$. Then, for any setting parameter $\mathbf{x}_0 \in \mathbb{R}^d$, $\mu_0 \in \mathbb{R}_{>0}$, $\mu_{\min} \leq \mu_0$, $c_0 \in \mathbb{R}$, $s_{\max} \in \mathbb{N}$, and $\gamma \in (0, 1)$, the following holds:

$$\mathbb{E}_{\zeta_{[0, k-1]}}[\delta_k^2] \leq \left(\frac{1}{2}\right)^k \delta_0^2 + \frac{\mu_{\min}^2 \mu_0^2 H_F^2 (d+6)^3}{2d} + 2\mu_{\min}^2 \frac{d+4}{d} L_F^2 + 8L_{\xi}^2 \alpha^2 \mu_0^2 d + \frac{5}{m_{\min}} \sigma^2,$$

where $\delta_k := F(\mathbf{x}_k) - c_k$, $k \in \{1, \dots, T\}$, and $\zeta_{[0, k-1]} := \{\mathbf{u}_0, \boldsymbol{\xi}_0, \dots, \mathbf{u}_{k-1}, \boldsymbol{\xi}_{k-1}\}$.

Proof. From $M = \frac{L_{\xi}^2 \alpha^2}{\sigma^2}$ and the definition of a_i in line 7 of Algorithm 1, we have $a_i = \frac{1}{b_i \sum_{j=k-s}^{k-1} \frac{1}{b_j}}$, where $\hat{b}_i := L_{\xi}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i}$. Then, from Lemma 5, the following holds for any $\{\theta_i\}_{i=k-s}^{k-1}$ such that $\sum_{i=k-s}^{k-1} \theta_i = 1$ and $\theta_i \geq 0$.

$$\sum_{i=k-s}^{k-1} a_i^2 \left(L_{\xi}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i} \right) \leq \sum_{i=k-s}^{k-1} \theta_i^2 \left(L_{\xi}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i} \right).$$

Thus, considering the case $\theta_{k-1} = 1$ and $\theta_i = 0$ for $i \neq k-1$, we have

$$\sum_{i=k-s}^{k-1} a_i^2 \left(L_{\xi}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_i - \mu_i \mathbf{u}_i\|^2 + \frac{\sigma^2}{m_i} \right) \leq L_{\xi}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_{k-1} - \mu_{k-1} \mathbf{u}_{k-1}\|^2 + \frac{\sigma^2}{m_{k-1}}.$$

Then, from Lemma 4, it yields that

$$\delta_k^2 \leq 2 \left(L_{\xi}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_{k-1} - \mu_{k-1} \mathbf{u}_{k-1}\|^2 + \frac{\sigma^2}{m_{k-1}} \right).$$

Taking the expectation with respect to $\zeta_{[0,k-1]}$ for both sides of the above inequality, the following holds for $k \geq 1$ by letting $\zeta_{[0,-1]} := \emptyset$.

$$\begin{aligned} \mathbb{E}_{\zeta_{[0,k-1]}} [\delta_k^2] &\leq \mathbb{E}_{\zeta_{[0,k-1]}} \left[2 \left(L_{\xi}^2 \alpha^2 \|\mathbf{x}_k - \mathbf{x}_{k-1} - \mu_{k-1} \mathbf{u}_{k-1}\|^2 + \frac{\sigma^2}{m_{k-1}} \right) \right] \\ &\leq \mathbb{E}_{\zeta_{[0,k-1]}} \left[2 \left(L_{\xi}^2 \alpha^2 (2\|\mathbf{x}_k - \mathbf{x}_{k-1}\|^2 + 2\|\mu_{k-1} \mathbf{u}_{k-1}\|^2) + \frac{\sigma^2}{m_{k-1}} \right) \right] \\ &= 4L_{\xi}^2 \alpha^2 \beta^2 \mathbb{E}_{\zeta_{[0,k-1]}} [\|\mathbf{g}_{k-1}\|^2] + 4L_{\xi}^2 \alpha^2 \mu_{k-1}^2 \mathbb{E}_{\zeta_{[0,k-1]}} [\|\mathbf{u}_{k-1}\|^2] + \frac{2\sigma^2}{m_{k-1}} \\ &\stackrel{(*)}{\leq} 4L_{\xi}^2 \alpha^2 \beta^2 \mathbb{E}_{\zeta_{[0,k-1]}} [\|\mathbf{g}_{k-1}\|^2] + 4L_{\xi}^2 \alpha^2 \mu_{k-1}^2 d + \frac{2\sigma^2}{m_{k-1}} \\ &\stackrel{(**)}{\leq} 4L_{\xi}^2 \alpha^2 \beta^2 \left(\frac{3}{2} \mu_{k-1}^2 H_F^2 (d+6)^3 + 6(d+4) \mathbb{E}_{\zeta_{[0,k-2]}} [\|\nabla F(\mathbf{x}_{k-1})\|^2] + \frac{3\sigma^2 d}{\mu_{k-1}^2 m_{k-1}} + \frac{3d \mathbb{E}_{\zeta_{[0,k-2]}} [\delta_{k-1}^2]}{\mu_{k-1}^2} \right) \\ &\quad + 4L_{\xi}^2 \alpha^2 \mu_{k-1}^2 d + \frac{2\sigma^2}{m_{k-1}} \\ &\stackrel{(***)}{\leq} \frac{12L_{\xi}^2 \alpha^2 \beta^2 d \mathbb{E}_{\zeta_{[0,k-2]}} [\delta_{k-1}^2]}{\mu_{k-1}^2} + 6L_{\xi}^2 \alpha^2 \beta^2 \mu_{k-1}^2 H_F^2 (d+6)^3 + 24L_{\xi}^2 \alpha^2 \beta^2 (d+4) L_F^2 + \frac{12L_{\xi}^2 \alpha^2 \beta^2 \sigma^2 d}{\mu_{k-1}^2 m_{k-1}} \\ &\quad + 4L_{\xi}^2 \alpha^2 \mu_{k-1}^2 d + \frac{2\sigma^2}{m_{k-1}} \\ &\stackrel{(***)}{\leq} \frac{12L_{\xi}^2 \alpha^2 \beta^2 d \mathbb{E}_{\zeta_{[0,k-2]}} [\delta_{k-1}^2]}{\mu_{\min}^2} + 6L_{\xi}^2 \alpha^2 \beta^2 \mu_0^2 H_F^2 (d+6)^3 + 24L_{\xi}^2 \alpha^2 \beta^2 (d+4) L_F^2 + \frac{12L_{\xi}^2 \alpha^2 \beta^2 \sigma^2 d}{\mu_{\min}^2 m_{\min}} \\ &\quad + 4L_{\xi}^2 \alpha^2 \mu_0^2 d + \frac{2\sigma^2}{m_{\min}}, \end{aligned}$$

where (*) follows from Lemma B.3, (**) comes from Lemma 3, (***) is due to Lemma B.1, and (****) follows from the fact that $\mu_{\min} \leq \mu_{k-1} \leq \mu_0$ and $m_{\min} \leq m_{k-1}$.

Then, since $\beta \leq \frac{\mu_{\min}}{2L_{\xi}\alpha\sqrt{6d}}$,

$$\begin{aligned} \mathbb{E}_{\zeta_{[0,k-1]}} [\delta_k^2] &\leq \frac{\mathbb{E}_{\zeta_{[0,k-2]}} [\delta_{k-1}^2]}{2} + \frac{\mu_{\min}^2 \mu_0^2 H_F^2 (d+6)^3}{4d} + \mu_{\min}^2 \frac{d+4}{d} L_F^2 + 4L_{\xi}^2 \alpha^2 \mu_0^2 d + \frac{5}{2m_{\min}} \sigma^2 \\ &\leq \left(\frac{1}{2} \right)^k \delta_0^2 + \sum_{i=0}^{k-1} \left(\frac{1}{2} \right)^i \left(\frac{\mu_{\min}^2 \mu_0^2 H_F^2 (d+6)^3}{4d} + \mu_{\min}^2 \frac{d+4}{d} L_F^2 + 4L_{\xi}^2 \alpha^2 \mu_0^2 d + \frac{5}{2m_{\min}} \sigma^2 \right) \\ &= \left(\frac{1}{2} \right)^k \delta_0^2 + \frac{1 - \frac{1}{2^k}}{1 - \frac{1}{2}} \left(\frac{\mu_{\min}^2 \mu_0^2 H_F^2 (d+6)^3}{4d} + \mu_{\min}^2 \frac{d+4}{d} L_F^2 + 4L_{\xi}^2 \alpha^2 \mu_0^2 d + \frac{5}{2m_{\min}} \sigma^2 \right) \\ &\leq \left(\frac{1}{2} \right)^k \delta_0^2 + \frac{\mu_{\min}^2 \mu_0^2 H_F^2 (d+6)^3}{2d} + 2\mu_{\min}^2 \frac{d+4}{d} L_F^2 + 8L_{\xi}^2 \alpha^2 \mu_0^2 d + \frac{5}{m_{\min}} \sigma^2. \end{aligned}$$

□

B.7 Proof of Theorem 1

Theorem 1. Suppose that Assumptions 1–4 hold. Let $\{\mathbf{x}_k\}$ be the sequence generated by Algorithm 1 with $m_k = \Theta(\varepsilon^{-2}d^2)$ for all $k \in \{0, \dots, T\}$, $M := \frac{L_\xi^2 \alpha^2}{\sigma^2}$, $\mu_{\min} = \Theta(\varepsilon d^{-\frac{3}{2}})$, $\mu_0 = \Theta(\varepsilon d^{-\frac{3}{2}})$ such that $\mu_0 \geq \mu_{\min}$, and $\beta := \min\left(\frac{1}{12(d+4)H_F}, \frac{1}{(T+1)^{\frac{1}{2}}d^{\frac{3}{4}}}, \frac{\mu_{\min}}{2L_\xi \alpha \sqrt{6d}}\right)$. Let $\hat{\mathbf{x}} := \mathbf{x}_{k'}$, where k' is chosen from a uniform distribution over $\{0, \dots, T\}$. Then, for any $\mathbf{x}_0 \in \mathbb{R}^d$, $c_0 \in \mathbb{R}$, $s_{\max} \in \{1, \dots, T\}$, and $\gamma \in (0, 1)$, the iteration complexity required to obtain $\mathbb{E}[\|\nabla F(\hat{\mathbf{x}})\|^2] \leq \varepsilon^2$ is $O(d^{\frac{5}{2}}\varepsilon^{-4})$. Moreover, the sample complexity is $O(d^{\frac{9}{2}}\varepsilon^{-6})$.

Proof. Let $\Delta_k := \mathbf{g}_k - \nabla F_{\mu_k}(\mathbf{x}_k)$. From Assumption 4 and Lemma B.4, F_{μ_k} is H_F -smooth. Then,

$$\begin{aligned} F_{\mu_k}(\mathbf{x}_{k+1}) &\leq F_{\mu_k}(\mathbf{x}_k) + \nabla F_{\mu_k}(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{H_F}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= F_{\mu_k}(\mathbf{x}_k) - \beta \nabla F_{\mu_k}(\mathbf{x}_k)^\top \mathbf{g}_k + \frac{H_F \beta^2}{2} \|\mathbf{g}_k\|^2 \\ &= F_{\mu_k}(\mathbf{x}_k) - \beta \|\nabla F_{\mu_k}(\mathbf{x}_k)\|^2 - \beta \nabla F_{\mu_k}(\mathbf{x}_k)^\top \Delta_k + \frac{H_F \beta^2}{2} \|\mathbf{g}_k\|^2. \end{aligned}$$

Rearrange the terms in the above inequality, the following holds.

$$\begin{aligned} \beta \|\nabla F_{\mu_k}(\mathbf{x}_k)\|^2 &\leq F_{\mu_k}(\mathbf{x}_k) - F_{\mu_k}(\mathbf{x}_{k+1}) - \beta \nabla F_{\mu_k}(\mathbf{x}_k)^\top \Delta_k + \frac{H_F \beta^2}{2} \|\mathbf{g}_k\|^2 \\ &= F_{\mu_k}(\mathbf{x}_k) - F_{\mu_{k+1}}(\mathbf{x}_{k+1}) + F_{\mu_{k+1}}(\mathbf{x}_{k+1}) - F_{\mu_k}(\mathbf{x}_{k+1}) + \beta \nabla F_{\mu_k}(\mathbf{x})^\top \Delta_k + \frac{H_F \beta^2}{2} \|\mathbf{g}_k\|^2 \\ &\stackrel{(*)}{\leq} F_{\mu_k}(\mathbf{x}_k) - F_{\mu_{k+1}}(\mathbf{x}_{k+1}) + L_F \sqrt{d} |\mu_{k+1} - \mu_k| + \beta \nabla F_{\mu_k}(\mathbf{x})^\top \Delta_k + \frac{H_F \beta^2}{2} \|\mathbf{g}_k\|^2, \end{aligned}$$

where $(*)$ holds from Lemma B.5. Let $F^* := \min_{\mathbf{x}} F(\mathbf{x})$. Summing up the above inequalities for $0 \leq k \leq$

T , we obtain

$$\begin{aligned}
\sum_{k=0}^T \beta \|\nabla F_{\mu_k}(\mathbf{x}_k)\|^2 &\leq F_{\mu_0}(\mathbf{x}_0) - F_{\mu_{T+1}}(\mathbf{x}_{T+1}) + L_F \sqrt{d} \sum_{k=0}^T |\mu_{k+1} - \mu_k| \\
&\quad + \beta \sum_{k=0}^T \nabla F_{\mu_k}(\mathbf{x}_k)^\top \Delta_k + \frac{H_F \beta^2}{2} \sum_{k=0}^T \|\mathbf{g}_k\|^2 \\
&\stackrel{(*)}{\leq} F(\mathbf{x}_0) + L_F \sqrt{d} \mu_0 - F^* + L_F \sqrt{d} \sum_{k=0}^T |\mu_{k+1} - \mu_k| \\
&\quad + \beta \sum_{k=0}^T \nabla F_{\mu_k}(\mathbf{x}_k)^\top \Delta_k + \frac{H_F \beta^2}{2} \sum_{k=0}^T \|\mathbf{g}_k\|^2 \\
&= F(\mathbf{x}_0) - F^* + L_F \sqrt{d} \left(\mu_0 + \sum_{k=0}^T |\mu_{k+1} - \mu_k| \right) \\
&\quad + \beta \sum_{k=0}^T \nabla F_{\mu_k}(\mathbf{x}_k)^\top \Delta_k + \frac{H_F \beta^2}{2} \sum_{k=0}^T \|\mathbf{g}_k\|^2 \\
&\stackrel{(**)}{\leq} F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} \\
&\quad + \beta \sum_{k=0}^T \nabla F_{\mu_k}(\mathbf{x}_k)^\top \Delta_k + \frac{H_F \beta^2}{2} \sum_{k=0}^T \|\mathbf{g}_k\|^2,
\end{aligned}$$

where $(*)$ comes since $F_{\mu_0}(\mathbf{x}_0) \leq F(\mathbf{x}_0) + L_F \sqrt{d} \mu_0$ from Lemma B.5 and $F_{\mu_{T+1}}(\mathbf{x}_{T+1}) = \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)}[F(\mathbf{x}_{T+1} + \mu_{T+1} \mathbf{u})] \geq \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)}[F^*] = F^*$. The inequality $(**)$ follows from the fact that $\sum_{k=0}^T |\mu_{k+1} - \mu_k| = \sum_{k=0}^T (\mu_k - \mu_{k+1}) = \mu_0 - \mu_{T+1} \leq \mu_0$. Here, let $\zeta_k := \{\mathbf{u}_k, \boldsymbol{\xi}_k\}$, $\zeta_{[0, T]} = \{\mathbf{u}_0, \boldsymbol{\xi}_0, \dots, \mathbf{u}_T, \boldsymbol{\xi}_T\}$, and $m_{\min} := \min_k m_k$. Taking the expectation with respect to the random vectors $\zeta_{[0, T]}$, we obtain

$$\begin{aligned}
&\beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0, T]}} [\|\nabla F_{\mu_k}(\mathbf{x}_k)\|^2] \\
&\leq F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0, T]}} [\nabla F_{\mu_k}(\mathbf{x}_k)^\top \Delta_k] + \frac{H_F \beta^2}{2} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0, T]}} [\|\mathbf{g}_k\|^2] \\
&\stackrel{(*)}{\leq} F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0, k-1]}} [\nabla F_{\mu_k}(\mathbf{x}_k)^\top \mathbb{E}_{\zeta_k} [\Delta_k \mid \zeta_{[0, k-1]}]] \\
&\quad + \frac{H_F \beta^2}{2} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0, k-1]}} \left[\frac{3}{2} \mu_k^2 H_F^2 (d+6)^3 + 6(d+4) \|\nabla F(\mathbf{x}_k)\|^2 + \frac{3\sigma^2 d}{\mu_k^2 m_k} + \frac{3\delta_k^2 d}{\mu_k^2} \right] \\
&\stackrel{(**)}{=} F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \frac{3\beta^2 H_F^3 (d+6)^3}{4} \sum_{k=0}^T \mu_k^2 \\
&\quad + 3(d+4) H_F \beta^2 \sum_{k=0}^T \mathbb{E}_{\zeta_{[0, k-1]}} [\|\nabla F(\mathbf{x}_k)\|^2] + \frac{3}{2} H_F \beta^2 \sigma^2 d \sum_{k=0}^T \frac{1}{\mu_k^2 m_k} + \frac{3}{2} H_F \beta^2 d \sum_{k=0}^T \frac{\mathbb{E}_{\zeta_{[0, k-1]}} [\delta_k^2]}{\mu_k^2} \\
&\stackrel{(***)}{\leq} F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \frac{3\mu_0^2 \beta^2 H_F^3 (d+6)^3 (T+1)}{4} + 3(d+4) H_F \beta^2 \sum_{k=0}^T \mathbb{E}_{\zeta_{[0, k-1]}} [\|\nabla F(\mathbf{x}_k)\|^2] + \frac{3H_F \beta^2 \sigma^2 d (T+1)}{2\mu_{\min}^2 m_{\min}} \\
&\quad + \frac{3H_F \beta^2 d}{2\mu_{\min}^2} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0, k-1]}} [\delta_k^2],
\end{aligned}$$

where (*) follows from Lemma 3, (**) holds from the fact that

$\mathbb{E}_{\zeta_k}[\Delta_k \mid \zeta_{[0,k-1]}] = \mathbb{E}_{\zeta_k}[\Delta_k] = \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)}[\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_k + \mu \mathbf{u})}[\Delta_k]] = 0$ from Lemma 2, and (***) follows from the fact that $\mu_{\min} \leq \mu_k \leq \mu_0$ and $m_{\min} \leq m_k$ for $k = 1, 2, \dots, T$. Then, from Lemma 6,

$$\begin{aligned}
& \beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F_{\mu_k}(\mathbf{x}_k)\|^2] \\
& \leq F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \frac{3\mu_0^2 \beta^2 H_F^3 (d+6)^3 (T+1)}{4} + 3(d+4) H_F \beta^2 \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,k-1]}} [\|\nabla F(\mathbf{x}_k)\|^2] + \frac{3H_F \beta^2 \sigma^2 d(T+1)}{2\mu_{\min}^2 m_{\min}} \\
& \quad + \frac{3H_F \beta^2 d}{2\mu_{\min}^2} \sum_{k=0}^T \left(\left(\frac{1}{2} \right)^k \delta_0^2 + \frac{\mu_{\min}^2 \mu_0^2 H_F^3 (d+6)^3}{2d} + 2\mu_{\min}^2 \frac{d+4}{d} L_F^2 + 8L_{\xi}^2 \alpha^2 \mu_0^2 d + \frac{5}{m_{\min}} \sigma^2 \right) \\
& = F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \frac{3\mu_0^2 \beta^2 H_F^3 (d+6)^3 (T+1)}{2} + 3(d+4) H_F \beta^2 \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,k-1]}} [\|\nabla F(\mathbf{x}_k)\|^2] + \frac{9H_F \beta^2 \sigma^2 d(T+1)}{\mu_{\min}^2 m_{\min}} \\
& \quad + \frac{3H_F \beta^2 d \delta_0^2}{2\mu_{\min}^2} \sum_{k=0}^T \frac{1}{2^k} + 3H_F \beta^2 (d+4) L_F^2 (T+1) + \frac{12H_F \beta^2 L_{\xi}^2 \alpha^2 \mu_0^2 d^2 (T+1)}{\mu_{\min}^2} \\
& \stackrel{(*)}{\leq} F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \frac{3\mu_0^2 \beta^2 H_F^3 (d+6)^3 (T+1)}{2} + 3(d+4) H_F \beta^2 \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,k-1]}} [\|\nabla F(\mathbf{x}_k)\|^2] + \frac{9H_F \beta^2 \sigma^2 d(T+1)}{\mu_{\min}^2 m_{\min}} \\
& \quad + \frac{3H_F \beta^2 d \delta_0^2}{\mu_{\min}^2} + 3H_F \beta^2 (d+4) L_F^2 (T+1) + \frac{12H_F \beta^2 L_{\xi}^2 \alpha^2 \mu_0^2 d^2 (T+1)}{\mu_{\min}^2}, \tag{9}
\end{aligned}$$

where (*) holds since $\sum_{k=0}^T \frac{1}{2^k} = \frac{1-2^{-(T+1)}}{1-2^{-1}} \leq 2$.

Here, from Lemma B.7,

$$\begin{aligned}
& \beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] \leq 2\beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F_{\mu_k}(\mathbf{x}_k)\|^2] + \frac{\beta H_F^2 (d+6)^3}{2} \sum_{k=0}^T \mu_k^2 \\
& \stackrel{(*)}{\leq} 2 \left(F(\mathbf{x}_0) - F^* + 2L_F \mu_0 \sqrt{d} \right) + 3\mu_0^2 \beta^2 H_F^3 (d+6)^3 (T+1) \\
& \quad + 6(d+4) H_F \beta^2 \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,k-1]}} [\|\nabla F(\mathbf{x}_k)\|^2] + \frac{18H_F \beta^2 \sigma^2 d(T+1)}{\mu_{\min}^2 m_{\min}} \\
& \quad + \frac{6H_F \beta^2 d \delta_0^2}{\mu_{\min}^2} + 6L_F^2 H_F \beta^2 (d+4)(T+1) + \frac{24H_F \beta^2 L_{\xi}^2 \alpha^2 \mu_0^2 d^2 (T+1)}{\mu_{\min}^2} \\
& \quad + \frac{\beta H_F^2 (d+6)^3 \mu_0^2 (T+1)}{2},
\end{aligned}$$

where (*) comes from (9) and the fact that $\mu_k \leq \mu_0$ for $k \in \{0, \dots, T\}$. Rearrange the terms in the above inequality, we obtain

$$\begin{aligned}
& (\beta - 6(d+4) H_F \beta^2) \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] \\
& \leq 2 \left(F(\mathbf{x}_0) - F^* + 2L_F \mu_0 \sqrt{d} \right) + 3\mu_0^2 \beta^2 H_F^3 (d+6)^3 (T+1) + \frac{18H_F \beta^2 \sigma^2 d(T+1)}{\mu_{\min}^2 m_{\min}} \\
& \quad + \frac{6H_F \beta^2 d \delta_0^2}{\mu_{\min}^2} + 6L_F^2 H_F \beta^2 (d+4)(T+1) + \frac{24H_F \beta^2 L_{\xi}^2 \alpha^2 \mu_0^2 d^2 (T+1)}{\mu_{\min}^2} + \frac{\beta H_F^2 (d+6)^3 \mu_0^2 (T+1)}{2}.
\end{aligned}$$

Dividing both sides of the above inequality by $(\beta - 6(d+4)H_F\beta^2)(T+1)$ yields

$$\begin{aligned}
& \frac{1}{T+1} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] \\
& \leq 2 \frac{F(\mathbf{x}_0) - F^* + 2L_F\mu_0\sqrt{d}}{(\beta - 6(d+4)H_F\beta^2)(T+1)} + \frac{3\mu_0^2\beta^2H_F^3(d+6)^3}{\beta - 6(d+4)H_F\beta^2} + \frac{18H_F\beta^2\sigma^2d}{\mu_{\min}^2m_{\min}(\beta - 6(d+4)H_F\beta^2)} \\
& \quad + \frac{6H_F\beta^2d\delta_0^2}{\mu_{\min}^2(\beta - 6(d+4)H_F\beta^2)(T+1)} + \frac{6L_F^2H_F\beta^2(d+4)}{\beta - 6(d+4)H_F\beta^2} \\
& \quad + \frac{24H_F\beta^2L_{\xi}^2\alpha^2\mu_0^2d^2}{\mu_{\min}^2(\beta - 6(d+4)H_F\beta^2)} + \frac{\beta H_F^2(d+6)^3\mu_0^2}{2(\beta - 6(d+4)H_F\beta^2)}.
\end{aligned}$$

Moreover, since $\beta = \min \left(\frac{1}{12(d+4)H_F}, \frac{1}{(T+1)^{\frac{1}{2}}d^{\frac{3}{4}}}, \frac{\mu_{\min}}{2L_{\xi}\alpha\sqrt{6d}} \right)$, we have

$$\begin{aligned}
\frac{1}{\beta - 6(d+4)H_F\beta^2} & \leq \frac{1}{\beta - 6(d+4)H_F\beta \cdot \frac{1}{12(d+4)H_F}} = \frac{2}{\beta}, \\
\frac{1}{\beta} & \leq 12(d+4)H_F + (T+1)^{\frac{1}{2}}d^{\frac{3}{4}} + \frac{2L_{\xi}\alpha\sqrt{6d}}{\mu_{\min}}.
\end{aligned}$$

Then,

$$\begin{aligned}
& \frac{1}{T+1} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] \\
& \leq 4 \frac{F(\mathbf{x}_0) - F^* + 2\mu_0L_F\sqrt{d}}{T+1} \left(12(d+4)H_F + (T+1)^{\frac{1}{2}}d^{\frac{3}{4}} + \frac{2L_{\xi}\alpha\sqrt{6d}}{\mu_{\min}} \right) \\
& \quad + 6\mu_0^2\beta H_F^3(d+6)^3 + \frac{36H_F\beta\sigma^2d}{\mu_{\min}^2m_{\min}} + \frac{12H_F\beta d\delta_0^2}{\mu_{\min}^2(T+1)} + 12L_F^2H_F\beta(d+4) \\
& \quad + \frac{48H_F\beta L_{\xi}^2\alpha^2\mu_0^2d^2}{\mu_{\min}^2} + H_F^2(d+6)^3\mu_0^2,
\end{aligned}$$

and therefore, we obtain the following.

$$\begin{aligned}
& \frac{1}{T+1} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] \\
& = O \left(\left((T+1)^{-1} + \mu_0d^{\frac{1}{2}}(T+1)^{-1} \right) \left(d + (T+1)^{\frac{1}{2}}d^{\frac{3}{4}} + d^{\frac{1}{2}}\mu_{\min}^{-1} \right) \right) \\
& \quad + O(\mu_0^2\beta d^3) + O(\beta d\mu_{\min}^{-2}m_{\min}^{-1}) + O(\beta d\mu_{\min}^{-2}(T+1)^{-1}) + O(\beta d) \\
& \quad + O(\beta\mu_0^2d^2\mu_{\min}^{-2}) + O(d^3\mu_0^2) \\
& \stackrel{(*)}{=} O \left(\left((T+1)^{-1} + \varepsilon d^{-1}(T+1)^{-1} \right) \left(d + (T+1)^{\frac{1}{2}}d^{\frac{3}{4}} + d^2\varepsilon^{-1} \right) \right) \\
& \quad + O(d^{-3/4}(T+1)^{-\frac{1}{2}}\varepsilon^2) + O(d^{5/4}(T+1)^{-\frac{1}{2}}) + O(d^{13/4}(T+1)^{-\frac{3}{2}}\varepsilon^{-2}) + O(d^{1/4}(T+1)^{-\frac{1}{2}}) \\
& \quad + O(d^{5/4}(T+1)^{-\frac{1}{2}}) + O(\varepsilon^2),
\end{aligned}$$

where (*) is due to the fact that $\mu_{\min} = \Theta(\varepsilon d^{-\frac{3}{2}})$, $\mu_0 = \Theta(\varepsilon d^{-\frac{3}{2}})$, $m_{\min} = \Theta(\varepsilon^{-2} d^2)$, and $\beta = O(d^{-\frac{3}{4}}(T+1)^{-\frac{1}{2}})$. Let $T := \Theta(d^{\frac{5}{2}} \varepsilon^{-4})$. Then,

$$\frac{1}{T+1} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] \leq O(\varepsilon^2).$$

Therefore, the iteration complexity is $O(d^{\frac{5}{2}} \varepsilon^{-4})$. Moreover, the sample complexity is $\sum_{k=1}^T m_k = T \cdot \Theta(d^2 \varepsilon^{-2}) = O(d^{\frac{9}{2}} \varepsilon^{-6})$. \square

B.8 Proof of Lemma 7

Lemma 7. For any $\mathbf{x} \in \mathbb{R}^d$ and $\mu \in \mathbb{R}_{>0}$, the following holds.

$$\mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\mathbb{E}_{\boldsymbol{\xi}^1 \sim D(\mathbf{x} + \mu \mathbf{u}), \boldsymbol{\xi}^2 \sim D(\mathbf{x} - \mu \mathbf{u})} [\mathbf{g}_2(\mathbf{x}, \mu, \mathbf{u}, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2)]] = \nabla F_\mu(\mathbf{x}).$$

Proof.

$$\begin{aligned} & \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\mathbb{E}_{\boldsymbol{\xi}^1 \sim D(\mathbf{x} + \mu \mathbf{u}), \boldsymbol{\xi}^2 \sim D(\mathbf{x} - \mu \mathbf{u})} [\mathbf{g}_2(\mathbf{x}, \mu, \mathbf{u}, \boldsymbol{\xi}^1, \boldsymbol{\xi}^2)]] \\ &= \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\boldsymbol{\xi}^1 \sim D(\mathbf{x} + \mu \mathbf{u}), \boldsymbol{\xi}^2 \sim D(\mathbf{x} - \mu \mathbf{u})} \left[\frac{f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}^1) - f(\mathbf{x} - \mu \mathbf{u}, \boldsymbol{\xi}^2)}{2\mu} \mathbf{u} \right] \right] \\ &= \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\frac{F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x} - \mu \mathbf{u})}{2\mu} \mathbf{u} \right] \\ &\stackrel{(*)}{=} \nabla F_\mu(\mathbf{x}), \end{aligned}$$

where (*) comes from [Nesterov and Spokoiny, 2017, Eq. (26)]. \square

B.9 Proof of Lemma 8

Lemma 8. Suppose that Assumptions 1 and 4 hold. For any $\mathbf{x} \in \mathbb{R}^d$, the following holds.

$$\begin{aligned} & \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m, \{\boldsymbol{\xi}^{2,j}\}_{j=1}^m} \left[\left\| \frac{1}{m} \sum_{j=1}^m \mathbf{g}_2(\mathbf{x}, \mu, \mathbf{u}, \boldsymbol{\xi}^{1,j}, \boldsymbol{\xi}^{2,j}) \right\|^2 \right] \right] \\ & \leq \frac{3}{2} \mu^2 H_F^2 (d+6)^3 + 6(d+4) \|\nabla F(\mathbf{x})\|^2 + \frac{3\sigma^2 d}{2\mu^2 m}, \end{aligned}$$

where $\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m \sim D(\mathbf{x} + \mu \mathbf{u})$ and $\{\boldsymbol{\xi}^{2,j}\}_{j=1}^m \sim D(\mathbf{x} - \mu \mathbf{u})$.

Proof.

$$\begin{aligned}
& \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m, \{\boldsymbol{\xi}^{2,j}\}_{j=1}^m} \left[\left\| \frac{1}{m} \sum_{j=1}^m \mathbf{g}_2(\mathbf{x}, \mu, \mathbf{u}, \boldsymbol{\xi}^{1,j}, \boldsymbol{\xi}^{2,j}) \right\|^2 \right] \right] \\
&= \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m, \{\boldsymbol{\xi}^{2,j}\}_{j=1}^m} \left[\left\| \frac{\frac{1}{m} \sum_{j=1}^m f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}^{1,j}) - \frac{1}{m} \sum_{j=1}^m f(\mathbf{x} - \mu \mathbf{u}, \boldsymbol{\xi}^{2,j})}{2\mu} \mathbf{u} \right\|^2 \right] \right] \\
&= \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\mathbb{E}_{\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m, \{\boldsymbol{\xi}^{2,j}\}_{j=1}^m} \left[\left\| \frac{F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x} - \mu \mathbf{u})}{2\mu} \mathbf{u} + \frac{\frac{1}{m} \sum_{j=1}^m f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}^{1,j}) - F(\mathbf{x} + \mu \mathbf{u})}{2\mu} \mathbf{u} \right. \right. \\
&\quad \left. \left. + \frac{-\frac{1}{m} \sum_{j=1}^m f(\mathbf{x} - \mu \mathbf{u}, \boldsymbol{\xi}^{2,j}) + F(\mathbf{x} - \mu \mathbf{u})}{2\mu} \mathbf{u} \right\|^2 \right] \right] \\
&\stackrel{(*)}{\leq} \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} \left[3 \left\| \frac{F(\mathbf{x} + \mu \mathbf{u}) - F(\mathbf{x} - \mu \mathbf{u})}{2\mu} \mathbf{u} \right\|^2 + 3 \mathbb{E}_{\{\boldsymbol{\xi}^{1,j}\}_{j=1}^m} \left[\left\| \frac{\frac{1}{m} \sum_{j=1}^m f(\mathbf{x} + \mu \mathbf{u}, \boldsymbol{\xi}^{1,j}) - F(\mathbf{x} + \mu \mathbf{u})}{2\mu} \mathbf{u} \right\|^2 \right] \right. \\
&\quad \left. + 3 \mathbb{E}_{\{\boldsymbol{\xi}^{2,j}\}_{j=1}^m} \left[\left\| \frac{-\frac{1}{m} \sum_{j=1}^m f(\mathbf{x} - \mu \mathbf{u}, \boldsymbol{\xi}^{2,j}) + F(\mathbf{x} - \mu \mathbf{u})}{2\mu} \mathbf{u} \right\|^2 \right] \right] \\
&\stackrel{(**)}{\leq} \frac{3}{2} \mu^2 H_F^2 (d+6)^3 + 6(d+4) \|\nabla F(\mathbf{x})\|^2 + \frac{3\sigma^2}{4\mu^2 m} \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\|\mathbf{u}\|^2] + \frac{3\sigma^2}{4\mu^2 m} \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\|\mathbf{u}\|^2] \\
&\stackrel{(***)}{\leq} \frac{3}{2} \mu^2 H_F^2 (d+6)^3 + 6(d+4) \|\nabla F(\mathbf{x})\|^2 + \frac{3\sigma^2 d}{2\mu^2 m},
\end{aligned}$$

where (*) follows from the fact that $\|\mathbf{x} + \mathbf{y} + \mathbf{z}\|^2 \leq 3\|\mathbf{x}\|^2 + 3\|\mathbf{y}\|^2 + 3\|\mathbf{z}\|^2$ for any $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^d$, and $\mathbf{z} \in \mathbb{R}^d$, as shown in the proof of Lemma 3, (**) follows from Assumption 1, Lemma B.6, and Lemma B.9. (***) comes from Lemma B.3. \square

B.10 Proof of Theorem 2

Theorem 2. Suppose that Assumptions 1–4 hold. Let $\{\mathbf{x}_k\}$ be the sequence generated by Algorithm 2 with $m_k = \Theta(\varepsilon^{-2} d^2)$ for all $k \in \{0, \dots, T\}$, $\mu_{\min} = \Theta(\varepsilon d^{-\frac{3}{2}})$, $\mu_0 = \Theta(\varepsilon d^{-\frac{3}{2}})$ such that $\mu_0 \geq \mu_{\min}$, and $\beta := \min \left(\frac{1}{12(d+4)H_F}, \frac{1}{(T+1)^{\frac{1}{2}} d^{\frac{3}{4}}} \right)$. Let $\hat{\mathbf{x}} := \mathbf{x}_{k'}$, where k' is chosen from a uniform distribution over $\{0, \dots, T\}$. Then, for any $\mathbf{x}_0 \in \mathbb{R}^d$ and $\gamma \in (0, 1)$, the iteration complexity required to obtain $\mathbb{E}[\|\nabla F(\hat{\mathbf{x}})\|^2] \leq \varepsilon^2$ is $O(d^{\frac{5}{2}} \varepsilon^{-4})$. Moreover, the sample complexity is $O(d^{\frac{9}{2}} \varepsilon^{-6})$.

Proof. Let $\Delta_k := \mathbf{g}_k - \nabla F_{\mu_k}(\mathbf{x}_k)$, $\zeta_k := \{\mathbf{u}_k, \boldsymbol{\xi}_k^1, \boldsymbol{\xi}_k^2\}$, $\zeta_{[0,T]} = \{\mathbf{u}_0, \boldsymbol{\xi}_0^1, \boldsymbol{\xi}_0^2, \dots, \mathbf{u}_T, \boldsymbol{\xi}_T^1, \boldsymbol{\xi}_T^2\}$, and $m_{\min} := \min_k m_k$. By the same arguments as in Theorem 1, we obtain the following:

$$\begin{aligned}
& \beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F_{\mu_k}(\mathbf{x}_k)\|^2] \\
& \leq F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\nabla F_{\mu_k}(\mathbf{x}_k)^\top \Delta_k] + \frac{H_F \beta^2}{2} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\mathbf{g}_k\|^2].
\end{aligned}$$

Let $[0, -1] := \emptyset$. Then, from Lemma 8, we have

$$\begin{aligned}
\beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F_{\mu_k}(\mathbf{x}_k)\|^2] &\leq F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,k-1]}} [\nabla F_{\mu_k}(\mathbf{x}_k)^\top \mathbb{E}_{\zeta_k} [\Delta_k \mid \zeta_{[0,k-1]}]] \\
&\quad + \frac{H_F \beta^2}{2} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,k-1]}} \left[\frac{3}{2} \mu_k^2 H_F^2 (d+6)^3 + 6(d+4) \|\nabla F(\mathbf{x}_k)\|^2 + \frac{3\sigma^2 d}{2\mu_k^2 m_k} \right] \\
&\stackrel{(*)}{=} F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \frac{3\beta^2 H_F^3 (d+6)^3}{4} \sum_{k=0}^T \mu_k^2 \\
&\quad + 3H_F \beta^2 (d+4) \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,k-1]}} [\|\nabla F(\mathbf{x}_k)\|^2] + \frac{3}{4} H_F \beta^2 \sigma^2 d \sum_{k=0}^T \frac{1}{\mu_k^2 m_k} \\
&\stackrel{(**)}{\leq} F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} + \frac{3\beta^2 H_F^3 (d+6)^3 (T+1) \mu_0^2}{4} \\
&\quad + 3H_F \beta^2 (d+4) \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,k-1]}} [\|\nabla F(\mathbf{x}_k)\|^2] + \frac{3H_F \beta^2 \sigma^2 d (T+1)}{4\mu_{\min}^2 m_{\min}}, \tag{10}
\end{aligned}$$

where $(*)$ holds since $\mathbb{E}_{\zeta_k} [\Delta_k \mid \zeta_{[0,k-1]}] = \mathbb{E}_{\zeta_k} [\Delta_k] = \mathbb{E}_{\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)} [\mathbb{E}_{\boldsymbol{\xi} \sim D(\mathbf{x}_k + \mu \mathbf{u})} [\Delta_k]] = 0$ from Lemma 7. The inequality $(**)$ holds since $\mu_{\min} \leq \mu_k$ and $m_{\min} \leq m_k$ for $k = 1, 2, \dots, T$. Then, from Lemma B.7,

$$\begin{aligned}
&\beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] \\
&\leq 2\beta \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F_{\mu_k}(\mathbf{x}_k)\|^2] + \frac{\beta H_F^2 (d+6)^3}{2} \sum_{k=0}^T \mu_k^2 \\
&\stackrel{(*)}{\leq} 2 \left(F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} \right) + \frac{3\beta^2 H_F^3 (d+6)^3 (T+1) \mu_0^2}{2} \\
&\quad + 6(d+4) H_F \beta^2 \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,k-1]}} [\|\nabla F(\mathbf{x}_k)\|^2] + \frac{3H_F \beta^2 d \sigma^2 (T+1)}{2\mu_{\min}^2 m_{\min}} + \frac{\beta H_F^2 (d+6)^3 (T+1) \mu_0^2}{2},
\end{aligned}$$

where $(*)$ comes from (10) and $\mu_k \leq \mu_0$. Rearrange the terms in the above inequality, we obtain

$$\begin{aligned}
&(\beta - 6(d+4)H_F \beta^2) \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] \\
&\leq 2 \left(F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d} \right) + \frac{3\beta^2 H_F^3 (d+6)^3 (T+1) \mu_0^2}{2} + \frac{3H_F \beta^2 d \sigma^2 (T+1)}{2\mu_{\min}^2 m_{\min}} + \frac{\beta H_F^2 (d+6)^3 (T+1) \mu_0^2}{2}.
\end{aligned}$$

Dividing both sides of the above inequality by $(\beta - 6(d+4)H_F \beta^2)(T+1)$ yields

$$\begin{aligned}
\frac{1}{T+1} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] &\leq 2 \frac{F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d}}{(\beta - 6(d+4)H_F \beta^2)(T+1)} + \frac{3\beta^2 H_F^3 (d+6)^3 \mu_0^2}{2(\beta - 6(d+4)H_F \beta^2)} \\
&\quad + \frac{3H_F \beta^2 d \sigma^2}{2\mu_{\min}^2 m_{\min} (\beta - 6(d+4)H_F \beta^2)} + \frac{\beta H_F^2 (d+6)^3 \mu_0^2}{2(\beta - 6(d+4)H_F \beta^2)}.
\end{aligned}$$

Here, since $\beta = \min \left(\frac{1}{12(d+4)H_F}, \frac{1}{(T+1)^{\frac{1}{2}} d^{\frac{3}{4}}} \right)$, we have

$$\begin{aligned}
\frac{1}{\beta - 6(d+4)H_F \beta^2} &\leq \frac{1}{\beta - 6(d+4)H_F \beta \cdot \frac{1}{12(d+4)H_F}} = \frac{2}{\beta}, \\
\frac{1}{\beta} &\leq 12(d+4)H_F + (T+1)^{\frac{1}{2}} d^{\frac{3}{4}}.
\end{aligned}$$

Then,

$$\begin{aligned} \frac{1}{T+1} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] &\leq 4 \frac{F(\mathbf{x}_0) - F^* + 2\mu_0 L_F \sqrt{d}}{T+1} \left(12(d+4)H_F + (T+1)^{\frac{1}{2}} d^{\frac{3}{4}} \right) \\ &\quad + 3\beta H_F^3 (d+6)^3 \mu_0^2 + \frac{3H_F \beta d \sigma^2}{\mu_{\min}^2 m_{\min}} + H_F^2 (d+6)^3 \mu_0^2. \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} \frac{1}{T+1} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] &= O((T+1)^{-1} + \mu_0 d^{\frac{1}{2}} (T+1)^{-1}) (d + (T+1)^{\frac{1}{2}} d^{\frac{3}{4}}) \\ &\quad + O(\beta d^3 \mu_0^2) + O(\beta d \mu_{\min}^{-2} m_{\min}^{-1}) + O(d^3 \mu_0^2) \\ &\stackrel{(*)}{=} O\left(((T+1)^{-1} + d^{-1} (T+1)^{-1} \varepsilon) \left(d + (T+1)^{\frac{1}{2}} d^{\frac{3}{4}} \right) \right) \\ &\quad + O(d^{-\frac{3}{4}} (T+1)^{-\frac{1}{2}} \varepsilon^2) + O(d^{\frac{5}{4}} (T+1)^{-\frac{1}{2}}) + O(\varepsilon^2), \\ &= O\left(d(T+1)^{-1} + d^{\frac{3}{4}} (T+1)^{-\frac{1}{2}} + (T+1)^{-1} \varepsilon + d^{-\frac{1}{4}} (T+1)^{-\frac{1}{2}} \varepsilon \right) \\ &\quad + O(d^{-\frac{3}{4}} (T+1)^{-\frac{1}{2}} \varepsilon^2) + O(d^{\frac{5}{4}} (T+1)^{-\frac{1}{2}}) + O(\varepsilon^2), \end{aligned}$$

where (*) follows from the fact that $\mu_{\min} = \Theta(\varepsilon d^{-\frac{3}{2}})$, $\mu_0 = \Theta(\varepsilon d^{-\frac{3}{2}})$, $\frac{1}{m_{\min}} = O(\varepsilon^2 d^{-2})$, and $\beta = O((T+1)^{-\frac{1}{2}} d^{-\frac{3}{4}})$. Let $T := \Theta(d^{\frac{5}{2}} \varepsilon^{-4})$. Then,

$$\frac{1}{T+1} \sum_{k=0}^T \mathbb{E}_{\zeta_{[0,T]}} [\|\nabla F(\mathbf{x}_k)\|^2] \leq O(\varepsilon^2).$$

Therefore, the iteration complexity is $O(d^{\frac{5}{2}} \varepsilon^{-4})$. Moreover, the sample complexity is $\sum_{k=1}^T m_k = T \cdot \Theta(d^2 \varepsilon^{-2}) = O(d^{\frac{9}{2}} \varepsilon^{-6})$.

□