Stochastic Zeroth-order Functional Constrained Optimization: Oracle Complexity and Applications

Anthony Nguyen*  Krishnakumar Balasubramanian†

October 11, 2022

Abstract

Functionally constrained stochastic optimization problems, where neither the objective function nor the constraint functions are analytically available, arise frequently in machine learning applications. In this work, assuming we only have access to the noisy evaluations of the objective and constraint functions, we propose and analyze stochastic zeroth-order algorithms for solving the above class of stochastic optimization problem. When the domain of the functions is $\mathbb{R}^n$, assuming there are $m$ constraint functions, we establish oracle complexities of order $O((m+1)n/\epsilon^2)$ and $O((m+1)n/\epsilon^3)$ respectively in the convex and nonconvex setting, where $\epsilon$ represents the accuracy of the solutions required in appropriately defined metrics. The established oracle complexities are, to our knowledge, the first such results in the literature for functionally constrained stochastic zeroth-order optimization problems. We demonstrate the applicability of our algorithms by illustrating its superior performance on the problem of hyperparameter tuning for sampling algorithms and neural network training.

1 Introduction

We develop and analyze stochastic zeroth-order algorithms for solving the following non-linear optimization problem with functional constraints:

$$\min_{x \in X} f_0(x) \quad \text{such that} \quad f_i(x) \leq 0, \quad i \in \{0, 1, \ldots, m\}, \quad (1)$$

where, for $i \in \{0, 1, \ldots, m\}$, $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are continuous functions which are not necessarily convex defined as $f_i(x) = \mathbb{E}_{\xi_i}[F_i(x, \xi_i)]$ with $\xi_i$ denoting the noise vector associated with function $f_i$, and $X \subseteq \mathbb{R}^n$ is a convex compact set that represents known constraints (i.e., constraints that are analytically available). In the stochastic zeroth-order setting, we neither observe the objective function $f_0$ nor the constraint functions $f_i$ analytically. We only have access to noisy function evaluations of them. The study of stochastic zeroth-order optimization algorithms for unconstrained optimization problems goes back to the early works of Kiefer and Wolfowitz (1952), Blum (1954), Hooke and Jeeves (1961), Spendley et al. (1962), Powell (1964), Nelder and Mead (1965), Nemirovski and Yudin (1983), Spall (1987). Such zeroth-order algorithms have proved to be extremely useful for hyperparameter tuning (Snoek et al. 2012, Hernández-Lobato et al. 2015, Gelbart et al. 2014, Ruan et al. 2019, Golovin et al. 2017), reinforcement learning (Mania et al. 2018, Salimans et al. 2017, Gao et al. 2020, Choromanski et al. 2020) and robotics (Jaquier et al. 2020, Jaquier and Rozo 2020). However, the study of zeroth-order algorithms and their oracle complexities for constrained problem as in (1) is limited, despite the fact that several real-world machine learning problems fall under the setting of (1). We now describe two such applications that serve as our main motivation for developing stochastic zeroth-order optimization algorithms for solving (1), and analyzing their oracle complexity.

*Department of Mathematics, University of California, Davis. antngu@ucdavis.edu.
†Department of Statistics, University of California, Davis kbala@ucdavis.edu.
1.1 Motivating application I:

Hamiltonian Monte Carlo (HMC) algorithm, proposed by Duane et al. (1987) and popularized in the statistical machine learning community by Neal (2011), is a gradient-based sampling algorithm that works by discretizing the continuous time degenerate Langevin diffusion (Leimkuhler and Matthews 2015). It has been used successfully as a state-of-the-art sampler or a numerical integrator in the Bayesian statistical machine learning community by Hoffman and Gelman (2014), Wang et al. (2013), Girolami and Calderhead (2011), Chen et al. (2014), Carpenter et al. (2017). However, in order to obtain successful performance in practice using HMC, several hyperparameters need to be tuned optimally. Typically, the functional relationship between the hyperparameters that need to be tuned and the performance measure used is not available in an analytical form. We can only evaluate the performance of the sampler for various settings of the hyperparameter. Furthermore, in practice several constraints, for example, constraints on running times and constraints that enforce the generated samples to pass certain standard diagnostic tests (Geweke 1991, Gelman and Rubin 1992), are enforced in the hyperparameter tuning process. The functional relationship between such constraints and the hyperparameters is also not available analytically. This makes the problem of optimally setting the hyperparameters for HMC a constrained zeroth-order optimization problem. As a preview, in Section 4.1, we show that our approach provides significant improvements over existing methods of Mahendran et al. (2012), Gelbart et al. (2014), Hernández-Lobato et al. (2015), which are based on Bayesian optimization techniques for tuning HMC, when we measure the performance adopting the widely used effective sample size metric (Kass et al. 1998).

1.2 Motivating application II:

Deep learning has achieved state-of-the-art performance in the recent years for various prediction tasks (Goodfellow et al. 2016). Among the various factors involved behind the success of deep learning, hyperparameter tuning is one of primary factors (Snoek et al. 2012, Bergstra and Bengio 2012, Li et al. 2017, Hazan et al. 2018, Elsken et al. 2019). However, most of the existing methods for tuning the hyperparameters do not enforce any constraints on the prediction time required on the validation set or memory constraints on the training algorithm. Such constraints are typically required to make deep learning methods widely applicable to problem arising in several consumer applications based on tiny devices (Perera et al. 2015, Latré et al. 2011, Yang et al. 2008). As in the above motivating application, the functional relationship between such constraints and the hyperparameters is not available analytically. As a preview, in Section 4.2, we show that our approach provides significant improvements over the existing works of Gelbart et al. (2014), Hernández-Lobato et al. (2015), Ariafar et al. (2019) that developed hyperparameter tuning techniques which explicitly take into account time/memory constraints.

1.3 Related works

In the operations research and statistics communities, zeroth-order optimization techniques are well-studied under the name of derivative-free optimization. Interested readers are referred to Conn et al. (2009) and Audet and Hare (2017). In the machine learning community, Bayesian optimization techniques have been developed for optimizing functions with only noisy function evaluations. We refer the reader to Mockus (1994), Kolda et al. (2003), Spall (2005), Conn et al. (2009), Mockus (2012), Brent (2013), Shahriari et al. (2015), Audet and Hare (2017), Larson et al. (2019), Frazier (2018), Archetti and Candelieri (2019), Liu et al. (2020) for more details. In what follows, we focus on relevant literature from zeroth-order optimization and Bayesian optimization literature for known constrained optimization problems (i.e., problems with constraints that are analytically available). When the constraint set is analytically available and only the objective function is not, Lewis and Torczon (2002) and Bueno et al. (2013) considered an augmented Lagrangian approach and an inexact restoration method respectively, and provided convergence analysis. Furthermore, Kolda et al.
Amaioua et al. (2018), Audet et al. (2015) extended the popular mesh adaptive direct search to this setting. Projection-free methods based on Frank-Wolfe methods have been considered in Balasubramanian and Ghadimi (2018), Sahu et al. (2019) for the case when the constraint set is a convex subset of $\mathbb{R}^n$. Furthermore, Li et al. (2020) considered the case when the constraint set is a Riemannian submanifold embedded in $\mathbb{R}^n$ (and the function is defined only over the manifold). None of the above works are directly applicable to the case of unknown constraints that we consider in this work.

We now discuss some existing methods for solving (variants of) problem (1) in the zeroth-order setting. For solving (1) in the deterministic setting (i.e., we could obtain exact evaluations of the objective and the constraint functions at a given point), filter methods which reduce the objective function while trying to reduce constraint violations were proposed and analyzed in Audet and Dennis Jr (2004), Echebest et al. (2017), Pourmohamad and Lee (2020). Barrier methods in the zeroth-order setting were considered in Audet and Dennis Jr (2006, 2009), Liuzzi and Lucidi (2009), Gratton and Vicente (2014), Fasano et al. (2014), Liuzzi et al. (2010), Dzahini et al. (2020), with some works also developing line search approaches for setting the tuning parameters. Model based approaches were considered in the works of Müller and Woodbury (2017), Tröltzsch (2016), Augustin and Marzouk (2014), Gramacy et al. (2016), Conn and Le Digabel (2013). Furthermore, Bürmen et al. (2006), Audet and Tribes (2018) developed extensions of Nelder–Mead algorithm to the constrained setting.

Several works in the statistical machine learning community also considered Bayesian optimization methods in the constrained setting, in both the noiseless and noisy setting. We refer the reader, for example, to Gardner et al. (2014), Gelbart et al. (2016), Ariafar et al. (2019), Balandat et al. (2020), Bachoc et al. (2020), Greenhill et al. (2020), Eriksson and Poloczek (2020), Letham et al. (2019), Hernández-Lobato et al. (2015), Lam and Willcox (2017), Picheny et al. (2016), Acerbi and Ma (2017). On one hand, the above works demonstrate the interest in the optimization and machine learning communities for developing algorithms for constrained zeroth-order optimization problems. On the other hand, most of the above works are not designed to handle stochastic zeroth-order constrained optimization that we consider. Furthermore, a majority of the above works are methodological, and the few works that develop convergence analysis do so only in the asymptotic setting. A recent work by Usmanova et al. (2019) considered the case when the constraints are linear functions (but unknown), and provided a Frank-Wolfe based algorithm with estimated constraints. However, the proposed approach is limited to only linear constraints, and the oracle complexities established are highly sub-optimal. To the best of our knowledge, there is no rigorous non-asymptotic analysis of the oracle complexity of stochastic zeroth-optimization when the constraints and the objective values are available only via noisy function evaluations.

1.4 Methodology and Main Contributions:

Our methodology is based on a novel constraint extrapolation technique developed for the zeroth-order setting, extending the work of Boob et al. (2022) in the first-order setting, and the Gaussian smoothing based zeroth-order stochastic gradient estimators. Specifically, we propose the Szo-ConEX method in Algorithm 1 for solving problems of the form in (1). We theoretically characterize how to set the tuning parameters of the algorithm so as to mitigate the issues caused by the bias in the stochastic zeroth-order gradient estimates and obtain the oracle complexity of our algorithm. More specifically, we make the following main contributions:

- When the functions $f_i$, $i = 0, \ldots, m$, are convex, in Theorem 3.1, we show that the number of calls to the stochastic zeroth-order oracle to achieve an appropriately defined $\epsilon$-optimal solution of (1) (see Definition 3.1) is of order $O((m + 1)n/\epsilon^2)$.

- When the functions are nonconvex, in Proposition 3.1, we show that the number of calls to the stochastic zeroth-order oracle to achieve an appropriately defined $\epsilon$-optimal KKT solution of (1) (see Definition 3.2) is of order $O((m + 1)n/\epsilon^3)$.
To our knowledge, these are the first non-asymptotic oracle complexity results for stochastic zeroth-order optimization with stochastic zeroth-order functional constraints. We illustrate the practical applicability of the developed methodology by testing its performance on hyperparameter tuning for HMC sampling algorithm (Section 4.1) and 3-layer neural network (Section 4.2).

2 Preliminaries and Methodology

Notations: Let 0 denote the vector of elements 0 and \([m] := \{1, \ldots, m\}\). Let \(f(x) := [f_1(x), \ldots, f_m(x)]^T\); then, the constraints in (1) can be expressed as \(f(x) \leq 0\). We use \(\xi := [\xi_1, \ldots, \xi_m]\) to denote the random vectors in the constraints. Furthermore, \(\|\cdot\|\) denotes a general norm and \(\|\cdot\|_*)\) denotes its dual norm defined as \(\|z\|_* := \sup\{z^T x : \|x\| \leq 1\}\). Furthermore, \([x]_+ := \max\{x, 0\}\) for any \(x \in \mathbb{R}\). For any vector \(x \in \mathbb{R}^k\), we define \([x]_+\) as element-wise application of \([\cdot]_+\).

We now describe the precise assumption made on the stochastic zeroth-order oracle in this work.

**Assumption 2.1.** Let \(\|\cdot\|\) be a norm on \(\mathbb{R}^n\). For \(i \in \{0, \ldots, m\}\) and for any \(x \in \mathbb{R}^n\), the zeroth-order oracle outputs an estimator \(F_i(x, \xi_i)\) of \(f_i(x)\) such that \(\mathbb{E}[F_i(x, \xi_i)] = f_i(x)\), \(\mathbb{E}[F_i(x, \xi_i)^2] \leq \sigma_i^2\), \(\mathbb{E}[\nabla F_i(x, \xi_i)] = \nabla f_i(x)\), \(\mathbb{E}[\|\nabla F_i(x, \xi_i) - \nabla f_i(x)\|_*^2] \leq \sigma_i^2,\) where \(\|\cdot\|_*\) denotes the dual norm.

The assumption above assumes that we have access to a stochastic zeroth-order oracle which provides unbiased function evaluations with bounded variance. It is worth noting that in the above assumption, we do not necessarily assume the noise \(\xi\) is additive. Furthermore, we allow for different noise models for the objective function and the \(m\) constraint functions, which is a significantly general model compared to several existing works such as [Digabel and Wild (2015)]. Our gradient estimator is then constructed by leveraging the Gaussian smoothing technique proposed in [Nemirovski and Yudin (1983)], [Nesterov and Spokoiny (2017)]. For \(\nu_i \in (0, \infty)\) we introduce the smoothed function \(f_{i, \nu_i}(x) = \mathbb{E}_{u_i}[f_i(x + \nu_i u_i)]\) where \(u_i \sim \mathcal{N}(0, I_n)\) and independent across \(i\). We can estimate the gradient of this smoothed function using function evaluations of \(f_i\). Specifically, we define the stochastic zeroth-order gradient of \(f_{i, \nu_i}(x)\) as

\[
G_{i, \nu_i}(x, \xi_i, u_i) = \frac{F_i(x + \nu_i u_i, \xi_i) - F_i(x, \xi_i)}{\nu_i},
\]

which is an unbiased estimator of \(\nabla f_{i, \nu_i}(x)\), i.e., we have \(\mathbb{E}_{u,\xi_i}[G_{i, \nu_i}(x, \xi_i, u)] = \nabla f_{i, \nu_i}(x)\). However, it is well-known that \(G_{i, \nu_i}(x, \xi_i, u_i)\) is a biased estimator of \(\nabla f_i(x)\). An interpretation of the gradient estimator in (2) as a consequence of Gaussian Stein’s identity, popular in the statistics literature ([Stein 1972]), was provided in [Balasubramanian and Ghadimi (2022)].

The gradient estimator in (2) is referred to as the two-point estimator in the literature. The reason is that, for a given random vector \(\xi_i\), it is assumed that the stochastic function in (2) could be evaluated at two points, \(F_i(x + \nu_i u_i, \xi_i)\) and \(F_i(x, \xi_i)\). Such an assumption is satisfied in several statistics, machine learning, simulation based optimization, and sampling problems; see for example [Spall (2005)], [Mokkadem and Pelletier (2007)], [Dippon (2003)], [Agarwal et al. (2010)], [Duchi et al. (2015)], [Ghadimi and Lan (2013)], [Nesterov and Spokoiny (2017)]. Yet another estimator in the literature is the one-point estimator, which assumes that for each \(\xi_i\), we observe only one noisy function evaluation \(F_i(x + \nu_i u_i, \xi_i)\). It is well-known that the one-point setting is more challenging than the two-point setting ([Shamir 2013]). From a theoretical point of view, the use of two-point evaluation based gradient estimator is primarily motivated by the sub-optimality (in terms of oracle complexity) of one-point feedback based stochastic zeroth-order optimization methods either in terms of the approximation accuracy or dimension dependency. For the rest of this work, we focus on the two-point setting and leave the question of obtaining results in the one-point setting as future work.

We now describe our assumptions on the objective and constraint functions.
Assumption 2.2. Function $F_i$ has Lipschitz continuous gradient with constant $L_i$, almost surely for any $\xi_i$, i.e., \(|\nabla F_i(y, \xi_i) - \nabla F_i(x, \xi_i)| \leq L_i|y - x|\), which consequently implies that \(|F_i(y, \xi_i) - F_i(x, \xi_i) - \langle \nabla F_i(x, \xi_i), y - x \rangle| \leq \frac{L_i}{2}|y - x|^2\) for $i \in \{0, 1, \ldots, m\}$.

Assumption 2.3. Function $F_i$ is Lipschitz continuous with constant $M_i$, almost surely for any $\xi_i$, i.e., \(|F_i(y, \xi_i) - F_i(x, \xi_i)| \leq M_i|y - x|\), for $i \in \{0, 1, \ldots, m\}$.

The above smoothness assumptions are standard in the literature on stochastic zeroth-order optimization and are made in several works [Nesterov and Spokoiny (2017), Ghadimi and Lan (2013), Balasubramanian and Ghadimi (2022)] for obtaining oracle complexity results. It is easy to see that Assumption 2.2 implies that for $i \in \{0, \ldots, m\}$, $f_i$ has Lipschitz continuous gradient with constant $L_i$ since \(|\nabla f_i(y) - \nabla f_i(x)|_* \leq \mathbb{E}[||\nabla F(y, \xi) - \nabla F(x, \xi)||_*] \leq L_i|y - x|\), due to Jensen’s inequality for the dual norm. By similar reasoning and Assumption 2.3, we also see that $f_i$ is Lipschitz continuous with constant $M_i$. Due to Assumptions 2.2 and 2.3, we also have \(|f(x_1) - f(x_2)|_2 \leq M_f|x_1 - x_2|\), \(|\nabla f(x_2)^T(x_1 - x_2)|_2 \leq M_f|x_1 - x_2|\) and \(|\nabla f(x_1) - \nabla f(x_2)^T(x_1 - x_2)|_2 \leq \frac{L_f}{2}|x_1 - x_2|^2\), for all $x_1, x_2 \in \mathbb{R}^n$, where $\nabla f(\cdot) := [\nabla f_1(\cdot), \ldots, \nabla f_m(\cdot)] \in \mathbb{R}^n \times \mathbb{R}^m$ and constants $M_f$ and $L_f$ are defined as

\[
M_f := \sqrt{\sum_{i=1}^m M_i^2} \quad \text{and} \quad L_f := \sqrt{\sum_{i=1}^m L_i^2}.
\]

We now state the definition of the $\text{prox}$-function and the $\text{prox}$-operator. The class of algorithms based on $\text{prox}$-operators are called proximal algorithms. Such algorithms have turned out to be particularly useful for efficiently solving various machine learning problems in the past. We refer the interested reader to [Parikh and Boyd (2014), Beck (2017)] for more details.

Definition 2.1. Let $\omega : X \to \mathbb{R}$ be continuously differentiable, $L_\omega$-Lipschitz gradient smooth, and 1-strongly convex with respect to $\| \cdot \|$ function. We define the $\omega$-prox-function associated with $\omega(\cdot)$, $\forall x, y \in \mathbb{R}^n$, as $W(y, x) := \omega(y) - \omega(x) - \langle \nabla \omega(x), y - x \rangle$. Based on the smoothness and strong convexity of $\omega(x)$, we have $W(y, x) \leq \frac{L_\omega}{2}\|x - y\|^2 \leq L_\omega W(x, y)$, $\forall x, y \in \mathbb{R}^n$. For any $v \in \mathbb{R}^n$, we define the following $\text{prox}$-operator as $\text{prox}(v, \bar{x}, \eta) := \arg \min_{x \in X} \{\langle v, x \rangle + \eta W(x, \bar{x})\}$.

The function $W$ is also called as Bregman divergence in the literature. A canonical example of $W$ is that of the Euclidean distance function $\|x - y\|^2$ which is useful when $X = \mathbb{R}^n$. We will see in Section 2.1 that our algorithm is based on the above $\text{prox}$-operator. Finally, we have the following results which will prove to be useful for subsequent calculations. Let $u := [u_1, \ldots, u_m]$ and $D_X := \sup_{x, y \in X} \|W(x, y)\|$ be the diameter of the set $X$.

Lemma 2.1. Let $\nu := [\nu_1, \ldots, \nu_m]$, $F_\nu(x, \xi, u) := [F_1(x + \nu_1 u_1, \xi_1), \ldots, F_m(x + \nu_m u_m, \xi_m)]^T$ and $f_\nu(x) := [f_1(x), \ldots, f_m(x)]^T$. Under assumption 2.3, we have $\mathbb{E}_{u, \xi}[||F_\nu(x, \xi, u) - f_\nu(x)||^2] \leq \sigma_{1,\nu}^2$, where $\sigma_{1,\nu}^2 := (\sum_{i=1}^m 4(n + 2)M_i^2 \nu_i^2 + L_i^2 \nu_i^2 n^2) + 2\sigma_{2,\nu}^2$, where $\sigma_{2,\nu}^2 := \sum_{i=1}^m \sigma_{2,i}^2$.

Lemma 2.2. Let $B_i := \frac{\nu_i^2}{2}L_i(n + 3)^2 2 + L_iD_X + M_i$. Under assumptions 2.4 and 2.2, we have

\[
\mathbb{E}_{u, \xi}[||G_{i,\nu_i}(x, \xi, u) - \nabla f_{i,\nu_i}(x)||^2] \leq \sigma_{i,\nu_i}^2,
\]

where $\sigma_{i,\nu_i}^2 := \nu_i^2 L_i^2(n + 6)^2 + 10(n + 4)[\sigma_i^2 + \tilde{B}_i^2]$.

2.1 Algorithmic Methodology

We now present the $\text{SZO-ConEX}$ algorithm for solving the stochastic zeroth-order functional constrained optimization problem [1]. The constraint extrapolation framework is a novel primal-dual method that proceeds
by (i) considering the Lagrangian formulation of (1), (ii) constructing linear approximations for the constraint functions, and (iii) constructing an extrapolation operation which enables acceleration. Such an approach has the advantage that: (i) it does not require the projection of Lagrangian multipliers onto a possibly unknown bounded set (which is required by several other primal-dual methods), (ii) it is a single-loop algorithm with a built-in acceleration step. It is worth remarking that Boob et al. (2022) and Hamedani and Aybat (2021) showed that such an approach helps achieve better rate of convergence than existing methods for solving Lagrangian problems (of the form in (5) below) in the stochastic first-order setting. However, their approach is not directly applicable to the zeroth-order setting where the estimated stochastic gradients are biased and have variances that are not uniformly bounded.

Recall the problem in (1) and notice that there are two types of constraints. The set \( X \) represents known constraints (i.e., constraints that are analytically available) and the inequality constraints defined by the functions \( f_i, i \in [m] \) are the unknown or zeroth-order constraints. The Lagrangian of (1) is given by

\[
\min_{x \in X} \max_{y \geq 0} \{ \mathcal{L}(x, y) := f_0(x) + \sum_{i=1}^{m} y_i f_i(x) \}. \tag{5}
\]

In other words, \((x^*, y^*)\) is a saddle point of the Lagrange function \( \mathcal{L}(x, y) \) such that

\[
\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*), \tag{6}
\]

for all \( x \in X, y \geq 0 \), whenever the optimal dual, \( y^* \), exists. Throughout this work, we assume the existence of \( y^* \) satisfying (6). In order to handle the zeroth-order setting, we also define Lagrangian with the smoothed functions as

\[
\mathcal{L}_\nu(x, y) := f_{0,\nu_0}(x) + \sum_{i=1}^{m} y_i f_{i,\nu_1}(x). \tag{7}
\]

Now, we describe the linearization in the context of the iterates directly as it will be easier to understand in the stochastic setting that we are in. Let \( x^{(t)} \) be the sequence produced by the algorithm (to be discussed later). The linearization of \( f(\cdot) \) at the point \( x^{(t)} \), with respect to the point \( x^{(t-1)} \), is given by

\[
\ell_f(x^{(t)}) := f_\nu(x^{(t-1)}) + \nabla f_\nu(x^{(t-1)})^T (x^{(t)} - x^{(t-1)}),
\]

where similar to \( \nabla f \), we define \( \nabla f_\nu(x^{(t-1)}) := [\nabla f_{1,\nu_1}(x^{(t-1)}), \ldots, \nabla f_{m,\nu_m}(x^{(t-1)})] \). For the implementation, we use the version of linearization with the Gaussian smoothing based stochastic zeroth-order gradients. In particular, we define \( \ell_F(x^{(t)}) := F_\nu(x^{(t-1)}, \xi^{(t-1)}, \bar{u}^{(t-1)}) + G_\nu(x^{(t-1)}, \xi^{(t-1)}, \bar{u}^{(t-1)})^T (x^{(t)} - x^{(t-1)}), \)

where \( G_\nu(x^{(t-1)}, \xi^{(t-1)}, \bar{u}^{(t-1)}) \in \mathbb{R}^{n \times m} \) is given by

\[
[G_{1,\nu_1}(x^{(t-1)}, \xi^{(t-1)}, \bar{u}^{(t-1)}), \ldots, G_{m,\nu_m}(x^{(t-1)}, \xi^{(t-1)}, \bar{u}^{(t-1)})].
\]

Here, by \( \xi^{(t-1)}, \bar{u}^{(t-1)} \) we mean an independent (of \( \xi^{(t-1)}, \bar{u}^{(t-1)} \), respectively) realization of random objects \( \xi, \bar{u} \), respectively.

Based on this, the overall procedure, termed as \text{SZO-ConEx} is provided in Algorithm 1. We now explain the individual steps in more detail.

- **Step 3:** This extrapolation step, considered by Boob et al. (2022) (see also, Hamedani and Aybat (2021)) for the stochastic first-order setting forms the main methodological innovation over existing primal-dual method. First, note that instead of working with constraint functions, we work with a stochastic linearization of them. The extrapolation or moving average is essentially a way to incorporate momentum in the \( s^{(t)} \) sequence. From the analysis, it turns out that the choice of constant \( \theta_t \) (which we set as \( \theta_t = 1 \) without any loss of generality) gives the best possible oracle complexity in our analysis.
We now present our main results on the oracle complexity of Algorithm 1 of unbiased stochastic gradient and uniformly bounded variance is not directly applicable. Furthermore, on Boob et al. (2022), which is for the stochastic first-order setting under the assumption of unbiased estimators of the smoothed functions $f$ by $L$ as using the constraint extrapolation method of Boob et al. (2022) for solving (5) with stochastic zeroth-order gradient estimators from (2). At a high-level, the algorithm could be interpreted as solving the problem in (1) in the non-noisy setting. However, they did not propose the above constraint extrapolation technique. In our experiments in Section 4, we show that our constraint extrapolation approach significantly outperforms the approach in Gramacy et al. (2016) in simulations and real-world problems.

### 3 Main results

We now present our main results on the oracle complexity of $\text{SZO-ConEx}$ algorithm. Recall the definition of the stochastic zeroth-order gradient estimators from (2). At a high-level, the algorithm could be interpreted as using the constraint extrapolation method of Boob et al. (2022) for solving (5) with $\mathcal{L}(x, y)$ replaced by $L_{\nu}(x, y)$ as defined in (7), as the stochastic zeroth-order gradients used in Algorithm are essentially unbiased estimators of the smoothed functions $f_{\nu, i}$ (for $i \in [m]$). However, they have unbounded variance. Hence, the analysis of Boob et al. (2022), which is for the stochastic first-order setting under the assumption of unbiased stochastic gradient and uniformly bounded variance is not directly applicable. Furthermore, on
the one hand as the smoothing parameters \( \nu_i \) (for \( i \in [m] \)) tend to zero, \( \mathcal{L}_\nu(x, y) \) converges to \( \mathcal{L}(x, y) \) defined in (5). However, on the other hand, the parameters \( \nu_i \) are in the denominator of the stochastic zeroth-order gradient estimators (see (2)). Hence, we cannot let them tend to zero at any arbitrary rate. Picking the tuning parameters \( \nu_i \) carefully to balance this tension and get the best possible oracle complexity forms the crux of our analysis. Finally, we also point out that general strategies for picking the smoothing parameters (as proposed in Beck and Teboulle (2012) for dealing with non-smooth stochastic first-order optimization problems) are also not directly applicable for analyzing stochastic zeroth-order algorithms and specialized approaches are often required – we refer the reader to Duchi et al. (2015), Nesterov and Spokoiny (2017), Ghadimi and Lan (2013), Balasubramanian and Ghadimi (2022) for several related techniques for analyzing unconstrained stochastic zeroth-order optimization algorithms.

### 3.1 Convex Setting

We first provide our theoretical results for the case when the functions \( f_i \), for \( i \in [m] \), are convex. We start by describing the measure of optimality we consider for solving (1).

**Definition 3.1.** A point \( \bar{x} \) is an \( \epsilon \)-approximately optimal solution in expectation, for (1), if it satisfies \( \mathbb{E}[f_0(\bar{x}) - f_0^*] \leq \epsilon \) and \( \mathbb{E}[\|f(\bar{x})\|_2] \leq \epsilon \), where \( f_0^* \) is the optimal value of (1) and the expectation is with respect to the randomness arising due to \( \xi_i \) and \( u_i \) across all iterations.

The first part of the above definition corresponds to the standard optimality condition for the convex problem. The next part corresponds to constraint violation. Our main result is described next. We define \( M_X := \sup_{x \in X} \|x\| \). Furthermore, we define \( \sigma_\nu := \left[\sigma_{1,\nu_1}, \ldots, \sigma_{m,\nu_m}\right] \), where \( \sigma_{i,\nu_i} \), for \( i \in [m] \) are as defined in Lemma 2.2, \( \sigma_{X,f} := (\nu^2 + D_X^2 \|\sigma_\nu\|^2_2)^{1/2} \) (where \( \nu^2 \) is as defined in Lemma 2.1).

**Theorem 3.1.** Suppose the functions \( f_i \), for \( i \in [m] \), are convex and satisfy Assumptions 2.1, 2.2 and 2.3. Define \( \mathcal{H}_* := (L_f D_X \|y^*\|_2)/2 \). Set \( y_0 = 0 \) and \( \{\gamma_t, \theta_t, \eta_t, \tau_t\} \) in Algorithm 1 according to the following:

\[
\begin{align*}
\eta := & \max \left\{ \frac{\sqrt{2T[H_x^2 + \sigma_0^2 + 48\|\sigma_\nu\|^2_2]}}{D_X}, \frac{6 \max \{2M_f, 4\|\sigma_\nu\|_2\}}{D_X} \right\}, \\
\tau := & \max \left\{ \sqrt{96T} \sigma_{X,f}, 2D_X \max \{M_f, 4\|\sigma_\nu\|_2\} \right\}.
\end{align*}
\]

Then, we have

\[
\begin{align*}
\mathbb{E}[f_0(\bar{x}_T) - f_0(x^*)] & \leq \frac{(L_0 + L_f)D_X^2}{T} + \max \{12M_f, 24\|\sigma_\nu\|_2\}D_X \\
& + \frac{1}{\sqrt{T}} \sqrt{2(H_x^2 + \sigma_0^2 + 48\|\sigma_\nu\|^2_2)D_X} \\
& + \frac{1}{\sqrt{T}} \left\{ \frac{\sqrt{2} \zeta^2 D_X}{\sqrt{H_x^2 + \sigma_0^2 + 48\|\sigma_\nu\|^2_2}} + \frac{\sqrt{3} \sigma_{X,f}}{\sqrt{2}} \right\} \\
& + [\nu_0^2 L_0 n + M_X n (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}],
\end{align*}
\] (8)
Theorem 3.1. We now consider the case when objective function

\[ \nu \]

construct a sequence of convex relaxations for the nonconvex problem, and (ii) leverage the algorithm

\[ \nu \]

following the standard analysis for selecting the tuning parameters for stochastic zeroth-order algorithms, Remark 3.2. A word is in order regarding the choice of the tuning parameters \( \nu \), instead of the linear dependency that we obtain now. A main part of our proof involves obtaining

\[ \nu \]

Although the parameter settings of Theorem 3.1 and the right hand side of Remark 3.1. \( \nu \)-approximately optimal solution of (1)

\[ \nu \]

is of the order \( \Omega \left( \left( (m+1)n \right) / \epsilon^2 \right) \).

Remark 3.1. Although the parameter settings of Theorem 3.1 and the right hand side of (8) and (9) appear complicated to parse, the important take away message is that the right hand side of (8) and (9) are of the order \( \Omega(1/\sqrt{T}) \) which leads to the oracle complexity described above. Furthermore, the order of \( \epsilon \) in the oracle complexity is of the same order as that in [Boob et al., 2022] for the stochastic first-order setting. The \( (m+1)n \) factor in the oracle complexity appears because we are required to estimate \( m+1 \) gradient vectors, each of dimension \( n \). The dimension dependency is unavoidable even in the unconstrained setting, as showed via lower bounds in [Jamieson et al., 2012], [Duchi et al., 2015]. For a fixed dimensionality \( n \), the oracle complexity in the zeroth-order setting is linear in the number of constraints \( m \).

Remark 3.2. A word is in order regarding the choice of the tuning parameters \( \nu_i, i \in [m] \) in (11). If one follows the standard analysis for selecting the tuning parameters for stochastic zeroth-order algorithms, which are predominantly developed for unconstrained problems, the \( m \) related factors appearing in the choice of \( \nu_i \) would be missed. This subsequently would lead to an increased dependency of the oracle complexity on \( m \), instead of the linear dependency that we obtain now. A main part of our proof involves obtaining the choice of the smoothing parameters \( \nu_i \) as in (11), that helps us to obtain oracle complexity as stated in Theorem 3.7.

3.2 Proximal-point based Meta-Algorithm for the Nonconvex Setting

We now consider the case when objective function \( f_0 \), and the constraint functions \( f_1, \ldots, f_m \) are nonconvex. In this case, [Boob et al., 2022] analyzed a two-step meta-algorithm, which is based on the standard proximal-method; see, for example [Drusvyatskiy, 2017] for a survey.

The basic idea behind the method (as stated in Algorithm 2) consists of the following two steps: (i) construct a sequence of convex relaxations for the nonconvex problem, and (ii) leverage the algorithm
We refer to the distance between two sets which is the standard approximate optimality condition for solving (1) in the nonconvex setting.

Algorithm 2 Meta-Algorithm for Nonconvex Setting

Input: Input $x_0$, parameters $\mu_0$, $\mu_i$, $i \in [m]$.
1: for $k = 1, \ldots, K$ do
2: 
3: Obtain an $\epsilon$-approximately optimal solution to the problem:

$$\arg \min_{x \in X} f_0(x; x_{k-1}) \quad \text{s.t.} \quad f_i(x; x_{k-1}) \leq 0, \quad i \in [m].$$

by using $SZO-\text{ConEx}$ in Algorithm 1. Denote it by $x_k$, for $k = 1, \ldots, K$.
4: Randomly choose $k \in \{1, \ldots, K\}$
5: return $x_k$.

developed for the convex setting. Given our Algorithm 1 we leverage this framework to solve (1) in the nonconvex setting.

We first define the exact Karush-Kuhn-Tucker (KKT) condition for (1) as follows. For a convex set $X$, we denote its interior as $\text{int} X$, the normal cone at $x \in X$ as $N_X(x)$, and its dual cone as $N_X^*(x)$. For convenience, we recall the definition of normal cone: For convex set $X$, we have $N_X^*(x) := \{v \in \mathbb{R}^n : \langle y, z - x \rangle \leq 0, \text{ for all } z \in X\}$; see [Rockafellar 2015, Part I and II] for additional properties and examples. Let $\oplus$ denote the Minkowski sum of two sets $A, B \subset \mathbb{R}^n$, defined as $A \oplus B = \{a + b : a \in A \text{ and } b \in B\}$. We refer to the distance between two sets $A, B \subset \mathbb{R}^n$ as $d(A, B) := \inf_{a \in A, b \in B} \|a - b\|$.

Definition 3.2. We say that $x^* \in X$ is a critical KKT point of (1) if $f_i(x^*) \leq 0$ and $\exists y^* := [y_1^*, \ldots, y_m^*]^T \geq 0$ such that

$$y_i^* f_i(x^*) = 0, \quad i \in [m],$$

$$d(\nabla f_0(x^*)) + \sum_{i=1}^m y_i^* \nabla f_i(x^*) + N_X(x^*), 0) = 0.$$

The parameters $\{y_i^*\}_{i \in [m]}$ are called Lagrange multipliers. For brevity, we use the notation $y^*$ and $[y_1^*, \ldots, y_m^*]^T$ interchangeably. With this definition, we also have the following approximate KKT condition which is the standard approximate optimality condition for solving (1) in the nonconvex setting.

Definition 3.3. We say that a point $\hat{x} \in X$ is an $(\epsilon, \delta)$-KKT point in expectation for (1) if there exists $(\bar{x}, \bar{y})$ such that $f(\bar{x}) \leq 0, \bar{y} \geq 0$ and

$$\mathbb{E}[\sum_{i=1}^m |\bar{y}_i f_i(\bar{x})|] \leq \epsilon, \mathbb{E}[\|\bar{x} - \hat{x}\|^2] \leq \delta$$

$$\mathbb{E}[d(\nabla f_0(\bar{x})) + \sum_{i=1}^m \bar{y}_i \nabla f_i(\bar{x}) + N_X(\bar{x}), 0^2] \leq \epsilon.$$

Proposition 3.1. Consider solving (1) with both the objective and the constraint function being nonconvex and satisfying Assumptions 2.1, 2.2 and 2.3 Then, by running Algorithm 2 with $K = O(1/\epsilon)$, we obtain $(\epsilon, 2\epsilon/2\mu_0\mu_{\text{max}})$-KKT point, where $\mu_{\text{max}} := \max\{\mu_1, \ldots, \mu_m\}$. Hence, the total number of calls to the stochastic zeroth-order oracle is given by $O\left((m + 1)n/\epsilon^3\right)$.

The proof of the above proposition follows immediately by Theorem 4.1 and Corollary 3.19 from [Boob et al. 2022] and is hence omitted. The parameters $\mu_0$ and $\mu_i$, $i \in [m]$ in Algorithm 2 are set according to the desired level of accuracy based on Proposition 3.1. To the best of our knowledge, we are not aware of a non-asymptotic result on the oracle complexity of stochastic zeroth-order optimization with stochastic zeroth-order functional constraints, in both the convex and nonconvex settings.
3.3 Detailed Comparison to Boob et al. (2022)

In this subsection, we highlight the main differences between our work and Boob et al. (2022). As mentioned previously, our methodological and theoretical results builds upon the work of Boob et al. (2022).

- **Methodological:** At a methodological level, our work focuses on the case when we only have noisy function evaluations, whereas Boob et al. (2022) focus on the case when we have access to noisy gradients. To deal with this, we use the Gaussian smoothing based zeroth-order gradient estimator in combination with the constraint extrapolation technique from Boob et al. (2022).

- **Biased gradients:** The use of the Gaussian smoothing based zeroth-order gradient estimator leads to stochastic gradients that are biased. Although Boob et al. (2022) consider noisy gradients, they assume their stochastic gradients are unbiased. This complicates the analysis of the zeroth-order setting we work with.

- **Non-uniform variance:** Apart from the unbiased stochastic gradient assumption, Boob et al. (2022) require the variance of their stochastic gradient to be uniformly bounded over the entire parameter space. However, the Gaussian smoothing based gradient estimator does not satisfy this assumption. A major technical part of our analysis involves dealing with stochastic gradients that are not uniformly bounded.

- **Smoothing parameters:** Our method requires dealing with the additional tuning parameters ($\nu_i$’s) that determine the level of smoothing in the zeroth-order gradient estimator. Dealing with this requires a careful analysis, as otherwise one would end up with worse oracle complexity than we have established in this work; see Remark 3.2 for details. In contrast, Boob et al. (2022) do not require dealing with any tuning parameters for their stochastic gradient, due to their generic set of assumptions.

- **Experiments:** Boob et al. (2022) do not provide any experimental verification of their algorithm. In contrast, in Section 4 that follows, we provide a detailed experimental evaluation, comparing to the existing state-of-the-art methods for constrained zeroth-order optimization, and demonstrate the advantages of the proposed approach.

4 Experimental Results

We compare the performance of our algorithm (Algorithm 1) with the following widely used algorithms for constrained zeroth-order optimization.

- **ALBO method** by Gramacy et al. (2016): This method takes a hybrid approach for constrained zeroth-order optimization, based on combining Bayesian optimization (i.e., Gaussian process based approaches) with Augmented Lagrangian methods. Specifically, the objective function of Augmented Lagrangian (which is similar in spirit to (5)) is estimated using Gaussian process priors. This method has various tuning parameters which makes the implementation a bit difficult. In fact, Gramacy et al. (2016) do not provide the full implementation details and mention that “many specifics have been omitted for space considerations”. We use the implementation provided in Gramacy (2016) as recommended by Gramacy et al. (2016).

- **Slack-AL method** by Picheny et al. (2016): This method builds upon the ALBO method and is also a hybrid method. Specifically, a particular step in estimating the objective function using Gaussian process technique (referred to as the Expected-Improvement step) is avoided by using slack variables. Similar to previous method, we use the implementation provided in Gramacy (2016).
• **ADMMBO** method by Ariafar et al. (2019): This method is also a hybrid method that uses Bayesian optimization methods. However, they use an ADMM-based approach to solve the augmented Lagrangian problem. We follow the recommendation in Section 5.1 of Ariafar et al. (2019) for the implementation.

• **PESC** method by Hernández-Lobato et al. (2015): This is a purely Bayesian optimization method that uses predictive entropy search for solving constrained zeroth-order optimization methods. As mentioned in Hernández-Lobato et al. (2015), “One disadvantage of PESC is that it is relatively difficult to implement”. Furthermore, all the implementation details are not provided in detail in Ariafar et al. (2019). Hence, we follow the implementation provided in Ariafar et al. (2019) for our experiments.

Compared to the above methods, our algorithm comes with a theoretical guarantee for setting the various tuning parameters of the proposed algorithm.

We first report simulation experiments on: (i) the oracle complexity of SZO-ConEX on 2 different test case objective and constraint functions, and (ii) the effect of the smoothing parameters (corresponding to the zeroth-order gradient estimation process) on the oracle complexity. For our experiments, we consider the following optimization problem (termed as Quadratically Constrained Quadratic Programing (QCQP) in the literature) where the objective function and the constraint function are quadratic functions:

$$\min_{x \in \mathbb{R}^n} f_0(x) := x^\top A_0 x + b_0^\top x + c_0$$

such that

$$f_1(x) := x^\top A_1 x + b_1^\top x + c_1 \leq 1.$$ 

Here, $A_0, A_1 \in \mathbb{R}^{n \times n}$, $b_0, b_1 \in \mathbb{R}^n$, and $c_0, c_1 \in \mathbb{R}$. When the matrices $A_0, A_1 \in \mathbb{R}^{n \times n}$ are further assumed to be symmetric and positive semidefinite, the above problem is a convex optimization problem with convex constraints. In the general case, nonconvex QCQPs form a rich class of optimization problems. For example, every polynomial optimization problem with polynomial constraint could be turned into a nonconvex QCQP at the expense of increasing the number of the optimization variables (d’Aspremont and Boyd 2003). Furthermore, it is also known that it is NP-hard to find global minimizers of nonconvex QCQP problem in the worst case.

**Convex setting:** We first consider the convex setting. Here, we set $A_0$ and $A_1$ to be random but fixed symmetric positive semidefinite matrices. Similarly $b_0, b_1, c_0$ and $c_1$ were generated randomly but fixed. Hence, the problem instance is fixed. In our experiments, we only use (noisy) function evaluations of both the objective and constraint functions. We used standard normal distribution and student $t$-distribution with degrees of freedom 5 for the noise in the function evaluations. For Algorithm 1, $\theta_t$ was set to 1 based on the theoretical result. Furthermore, $\tau$ and $\eta$, the parameters corresponding to the ascent step and the descent step were set based on trial and error to achieve the best performance. We remark that one could potentially use principled approaches like line-search for setting the step-size parameters (Berahas et al. 2019). As we are working in the zeroth-order setting, in our experimentation we provide additional attention to the smoothing parameters ($\nu_0$ and $\nu_1$) corresponding to the zeroth-order gradient estimators. We set them both to 0.05, 0.1 and 2 and report our performance.

In figure 1, we report the function value difference (corresponding to Theorem 3.1) versus number of calls to the (noisy) zeroth-order oracle, for various algorithms and our algorithm with the three choices of smoothing parameters. We work with dimensions $n = 200$ and $n = 500$ for our problem. Note here that it is easy to obtain the function value at the optimal solution for convex QCQP by using standard solvers (we use CVXPy to calculate it). The curves in figure 1 correspond to average over 100 trials. We notice that the performance of our algorithm is uniformly better than the compared algorithms in terms of number of function calls required to obtain a prescribe accuracy. Furthermore, we notice that our algorithm is robust to the choice of smoothing parameters: as long as it is small enough, we have fast convergence, but the iterates diverge when the smoothing parameter value is large.
Nonconvex setting: We now list the changes we make for the nonconvex setting. First, while the matrices are still random but fixed, we make them non-positive-definite. Furthermore, for Algorithm 2, we set $K = 50$. In figure 2 (bottom two rows), we report the norm of the gradient of the objective function (corresponding to Theorem 3.1) versus number of calls to the (noisy) zeroth-order oracle, for various algorithms and our algorithm with the three choices of smoothing parameters. The curves in figure 2 correspond to average over 100 trials. We notice that similar to the convex case, the performance of our algorithm is uniformly better than the compared algorithms in terms of number of function calls required to obtain a prescribe accuracy.

A brief summary of the observations are: (i) the oracle complexity of SZO-ConEX method is consistently lower than other existing techniques including ALBO Gramacy et al. (2016), Slack-AL Picheny et al. (2016), ADMMBO Ariafar et al. (2019), and PESC Hernández-Lobato et al. (2015), highlighting the benefit of constraint extrapolation step, and (ii) the SZO-ConEX method is robust to the smoothing parameters as long as it is less than a particular threshold. Next, we report the performance of our algorithm on the two motivating examples from Section 1.

4.1 Application I: Tuning HMC Algorithm

We now consider the problem of optimizing the hyperparameters of the HMC algorithm. A brief description of the HMC algorithm is provided in Section 6 for completeness. We follow Gelbart et al. (2014), Hernández-Lobato et al. (2015) closely for the experimental setup. The specific hyperparameters that we consider for this experiment are: (i) the number of leapfrog steps, denoted by $\tau$, (ii) step-size parameter, denoted by $\eta$. 
Figure 2: Performance comparison on simulation experiment: Plot of number of queries versus norm of the gradient. The plots represent average curves over 100 trials and the shaded region corresponds to standard error. In the legend the curves corresponding to \( \nu \) correspond to \( \text{SZO-ConEX} \) algorithm.

(iii) scalar coefficient of the mass matrix, denoted by \( \kappa \) (here, following [Neal 2011], we parametrize the mass matrix as \( \kappa \) times an identity matrix), and (iv) the fraction of the allotted time the algorithm spends in the burning phase. Hence, the optimization variables are given by \( x \in \mathbb{R}^4 \). We remark that while the number of leap-frog steps is an integer, for our experiments, we consider it to be real-valued number. In practice, we round it off to the closest integer, with ties broken randomly.

The objective function we maximize is the number of effective samples in a fixed computation time. This is a widely used diagnostic metric for measuring the performance of sampling algorithms in Bayesian statistical machine learning (Kass et al. 1998, Lenth 2001). For sampling problems, effective sample size is defined as follows. First note that the samples outputted by a sampling algorithm are typically correlated. The effective sample size is defined as the number of independent samples from the target density that achieves the same performance as the correlated samples outputted by the sampling algorithm. However, there is no closed-form analytical relationship between this performance measure and the optimization variable \( x \). For our experiments, we use the CODA package (Plummer et al. 2006) for calculating the effective sample size. The constraint functions that we use are: (i) the generated samples must pass the Geweke diagnostics (Geweke 1991); the worst Geweke test score across all variables and chains could be at most 2.0, (ii) the generated samples must pass the Gelman-Rubin convergence diagnostics (Gelman and Rubin 1992); the worst Gelman-Rubin score between variables and chains could be at most be 1.2. The analytical form of the above convergence diagnostics and the optimization variable \( x \) is also not available in closed-form. We use PyMC package (Patil et al. 2010) for evaluating the above diagnostic metrics.

We tune the HMC sampling algorithm with the above setup for the problem of sampling from the
Table 1: Effective Sample Size (ESS) of Hamiltonian Monte Carlo sampling algorithm tuned by various methods, along with their standard error.

| Algorithm      | ALBO   | Slack-AL | ADMMBO | PESC   | SZO-ConEx |
|----------------|--------|----------|--------|--------|-----------|
| ESS            | $9.4 \times 10^4 \pm 924$ | $9.3 \times 10^4 \pm 982$ | $9.4 \times 10^4 \pm 884$ | $9.9 \times 10^4 \pm 998$ | $10.8 \times 10^4 \pm 992$ |

Table 2: Validation Error (VE) along the standard error of 3-layer neural network trained using SGD with momentum for 5000 iterations on MNIST and CIFAR-10 datasets by picking hyperparameters tuned by various methods. The numbers reported are related to the constraint that the prediction time is not greater than 0.050 seconds on a Nvidia Tesla K20 GPU.

| Algorithm      | ALBO   | Slack-AL | ADMMBO | PESC   | SZO-ConEx |
|----------------|--------|----------|--------|--------|-----------|
| VE on MNIST    | $3.4 \pm 0.05$ | $3.1 \pm 0.08$ | $3.0 \pm 0.05$ | $2.9 \pm 0.03$ | $1.9 \pm 0.04$ |
| VE on CIFAR-10 | $4.7 \pm 0.02$ | $4.0 \pm 0.03$ | $3.9 \pm 0.05$ | $3.4 \pm 0.03$ | $2.2 \pm 0.02$ |

posterior distribution of a logistic regression binary classification problem on the German credit data set from UCI machine learning repository (Dua and Graff 2017). The data set contains 1000 observations that are normalized to have unit variance. We initialize each chain randomly with independent draws from a Gaussian distribution with mean zero and standard deviation $10^{-3}$. For each set of inputs, we compute two chains, each with 5 minutes of computation time. As mentioned previously, all our simulation settings are following that of Gelbart et al. (2014), Hernández-Lobato et al. (2015). We conduct our experiments by sub-sampling data sets of size 800 from the original dataset and repeating the procedure for 100 trials. We compare the performance of our algorithm (with $K = 50$) with that of ALBO method by Gramacy et al. (2016), Slack-AL method by Picheny et al. (2016), ADMMBO method by Ariafar et al. (2019), and PESC method by Hernández-Lobato et al. (2015). The tuning parameters of the respective methods were set according to the guidelines provided in the papers. For our algorithm, we found the performance was robust to the choice of the smoothing parameters, as long as it was sufficiently small. For the performance reported in Table 1, we set it to $\nu_i = 0.05$. In Table 1, we report the average Effective Sample Size (ESS) for the various methods, along with the standard deviation. We notice that the performance of SZO-ConEx is significantly better than that of the other methods, thereby demonstrating the effectiveness of our method for the problem of hyperparameter tuning for HMC sampling algorithm.

### 4.2 Application II: Tuning a 3-Layer Neural Network

Next, we turn to the problem of tuning the hyperparameters of a 3-layer neural network with ReLU activation function trained by stochastic gradient descent algorithm with momentum (Sutskever et al. 2013) for 5000 iterations. We follow Hernández-Lobato et al. (2015), Ariafar et al. (2019) closely for the experimental setup. The specific hyperparameters that we consider for this experiment are: (i) two learning rate parameters (initial and decay rate), (ii) momentum parameters (initial and final), (iii) dropout parameters (input layer and hidden layers), (iv) regularization parameters corresponding to weight decay and max weight norm, and (v) the number of hidden units in each of the 3 hidden layers. Hence, the optimization variables are given by $x \in \mathbb{R}^{11}$. Similar to the previous experiment, we treat the number of hidden layers as a real-valued variable and use the same rounding technique in practice.

The objective function we minimize is the classification error on the validation set (which we call Validation Error (VE)). Indeed, there is no good closed form expression connecting the above mentioned hyperparameters and the VE. The constraint function that we use is that the prediction time must not exceed 0.050 seconds. Hence, we compute the prediction time as the average time of 1000 predictions, over a batch of size 128 (Hernández-Lobato et al. 2015, Ariafar et al. 2019). The number 0.050 seconds is set based on
the computing resource we use (Nvidia Tesla K20 GPU) so that we can see an active trade off between the objective function (the VE) and the constraint function (prediction time). As highlighted by Hernández-Lobato et al. (2015), Ariafar et al. (2019), this specific choice is highly dependent on the computing resource used. Clearly, there is no analytical form for the function describing the relationship between the hyperparameters and the constraint function. All our implementations for this experiment were based on PyTorch open source machine learning library (Paszke et al. 2019).

We tune the SGD algorithm with momentum with the above setup for the problem of classification on MNIST (LeCun and Cortes 2010) and CIFAR-10 datasets (Krizhevsky 2009). For both datasets, we conduct our experiments by sub-sampling 90% of the training data and report our error over 100 trials. Similar to the previous case, we compare the performance of our algorithm (with $K = 50$) with that of ALBO method by Gramacy et al. (2016), Slack–AL method by Picheny et al. (2016), ADMMBO method by Ariafar et al. (2019), and PESC method by Hernández-Lobato et al. (2015). The tuning parameters of the respective methods were set as suggested in the respective papers. The smoothing parameter for our algorithm was set as $\nu = 0.03$. In Table 2, we report the validation error achieved such that the constraint on the prediction time is respected for the various algorithms. From the results, we notice that the SZO-ConEX method outperforms the other methods on both the MNIST and CIFAR-10 datasets.

5 Conclusion

In this paper, we proposed and analyzed stochastic zeroth-order optimization algorithms for nonlinear optimization problems with functional constraints. We consider the case when both the objective function and the constraint functions are observed only via noisy function queries. Our algorithm is based on leveraging the constraint extrapolation technique proposed by Boob et al. (2022) and the Gaussian smoothing technique. We characterize the oracle complexity of the proposed algorithm in both the convex and nonconvex setting. We also apply our methodology to the problem of hyperparameter tuning for the HMC algorithm and 3-Layer neural networks trained using SGD with momentum, and demonstrate its superior performance.

For future work, we plan to develop parallel versions of our algorithm for the case when the objective functions and the constraint functions are available only locally in different machines. We also plan to develop lower bounds on the oracle complexity of stochastic zeroth-order optimization algorithms in the constrained setting. It is of great interest to find other applications of the proposed methodology in statistical machine learning, reinforcement learning, and other scientific and engineering fields. Finally, it is also interesting to extend our methodology to the case of mixed constraints (i.e., equality and inequality constraint), and to develop novel methodology and analysis for constrained zeroth-order optimization with both binary and real-valued decision variables.

References

Luigi Acerbi and Wei Ji Ma. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. In Advances in neural information processing systems, pages 1836–1846, 2017.

Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In Conference on Learning Theory, pages 28–40, 2010.

Nadir Amaioua, Charles Audet, Andrew R Conn, and Sébastien Le Digabel. Efficient solution of quadratically constrained quadratic subproblems within the mesh adaptive direct search algorithm. European Journal of Operational Research, 268(1):13–24, 2018.

Francesco Archetti and Antonio Candelieri. Bayesian optimization and data science. Springer, 2019.

Setareh Ariafar, Jaume Coll-Font, Dana H Brooks, and Jennifer G Dy. ADMMBO: Bayesian Optimization with Unknown Constraints using ADMM. Journal of Machine Learning Research, 20(123):1–26, 2019.
Charles Audet and John E Dennis Jr. A pattern search filter method for nonlinear programming without derivatives. *SIAM Journal on Optimization*, 14(4):980–1010, 2004.

Charles Audet and John E Dennis Jr. Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on Optimization*, 17(1):188–217, 2006.

Charles Audet and John E Dennis Jr. A progressive barrier for derivative-free nonlinear programming. *SIAM Journal on Optimization*, 20(1):445–472, 2009.

Charles Audet and Warren Hare. Derivative-free and blackbox optimization. 2017.

Charles Audet and Christophe Tribes. Mesh-based Nelder–Mead algorithm for inequality constrained optimization. *Computational Optimization and Applications*, 71(2):331–352, 2018.

Charles Audet, Sébastien Le Digabel, and Mathilde Peyrega. Linear equalities in blackbox optimization. *Computational Optimization and Applications*, 61(1):1–23, 2015.

F Augustin and YM Marzouk. NOWPAC: A provably convergent derivative-free nonlinear optimizer with path-augmented constraints. *arXiv preprint arXiv:1403.1931*, 2014.

François Bachoc, Céline Helbert, and Victor Picheny. Gaussian process optimization with failures: Classification and convergence proof. *Journal of Global Optimization*, 78(3):483–506, 2020.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3459–3468, 2018.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order Nonconvex Stochastic Optimization: Handling Constraints, High-Dimensionality and Saddle-Points. *Foundations of Computational Mathematics*, 2022.

Krishnakumar Balasubramanian, Saeed Ghadimi, and Anthony Nguyen. Stochastic multilevel composition optimization algorithms with level-independent convergence rates. *SIAM Journal on Optimization*, 32(2):519–544, 2022.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

Albert S Berahas, Liyuan Cao, and Katya Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *arXiv preprint arXiv:1910.04055*, 2019.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Julius R Blum. Multidimensional stochastic approximation methods. *The Annals of Mathematical Statistics*, pages 737–744, 1954.

Digvijay Boob, Qi Deng, and Guanghui Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, pages 1–65, 2022.

Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

Luis Felipe Bueno, Ana Friedlander, José Mario Martinez, and FNC Sobral. Inexact restoration method for derivative-free optimization with smooth constraints. *SIAM Journal on Optimization*, 23(2):1189–1213, 2013.

Árpád Bürmen, Janez Puhan, and Tadej Tuma. Grid restrained Nelder-Mead algorithm. *Computational optimization and applications*, 34(3):359–375, 2006.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. STAN: A Probabilistic Programming Language. *Journal of Statistical Software*, 76, 2017.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.
Krzysztof Choromanski, Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, Deepali Jain, Yuxiang Yang, Atil Iscen, Jasmine Hsu, and Vikas Sindhwani. Provably robust blackbox optimization for reinforcement learning. In Conference on Robot Learning, pages 683–696. PMLR, 2020.

Andrew Conn, Katya Scheinberg, and Luis Vicente. Introduction to Derivative-Free Optimization, volume 8. SIAM, 2009.

Andrew R Conn and Sébastien Le Digabel. Use of quadratic models with mesh-adaptive direct search for constrained black box optimization. Optimization Methods and Software, 28(1):139–158, 2013.

Alexandre d’Aspremont and Stephen Boyd. Relaxations and randomized methods for nonconvex qcpqs. EE392o Class Notes, Stanford University, 1:1–16, 2003.

Sébastien Le Digabel and Stefan M Wild. A taxonomy of constraints in simulation-based optimization. arXiv preprint arXiv:1505.07881, 2015.

Jürgen Dippon. Accelerated randomized stochastic optimization. The Annals of Statistics, 31(4):1260–1281, 2003.

Dmitriy Drusvyatskiy. The proximal point method revisited. arXiv preprint arXiv:1712.06038, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. Physics letters B, 195(2):216–222, 1987.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. IEEE Transactions on Information Theory, 61(5):2788–2806, 2015.

Kwassi Joseph Dzahini, Michael Kokkolaras, and Sébastien Le Digabel. Constrained stochastic blackbox optimization using a progressive barrier and probabilistic estimates. arXiv preprint arXiv:2011.04225, 2020.

N Echebest, María Laura Schuverdt, and Raúl Pedro Vignau. An inexact restoration derivative-free filter method for nonlinear programming. Computational and Applied Mathematics, 36(1):693–718, 2017.

Thomas Elsken, Jan Hendrik Metzen, Frank Hutter, et al. Neural architecture search: A survey. J. Mach. Learn. Res., 20(55):1–21, 2019.

David Eriksson and Matthias Poloczek. Scalable Constrained Bayesian Optimization. arXiv preprint arXiv:2002.08526, 2020.

Giovanni Fasano, Giampaolo Liuzzi, Stefano Lucidi, and Francesco Rinaldi. A line-search based derivative-free approach for nonsmooth constrained optimization. SIAM journal on optimization, 24(3):959–992, 2014.

Peter I Frazier. A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811, 2018.

Wenbo Gao, Laura Graesser, Krzysztof Choromanski, Xingyou Song, Nevena Lazić, Pannag Sanketi, Vikas Sindhwani, and Navdeep Jaitly. Robotic table tennis with model-free reinforcement learning. arXiv preprint arXiv:2003.14398, 2020.

Jacob Gardner, Matt Kusner, Kilian Weinberger, and John Cunningham. Bayesian optimization with inequality constraints. In International Conference on Machine Learning, pages 937–945, 2014.

Michael A Gelbart, Jasper Snoek, and Ryan P Adams. Bayesian optimization with unknown constraints. In 30th Conference on Uncertainty in Artificial Intelligence, UAI 2014, pages 250–259, 2014.

Michael A Gelbart, Ryan P Adams, Matthew W Hoffman, and Zoubin Ghahramani. A general framework for constrained bayesian optimization using information-based search. Journal of Machine Learning Research, 17 (160):1–53, 2016.

Andrew Gelman and Donald B Rubin. A single series from the Gibbs sampler provide a false sense of security. Bayesian Statistics, 4, 1992.

John Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, volume 196. 1991.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 23(4):2341–2368, 2013.

Saeed Ghadimi, Andrzejs Ruszcynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. SIAM Journal on Optimization, 30(1):960–979, 2020.
Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D Sculley. Google vizier: A service for black-box optimization. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1487–1495, 2017.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Robert B Gramacy. lagp: large-scale spatial modeling via local approximate gaussian processes in r. *Journal of Statistical Software*, 72:1–46, 2016.

Robert B Gramacy, Genetha A Gray, Sébastien Le Digabel, Herbert Lee, Pritam Ranjan, Garth Wells, and Stefan M Wild. Modeling an augmented Lagrangian for blackbox constrained optimization. *Technometrics*, 58(1):1–11, 2016.

Serge Gratton and Luís Nunes Vicente. A merit function approach for direct search. *Siam journal on optimization*, 24(4):1980–1998, 2014.

Stewart Greenhill, Santu Rana, Sunil Gupta, Pratibha Vellanki, and Svetla Venkatesh. Bayesian optimization for adaptive experimental design: A review. *IEEE Access*, 8:13937–13948, 2020.

Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

Elad Hazan, Adam Klivans, and Yang Yuan. Hyperparameter optimization: A spectral approach. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

José Miguel Hernández-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive entropy search for Bayesian optimization with unknown constraints. In *International conference on machine learning*, pages 1699–1707. PMLR, 2015.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Robert Hooke and Terry A Jeeves. Direct search solution of numerical and statistical problems. *Journal of the ACM (JACM)*, 8(2):212–229, 1961.

Kevin G Jamieson, Robert D Nowak, and Benjamin Recht. Query complexity of derivative-free optimization. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pages 2672–2680, 2012.

Noëmie Jaquier and Leonel Rozo. High-Dimensional Bayesian Optimization via Nested Riemannian Manifolds. *Advances in Neural Information Processing Systems*, 33, 2020.

Noëmie Jaquier, Leonel Rozo, Sylvain Calinon, and Mathias Bürger. Bayesian optimization meets Riemannian manifolds in robot learning. In *Conference on Robot Learning*, pages 233–246. PMLR, 2020.

Robert E Kass, Bradley P Carlin, Andrew Gelman, and Radford M Neal. Markov chain monte carlo in practice: A roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.

Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

Tamara G Kolda, Robert Michael Lewis, and Virginia Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM review*, 45(3):385–482, 2003.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master’s thesis, University of Toronto*, 2009.

Remi Lam and Karen Willcox. Lookahead Bayesian optimization with inequality constraints. In *Advances in Neural Information Processing Systems*, pages 1890–1900, 2017.

Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

Benoît Latré, Bart Braem, Ingrid Moerman, Chris Blondia, and Piet Demeester. A survey on wireless body area networks. *Wireless networks*, 17(1):1–18, 2011.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/
Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, volume 39. Springer, 2015.

Russell V Lenth. Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193, 2001.

Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2):495–519, 2019.

Robert Michael Lewis and Virginia Torczon. A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Optimization*, 12(4):1075–1089, 2002.

Jiaxiang Li, Krishnakumar Balasubramanian, and Shiqian Ma. Stochastic zeroth-order Riemannian derivative estimation and optimization. *arXiv preprint arXiv:2003.11238*, 2020.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

Giampaolo Liuzzi and Stefano Lucidi. A derivative-free algorithm for inequality constrained nonlinear programming via smoothing of an \( \ell_\infty \) penalty function. *SIAM Journal on Optimization*, 20(1):1–29, 2009.

Giampaolo Liuzzi, Stefano Lucidi, and Marco Sciandrone. Sequential penalty derivative-free methods for nonlinear constrained optimization. *SIAM Journal on Optimization*, 20(5):2614–2635, 2010.

Nimalan Mahendran, Ziyu Wang, Firas Hamze, and Nando De Freitas. Adaptive MCMC with Bayesian optimization. In *Artificial Intelligence and Statistics*, pages 751–760. PMLR, 2012.

Horia Mania, Aurélia Guy, and Benjamin Recht. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

Jonas Mockus. Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4(4):347–365, 1994.

Jonas Mockus. *Bayesian approach to global optimization: Theory and applications*, volume 37. Springer Science & Business Media, 2012.

Abdelkader Mokkadem and Mariane Pelletier. A companion for the Kiefer–Wolfowitz–Blum stochastic approximation algorithm. *The Annals of Statistics*, 35(4):1749–1772, 2007.

Juliane Müller and Joshua D Woodbury. GOSAC: global optimization with surrogate approximation of constraints. *Journal of Global Optimization*, 69(1):117–136, 2017.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

Arkadij Semenovič Nemirovski and David Borisovich Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience*, 1983.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

Anand Patil, David Huard, and Christopher J Fonnesbeck. PyMC: Bayesian stochastic modelling in Python. *Journal of statistical software*, 35(4):1, 2010.

Charith Perera, Chi Harold Liu, and Srimal Jayawardena. The emerging internet of things marketplace from an industrial perspective: A survey. *IEEE Transactions on Emerging Topics in Computing*, 3(4):585–598, 2015.
Victor Picheny, Robert B Gramacy, Stefan Wild, and Sebastien Le Digabel. Bayesian optimization under mixed constraints with a slack-variable augmented Lagrangian. In *Advances in neural information processing systems*, pages 1435–1443, 2016.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: convergence diagnosis and output analysis for MCMC. *R news*, 6(1):7–11, 2006.

Tony Pourmohamad and Herbert Lee. The statistical filter approach to constrained optimization. *Technometrics*, 62(3):303–312, 2020.

Michael JD Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

Yangjun Ruan, Yuanhao Xiong, Sashank Reddi, Sanjiv Kumar, and Cho-Jui Hsieh. Learning to learn by zeroth-order oracle. *arXiv preprint arXiv:1910.09464*, 2019.

Andrzej Ruszczyński. A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization. *SIAM Journal on Control and Optimization*, 59(3):2301–2320, 2021.

Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. Towards gradient free and projection free stochastic optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3468–3477. PMLR, 2019.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24. PMLR, 2013.

Jasper Snoek, Hugo Larochelle, and Ryan Prescott Adams. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 2012.

James C Spall. A stochastic approximation technique for generating maximum likelihood parameter estimates. In *1987 American control conference*, pages 1161–1167. IEEE, 1987.

James C Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, volume 65. John Wiley & Sons, 2005.

WGRFR Spendley, George R Hext, and Francis R Himsworth. Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics*, 4(4):441–461, 1962.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

Anke Tröltzsch. A sequential quadratic programming algorithm for equality-constrained optimization without derivatives. *Optimization Letters*, 10(2):383–399, 2016.

Ilıura Usmanova, Andreas Krause, and Maryam Kamgarpour. Safe convex learning under uncertain constraints. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2106–2114. PMLR, 2019.

Ziyu Wang, Shakir Mohamed, and Nando Freitas. Adaptive Hamiltonian and Riemann Manifold Monte Carlo. In *International conference on machine learning*, pages 1462–1470. PMLR, 2013.

Allen Y Yang, Sameer Iyengar, Shankar Sastry, Ruzena Bajcsy, Philip Kuryloski, and Roozbeh Jafari. Distributed segmentation and classification of human actions using a wearable motion sensor network. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.
6 Basics of Hamiltonian Monte Carlo sampling

For the sake of completeness, we give a brief description of the Hamiltonian Monte Carlo sampling algorithm used in Section 4.1. The presentation below follows Neal (2011) for the most part. Suppose the problem is to sample from the distribution \( \pi(q) : \mathbb{R}^d \to \mathbb{R} \) whose potential function is given by \( f(q) : \mathbb{R}^d \to \mathbb{R} \). First consider the Hamiltonian form, given by

\[
H(q,p) = f(q) + K(p) = f(q) + p^\top M^{-1}p,
\]

where \( M \in \mathbb{R}^{d \times d} \) is the ‘mass matrix’. Following Neal (2011), we assume a diagonal parametrization for \( M \), i.e., we have \( M = \kappa I \). The Hamiltonian dynamics of the position vector \( q \) and the momentum vector \( p \) is determined by the equation given by

\[
\frac{dz}{dt} = J \nabla H(z), \quad \text{where} \quad J = \begin{pmatrix} 0_{d \times d} & I_{d \times d} \\ -I_{d \times d} & 0_{d \times d} \end{pmatrix}
\]

and \( z := (q, p) \in \mathbb{R}^{2d} \) and \( \nabla H \) is the gradient of the Hamiltonian function in (13). The HMC sampling algorithm is based on performing \( \tau \) leapfrog steps for discretizing the above equation. Here, a leapfrog (or symplectic integrator) step, for a given step-size \( \eta \), is given by

\[
p_{n+1/2} = p_n - \frac{\eta}{2} \frac{dH}{dq}(q_n)
\]

\[
q_{n+1} = q_n + \frac{\eta}{\kappa} p_{n+1/2}
\]

\[
p_{n+1} = p_{n+1/2} - \frac{\eta}{2} \frac{dH}{dq}(q_{n+1}),
\]

where \( n \) is the index of the number of steps. More details regarding HMC could also be found in Betancourt (2017).

7 Proofs for Section 2

We start with the following well-known result on the stochastic zeroth-order gradient estimator in (2).

**Theorem 7.1 (Nesterov and Spokoiny (2017)).** For a Gaussian random vector \( u \sim N(0, I_n) \) we have

\[
\mathbb{E}[\|u\|^k] \leq (n + k)^{k/2}
\]

for any \( k \geq 2 \). Moreover, the following statements hold for any function \( \psi \) whose gradient is Lipschitz continuous with constant \( L \)

- The gradient of \( \psi_{\nu}(x) := \mathbb{E}_u[\psi(x + \nu u)] \) is Lipschitz continuous with constant \( L_{\psi_{\nu}} \) such that \( L_{\psi_{\nu}} \leq L \).
- For any \( x \in \mathbb{R}^n \), we have

\[
|\psi_{\nu}(x) - \psi(x)| \leq \frac{\nu^2}{2} L n,
\]

\[
\|\nabla \psi_{\nu}(x) - \nabla \psi(x)\| \leq \frac{\nu}{2} L (n + 3)^{3/2}.
\]
For any \( x \in \mathbb{R}^n \), we have

\[
\frac{1}{\nu_2^2} \mathbb{E}_u \{ |\psi(x + \nu u) - \psi(x)|^2 \} \leq \frac{\nu_2^2}{2} L^2(n + 6)^3 + 2(n + 4)\|\nabla \psi(x)\|^2.
\] (18)

**Proof of Lemma 2.1** Note that

\[
\|F_{i,\nu}(x, \xi, u) - f_{i,\nu}(x)\|^2 = \sum_{\nu \in I} (f_{i,\nu}(x) - F_{i}(x + \nu u, \xi))^2.
\]

By Young’s inequality, we have

\[
|F_{i}(x + \nu u, \xi) - f_{i,\nu}(x)|^2 = |F_{i}(x + \nu u, \xi) - F_{i}(x, \xi)| + |F_{i}(x, \xi) - f_{i}(x)| + |f_{i}(x) - f_{i,\nu}(x)|^2
\]

\[
\leq 4|F_{i}(x + \nu u, \xi) - F_{i}(x, \xi)|^2 + 4|f_{i}(x) - f_{i,\nu}(x)|^2 + 2|F_{i}(x, \xi) - f_{i}(x)|^2
\]

\[
\leq 4M_i^2 \nu_2^2 \|u\|^2 + 4 \left( \frac{\nu_2^2}{2} L_i n \right)^2 + 2|F_{i}(x, \xi) - f_{i}(x)|^2.
\]

Now, by Assumption 2.3 and Theorem 7.1 we have

\[
\mathbb{E}|f_{i,\nu}(x) - F_{i}(x + \nu u, \xi)|^2 \leq 4M_i^2 \nu_2^2 (n + 2) + 2\sigma_{f,i}^2 + L_i^2 \nu_4^4 n^2.
\]

Consequently, we obtain

\[
\mathbb{E}\|F_{i}(x, \xi, u) - f_{i,\nu}(x)\|^2 \leq (\sum_{\nu \in I} 4M_i^2 \nu_2^2 (n + 2) + L_i^2 \nu_4^4 n^2) + 2\sigma_{f,i}^2 =: \sigma_{f,\nu}^2.
\]

**Proof of Lemma 2.2** First note that by Theorem 7.1 we have

\[
\frac{1}{\nu_2^2} \mathbb{E}_u \{ |F_{i}(x + \nu u, \xi) - F_{i}(x, \xi)|^2 \} \|u\|^2
\]

\[
\leq \frac{\nu_2^2}{2} L_i^2 (n + 6)^3 + 2(n + 4)\|\nabla F_{i}(x, \xi)\|^2
\]

\[
\leq \frac{\nu_2^2}{2} L_i^2 (n + 6)^3 + 4(n + 4)|\nabla F_{i}(x, \xi) - \nabla f_{i}(x)|^2 + |\nabla f_{i}(x)|^2.
\] (19)

Next note that

\[
\|\nabla f_{i,\nu}(x)\| \leq \|\nabla f_{i,\nu}(x) - \nabla f_{i}(x)\| + \|\nabla f_{i}(x)\|
\]

\[
\leq \frac{\nu_2}{2} L_i (n + 3)^{3/2} + L_i D_X + \|\nabla f_{i}(x^*)\|
\]

\[
\leq \frac{\nu_2}{2} L_i (n + 3)^{3/2} + L_i D_X + M_i =: \tilde{B}_i,
\]

where \( M_i \) is from Assumption 2.3. Taking the expectation with respect to \( \xi \) on both sides of (19), we have

\[
\mathbb{E}[|G_{i,\nu}(x, \xi, u)|^2] \leq \frac{\nu_2^2}{2} L_i^2 (n + 6)^3 + 4(n + 4)[\sigma_{f,i}^2 + \tilde{B}_i^2].
\]

From the above inequalities, using Assumptions 2.2 and 2.3, Theorem 7.1 and Young’s inequality, we have

\[
\mathbb{E}[|G_{i,\nu}(x, \xi, u) - \nabla f_{i,\nu}(x)|^2] \leq 2\mathbb{E}[|G_{i,\nu}(x, \xi, u)|^2] + 2|\nabla f_{i,\nu}(x)|^2
\]

\[
\leq \nu_2^2 L_i^2 (n + 6)^3 + 8(n + 4)[\sigma_{i}^2 + \tilde{B}_i^2] + 2\tilde{B}_i^2
\]

\[
\leq \nu_2^2 L_i^2 (n + 6)^3 + 10(n + 4)[\sigma_{i}^2 + \tilde{B}_i^2],
\]

which completes the proof.

\[\square\]
8 Proofs for Section 3

In the proofs below, to avoid notational clutter, we use $x_t$ instead of using $x^{(t)}$, and we use $G_{i,\nu}(x_t, \xi_t, u_t)$ instead of $G_{i,\nu}(x^{(t)}, \xi^{(t)}, u^{(t)}_i)$. Next, in order to obtain the oracle complexity of Algorithm 1 we define a primal-dual gap function for the equivalent saddle point problem \(5\). In particular, given a pair of feasible solution $z = (x, y)$ and $\bar{z} = (\bar{x}, \bar{y})$ of \(5\), we define the primal-dual gap function $Q(z, \bar{z})$ as

$$Q(z, \bar{z}) := \mathcal{L}(x, \bar{y}) - \mathcal{L}(\bar{x}, y).$$

(20)

For the remainder of the paper, we denote $Q_\nu(z, \bar{z}) = \mathcal{L}_\nu(x, \bar{y}) - \mathcal{L}_\nu(\bar{x}, y)$. Now we establish the error between these two functions.

**Lemma 8.1.** Under Assumptions 2.1, 2.2 and 2.3 we have

$$|Q(z, \bar{z}) - Q_\nu(z, \bar{z})| \leq \nu_0^2 L_0 n + M_X n (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2},$$

(21)

where $M_X = \sup_{x \in X} \|x\|$.

**Proof of Lemma 8.1** First, we claim that the following is true:

$$\|f(x) - f_\nu(x)\| = \frac{n}{2} (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}. \quad (22)$$

To see that, note that since the components $f_i$ of $f$ have continuous Lipschitz gradient and using theorem 7.1 we have

$$\|f(x) - f_\nu(x)\| = \left(\sum_{i=1}^m (f_i(x) - f_i(\nu))(x))^2\right)^{1/2}$$

\[= \left(\sum_{i=1}^m \frac{\nu_i^2 L_i n^2}{2}\right)^{1/2} \]

\[= \left(\sum_{i=1}^m \frac{\nu_i^4 L_i^2 n^2}{4}\right)^{1/2} \]

\[= \frac{n}{2} (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2} \]

Utilizing this relation, using Theorem 7.1 and Cauchy-Schwartz inequality, we have

$$|Q(z, \bar{z}) - Q_\nu(z, \bar{z})| = |\mathcal{L}(x, \bar{y}) - \mathcal{L}(\bar{x}, y) - \mathcal{L}_\nu(x, \bar{y}) + \mathcal{L}_\nu(\bar{x}, y)|$$

\[= |f_0(x) + \bar{y}^T f(x) - f_0(\bar{x}) - y^T f(\bar{x}) - f_0,\nu_0(x) - \bar{y}^T f_\nu(x) + f_0,\nu_0(\bar{x}) + y^T f_\nu(\bar{x})| \leq |f_0(x) - f_0,\nu_0(x)| + |f_0(\bar{x}) - f_0,\nu_0(\bar{x})| + |\bar{y}^T f(x) - f_\nu(x)| + |y^T f(\bar{x}) - f_\nu(\bar{x})| \leq |f_0(x) - f_0,\nu_0(x)| + |f_0(\bar{x}) - f_0,\nu_0(\bar{x})| + \|\bar{y}\| ||f(x) - f_\nu(x)|| + \|y\| ||f(\bar{x}) - f_\nu(\bar{x})|| \leq |f_0(x) - f_0,\nu_0(x)| + |f_0(\bar{x}) - f_0,\nu_0(\bar{x})| + M_X ||f(x) - f_\nu(x)|| + ||f(\bar{x}) - f_\nu(\bar{x})|| \leq \nu_0^2 L_0 n + M_X n (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2} \]

This concludes the proof.
Lemma 8.2. Suppose Assumptions \[2.1\], \[2.2\], and \[2.3\] are satisfied. Then, for all \( T \geq 1 \), we have

\[
\mathbb{E}[f_0(\bar{x}_T) - f_0(x^*)] \leq \frac{1}{\Gamma_T} \left[ \gamma T \theta_0 W(x^*, x_0) \right] + \sum_{t=0}^{T-1} \frac{2\gamma T \sigma_t}{T - t} \mathbb{E}[\delta_t^G]^2 \right] \right] + \left[ \sigma_t^G \mathbb{E} \left[ \frac{\bar{n}}{T} \right] + M_X n (\sum_{i=1}^{n} \nu_i L_i^2)^{1/2} \right]
\]

(23)

\[
\mathbb{E}[\|f(\bar{x}_T)\|_2] \leq \frac{1}{\Gamma_T} \left[ \gamma T \theta_0 \|y_0\|_2^2 + 3 \|y_2\|_2 + 1 \right] ^2 \gamma T \theta_0 W(x^*, x_0) \]

(24)

where \( \Gamma_T := \sum_{t=0}^{T-1} \gamma_t \) and \( \nu_{1, \nu_{i}, \ldots, \nu_{2m+1}, \nu_{m+1}} \) with \( \sigma_{1, \nu_{i}} \) as defined in \[\text{(4)}\], and \( \delta_t^G := G_{0, \nu_{0}}(x_t, \xi_t, u_t) - f_{0, \nu_{0}}(x_t) + \sum_{i=1}^{m} y_{t+1}^{(i)} \left( G_{i, \nu_{i}}(x_t, \xi_t, u_t) - f_{i, \nu_{i}}^{(i)}(x_t) \right) \).

Proof of Lemma 8.2. First, observe that \( y_{t+1} \) is a constant conditioned on random variable \( \xi_{[t-1]}, \bar{u}_{[t-1]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]} \).

In particular,

\[
\mathbb{E}[\langle \delta_t^G, x_t - x \rangle] = \mathbb{E}(\mathbb{E}[\xi_{[t-1]}], u_{[t-1]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}[\delta_t^G], x_t - x) = 0
\]

(25)

for any non-random \( x \). This follows due to the following relation

\[
\mathbb{E}[\xi_{[t-1]}], u_{[t-1]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}[\delta_t^G]
\]

\[
= \mathbb{E}[\xi_{[t-1]}], u_{[t-1]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}[G_{0, \nu_{0}}(x_t, \xi_t, u_t) - f_{0, \nu_{0}}(x_t)]
\]

\[
+ \sum_{i=1}^{m} y_{t+1} \mathbb{E}[\xi_{[t-1]}, u_{[t-1]}], \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}[G_{i, \nu_{i}}(x_t, \xi_t, u_t) - f_{i, \nu_{i}}^{(i)}(x_t)]
\]

\[
= 0.
\]

Similarly, we have

\[
\mathbb{E}[\langle \delta_{t+1}^F, y_{t+1} - y \rangle] = \mathbb{E}(\mathbb{E}[\xi_{[t-1]}], u_{[t-1]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}[\delta_{t+1}^F], y_{t+1} - y) = 0,
\]

(26)

for any non-random \( y \). Here, we note that

\[
\mathbb{E}[\xi_{[t]}, u_{[t]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}[\delta_{t+1}^F] = \mathbb{E}[\xi_{[t]}, u_{[t]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}[F_{\nu}(x_t, \xi_t, u_t)] - f_{\nu}(x_t)
\]

\[
+ \mathbb{E}[\xi_{[t]}, u_{[t]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}[G_{\nu}(x_t, \xi_t, u_t)] - f_{\nu}'(x_t)]
\]

\[
= 0,
\]

(27)

where the first term in RHS is \( 0 \) due to \( \mathbb{E}_{\xi, u} F_{\nu}(x, \xi, u) = f_{\nu}(x) \), the second term is \( 0 \) due to the \( \mathbb{E}_{\xi, u} G_{\nu}(x, \xi, u) = f_{\nu}'(x) \) and the common fact for both the terms that \( x_t, x_{t+1} \) are constants for given \( \xi_{[t]}, u_{[t]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]} \). We now note that

\[
\mathbb{E}[\| \delta_t^F \|^2] \leq 2 \mathbb{E}[\| F_{\nu}(x_{t-1}, \bar{\xi}_{t-1}, \bar{u}_{t-1}) - f_{\nu}(x_{t-1}) \|^2] + 2 \mathbb{E}[\| G_{\nu}(x_{t-1}, \bar{\xi}_{t-1}, \bar{u}_{t-1}) - f_{\nu}'(x_{t-1}) \|^2]
\]

\[
\leq 2 \sigma_{\nu}^2 + 2 \mathbb{E}\left[ \sum_{i=1}^{m} \{(G_{i, \nu_{i}}(x_{t-1}, \bar{\xi}_{t-1}, \bar{u}_{t-1}) - f_{i, \nu_{i}}^{(i)}(x_{t-1}))^T (x_t - x_{t-1}) \}^2 \right]
\]

\[
\leq 2 \sigma_{\nu}^2 + 2 \mathbb{E}\left[ \sum_{i=1}^{m} \| G_{i, \nu_{i}}(x_{t-1}, \bar{\xi}_{t-1}, \bar{u}_{t-1}) - f_{i, \nu_{i}}^{(i)}(x_{t-1}) \|^2 \| x_t - x_{t-1} \|^2 \right]
\]

\[
\leq 2 \sigma_{\nu}^2 + 2 D_{\nu}^2 \| \sigma_{\nu} \|^2.
\]

(28)
Then, in view of above relation and definitions of $q_t, \tilde{q}_t$, and by defining $\delta^F_t := \ell_F(x_t) - \ell_f(x_t)$, we have
\begin{equation}
\mathbb{E}[\|q_t - \tilde{q}_t\|^2_2] = \mathbb{E}[\|\ell_F(x_t) - \ell_f(x_t) - \ell_F(x_{t-1}) + \ell_f(x_{t-1})\|^2_2] \\
\quad \leq 2\mathbb{E}[\|\delta^F_t\|^2_2] + 2\mathbb{E}[\|\tilde{\delta}^F_{t-1}\|^2_2] \leq 8(\sigma^2_{f,\nu} + D^2_X \|\sigma_{\nu}\|^2_2).
\end{equation}
(29)

Taking the expectation on both sides of (68) and using relation (25), (26) and (29), we have for all non-random $z \in \{(x, y) : x \in X, y \geq 0\},$
\begin{equation}
\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q_{\nu}(z_{t+1}, z)\right] \\
\quad \leq \gamma_0 \eta_0 W(x, x_0) - \gamma_{T-1} \eta_{T-1} \mathbb{E}[W(x, x_T)] + \frac{7\gamma_0 T_0}{2} \|y - y_0\|^2_2 \\
\quad + \sum_{t=0}^{T-1} \frac{2\gamma_t}{\eta_t - L_0 - L_f} \left[\mathbb{E}[\|\delta^G\|^2_2] + \left(\frac{L_f D_X}{2} \|y\|_2 - 1\right)^2\right] \\
\quad + \left(\frac{T-1 \gamma_{t-1}}{\eta_{t-1}} + \frac{T \gamma_{T-1}}{\eta_{T-1}}\right) (\sigma^2_{f,\nu} + D^2_X \|\sigma_{\nu}\|^2_2)
\end{equation}
(30)
where we dropped $\|y - y_T\|^2_2$. By Lemma 8.1 we have
$$Q(z_{t+1}, z) - [\nu^2_0 L_0 n + M_X n (\sum_{i=1}^m \nu^i L^2_i)^{1/2}] \leq Q_{\nu}(z_{t+1}, z).$$

Using this relation, multiplying both sides by $\gamma_t$, summing from $t = 0, \ldots, T - 1$, and taking expectation on both sides, we have
\begin{equation}
\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q(z_{t+1}, z)\right] - [\nu^2_0 L_0 n + M_X n (\sum_{i=1}^m \nu^i L^2_i)^{1/2}] \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q_{\nu}(z_{t+1}, z)\right] \leq \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q_{\nu}(z_{t+1}, z)\right]
\end{equation}
(31)

Using this relation, the convexity of $f_0(\cdot)$ and $f(\cdot)$, and noting the definition of $\Gamma_T$, we have for all non-random $y \geq 0$ and $x \in X,$
$$\Gamma_T \mathbb{E}[f_0(\bar{x}_T) + \langle y, f(\bar{x}_T) \rangle - f_0(x) - \langle \bar{y}_T, f(x) \rangle] - \nu^2_0 L_0 n + M_X n (\sum_{i=1}^m \nu^i L^2_i)^{1/2} \Gamma_T
\leq \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q(z_{t+1}, z)\right] - [\nu^2_0 L_0 n + M_X n (\sum_{i=1}^m \nu^i L^2_i)^{1/2}] \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q_{\nu}(z_{t+1}, z)\right]
\leq \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q_{\nu}(z_{t+1}, z)\right].
\end{equation}
(32)

Combining (30), (31) and (32), then choosing $x = x^*$, $y = 0$ (which are non-random) throughout the combined relation, observing that $|0 - 1| = + 0$, we have
\begin{equation}
\Gamma_T \mathbb{E}[f_0(\bar{x}_T) - f_0(x^*) - \langle \bar{y}_T, f(x^*) \rangle] - \nu^2_0 L_0 n + M_X n (\sum_{i=1}^m \nu^i L^2_i)^{1/2} \Gamma_T
\leq \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q_{\nu}(z_{t+1}, (x^*, 0))\right]
\leq \gamma_0 \eta_0 W(x^*, x_0) - \gamma_{T-1} \eta_{T-1} \mathbb{E}[W(x^*, x_T)] + \frac{7\gamma_0 T_0}{2} \|y_0\|^2_2 + \sum_{t=0}^{T-1} \frac{2\gamma_t}{\eta_t - L_0 - L_f} \mathbb{E}[\|\delta^G\|^2_2] \\
\quad + \left(\frac{T-1 \gamma_{t-1}}{\eta_{t-1}} + \frac{T \gamma_{T-1}}{\eta_{T-1}}\right) (\sigma^2_{f,\nu} + D^2_X \|\sigma_{\nu}\|^2_2)
\end{equation}
(33)
Ignoring the $\mathbb{E}[W(x^*, x_T)]$ term and noting that $f(x^*) \leq 0$ and $\bar{y}_T \geq 0$ implies $\langle \bar{y}_T, f(x^*) \rangle \leq 0$, we have
(23).

Now, we focus our attention to the infeasibility bound. First, we define $R := \|y^*\|_2 + 1$. Second, define an auxiliary sequence $\{y^*_t\}$ in the following way: $y^*_0 = y_0$ and for all $t \geq 0$, define
$$y^*_{t+1} := \arg \min_{y \in B^2_t(R)} \frac{1}{\tau_{t-1}} \langle \delta^F_t, y \rangle + \frac{1}{2} \|y - y^*_t\|^2_2,$$
where we recall that \( \mathcal{B}_+^2(R) = \{ x \in \mathbb{R}^n : ||x||_2 \leq R, x \geq 0 \} \). Then in view of Lemma 9.2, in particular relation (64), for all \( y \in \mathcal{B}_+^2(R) \) we have

\[
\frac{1}{\tau_t} \langle \delta_{t+1}^F, y_t^{x+1} - y \rangle \leq \frac{1}{2} ||y - y_t^{x+1}||_2^2 - \frac{1}{2} ||y - y_t^{x+2}||_2^2 + \frac{1}{2\tau_t} ||\delta_{t+1}^F||_2^2. \tag{34}
\]

Multiplying (34) by \( \gamma_t \tau_t \), taking a sum from \( t = 0 \) to \( T - 1 \) and noting the second relation in (66), we obtain

\[
\sum_{t=0}^{T-1} \gamma_t \langle \delta_{t+1}^F, y_t^{x+1} - y \rangle \leq \frac{\gamma_0 T_0}{2} ||y - y_1^*||_2^2 + \sum_{t=0}^{T-1} \frac{\gamma_t}{\tau_t} ||\delta_{t+1}^F||_2^2. \tag{35}
\]

for all \( y \in \mathcal{B}_+^2(R) \). Summing (35) and (68), we obtain

\[
\sum_{t=0}^{T-1} \gamma_t \langle \delta_{t+1}^F, y_t^{x+1} - y \rangle \leq \frac{\gamma_0 T_0}{2} ||y - y_1^*||_2^2 + \sum_{t=0}^{T-1} \frac{\gamma_t}{\tau_t} ||\delta_{t+1}^F||_2^2. \tag{36}
\]

for all \( z \in \{(x, y) : x \in X, y \in \mathcal{B}_+^2(R) \} \). Note that given \( \xi_{[t]}, u_{[t]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]} \), we have \( y_t+1, y_t^{x+1}, x_{t+1}, x_t \) are constants. Hence, we have

\[
\mathbb{E}[\langle \delta_{t+1}^F, y_t^{x+1} - y_t^{x+1} \rangle] = \mathbb{E}[\mathbb{E}[\mathbb{E}[\xi_{[t]}, u_{[t]}, \bar{\xi}_{[t-1]}, \bar{u}_{[t-1]}][\delta_{t+1}^F, y_t+1 - y_t^{x+1}]] = 0, \tag{37}
\]

where second equality follows from (27). Choosing \( z = \bar{z} := (x^*, \bar{y}) \) in (36) where \( \bar{y} := (||y^*||_2 + 1)[f(\bar{x}_T)]_+ ||f(\bar{x}_T)||_+^{-1} \in \mathcal{B}_+^2(R) \), taking expectation on both sides and noting (37), (28), (29), first relation in (25), we have

\[
\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q_{\nu}(z_{t+1}, \bar{z})\right] \leq \frac{\gamma_0 T_0}{2} \mathbb{E}[||\bar{y} - y_0||_2^2 + ||\bar{y} - y_1^*||_2^2 + \gamma_0 \eta_0 W(x^*, x_0)]
+ \sum_{t=0}^{T-1} \frac{2\gamma_t}{\tau_t} \left\{ \mathbb{E}[||\delta_t^F||_2^2] + \left(\frac{L_f}{2} ||f(x)||_2 - 1\right)_+ \right\}^2 + \sum_{t=0}^{T-1} \frac{\gamma_t}{\tau_t} ||\delta_{t+1}^F||_2^2.
\]

By Lemma 8.1, we then have \( Q(z_{t+1}, \bar{z}) - [\nu^2 L_0 n + M_X n (\sum_{t=1}^{m} \nu^4 (L_0^2)^2)^{1/2} ||x_{t+1}, \bar{z}||] \). Multiplying both sides by \( \gamma_t \), summing from \( t = 0 \) to \( T - 1 \), taking expectation of both sides and dividing by \( \Gamma_T \), we have

\[
\frac{1}{\Gamma_T} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q(z_{t+1}, \bar{z})\right] - [\nu^2 L_0 n + M_X n (\sum_{t=1}^{m} \nu^4 (L_0^2)^2)^{1/2} \leq \frac{1}{\Gamma_T} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q(z_{t+1}, \bar{z})\right]. \tag{39}
\]

Noting the convexity of \( Q \) in the first argument, we obtain

\[
\mathbb{E}[Q(\bar{z}, \bar{z})] \leq \frac{1}{\Gamma_T} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma_t Q(z_{t+1}, \bar{z})\right]. \tag{40}
\]

Now observe that we have \( \mathcal{L}(\bar{x}_T, y^*) - \mathcal{L}(x^*, y^*) \geq 0 \) which implies that \( f_0(\bar{x}_T) + \langle y^*, f(\bar{x}_T) \rangle - f_0(x^*) \geq 0 \), which follows from complementary slackness. In view of the relation

\[
\langle y^*, f(\bar{x}_T) \rangle \leq \langle y^*, f(\bar{x}_T) \rangle_+ \leq ||y^*||_2 [||f(\bar{x}_T)||_+],
\]

27
the above inequality implies that
\[ f_0(\bar{x}_T) + \|y^*\|_2 \|f(\bar{x}_T)\|_2^2 - f_0(x^*) \geq 0. \] (41)

Moreover, we have that
\[ Q(\bar{x}_T, \hat{\bar{z}}) = L(\bar{x}_T, \hat{\bar{y}}) - L(x^*, \hat{y}^T) \geq L(\bar{x}_T, \hat{\bar{y}}) - L(x^*, y^*) = f_0(\bar{x}_T) + (\|y^*\|_2 + 1)\|f(\bar{x}_T)\|_2 - f_0(x^*), \]
which along with (41) implies that
\[ Q(\bar{x}_T, \hat{\bar{z}}) \geq \|f(\bar{x}_T)\|_2. \]

The above relation, (38), (39) and (40) together yield
\[
\mathbb{E}[\|f(\bar{x}_T)\|_2] \leq \frac{1}{T} \left[ \frac{\gamma_0 T_0}{2} \mathbb{E}[\|y - y_0\|_2^2 + \|\hat{y} - \bar{y}_v\|_2^2] + \gamma_0 \eta_0 W(x^*, x_0) \right. \\
+ \sum_{t=0}^{T-1} \frac{12\eta_1^2 t}{2T - L_t - L_f} \left\{ \mathbb{E}[\|\delta_t^G\|_2^2] + \left( \frac{L_f D_X}{2} \|y^*\|_2^2 \right) \right\} \\
+ \left( \sum_{t=1}^{T-1} \frac{12\gamma_1^2 t}{2T - L_t - L_f} + \sum_{t=0}^{T-1} \frac{12\gamma_{T-1}^2 t}{2T - L_t - L_f} \right) \left( \sigma_{f,\nu}^2 + D_X^2 \|\sigma_v\|_2^2 \right) \\
+ \left( \nu_0^2 L_0 n + M_X n \left( \sum_{i=1}^{m} \nu_i^4 L_i^2 \right)^{1/2} \right).
\]

Noting the bound \(\|\hat{y} - y_0\|_2 \leq 2R\) and \(\|\hat{y} - y_0\|_2^2 \leq 2\|y_0\|_2^2 + 2\|\hat{y}\|_2^2 \leq 2\|y_0\|_2^2 + 2R^2\) in the above relation and recalling that \(R = \|\bar{y}\|_2 / 2\), we obtain (44). Hence, we conclude the proof. \(\Box\)

We next bound the term \(\mathbb{E}[\|\delta_t^G\|_2^2]\) appearing in the previous result in the zeroth-order setting. This result is crucial for obtaining a linear dependency on the number of constraints \(m\) for our oracle complexity results and is based on our Lemma 5.1.

**Lemma 8.3.** Assume that \(\{\gamma_t, \tau_t, \eta_t\}\) satisfy
\[
\frac{96\|\sigma_v\|_2^2}{\gamma_t (\eta_t - L_0 - L_f)} < 1,
\] (42)
for all \(t \leq T - 1\) and constants \(R_1\) and \(R_2\) satisfying the following conditions exist:

\[
R_1 \geq \left( 1 - \frac{96\|\sigma_v\|_2^2}{\tau_t (\eta_t - L_0 - L_f)} \right)^{-1} \left[ 2\sigma_{0,v_0}^2 + \frac{48\|\sigma_v\|_2^2}{\gamma_t \tau_t} \left( \frac{\gamma_0 \eta_0 W(x^*, x_0)}{2} + \frac{\gamma_0 \tau_t}{12} \|y^*\|_2^2 \right) \right. \\
+ \sum_{t=0}^{T-1} \frac{12\gamma_t^2 t}{2T - L_t - L_f} \left( \|y^*\|_2^2 \right)^2 + \left( \sum_{t=1}^{T-1} \frac{12\gamma_{T-1}^2 t}{2T - L_t - L_f} + \frac{12\gamma_{T-1}^2 \eta_t}{12} \right) \left( \sigma_{f,\nu}^2 + D_X^2 \|\sigma_v\|_2^2 \right) \\
+ \left[ \nu_0^2 L_0 n + M_X n \left( \sum_{i=1}^{m} \nu_i^4 L_i^2 \right)^{1/2} \right] \right] \\
\] (43)

for all \(t \leq T - 1\) and

\[
R_2 \geq \left( 1 - \frac{96\|\sigma_v\|_2^2}{\tau_t (\eta_t - L_0 - L_f)} \right)^{-1} \frac{96\|\sigma_v\|_2^2 \gamma_i}{\gamma_t \tau_t (\eta_t - L_0 - L_f)} \\
\] (44)

for all \(t \leq T - 1\) and \(i \leq t - 1\). Then, we have
\[
\mathbb{E}[\|\delta_t^G\|_2^2] \leq R_1 (1 + R_2)^t, \quad (45)
\]
for all \(t \leq T - 1\). In particular, if \(\|\sigma_v\|_2 = 0\) then we can set \(R_1 = 2\sigma_{0,v_0}^2\) and \(R_2 = 0\) implying
\[
\mathbb{E}[\|\delta_t^G\|_2^2] \leq 2\sigma_{0,v_0}^2.
\]
Proof of Lemma 8.3 First note that by Lemma 8.1, we have
\[
Q(z_{i+1}, z) - [\nu_0^2 L_0 n + M_X n (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}] \leq Q_\nu(z_{i+1}, z)
\]
Multiplying the above by \(\gamma_i\) and summing up \(i = 0\) to \(t\), we have
\[
\sum_{i=0}^t \gamma_i Q_\nu(z_{i+1}, z) - [\nu_0^2 L_0 n + M_X n (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}] \Gamma_{t+1} \leq \sum_{i=0}^t \gamma_i Q_\nu(z_{i+1}, z)
\]
Replacing \(\Gamma\) for \(t + 1 (\geq 1)\) in (68), we have
\[
\sum_{i=0}^t \gamma_i Q_\nu(z_{i+1}, z) + \sum_{i=0}^t \gamma_i [\langle \delta^G_i, x_i - x \rangle - \langle \delta^E_i, y_{i+1} - y \rangle]
\leq \gamma_0 \eta_0 W(x, x_0) - \gamma_0 \eta_0 W(x, x_{t+1}) + \frac{\gamma_0 \eta_0}{2} \| y_{t+1} \|_2^2 - \gamma_0 \eta_0 \| y - y_{t+1} \|_2^2
\]
\[
+ \sum_{i=0}^t \frac{2 \gamma_i}{\eta_0 - L_f} \| q_i - q_i \|_2^2 + \frac{3 \gamma_i}{\eta_0} \| q_{t+1} - q_{t+1} \|_2^2.
\]
(46)
Observe that \(Q(z_{i+1}, z^*) \geq 0 \) for \(i = 0, \ldots, t\) by our saddle point assumption where \(z^* = (x^*, y^*)\). Choosing \(z = z^*\) (both non-random) in the above relations, taking expectation, using (25) with \(x = x^*\) and (26) with \(y = y^*\), disregarding the term \(-\gamma_0 \eta_0 E[W(x^*, x_{t+1})]\) and noting (29), we have the following inequality
\[
- \nu_0^2 L_0 n + M_X n (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2} \Gamma_{t+1} + \frac{\gamma_0 \eta_0}{12} E \| y^* - y_{t+1} \|_2^2
\]
\[
\leq \gamma_0 \eta_0 W(x^*, x_0) + \frac{\gamma_0 \eta_0}{2} \| y^* - y_0 \|_2^2
\]
\[
+ \sum_{i=0}^t \frac{2 \gamma_i}{\eta_0 - L_f} \left[ E[\| \delta^G_i \|_2^2] + \left( \frac{L_f D_X}{2} \| y^* \|_2 - 1 \right) \right]
\]
\[
+ \left( \sum_{i=1}^t \frac{12 \gamma_i \eta^2}{\tau_i} + \frac{12 \gamma_i}{\tau_i} \right) \sigma_{f, \nu}^2 + D_X^2 \| \sigma_{\nu} \|_2^2.
\]
(48)
Now, let us define \(\delta^G_{t,i} := G_{t,i,\nu}(x_t, \xi_t, u_t) - f'_{i,\nu}(x_t)\) for \(i = 0, \ldots, m\). As a consequence, we have
\[
\delta_t^G = \delta^G_{t,0} + \sum_{i=1}^m y_{t+1}^{(i)} \delta^G_{t,i}.
\]
Then, we have
\[
E[\| \delta^G_{t,i} \|_2^2] = E[\| \delta^G_{t,0} + \sum_{i=1}^m y_{t+1}^{(i)} \delta^G_{t,i} \|_2^2]
\]
\[
\leq 2 E[\| \delta^G_{t,0} \|_2^2] + 2 E[\| \sum_{i=1}^m y_{t+1}^{(i)} \delta^G_{t,i} \|_2^2]
\]
\[
\leq 2 E[\| \delta^G_{t,0} \|_2^2] + 2 E[\| \sum_{i=1}^m \| y_{t+1}^{(i)} \delta^G_{t,i} \|_2^2]
\]
\[
\leq 2 [\sigma^2_{0,\nu_0} + E[\| y_{t+1} \|_2^2 \sum_{i=1}^m \| \delta^G_{t,i} \|_2^2]]
\]
\[
\leq 2 [\sigma^2_{0,\nu_0} + E[\| y_{t+1} \|_2^2 \sum_{i=1}^m \| \delta^G_{t,i} \|_2^2]]
\]
\[
\leq 2 [\sigma^2_{0,\nu_0} + E[\| y_{t+1} \|_2^2 \sum_{i=1}^m \| \delta^G_{t,i} \|_2^2]]
\]
\[
\leq 2 [\sigma^2_{0,\nu_0} + E[\| y_{t+1} \|_2^2 \sum_{i=1}^m \| \delta^G_{t,i} \|_2^2]]
\]
\[
= 2 \sigma^2_{0,\nu_0} + 2 \sigma_{\nu}^2 \| E[\| y_{t+1} \|_2^2] \|_2^2
\]
\[
\leq 2 \sigma^2_{0,\nu_0} + 4 \| \sigma_{\nu} \|_2^2 \| E[\| y_{t+1} \|_2^2 - y^* \|_2^2] \|_2^2.
\]
(49)
conditioned on random variables \( \xi_{[t-1]}, u_{[t-1]}, \tilde{\xi}_{[t-1]}, \tilde{u}_{[t-1]} \).

Adding \( \frac{\gamma t \tau_t}{12} \| y^* \|^2_t \) to both sides of (48), then multiplying it by \( \frac{48 \| \sigma_v \|^2}{\gamma t \tau_t} \) and observing (49), we have

\[
\mathbb{E}[\| \delta_t^G \|^2] \leq 2 \sigma_{0,0}^2 + \frac{48 \| \sigma_v \|^2}{\gamma t \tau_t} \left\{ \gamma_0 \eta_0 W(x^*, x_0) + \frac{\gamma_0 \tau_0}{2} \| y^* - y_0 \|^2 + \frac{\gamma t \tau_t}{12} \| y^* \|^2 \right. \\
+ \sum_{i=0}^{t} \frac{2 \gamma_i}{\eta_i - L_0 - L_f} \left( \frac{L_i D_X}{2} \| y^* \|^2 - 1 \right) + \left( \sum_{i=1}^{t} \frac{12 \gamma_i \theta_i^2}{\tau_i} + \frac{12 \gamma_i}{\tau_i} \right) (\sigma_{f,v}^2 + D_X^2 \| \sigma_v \|^2) + \left[ (2 \| \bar{\nu}_0 \| x_0 + M_X n (\sum_{i=1}^{m} \nu_i^t L_i^2)^{1/2}) \right] \Gamma_1 + \sum_{i=0}^{t} \frac{96 \| \sigma_v \|^2}{\gamma_i \tau_i (\eta_i - L_0 - L_f)} \mathbb{E}[\| \delta_t^G \|^2] \right).
\]

In view of (42), we have that the coefficient of the \( \delta_t^G \) term on the right hand side of the above relation is strictly less than 1. Moving the \( \delta_t^G \) term to the left hand side and noting the conditions imposed on constants \( R_1, R_2 \), we have

\[
\mathbb{E}[\| \delta_t^G \|^2] \leq R_1 + R_2 \sum_{i=1}^{t-1} \mathbb{E}[\| \delta_t^G \|^2],
\]

for all \( t \leq T - 1 \). Using Lemma 9.3 for the above relation, we have (45). Hence we conclude the proof.

\[ \square \]

We are now ready to prove Theorem 3.1. Before we proceed, we remark that we the results from Lemma 9.4 in Section 9 for the proof.

**Proof of Theorem 3.1** It is easy to verify that \( \{ \gamma_t, \theta_t, \eta_t, \tau_t \} \) set according to Theorem 3.1 satisfies (66). Note that (67) is satisfied if \( 4 M_f^2 \leq \frac{\tau_i (\eta - \eta_0 - L_0 - L_f)}{12} \). This follows due to the fact that \( \{ \eta_t \} \) is a non-decreasing sequence, \( \theta_t = 1 \) for all \( t \geq 0 \). Then we have

\[
\frac{\tau_i (\eta - \eta_0 - L_0 - L_f)}{12} \geq \frac{4 M_f}{D_X} 12 M_f D_X \times \frac{1}{12} = 4 M_f^2
\]

Also, since \( (\eta - \eta_0 - L_f) \geq \frac{24 \| \sigma_v \|}{D_X} \) and \( \tau_t \geq 8 D_X \| \sigma_v \|_2 \), we have

\[
\tau_t (\eta - \eta_0 - L_f) \geq 192 \| \sigma_v \|^2_t
\]

for all \( t \geq 0 \). In view of the above relation, we have

\[
\frac{96 \| \sigma_v \|^2}{\tau_i (\eta - \eta_0 - L_f)} \leq \frac{1}{2},
\]

hence (42) is satisfied. We also need to show the existence of \( R_1 \) and \( R_2 \) satisfying (43) and (44), respectively. Using the fact that \( \gamma_t, \eta_t \) and \( \tau_t \) are constants for all \( t \geq 0 \), \( \tau \eta \geq \frac{96 T \sigma_{X,f} \| \sigma_v \|_2}{D_X} \) and noting (50), we obtain

\[
\left( 1 - \frac{96 \| \sigma_v \|^2}{\tau_i (\eta - \eta_0 - L_f)} \right)^{-1} \frac{96 \| \sigma_v \|^2}{\gamma_i \tau_i (\eta - \eta_0 - L_f)} \leq 2 \frac{96 \| \sigma_v \|^2}{\tau \eta} \leq 2 \frac{\| \sigma_v \|^2}{T \sigma_{X,f}} \leq \frac{2}{T},
\]

where in the last relation, we used the fact that \( \sigma_{X,f} \geq D_X \| \sigma_v \|_2 \). In view of the above relation and (44), we can set

\[
R_2 := \frac{2}{T},
\]

30
Noting (43) along with the fact that $\mathcal{H}_* \geq L_f D_X (\|y^*\|_2 - 1)^+$, setting $y_0 = 0$, using (50), (12), $\gamma_t r_t = \tau \geq \sqrt{96T \sigma_{X,f}} \sum_{i=0}^t \eta_i - L_0 - L_f = \frac{t + 1}{\eta} \leq \sqrt{T D_X^{2} \mathcal{H}_*^{2}} \left( \frac{\sqrt{2 T D_X^{2} \mathcal{H}_*^{2}}}{\sqrt{\mathcal{H}_*^{2} + \sigma_{\nu,0}^{2} + 48 \|\sigma_{\nu}\|_{2}^{2}}} \right)_{2}$, and $\sum_{i=1}^t \frac{\gamma_i r_i^{2}}{r_i} + \frac{\gamma_i}{r_i} = \frac{t + 1}{\tau} \leq \frac{T}{\tau}$ for all $t \leq T - 1$, we can see that the RHS of (43) is at most

$$2 \left[ 2 \sigma_{0,\nu}^{2} + 48 \|\sigma_{\nu}\|_{2}^{2} \left\{ \frac{7}{12} \|y^*\|_{2}^{2} + \frac{\eta}{\tau} D_X^{2} + \frac{\sqrt{2 T D_X^{2} \mathcal{H}_*^{2}}}{\sqrt{\mathcal{H}_*^{2} + \sigma_{\nu,0}^{2} + 48 \|\sigma_{\nu}\|_{2}^{2}}} \frac{1}{\sqrt{96T \sigma_{X,f}}} + 12 \sigma_{X,f}^{2} \frac{T}{\tau^{2}} \right\} \right]^{1/2} \left[ \nu_0^{2} L_{0} n + M_{X} (\sum_{i=1}^{m} \nu_{i} L_{i}^{2})^{1/2} \right] \right]$$

$$\leq 2 \left[ 2 \sigma_{0,\nu}^{2} + 48 \|\sigma_{\nu}\|_{2}^{2} \left\{ \frac{7}{12} \|y^*\|_{2}^{2} + \frac{\eta}{\tau} D_X^{2} + \frac{D_X H_*}{\sqrt{48 \sigma_{X,f}}} + 12 T \mathcal{H}_* \frac{1}{\sqrt{96 T \sigma_{X,f}^{2}}} + \frac{[\nu_0^{2} L_{0} n + M_{X} (\sum_{i=1}^{m} \nu_{i} L_{i}^{2})^{1/2} \sqrt{T}]}{4 \sqrt{6 \sigma_{X,f}}} \right\} \right]^{1/2}$$

$$\leq 2 \left( \left[ 2 \sigma_{0,\nu}^{2} + 28 \|\sigma_{\nu}\|_{2}^{2} \|y^*\|_{2}^{2} + 150 \|\sigma_{\nu}\|_{2}^{2} + \sqrt{48} \|\sigma_{\nu}\|_{2}^{2} [2 \mathcal{H}_* + (\sigma_{\nu,0} + \sqrt{48} \|\sigma_{\nu}\|_{2})] + 2 \sqrt{6} D_{X}^{-1} \|\sigma_{\nu}\|_{2} [\nu_0^{2} L_{0} n + M_{X} (\sum_{i=1}^{m} \nu_{i} L_{i}^{2})^{1/2} \sqrt{T}] \right]$$

$$\leq: R_{1} (52)$$

where in the last inequality, we used the fact that $\|\sigma_{\nu}\|_{2}^{2} D_{X} \mathcal{H}_* \sigma_{X,f}^{2} \leq 1$. Note that the last term in the above sequence of relations is a constant satisfying the requirement in (43). Hence, we can set

$$R_{1} := 2 \left( \left[ 2 \sigma_{0,\nu}^{2} + 28 \|\sigma_{\nu}\|_{2}^{2} \|y^*\|_{2}^{2} + 150 \|\sigma_{\nu}\|_{2}^{2} + \sqrt{48} \|\sigma_{\nu}\|_{2}^{2} [2 \mathcal{H}_* + (\sigma_{\nu,0} + \sqrt{48} \|\sigma_{\nu}\|_{2})] + 2 \sqrt{6} D_{X}^{-1} \|\sigma_{\nu}\|_{2} [\nu_0^{2} L_{0} n + M_{X} (\sum_{i=1}^{m} \nu_{i} L_{i}^{2})^{1/2} \sqrt{T}] \right)$$

$$\leq R_1 (1 + \frac{2}{T})^{T-1} \leq R_1 \epsilon^2$$

(53)

Then using Lemma (8.3) and noting (51), we have for all $t \leq T - 1$

$$E[||\delta_{t}^{G}||_{2}^{2}] \leq \begin{cases} 4 \sigma_{0,\nu}^{2} & \text{if } ||\sigma_{\nu}\|_{2} = 0; \\ R_1 (1 + \frac{2}{T})^{T-1} \leq R_1 \epsilon^2 & \text{otherwise.} \end{cases}$$

(54)

Noting the above relation, (53) and the definition of $\zeta$, we have

$$E[||\delta_{t}^{G}||_{2}^{2}] \leq \epsilon^2, \quad \forall t \leq T - 1.$$

Hence, according to (23) with $y_0 = 0$ and using (54), we have

$$E[f_{0}(\bar{x}_{T}) - f_{0}(x^{*})] \leq \frac{1}{T} \left[ (\eta + L_0 + L_f) W(x^{*}, x_0) + \frac{2T \zeta^2}{\eta} + 12 \sigma_{X,f}^{2} \frac{T}{\tau} \right]$$

$$+ [\nu_0^{2} L_{0} n + M_{X} (\sum_{i=1}^{m} \nu_{i} L_{i}^{2})^{1/2}].$$

31
Using the bound $W(x^*, x_0) \leq D_X^2$, we obtain (8). From (24) and (54), we have for $T \geq 1$

$$
\mathbb{E}[\| f(\bar{x}_T) \|_2] \leq \frac{1}{T} \left[ 3(\| y^* \|_2 + 1)^2 \tau + (\eta + L_0 + L_f) W(x^*, x_0) + \frac{2(\zeta^2 + H_*^2)T}{\eta} + \frac{13\sigma_{X,f}^2 T}{\tau} \right] + [\nu_0^2 L_0 n + M_X m (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2}].
$$

Using bounds $W(x^*, x_0) \leq D_X^2$, we obtain (9). Define

$$
\sigma_f^2 := 2(1 + \sigma_f^2),
$$

$$
\sigma_0^2 := 1 + 10(n + 4)[\sigma_0^2 + [L_0(1 + D_X) + M_0]^2],
$$

$$
\sigma_i^2 := \frac{1}{m} + 10(n + 4)[\sigma_i^2 + [L_i(1 + D_X) + M_i]^2] \quad \text{for } i \in \{1, \ldots, m\},
$$

$$
\sigma^2 = 1 + 10(n + 4)[\|x\|_2^2 + 2L_f^2(1 + D_X)^2 + 2M_f^2],
$$

$$
\sigma_{X,f}^2 = (2(1 + \sigma_f^2) + D_X^2 \sigma^2)^{1/2},
$$

$$
\zeta := 2e \left\{ \sigma_0^2 + \sigma^2(14\| y^* \|_2^2 + 75) + 2\sqrt{3\sigma(2H_* + \sigma_0 + \sqrt{48\sigma}) + \sqrt{6D_X^{-1}} \sigma} \right\}^{1/2}.
$$

By choice of $\nu_0, \nu_i$ for $i \in [m]$, definition of $\sigma_{f,\nu}, \bar{B}_i, \sigma_{i,\nu_i}^2$, and $\sigma_{\nu}$, we have

$$
\sigma_{f,\nu}^2 \leq 2 + 2\sigma_f^2 := \sigma_f^2,
$$

$$
\bar{B}_i \leq L_i(1 + D_X) + M_i,
$$

$$
\sigma_{0,\nu_0}^2 \leq 1 + 10(n + 4)[\sigma_0^2 + [L_0(1 + D_X) + M_{f,0}]^2],
$$

$$
\sigma_{i,\nu_i}^2 \leq \frac{1}{m} + 10(n + 4)[\sigma_i^2 + [L_i(1 + D_X) + M_{f,i}]^2] := \sigma_i^2 \quad \text{for } i \in [m],
$$

$$
\|\sigma_{\nu}\|_2^2 \leq 1 + 10(n + 4)[\|\sigma\|_2^2 + 2L_f^2(1 + D_X)^2 + 2M_f^2] := \sigma^2.
$$

Furthermore, we also have that $\nu_0^2 L_0 n + M_X m (\sum_{i=1}^m \nu_i^4 L_i^2)^{1/2} \leq \frac{1}{\sqrt{T}}$. Using these relations, we see that $\sigma_{X,f} \leq \sigma_{X,f}^2$ and $\zeta \leq \zeta$. Hence, we have

$$
\mathbb{E}[f_0(x_T) - f_0(x^*)] \leq \frac{(L_0 + L_f)D_X^2}{T} + \frac{\max\{12M_f, 24\sigma\} D_X}{T} + \frac{1}{\sqrt{T}} \left\{ \sqrt{2D_X \zeta^2} \frac{\sqrt{2D_X \zeta^2}}{\sqrt{H_*^2 + \sigma_0^2 + 48\|\sigma\|_2^2}} + \sqrt{3\sigma_{X,f}^2} \right\},
$$

and

$$
\mathbb{E}[\|f(\bar{x}_T)\|_2] \leq \frac{1}{\sqrt{T}} \left\{ 12\sqrt{6}(\|y^*\|_2 + 1)^2 + \frac{13}{4\sqrt{6}} \sigma_{X,f} + \sqrt{2D_X \left[ H_*^2 + \sigma_0^2 + 48\|\sigma\|_2^2 \right] + \frac{\zeta^2 + H_*^2}{\sqrt{H_*^2 + \sigma_0^2 + 48\|\sigma\|_2^2}}} \right\}
$$

As a consequence, to obtain an $(\varepsilon, \varepsilon)$-optimal solution with Algorithm 1, we need the number of iterations to
Then for any 

We highlight that the definition of the terms (Boob et al. 2022, Lemma 2.8) Let 

Lemma 9.3.

so by combining an argument from Nesterov and Spokoiny (2017) on the analysis of stochastic zeroth-order arguments of the smoothed functions rather than the original functions as done in Boob et al. (2022). We do deals with stochastic first-order optimization. This necessitates dealing with the Lipschitz continuity based on the stochastic zeroth-order gradient estimator (defined in (2)). Whereas, the corresponding terms from the following result, which is an adaptation of Lemma 2.5 in Boob et al. (2022) to the zeroth-order setting.

such that

Now, by the choice of \( \nu_0 \) and \( \nu_1 \) in (10) and (11) respectively, we see that the oracle complexity is given by \( O((m + 1)n)/\epsilon^2) \).

\[ \boxed{50} \frac{\varepsilon}{\varepsilon^2} \left[ D_X \sqrt{\mathcal{H}^2 + \bar{\sigma}^2 + 48\tilde{\sigma}^2} + \frac{D_X (\xi^2 + \mathcal{H}^2)}{\sqrt{\mathcal{H}^2 + \bar{\sigma}^2 + 48\tilde{\sigma}^2}} \right] \]

9 Auxiliary results

In this subsection, we state some auxiliary results from Boob et al. (2022), which we used in the proofs above.

**Lemma 9.1.** (Boob et al. 2022) Lemma 2.4) Assume that \( g : S \rightarrow \mathbb{R} \) satisfies

\[ g(y) \geq g(x) + \langle g'(x), y - x \rangle + \mu W(y, x), \quad \forall x, y \in S \]

for some \( \mu \geq 0 \), where \( S \) is convex set in \( \mathbb{R}^n \). If \( \bar{x} = \arg \min_{x \in S} \{ g(x) + W(x, \bar{x}) \} \), then \( g(\bar{x}) + W(\bar{x}, \bar{x}) + (\mu + 1)W(x, \bar{x}) \leq g(x) + W(x, \bar{x}), \quad \forall x \in S. \)

**Lemma 9.2.** (Boob et al. 2022) Lemma 2.6) Let \( \rho_0, \ldots, \rho_j \) be a sequence of elements in \( \mathbb{R}^n \) and let \( S \) be a convex set in \( \mathbb{R}^n \). Define the sequence \( v_t, t = 0, 1, \ldots, \) as follows: \( v_0 \in S \) and

\[ v_{t+1} = \arg \min_{x \in S} \langle \rho_t, x \rangle + \frac{1}{2} \| x - v_t \|^2. \]

Then for any \( x \in S \) and \( t \geq 0 \), the following inequalities hold

\[ \langle \rho_t, v_t - x \rangle \leq \frac{1}{2} \| x - v_t \|^2 - \frac{1}{2} \| x - v_{t+1} \|^2 + \frac{1}{2} \| \rho_t \|^2, \]

\[ \sum_{t=0}^{j} \langle \rho_t, v_t - v_0 \rangle \leq \frac{1}{2} \| x - v_0 \|^2 + \frac{1}{2} \sum_{t=0}^{j} \| \rho_t \|^2. \]

**Lemma 9.3.** (Boob et al. 2022) Lemma 2.8) Let \( \{ a_t \}_{t=0} \) be a nonnegative sequence, \( m_1, m_2 \geq 0 \) be constants such that \( a_0 \leq m_1 \) and the following relation holds for all \( t \geq 1 \):

\[ a_t \leq a_0 + m_1 \sum_{k=0}^{t-1} a_k. \]

Then we have \( a_t \leq m_1 (1 + m_2)^t \).

The proof of the above three lemmas could be found in Boob et al. (2022). We also state and prove the following result, which is an adaptation of Lemma 2.5 in Boob et al. (2022) to the zeroth-order setting. We highlight that the definition of the terms \( q_t, \bar{q}_t, \delta_t^F \) and \( \delta_t^G \) appearing in Lemma 9.4 below are based on the stochastic zeroth-order gradient estimator (defined in (2)). Whereas, the corresponding terms from Lemma 2.5 in Boob et al. (2022) are based on the stochastic first-order gradients (as Boob et al. (2022) deals with stochastic first-order optimization). This necessitates dealing with the Lipschitz continuity based arguments of the smoothed functions rather than the original functions as done in Boob et al. (2022). We do so by combining an argument from Nesterov and Spokoiny (2017) on the analysis of stochastic zeroth-order method, along with the proof of Lemma 2.5 from Boob et al. (2022). Hence, we provide a full proof of Lemma 9.4 below for the convenience of readers who might be unfamiliar with the analysis of stochastic zeroth-order optimization algorithms.
Lemma 9.4. (Boob et al. 2022) Lemma 2.5 adapted to the stochastic zeroth-order setting: Suppose Assumptions 2.1, 2.2, and 2.3 are satisfied. Assume that \( \{\gamma_t, \eta_t, \tau_t, \theta_t\} \) is a non-negative sequence satisfying
\[
\gamma_t \theta_t = \gamma_{t-1}, \quad \gamma_t \tau_t \leq \gamma_{t-1} \tau_{t-1}, \quad \tau_t \eta_t \leq \gamma_{t-1} \eta_{t-1},
\]
and
\[
(2M_f)^2 \frac{\theta_t}{\theta_{t-1}} \leq \frac{\tau_t (\eta_{t-2} - L_0 - L_f)}{12}, \quad \theta_t (M_f)^2 \leq \frac{\tau_t (\eta_{t-1} - L_0 - L_f)}{12},
\]
\[
(2M_f)^2 \frac{1}{\theta_{t-1}} \leq \frac{\tau_{t-1} (\eta_{t-2} - L_0 - L_f)}{12}, \quad M_f^2 \leq \frac{\tau_{t-1} (\eta_{t-1} - L_0 - L_f)}{12},
\]
where \( M_f, L_f \) are defined in (3). Then, for all \( T \geq 1 \) and \( z \in \{(x, y) : x \in X, y \geq 0\} \), we have
\[
\sum_{t=0}^{T-1} \gamma_t Q_{\nu_t} (z_{t+1}, z) + \sum_{t=0}^{T-1} \gamma_t \left[ (\delta^G_t, x_t - x) - \langle \delta^F_{t+1}, y_{t+1} - y \rangle \right] \leq \gamma_0 \eta_0 W(x, x_0) - \gamma_{t-1} \eta_{t-1} W(x, x_T) + \frac{\gamma_0 \eta_0 T_0}{2} \|y - y_0\|_2^2 - \frac{\gamma_{t-1} \tau_{t-1}}{12} \|y - y_T\|_2^2
\]
\[
+ \sum_{t=0}^{T-1} \frac{2 \gamma_t}{m - L_0 - L_f} \left[ \|\delta^G_t\|_2^2 + \left( \frac{L_f D x}{2} \|y\|_2 - 1 \right)^2 \right]
\]
\[
+ \sum_{t=0}^{T-1} \frac{3 \gamma_t \theta_t^2}{2 r_t^2} \|q_t - \tilde{q}_t\|_2^2 + \frac{3 \gamma_{t+1} T_1}{2 r_{t+1}} \|q_T - \tilde{q}_T\|_2^2.
\]
Here \( q_t := \ell_F (x_t) - \ell_F (x_{t-1}), \tilde{q}_t := \ell_f (x_t) - \ell_f (x_{t-1}), \delta^F_t := \ell_F (x_t) - \ell_f (x_t) \) and \( \delta^G_t := G_{0,0} (x_t, \xi_t, u_t) + \sum_{i \in [m]} G_{i, \nu_t} (x_t, \xi_t, u_t) y_{i, t+1} - f'_{0, 0} (x_t) - \sum_{i \in [m]} f'_{1, \nu_t} (x_t) y_{i, t+1} \). where \( G_{0,0} \) and \( G_{i, \nu_t}, i \in [m] \) are the stochastic zeroth-order gradients defined in (2).

Proof of Lemma 9.4 Note that \( y_{t+1} = \arg \min_{y \geq 0} \langle -s_t, y \rangle + \frac{\gamma_t}{2} \|y - y_t\|_2^2 \). Hence, using Lemma 9.1 with \( y \mapsto \langle -s_t, y \rangle \) and \( \mu = 0 \), we have for all \( y \geq 0 \),
\[
-\langle s_t, y_{t+1} - y \rangle \leq \frac{\gamma_t}{2} \|y - y_t\|_2^2 - \|y_{t+1} - y_t\|_2^2 - \|y - y_{t+1}\|_2^2.
\]
Let us denote \( \nu_t := f'_{0,0} (x_t) + \sum_{i \in [m]} f'_{1, \nu_t} (x_t) y_{i, t+1} \) and \( V_t := G_{0,0} (x_t, \xi_t, u_t) + \sum_{i \in [m]} G_{i, \nu_t} (x_t, \xi_t, u_t) y_{i, t+1} \). Then using Lemma 9.1 with \( x \mapsto \langle V_t, x \rangle \) and the optimality of \( x_{t+1} \), we have for all \( x \in X \),
\[
\langle V_t, x_{t+1} - x \rangle \leq \eta_t [W(x, x_t) - W(x_{t+1}, x_t)] - \eta_t W(x, x_{t+1}).
\]

Due to the convexity of \( f_{0,0} \) and \( f_{i, \nu_t} \), and since \( f_0, f_i \) are Lipschitz, and by the definition of \( \ell_f \), and the fact that \( y_{t+1} \geq 0 \), we have
\[
\langle \nu_t, x_{t+1} - x \rangle = \langle f'_{0,0} (x_t) + \sum_{i \in [m]} f'_{1, \nu_t} (x_t) y_{i, t+1}, x_{t+1} - x \rangle
\]
\[
= \langle f'_{0,0} (x_t), x_{t+1} - x \rangle + \langle f'_{1, \nu_t} (x_t) y_{i, t+1}, x_{t+1} - x \rangle
\]
\[
\geq f_{0,0} (x_t) - f_{0,0} (x_t) + f_{0,0} (x_{t+1}) - f_{0,0} (x_t) - \frac{L_0}{2} \|x_{t+1} - x_t\|^2
\]
\[
+ \langle y_{t+1}, \ell_f (x_{t+1}) - f_{\nu} (x_t) \rangle + \langle y_{t+1}, f_{\nu} (x_t) - f_{\nu} (x) \rangle
\]
\[
= f_{0,0} (x_{t+1}) - f_{0,0} (x_t) + \langle \ell_f (x_{t+1}) - f_{\nu} (x), y_{t+1} \rangle - \frac{L_0}{2} \|x_{t+1} - x_t\|^2.
\]
Combining (70), (71), noting that $\delta_t^G = V_t - v_t$, we have
\[
f_0(x_{t+1}) - f_0(x) + \langle \ell_f(x_{t+1}) - f_v(x), y_{t+1} \rangle + \langle \delta_t^G, x_{t+1} - x \rangle \\
\leq \eta_t W(x, x_t) - \eta_t W(x, x_{t+1}) + O_{t+1}.
\] (72)

Noting the definition of $Q_v(\cdot, \cdot)$ (see (20)) and, adding (69) and (72), we obtain
\[
Q_v(z_{t+1}, z) - \langle f_v(x_{t+1}), y \rangle + \langle \ell_f(x_{t+1}), y_{t+1} \rangle - \langle s_t, y_{t+1} - y \rangle + \langle \delta_t^G, x_{t+1} - x \rangle \\
\leq \frac{\eta_t}{2} \left[ \| y - y_t \|^2 - \| y_{t+1} - y \|^2 - \| y - y_t \|^2 \right] \\
+ \eta_t W(x, x_t) - \eta_t W(x, x_{t+1}) + O_{t+1}.
\] (73)

Note that we also have $f_i(x_{t+1}) - \ell_f(x_{t+1}) \leq L_f^2 \| x_{t+1} - x_t \|^2$. Then, using Cauchy-Schwarz inequality and noting definitions of $L_f$, we have
\[
\langle y, f_v(x_{t+1}) - \ell_f(x_{t+1}) \rangle \leq \| y \|_2 \frac{L_f}{2} \| x_{t+1} - x_t \|^2.
\]

Noting the above relation and definitions of $q_t$ and $\delta_t^{F_{t+1}}$, we have
\[
\langle \ell_f(x_{t+1}), y_{t+1} \rangle - \langle f_v(x_{t+1}), y \rangle - \langle s_t, y_{t+1} - y \rangle \\
\geq \langle \ell_f(x_{t+1}), y_{t+1} \rangle - \langle \ell_f(x_{t+1}), y \rangle - \langle s_t, y_{t+1} - y \rangle - \| y \|_2 C_{t+1} \\
= \langle \ell_f(x_{t+1}) - s_t, y_{t+1} - y \rangle - \| y \|_2 C_{t+1} \\
= \langle \ell_f(x_{t+1}) - \ell_f(x_t) - \theta_t q_{t}, y_{t+1} - y \rangle - \| y \|_2 C_{t+1} \\
= \langle q_{t+1}, y_{t+1} - y \rangle - \theta_t \langle q_t, y_t - y \rangle - \theta_t \langle q_t, y_{t+1} - y_t \rangle - \langle \delta_t^{F_{t+1}}, y_{t+1} - y \rangle - \| y \|_2 C_{t+1}.
\] (74)

Then
\[
\| y \|_2 C_{t+1} = \frac{L_f}{2} (\| y \|_2 - 1) \| x_{t+1} - x_t \|^2 + \frac{L_f}{2} \| x_{t+1} - x_t \|^2 \\
\leq \frac{L_f}{2} (\| y \|_2 - 1) \| x_{t+1} - x_t \|^2 + \frac{L_f}{2} \| x_{t+1} - x_t \|^2 \\
\leq \frac{L_f}{2} \| x_{t+1} - x_t \|^2 + \frac{L_f D_X}{2} (\| y \|_2 - 1) \| x_{t+1} - x_t \|.\] (75)

By (73), (74), and (75), noting the definition of $O_{t+1}$ and using the relation $\frac{1}{2} \| a - b \|^2 \leq W(a, b)$, we have
\[
Q_v(z_{t+1}, z) - \theta_t \langle q_{t+1}, y_{t+1} - y \rangle - \langle \delta_t^{F_{t+1}}, y_{t+1} - y \rangle \\
\leq \theta_t \langle q_{t+1}, y_{t+1} - y \rangle - \langle \delta_t^{G}, x_{t+1} - x_t \rangle \\
+ \eta_t W(x, x_t) - \eta_t W(x, x_{t+1}) + \frac{\eta_t}{2} \left[ \| y - y_t \|^2 - \| y_{t+1} - y_t \|^2 - \| y - y_t \|^2 \right] \\
- (\eta_t - L_0 - L_f) W(x_{t+1}, x_t) + \frac{L_f D_X}{2} (\| y \|_2 - 1) \| x_{t+1} - x_t \|.
\] (76)
Multiplying (76) by $\gamma_t$, summing them up from $t = 0$ to $T - 1$ with $T \geq 1$, we obtain

$$
\sum_{t=0}^{T-1} \gamma_t q_t(z_{t+1}, z) + \sum_{t=0}^{T-1} \left[ \gamma_t \langle q_{t+1}, y_{t+1} - y_t \rangle - \gamma_t \theta_t(q_t, y_t) \right] + \sum_{t=0}^{T-1} \gamma_t \langle \delta_t^G, x_t - x \rangle - \langle \delta_{t+1}^F, y_{t+1} - y_t \rangle \\
\leq \sum_{t=0}^{T-1} \gamma_t \theta_t(q_t - q_t, y_{t+1} - y_t) + \gamma_t \theta_t(q_t, y_{t+1} - y_t) + \langle \gamma_t \delta_t^G, x_t - x_{t+1} \rangle \\
+ \sum_{t=0}^{T-1} \left[ \frac{\tau_t}{2} \| y_t \|^2 - \frac{\tau_t}{2} \| y_{t+1} \|^2 \right] - \sum_{t=0}^{T-1} \gamma_t \delta_t^G \| y_{t+1} - y_t \|^2 \\
+ \sum_{t=0}^{T-1} \gamma_t \eta_t W(x, x_t) - \gamma_t \eta_t W(x, x_{t+1})
$$

$$
- \sum_{t=0}^{T-1} \left[ \gamma_t (\eta_t - L_0 - L_f) W(x_{t+1}, x_t) - \gamma_t \left( \frac{L_f D}{2} \| y_t \|_2 \right) \| x_{t+1} - x_t \| \right] \right] + \mathcal{H}(y) (77),
$$

where $\mathcal{H}(y) := \frac{L_f D}{2} \| y_2 \|_2 - 1 + \|$. Now we focus our attention to handle the inner product terms of (77). Noting the definition of $\eta_t$, we have

$$
\| \tilde{q}_t \|_2 = \| \ell_f(x_t) - \ell_f(x_{t-1}) \|_2 \\
= \| f_\nu(x_{t-1}) + f_\nu(x_{t-1})^T(x_t - x_{t-1}) - f_\nu(x_t) - f_\nu(x_t)^T(x_t - x_{t-1}) \|_2 \\
\leq \| f_\nu(x_{t-1}) - f_\nu(x_t) \|_2 + \| f_\nu(x_{t-1})^T(x_t - x_{t-1}) \|_2 + \| f_\nu(x_t)^T(x_t - x_{t-1}) \|_2 \\
\leq 2M_f \| x_t - x_{t-1} \| + M_f \| x_t - x_{t-1} \|, \; (78)
$$

where we used the fact that $\| f_\nu(x) - f_\nu(y) \| \leq M_f \| x - y \|$ and $\| f_\nu(x)^T(y - x) \| \leq M_f \| y - x \|$, which follows from the Assumptions [2.2 and 2.3 and Theorem 7.1 see Nesterov and Spokoiny (2017)] for a similar argument.

Using the above relation for $\| \tilde{q}_t \|_2$, we now obtain

$$
\gamma_t \theta_t(q_t, y_{t+1} - y_t) - \frac{\gamma_t \tau_t}{3} \| y_{t+1} - y_t \|^2 - \frac{\gamma_t \tau_t}{4} \| \eta_t - L_0 - L_f \| W(x_{t+1}, x_t) \\
- \frac{\gamma_t \tau_t}{4} \| \eta_t - L_0 - L_f \| W(x_t, x_{t-1}) \\
\leq \gamma_t \theta_t \| \tilde{q}_t \|_2 \| y_{t+1} - y_t \|^2 - \frac{\gamma_t \tau_t}{3} \| y_{t+1} - y_t \|^2 \\
- \frac{\gamma_t \tau_t}{4} \| \eta_t - L_0 - L_f \| W(x_{t+1}, x_t) - \frac{\gamma_t \tau_t}{4} \| \eta_t - L_0 - L_f \| W(x_t, x_{t-1}) \\
\leq 2M_f \gamma_t \theta_t \| x_t - x_{t-1} \| \| y_{t+1} - y_t \|^2 - \frac{\gamma_t \tau_t}{6} \| y_{t+1} - y_t \|^2 - \frac{\gamma_t \tau_t}{4} \| \eta_t - L_0 - L_f \| W(x_{t+1}, x_t) \\
+ M_f \gamma_t \theta_t \| x_t - x_{t-1} \| \| y_{t+1} - y_t \|^2 - \frac{\gamma_t \tau_t}{6} \| y_{t+1} - y_t \|^2 - \frac{\gamma_t \tau_t}{4} \| \eta_t - L_0 - L_f \| W(x_t, x_{t-1}) \\
\leq 0, \; (80)
$$

where the last inequality follows by applying the relation $W(x, y) \geq \frac{1}{2} \| x - y \|$, Young’s inequality ($2ab \leq a^2 + b^2$) applied twice, once with

$$
a = \left( \frac{\gamma_t \tau_t}{6} \right)^{1/2} \| y_{t+1} - y_t \|, \; b = \left( \frac{\gamma_t \tau_t}{8} \right)^{1/2} \| x_{t+1} - x_{t-2} \|
$$

and second time with

$$
a = \left( \frac{\gamma_t \tau_t}{6} \right)^{1/2} \| y_{t+1} - y_t \|, \; b = \left( \frac{\gamma_t \tau_t}{8} \right)^{1/2} \| x_{t} - x_{t-1} \|, \; (80)
$$

36
and the fact that
\[
2M_f \gamma_t \theta_t \leq \left( \frac{\gamma_t \gamma_{t-2} \tau_t (\eta_{t-2} - L_0 - L_f)}{12} \right)^{1/2} \quad \Leftrightarrow \quad (2M_f)^2 \frac{\theta_t}{\theta_{t-1}} \leq \frac{\tau_t (\eta_{t-2} - L_0 - L_f)}{12},
\]
\[
M_f^2 \gamma_t^2 \theta_t^2 \leq \left( \frac{\gamma_t \eta_{t-2} \tau_t (\eta_{t-1} - L_0 - L_f)}{12} \right)^{1/2} \quad \Leftrightarrow \quad M_f^2 \theta_t \leq \frac{\tau_t (\eta_{t-1} - L_0 - L_f)}{12},
\]
where the equivalences follow due to (66). Using Young’s inequality, Cauchy-Schwarz inequality and the relation \( u^T v \leq \|u\| \|v\| \), we have
\[
\gamma_t \theta_t (q_t - \tilde{q}_t, y_{t-1} + y_t) - \frac{\gamma_t \tau_t}{6} \| y_{t-1} - y_t \|_2^2 \leq \frac{3 \gamma_t \theta_t^2}{2 \tau_t} \| q_t - \tilde{q}_t \|_2^2,
\]
\[
\langle \gamma_t \delta^G_{t}, x_t - x_{t+1} \rangle - \frac{\gamma_t (\eta_t - L_0 - L_f)}{4} W(x_{t+1}, x_t) \leq \frac{2 \gamma_t}{\eta_t - L_0 - L_f} \| \delta^G_{t} \|, \quad \gamma_t \mathcal{H}(y) \| x_{t+1} - x_t \| - \frac{\gamma_t (\eta_t - L_0 - L_f)}{4} W(x_{t+1}, x_t) \leq \frac{2 \gamma_t}{\eta_t - L_0 - L_f} \mathcal{H}(y). \quad (81)
\]
Using (80) and (81) for \( t = 0, \ldots, T - 1 \) inside (77) and noting (66), we have
\[
\sum_{t=0}^{T-1} \gamma_t Q_{\nu_t}(z_{t+1}, z) + \gamma_{T-1} \langle q_T, y_T - y \rangle + \gamma_{T-1} \eta_{T-1} W(x_{T-1}, x_T) + \sum_{t=0}^{T-1} \gamma_t \| \delta^G_{t} \| - \| \delta_{t+1}^F, x_t - y \| - \| \delta_{t+1}^F, y_{t+1} - y \| \leq \gamma_0 \gamma_0 W(x, x_0) - \gamma_{t-1} \eta_{t-1} W(x, x_T) + \sum_{t=0}^{T-1} \gamma_t \| y - y_0 \|_2 - \frac{\gamma_{T-1} \tau_{T-1}}{2} \| y - y_T \|_2^2
\]
\[
+ \frac{3 \gamma_t \theta_t^2}{2 \tau_t} \| q_t - \tilde{q}_t \|_2^2 + \frac{2 \gamma_t}{\eta_t - L_0 - L_f} \| \delta^G_{t} \|_2 + \frac{2 \gamma_t}{\eta_t - L_0 - L_f} \mathcal{H}(y) \]
\[
- \frac{\gamma_{T-2} (\eta_{T-2} - L_0 - L_f)}{4} W(x_{T-1}, x_{T-2}) - \frac{\gamma_{T-1} (\eta_{T-1} - L_0 - L_f)}{2} W(x_T, x_{T-1}), \quad (82)
\]
where in the left hand side of the above relation, we used the fact that \( q_0 = \ell_F(x_0) - \ell_F(x_{-1}) = 0 \). Similarly, we see that \( \tilde{q}_0 = 0 \). Hence, we can ignore \( \| q_0 - \tilde{q}_0 \|_2^2 \) term in the right hand side of the above relation, after which we obtain
\[
- \gamma_{t-1} \langle q_T, y_T - y \rangle - \frac{\gamma_{T-1} \tau_{T-1}}{3} \| y - y_T \|_2^2
\]
\[
- \frac{\gamma_{T-2} (\eta_{T-2} - L_0 - L_f)}{4} W(x_{T-1}, x_{T-2}) - \frac{\gamma_{T-1} (\eta_{T-1} - L_0 - L_f)}{2} W(x_T, x_{T-1}) \leq M_f \gamma_{T-1} \| x_T - x_{T-1} \| \| y_T - y \|_2 - \frac{\gamma_{T-1} \tau_{T-1}}{12} \| y - y_T \|_2^2 - \frac{\gamma_{T-1} (\eta_{T-1} - L_0 - L_f)}{2} W(x_T, x_{T-1}) + 2 M_f \gamma_{T-1} \| x_{T-1} - x_T - 2 \| \| y_T - y \|_2 - \frac{\gamma_{T-1} \tau_{T-1}}{6} \| y - y_T \|_2^2 - \frac{\gamma_{T-2} (\eta_{T-2} - L_0 - L_f)}{4} W(x_{T-1}, x_{T-2})
\]
\[
- \frac{\gamma_{T-1} \tau_{T-1}}{12} \| y_T - y \|_2^2 \leq - \frac{\gamma_{T-1} \tau_{T-1}}{12} \| y_T - y \|_2^2, \quad (83)
\]
where the last relation follows from (67). Young’s inequality and the fact that
\[
2M_f \gamma_{T-1} \leq \left( \frac{\gamma_{T-2} \tau_{T-1} (\eta_{T-2} - L_0 - L_f)}{12} \right)^{1/2} \quad \Leftrightarrow \quad (2M_f)^2 \frac{1}{\theta_{T-1}} \leq \frac{\tau_{T-1} (\eta_{T-2} - L_0 - L_f)}{12},
\]
\[
M_f \gamma_{T-1} \leq \left( \frac{\gamma_{T-2} \tau_{T-1} (\eta_{T-1} - L_0 - L_f)}{12} \right)^{1/2} \quad \Leftrightarrow \quad M_f^2 \leq \frac{\tau_{T-1} (\eta_{T-1} - L_0 - L_f)}{12}.
\]
Moreover, again using Young’s inequality and Cauchy-Schwarz inequality, we have

\[-\gamma_{T-1}(q_T - \bar{q}_T, y_T - y) - \frac{\gamma_{T-1}\tau_{T-1}}{6}\|y - y_T\|_2^2 \leq \frac{3\gamma_{T-1}}{2\tau_{T-1}}\|q_T - \bar{q}_T\|_2^2.\] (84)

Using (83) and (84) in relation (82), noting that \(q_0 - \bar{q}_0 = 0\) and replacing the definition of \(H(y)\), we obtain (68), which completes the proof. \(\square\)