

# Stochastic Sequential Quadratic Programming for Optimization with Functional Constraints

Panchajanya Sanyal\*, Srujan Teja Thomdapu\*, and Ketan Rajawat

**Abstract**—Stochastic convex optimization problems with non-linear functional constraints are ubiquitous in machine learning applications, including multi-task learning, structured prediction, and multi-view learning. The presence of non-linear functional constraints renders the traditional projected stochastic gradient descent and related projection-based methods inefficient, and motivates the use of first-order methods. However, existing first-order methods, including primal and primal-dual algorithms, typically rely on a bounded (sub-)gradient assumption, which may be too restrictive in many settings.

We propose a stochastic sequential quadratic programming (SSQP) algorithm that works entirely in the primal domain, avoids projecting onto the feasible region, obviates the need for bounded gradients, and achieves state-of-the-art oracle complexity under standard smoothness and convexity assumptions. A faster version, namely SSQP-Skip, is also proposed where the quadratic subproblems can be skipped in most iterations. Finally, we develop an accelerated variance-reduced version of SSQP (VARAS), whose oracle complexity bounds match those for solving unconstrained finite-sum convex optimization problems. The superior performance of the proposed algorithms is demonstrated via numerical experiments on real datasets.

**Index Terms**—Stochastic optimization, functional constraints, sequential quadratic programming, first-order methods, variance reduction.

## I. INTRODUCTION

We consider the constrained optimization problem

$$\begin{aligned} \mathbf{x}_\star &= \arg \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) + h(\mathbf{x}), \\ \text{s. t.} \quad &g_k(\mathbf{x}) \leq 0, \quad 1 \leq k \leq m \end{aligned} \quad (\mathcal{P})$$

where  $f(\mathbf{x}) := \mathbb{E}_t[f_{i_t}(\mathbf{x})]$  and  $\mathbb{E}_t[\cdot]$  denotes the expectation with respect to the random index  $i_t$ . We will also consider a finite-sum case, which arises when the index  $i_t$  is sampled uniformly from  $\{1, \dots, n\}$ , so that  $f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$ . The functions  $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$  and  $g_k : \mathbb{R}^d \rightarrow \mathbb{R}$  are proper, closed, convex, and  $L$ -smooth. The regularization function  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  is convex but possibly non-smooth, and may include an indicator function corresponding to a set-inclusive constraint of the form  $\mathbf{x} \in \mathcal{K}$  for a closed convex set  $\mathcal{K}$ . The stochastic objective function in  $(\mathcal{P})$  commonly arises in the context of stochastic approximation and online learning. The finite-sum structure in particular is widely used in supervised learning tasks where the goal is to minimize the empirical risk over  $n$  training samples. Non-linear functional constraints similarly arise in many learning tasks, such as *multi-task learning*, where the constraints capture the relationships between tasks [1], *structured prediction*, where the constraints may encode structural prior knowledge about the feature space [2],

and *semi-supervised multi-view object recognition*, where the constraints model the relationship between views [3].

We consider the high-dimensional setting, where for a given  $\mathbf{x}$ , a stochastic first-order oracle (SFO) provides us with  $\nabla f_i(\mathbf{x})$  for a randomly selected index  $i$  as well as  $\{g_k(\mathbf{x}), \nabla g_k(\mathbf{x})\}_{k=1}^m$ . In such settings, classical methods such as the projected stochastic gradient descent (SGD) lose their efficacy, since each iteration requires projection onto a feasible region defined by functional constraints, which itself is a complicated operation. Instead, efficient and scalable algorithms for solving  $(\mathcal{P})$  must rely only on the first-order information provided by the oracle. The *SFO complexity* of an algorithm is defined as the number of SFO calls required to achieve an  $\epsilon$ -optimal solution, which may be a random vector  $\mathbf{x} \in \mathbb{R}^d$  such that

$$\mathbb{E}[f(\mathbf{x}) + h(\mathbf{x})] - f(\mathbf{x}_\star) - h(\mathbf{x}_\star) \leq \epsilon, \quad (1)$$

$$\sum_{k=1}^m \mathbb{E}[(g_k(\mathbf{x}))_+] \leq \epsilon. \quad (2)$$

For strongly convex objectives, we will directly characterize the complexity as the number of SFO calls required to ensure that  $\mathbb{E} \|\mathbf{x} - \mathbf{x}_\star\|^2 \leq \epsilon$  for a random  $\mathbf{x}$ .

State-of-the-art first-order algorithms for solving  $(\mathcal{P})$  and its variants, like constrained online convex optimization (COCO), include primal algorithms [4]–[8] as well as primal-dual algorithms [9]–[15]. Primal algorithms can be classified into two main categories: (a) stochastic versions of Polyak’s subgradient method [5], [7], [8] that switch between either minimizing the objective function or reducing the infeasibility at every iteration; and (b) composite approaches [4], [6], that involve both a proximal gradient step to reduce the objective and a sub-gradient step to reduce the constraint violation at each iteration. On the other hand, the primal-dual methods in [9]–[15] seem to follow a common template of updating both a primal and a dual variable at every iteration. When applied to solve  $(\mathcal{P})$ , these approaches achieve an SFO complexity of  $\mathcal{O}(1/\epsilon^2)$  for the convex case and  $\mathcal{O}(1/\epsilon)$  for the strongly convex case. Additionally, all existing results require the boundedness of the (sub-)gradients of the objective function, a condition which may fail for common objectives such as the least-squares loss function or may be difficult to verify in practice. To the best of our knowledge, no existing method attains optimal SFO complexity for solving  $(\mathcal{P})$  while simultaneously avoiding both projection steps and bounded (sub-)gradient assumptions.

This work puts forth a new class of stochastic sequential quadratic programming (SSQP) algorithms for solving  $(\mathcal{P})$ . The key idea behind the proposed approach is to reformulate

\*These authors contributed equally.

( $\mathcal{P}$ ) as an unconstrained non-smooth optimization problem using the exact penalty method [16], [17] and to solve it via the prox-linear algorithm [18] using only stochastic first-order information. In the present case, each iteration reduces to solving a diagonal quadratic program (QP), instead of a full projection or general convex subproblems, while still achieving (near-)optimal SFO complexity and avoiding any bounded (sub-)gradient assumptions. Our main algorithmic contributions are: (i) a vanilla SSQP algorithm with SFO complexity matching state-of-the-art primal–dual algorithms for solving ( $\mathcal{P}$ ); (ii) an SSQP-Skip algorithm that solves the proximal and QP subproblems only infrequently, while retaining the same SFO complexity guarantees; and (iii) a first accelerated variance-reduced algorithm, VARAS, for finite-sum constrained optimization problems, whose SFO complexity is nearly on par with the best known methods in the unconstrained setting. Central to these algorithms is a novel one-step inequality that leads to the required optimality gap and constraint violation bounds. The numerical performance of the proposed variants is also tested on two real-world problems and substantially outperforms representative baselines for solving ( $\mathcal{P}$ ). The proposed algorithms differ fundamentally from classical sequential QP methods, which require full gradient and Hessian information at each iteration [19, Sec. 4.3.1] and are therefore ill-suited to large-scale stochastic settings; in contrast, the proposed SSQP methods operate with only stochastic first-order information.

#### A. Related Work

Sequential quadratic programming (SQP) methods are among the most effective approaches for solving nonlinear optimization problems of the form ( $\mathcal{P}$ ) [20]. Conventional SQP methods rely on second-order derivatives of  $f_i$  and  $\{g_k\}_{k=1}^m$  to solve a sequence of QP problems subject to linearized constraints [16], and have been widely applied to mixed-integer nonlinear programming and nonlinear optimization with nonlinear equality constraints; see e.g. [20], [21].

The exact penalty reformulation of constrained problems has been well studied in convex optimization, and a sequential QP approach is detailed in [19, Sec. 4.3.1]. The objective of the exact-penalty reformulation can be viewed as a compositional optimization problem, and in the deterministic case it can be solved using the approach of [22]. The corresponding stochastic problems can be addressed using the model-based framework of [23] or the prox-linear method of [18]. A different exact-penalty method was proposed in [24], where the focus is on non-convex equality constraints, leading to an oracle complexity of  $\mathcal{O}(\epsilon^{-3.5})$  for finding an  $\epsilon$ -stationary point. Using smooth approximations of exact penalties, so as to work with their gradients, is another alternative but typically yields weaker rates [25], [26]. Our work can be viewed as bringing these exact-penalty and prox-linear ideas into a stochastic SQP framework with explicit oracle-complexity guarantees.

Convex optimization problems with functional constraints have been extensively studied; see [27] and references therein. When the constraint sets are difficult to project onto, projected

SGD variants can be computationally expensive. The number of projections has been reduced to  $\mathcal{O}(\log T)$  in [28], [29] and to a single projection in [30], though these schemes can still be impractical when  $m$  is large. Subsequent works that completely avoid projections include primal–dual methods [9]–[15] and primal methods [4]–[8]. To the best of our knowledge, these projection-free methods have not been accelerated to attain the optimal rates that are possible in the unconstrained setting. The proposed work can be viewed as lying between these two classes: it avoids projections, instead solving a diagonal QP with linear constraints, while achieving (near-)optimal rates at par with accelerated variance-reduced projected-SGD [31]. In addition, and unlike many of the above approaches, our analysis does not require boundedness of the gradients of the constituent functions.

Convex optimization problems with linear inclusive constraints have been studied in [32]–[34]; in contrast, we focus here on nonlinear functional constraints. Our formulation also differs from those in [5], [7], [11], [12], [14], [15], [35], [36], where the constraints are only required to hold on average, i.e.,  $\mathbb{E}[f_i(\mathbf{x}, \xi)] \leq 0$ . Related stochastic formulations with more general stochastic function classes are considered in [26], [37], [38]. Other works address problems with infinitely many functional constraints; see [4], [6], [9], [10], [39]. The best convergence rates obtained in these papers are of order  $\mathcal{O}(1/\epsilon^2)$  for convex objectives and  $\mathcal{O}(1/\epsilon)$  for strongly convex objectives. These rates have only been improved in [40], [41] for a specific formulation imposing additional structure on the constraint functions. Finally, [42], [43] study related problems from the perspective of constrained online convex optimization (COCO), where the lack of stationarity assumptions leads to more conservative bounds. In contrast, in this paper we propose a new method for problems of the form ( $\mathcal{P}$ ) and establish (near-)optimal oracle-complexity rates for both convex and strongly convex objectives under our setting.

We remark that among these, [4], [6], [9], [39] adopt a different SFO model for ( $\mathcal{P}$ ), wherein at iteration  $t$ , only  $\{g_{j_t}(\mathbf{x}), \nabla g_{j_t}(\mathbf{x})\}$  for random index  $j_t$ , is revealed. While this greatly reduces the per-iteration cost, these algorithms require strong regularity assumptions that couple individual constraint violations with the distance to the full feasible set (see, e.g., [4, Assumption 4]), as well as bounded subgradients of both the objective and the constraints. Their resulting convergence rates match those of stochastic subgradient methods and depend explicitly on the regularity constant of the constraint system. In contrast, the proposed SSQP method uses all functional constraint gradients at every iteration, and therefore does not rely on such regularity conditions, while attaining (near-)optimal oracle complexity. Nevertheless, a single-constraint variant of SSQP with improved rates may be an interesting direction for future work.

A comparison of the proposed methods with the most relevant state-of-the-art algorithms is summarized in Table I. The table excludes works such as [21], which provide only asymptotic convergence guarantees, but includes bounds for online algorithms adapted to the present setting. We also omit the non-convex, distributed extension of our approach presented in [44].

TABLE I  
RELATED WORKS SOLVING  $(\mathcal{P})$  WITH STATE-OF-THE-ART COMPLEXITY

Ref	Method class	SFO Complexity	
		Convex	Strongly convex
[30]	primal-dual	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right)$
[11], [12]	primal-dual	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	-
[13]	primal-dual (online)	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right)$
[15]	primal	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	-
[5]	primal	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
[8]	mirror descent	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
[42]	SQP (online)	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{1}{\epsilon}\right)$
SSQP (Ours)	SQP	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
SSQP-Skip (Ours)	SQP	-	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
VARAS (Ours)	SQP	$\mathcal{O}\left(\sqrt{\frac{n}{\epsilon}}\right)$	$\mathcal{O}\left(n \log n + \sqrt{n} \log \frac{1}{\epsilon}\right)$

## B. Organization

This paper is organized as follows. Sec. II provides some preliminaries, including the assumptions, background, and basic inequalities. Sec. III details the proposed SSQP and SSQP-Skip algorithms for solving the general stochastic version of  $(\mathcal{P})$ , and provides their oracle complexity bounds. Sec. IV develops the accelerated variance-reduced SSQP algorithm for the finite-sum version of  $(\mathcal{P})$  and provides the corresponding oracle complexity bounds. The numerical performance of the proposed class of algorithms is provided in Sec. V and finally, Sec. VI concludes the paper.

## C. Notation

We use regular (bold)-faced letters to represent scalars (column vectors). We let  $[v]_+ := \max\{0, v\}$ , so that  $\max\{[v_k]_+\} = \max\{[v_1]_+, \dots, [v_m]_+\} = \max\{0, v_1, \dots, v_m\}$ . The Euclidean norm of a vector  $\mathbf{x}$  is denoted by  $\|\mathbf{x}\|$  and  $\mathbb{E}[\cdot]$  denotes the expectation operator.

## II. PRELIMINARIES

This section contains the assumptions and some technical claims that we use throughout the analysis. Other basic mathematical inequalities that are used throughout the text are listed in Appendix A.

### A. Assumptions

**A1.** The Slater condition holds for  $(\mathcal{P})$ , i.e., there exists a feasible  $\tilde{\mathbf{x}}$  such that

$$g_k(\tilde{\mathbf{x}}) \leq -\nu < 0, \quad 1 \leq k \leq m. \quad (3)$$

Additionally, we assume that the optimality gap at the Slater point is bounded, i.e.,  $f(\tilde{\mathbf{x}}) + h(\tilde{\mathbf{x}}) - f(\mathbf{x}_*) - h(\mathbf{x}_*) \leq \tilde{B}$ .

In the context of constrained optimization, the Slater constraint qualification (CQ) is one of the simplest and most widely used CQs that imply strong duality and existence of a primal-dual optimum pair  $(\mathbf{x}_*, \boldsymbol{\lambda}_*)$ . Since  $(\mathcal{P})$  is convex,

the optimum pair satisfies the Karush-Kuhn-Tucker (KKT) conditions, so that

$$f(\mathbf{x}_*) + h(\mathbf{x}_*) = \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) + h(\mathbf{x}) + \sum_{k=1}^m \lambda_{k,*} g_k(\mathbf{x}) \quad (4)$$

$$\leq f(\tilde{\mathbf{x}}) + h(\tilde{\mathbf{x}}) - \|\boldsymbol{\lambda}_*\|_1 \nu \quad (5)$$

for the Slater point  $\tilde{\mathbf{x}}$ . Rearranging and using Assumption **A1**, we obtain the bound  $\|\boldsymbol{\lambda}_*\|_1 \leq \frac{\tilde{B}}{\nu}$ . In practice,  $\nu$  and  $\tilde{B}$  are problem parameters that must be found by parameter tuning. For example, suppose that we have found a Slater point  $\tilde{\mathbf{x}}$  so that  $\nu = -\max_k g_k(\tilde{\mathbf{x}})$ . Then if we can also find the unconstrained minimum  $\mathbf{x}_u = \arg \min_{\mathbf{x} \in \mathbb{R}^d} (f(\mathbf{x}) + h(\mathbf{x}))$ , we can set  $\tilde{B} = f(\tilde{\mathbf{x}}) + h(\tilde{\mathbf{x}}) - f(\mathbf{x}_u) - h(\mathbf{x}_u)$ .

The next few assumptions define the other problem parameters used for the analysis. The first set of assumptions is standard.

**A2.** The following assumptions hold:

- 1) The functions  $f_i$  and  $g_k$  are proper, closed, and convex.
- 2) The functions  $f_i$  are  $L_f$ -smooth while the functions  $g_k$  are  $L_g$ -smooth.
- 3) The function  $f$  is  $\mu$ -strongly convex for  $\mu \geq 0$ .

The strong convexity assumption may not be invoked for some of the proposed algorithms, for which we will simply set  $\mu = 0$ . The following assumption will be required for the general stochastic optimization problem, but will be dropped for the finite-sum case.

**A3.** The gradient noise at the optimum  $\mathbf{x}_*$  is bounded as  $\mathbb{E}_t[\|\nabla f(\mathbf{x}_*) - \nabla f_{i_t}(\mathbf{x}_*)\|^2] \leq \sigma^2$ .

The bounded gradient noise assumption is again standard in the literature and is always required for SGD-like algorithms in general. Note that we only require the gradient noise to be bounded at a specific point  $\mathbf{x}_*$  rather than for the entire domain  $\mathcal{K}$ . Assumption **A3** along with the smoothness of  $f_{i_t}$  implies that for a given  $\mathbf{x}$ ,

$$\mathbb{E}_t[\|\nabla f_{i_t}(\mathbf{x}) - \nabla f(\mathbf{x})\|^2] \quad (6)$$

$$= \mathbb{E}_t[\|\nabla f_{i_t}(\mathbf{x}) - \nabla f(\mathbf{x}_*)\|^2] - \mathbb{E}_t[\|\nabla f(\mathbf{x}_*) - \nabla f(\mathbf{x})\|^2]$$

$$\stackrel{(45)}{\leq} 2\mathbb{E}_t[\|\nabla f_{i_t}(\mathbf{x}_*) - \nabla f_{i_t}(\mathbf{x})\|^2]$$

$$+ 2\mathbb{E}_t[\|\nabla f(\mathbf{x}_*) - \nabla f_{i_t}(\mathbf{x}_*)\|^2]$$

$$\stackrel{(A3)}{\leq} 2\mathbb{E}_t[\|\nabla f_{i_t}(\mathbf{x}_*) - \nabla f_{i_t}(\mathbf{x})\|^2] + 2\sigma^2 \quad (7)$$

$$\stackrel{(42)}{\leq} 4L_f \mathbb{E}_t[f_{i_t}(\mathbf{x}_*) - f_{i_t}(\mathbf{x}) - \langle \nabla f_{i_t}(\mathbf{x}), \mathbf{x}_* - \mathbf{x} \rangle] + 2\sigma^2$$

$$\leq 4L_f D_f(\mathbf{x}_*, \mathbf{x}) + 2\sigma^2, \quad (8)$$

where  $D_f(\mathbf{u}, \mathbf{v}) := f(\mathbf{u}) - f(\mathbf{v}) - \langle \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle$  and we have used some basic inequalities from Appendix A. The inequality in (8) bounds the gradient noise at arbitrary  $\mathbf{x}$  in terms of the Bregman divergence between  $\mathbf{x}$  and  $\mathbf{x}_*$  (with respect to  $f$ ) and the gradient noise at  $\mathbf{x}_*$ , and will turn out to be useful later on. Finally, we have the following initialization condition.

**A4.** All algorithms can be initialized with arbitrary, possibly infeasible  $\mathbf{x}_0 \in \mathcal{K}$  which satisfies  $\|\mathbf{x}_0 - \mathbf{x}_*\| \leq B_x$  and

$$f(\mathbf{x}_0) + h(\mathbf{x}_0) + \gamma \sum_{k=1}^m [g_k(\mathbf{x}_0)]_+ - f(\mathbf{x}_*) - h(\mathbf{x}_*) \leq B_\gamma$$

for a given  $\gamma > 0$ .

### B. Exact Penalty Reformulation

We reformulate the problem using the exact penalty method [19, Sec. 4.3.1], so as to obtain:

$$\begin{aligned} \mathbf{x}_* &= \arg \min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) := f(\mathbf{x}) + h(\mathbf{x}) + \gamma \max\{[g_k(\mathbf{x})]_+\} \\ &= \arg \min_{\mathbf{x} \in \mathbb{R}^d, v \geq 0} f(\mathbf{x}) + h(\mathbf{x}) + \gamma v \\ \text{s. t. } &g_k(\mathbf{x}) \leq v, \quad 1 \leq k \leq m \end{aligned} \quad (\mathcal{P}_1)$$

where recall that  $\max\{[v_k]_+\} = \max\{0, v_1, v_2, \dots, v_m\}$ . In general, the solution of  $(\mathcal{P}_1)$  is the same as that of  $(\mathcal{P})$  for sufficiently large  $\gamma$ . Specifically, under Assumption **A1**, it suffices to set  $\gamma \geq \frac{\tilde{B}}{\nu}$ . To see this, associate dual variables  $\mu_k \geq 0$  with the  $k$ -th constraint in  $(\mathcal{P}_1)$ , so that the Lagrangian becomes:

$$L(\mathbf{x}, v, \boldsymbol{\mu}) = f(\mathbf{x}) + h(\mathbf{x}) + \gamma v + \sum_{k=1}^m \mu_k (g_k(\mathbf{x}) - v) \quad (10)$$

$$= f(\mathbf{x}) + h(\mathbf{x}) + \sum_{k=1}^m \mu_k g_k(\mathbf{x}) + v(\gamma - \|\boldsymbol{\mu}\|_1) \quad (11)$$

where  $\boldsymbol{\mu} \in \mathbb{R}_+^m$  collects the dual variables  $\{\mu_k\}_{k=1}^m$ . Since the Slater CQ is satisfied by  $(\mathcal{P})$ , it is also satisfied by  $(\mathcal{P}_1)$ . Therefore, the first order KKT point  $(\mathbf{x}_*, v_*, \boldsymbol{\mu}_*)$  is such that

$$\begin{aligned} (\mathbf{x}_*, v_*) &= \arg \min_{\mathbf{x} \in \mathbb{R}^d, v \geq 0} f(\mathbf{x}) + h(\mathbf{x}) + \sum_{k=1}^m \mu_{k,*} g_k(\mathbf{x}) \\ &\quad + v(\gamma - \|\boldsymbol{\mu}_*\|_1) \end{aligned} \quad (12)$$

Hence, for  $\gamma = \frac{\tilde{B}}{\nu} \geq \|\boldsymbol{\mu}_*\|_1$ , it follows that  $v_* = 0$  and consequently  $(\mathbf{x}_*, \boldsymbol{\mu}_*)$  is KKT-optimal for  $(\mathcal{P})$ .

In addition to characterizing the SFO complexity, we observe that the SQP methods require solving a QP with  $m$  linear constraints at every iteration. Hence for this class of algorithms, we assume access to the quadratic minimization oracle (QMO) which can provide the solution to a given QP with  $m$  linear constraints. In this case, for general convex objectives, we will characterize the performance of the algorithms in terms of the number of SFO and QMO calls required to achieve an  $\epsilon$ -optimal solution.

The exact penalty reformulation confirms that for any solution  $\mathbf{x}_*$  of  $(\mathcal{P})$ ,  $\Delta_t := \mathbb{E}[F(\mathbf{x}_t)] - F(\mathbf{x}_*)$  is a non-negative quantity. All subsequent theorems will involve an intermediate step of upper bounding  $\Delta_t$  or related quantities, which will then lead to the required SFO and QMO complexity results for (1). Finally, observe that from the Markov's inequality, the  $\epsilon$ -optimal point  $\mathbf{x}$  is such that  $\sum_{k=1}^m \mathbb{P}(g_k(\mathbf{x}) > \kappa) \leq \frac{\epsilon}{\kappa}$  for any  $\kappa > 0$ . For instance, if we set  $\kappa = \sqrt{\epsilon}$  for some  $\epsilon < 1$ , it follows that the total probability of  $\sqrt{\epsilon}$ -constraint violations is at most  $\sqrt{\epsilon}$ . Further, it can be seen that if we run the algorithm several times, the smallest of these violations will be small with high probability.

## III. STOCHASTIC SEQUENTIAL QUADRATIC PROGRAMMING METHOD

In this section, we consider the general stochastic problem in  $(\mathcal{P})$ . Reformulating the problem as  $(\mathcal{P}_1)$  makes it amenable to the application of stochastic proximal gradient methods. Specifically, we utilize the stochastic prox-linear algorithm from [18], [23], [45], [46] to develop the proposed SSQP algorithm. Throughout this section, we will focus on obtaining state-of-the-art rates but ignore constants or higher-order terms. Future work may target sharper constants and lower bounds under this oracle.

### A. SSQP Algorithm

The SSQP algorithm entails performing a partial linearization of the objective in  $(\mathcal{P}_1)$ , adding a proximal penalty, and minimizing the resulting quadratic form. Specifically, the objective and constraint functions are linearized, but the regularizer, as well as the  $\max\{[\cdot]_+\}$  operator are not disturbed. Starting at an arbitrary  $\mathbf{x}_0$ , the updates of the proposed SSQP algorithm take the form:

$$\begin{aligned} \mathbf{x}_{t+1} &= \arg \min_{\mathbf{u} \in \mathbb{R}^d} \langle \nabla f_{i_t}(\mathbf{x}_t), \mathbf{u} \rangle + h(\mathbf{u}) + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{u}\|^2 \\ &\quad + \gamma \max\{[g_k(\mathbf{x}_t) + \langle \nabla g_k(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_t \rangle]_+\} \\ &= \arg \min_{\mathbf{u} \in \mathbb{R}^d, v \geq 0} \langle \nabla f_{i_t}(\mathbf{x}_t), \mathbf{u} \rangle + h(\mathbf{u}) + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{u}\|^2 + \gamma v \\ \text{s. t. } &g_k(\mathbf{x}_t) + \langle \nabla g_k(\mathbf{x}_t), \mathbf{u} - \mathbf{x}_t \rangle \leq v, \quad 1 \leq k \leq m. \end{aligned} \quad (14)$$

for  $t \geq 0$ , where  $i_t$  is a random index. Observe that when  $\mathcal{K} = \mathbb{R}^d$ , the updates bear resemblance to the sequential quadratic programming approach proposed in [47, Sec. 4.3.1], and hence we refer to our algorithm, summarized in Algorithm 1, as Stochastic SQP (SSQP).

---

#### Algorithm 1 SSQP

---

- 1: **Input:**  $\mathbf{x}_0 \in \mathcal{K}$ ,  $\gamma = \tilde{B}/\nu$ , and  $\eta_t \in (0, 1]$ .
  - 2: **for**  $t = 0, 1, \dots, T-1$
  - 3:   Sample  $i_t$  randomly
  - 4:   Update  $\mathbf{x}_{t+1}$  using (14)
  - 5: **end for**
  - 6: **Output:**  $\bar{\mathbf{x}}_T = \frac{\sum_{t=1}^T \eta_t \mathbf{x}_t}{\sum_{t=1}^T \eta_t}$ .
- 

Since SSQP is a special case of the stochastic prox-linear algorithm, the  $\mathcal{O}(1/\sqrt{t})$  convergence result for convex objectives follows from [23]. However, the generality of the prox-linear algorithm leads to relatively weaker bounds and requires stronger assumptions. Below, we provide a tighter bound which does not require the bounded gradients assumption commonly required for analyzing prox-linear algorithms.

The convergence of Algorithm 1 is established in the statement of the following theorem, whose proof is provided in Appendix B.

**Theorem 1.** Under Assumptions **A1-A4**,  $L = \max\{\gamma L_g, L_f\}$ , and  $\delta_0 = \|\mathbf{x}_0 - \mathbf{x}_*\|^2$ , we have the following SFO and QMO complexity bounds.

1) For a convex objective, using the stepsize  $\eta_t = \frac{\eta_0}{\sqrt{T}}$ , where  $\eta_0 = \min\{\frac{\sqrt{\delta_0}}{2\sigma}, \frac{1}{4L}\}$ , we obtain

$$\begin{aligned} \mathbb{E}[f(\bar{\mathbf{x}}_T) + h(\bar{\mathbf{x}}_T)] - f(\mathbf{x}_*) - h(\mathbf{x}_*) \\ \leq \frac{2}{\sqrt{T}} \max\{2L\delta_0, \sigma\sqrt{\delta_0}\} \end{aligned} \quad (16)$$

$$\mathbb{E}\left[\max_k\{[g_k(\bar{\mathbf{x}}_T)]_+\}\right] \leq \frac{2}{(\gamma - \frac{\bar{B}}{\nu})\sqrt{T}} \max\{2L\delta_0, \sigma\sqrt{\delta_0}\} \quad (17)$$

and an SFO/QMO complexity of  $\mathcal{O}\left(\frac{\max\{\delta_0^2 L^2, \sigma^2 \delta_0\}}{\epsilon^2}\right)$ .

2) For a  $\mu$ -strongly convex objective, the stepsize  $\eta_t = \frac{2}{\mu(t+16\kappa)+1}$  with  $\kappa = L/\mu$  results in the bound

$$\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}_*\|^2\right] \leq \frac{8\sigma^2}{\mu^2 T} + \frac{(16\kappa + 2)^3 \delta_0}{T^3} \quad (18)$$

and an SFO/QMO complexity of  $\mathcal{O}\left(\frac{\sigma^2}{\mu^2 \epsilon} + \frac{\kappa \delta_0^{1/3}}{\epsilon^{1/3}}\right)$ .

Proof of Theorem 1 relies on an important one-step inequality (Lemma 2) that is reminiscent of but different from corresponding inequality in the proximal SGD setting. As in proximal SGD, using a diminishing stepsize in the convex case yields a slightly worse bound of  $\mathcal{O}\left(\frac{\log(T)}{\sqrt{T}}\right)$  in (16). Ignoring the constant terms, the SFO complexity results in Theorem 1 also match the best known bounds for proximal SGD [48], [49] for both convex and strongly convex objectives. The obtained rates also match the best known rates achieved for the functional constrained problems in [5], [7], [8], [11], [12], [14], [15], [30] while not requiring any bounded gradient assumptions.

The bounds and proof of Theorem 1 reveal a deeper connection to proximal SGD, suggesting that recent advances in the proximal SGD literature can be leveraged to design even faster algorithms for  $(\mathcal{P})$ . To demonstrate this idea in practice, we next present the algorithm that skips the step of solving QP in the intermittent iterative steps. In the next section, we will consider the finite-sum variant of  $(\mathcal{P})$  and develop an accelerated and variance-reduced version of SSQP. Before concluding this subsection, we remark that while Algorithm 1 specifies a fixed and worst-case value of  $\gamma$ , such a choice does result in a poorer SFO complexity. In the SQP literature, it is common to choose the exact penalty parameter adaptively, i.e., we start with a small  $\gamma$  and increase it based on the dual value of the per-iteration subproblem. Similar adaptive variants can be developed for the more challenging SSQP case, but are left as future work.

### B. SSQP-Skip algorithm

We now consider a situation when solving the QP with  $m$  linear constraints is more expensive than evaluating  $\nabla f_{i_t}(\cdot)$  for a random  $i_t \in \{1, \dots, n\}$ . This may be the case, for instance, when  $m$  is large, or if the proximal operator with respect to the regularizer  $h$  is complicated, e.g., when  $h$  is an indicator function corresponding to complicated set constraints. In such situations, it may be desirable to have an algorithm that allows one to skip solving the QP at most iterations. To this end,

we put forth the SSQP-Skip algorithm which, for smooth and strongly convex functions, requires only  $\mathcal{O}(1/\sqrt{\epsilon})$  calls to the QMO, as opposed to the  $\mathcal{O}(1/\epsilon)$  calls required by Algorithm 1 as per Theorem 1.

The proposed SSQP-Skip algorithm builds upon a similar SProxSkip algorithm from [50], but incorporates constraints and yields slightly better bounds. Specifically, we maintain an auxiliary variable  $\mathbf{y}_t$  that is used in place of  $\nabla f_{i_t}(\mathbf{x}_t)$  in (14), resulting in the updates:

$$\begin{aligned} \hat{\mathbf{x}}_{t+1} = \arg \min_{\mathbf{u} \in \mathbb{R}^d} \langle \mathbf{y}_t, \mathbf{u} \rangle + h(\mathbf{u}) + \frac{p_t}{2\eta_t} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{u}\|^2 \\ + \gamma \max\{[g_k(\tilde{\mathbf{x}}_{t+1}) + \langle \nabla g_k(\tilde{\mathbf{x}}_{t+1}), \mathbf{u} - \tilde{\mathbf{x}}_{t+1} \rangle]_+\} \end{aligned} \quad (19)$$

$$= \arg \min_{\mathbf{u} \in \mathbb{R}^d, v \geq 0} \langle \mathbf{y}_t, \mathbf{u} \rangle + h(\mathbf{u}) + \frac{p_t}{2\eta_t} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{u}\|^2 + \gamma v$$

$$\text{s. t. } g_k(\tilde{\mathbf{x}}_{t+1}) + \langle \nabla g_k(\tilde{\mathbf{x}}_{t+1}), \mathbf{u} - \tilde{\mathbf{x}}_{t+1} \rangle \leq v, \quad 1 \leq k \leq m.$$

which are carried out with probability  $p_t \ll 1$ . Observe that compared to (14), the update in (19) also utilizes a modified stepsize parameter  $\eta_t/p_t$  and entails linearizing  $g_k$  around  $\tilde{\mathbf{x}}_{t+1}$ , which is given by

$$\tilde{\mathbf{x}}_{t+1} = \mathbf{x}_t - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_t) \quad (20)$$

For the subsequent iteration, we update  $\mathbf{x}_{t+1} = \hat{\mathbf{x}}_{t+1}$  when (19) is evaluated (with probability  $p_t$ ) and keep  $\mathbf{x}_{t+1} = \tilde{\mathbf{x}}_{t+1}$  otherwise (with probability  $1 - p_t$ ). Finally, the auxiliary variable  $\mathbf{y}_{t+1}$  is kept the same when the QP is not solved, but updated whenever the QP is solved. The proposed approach is summarized in Algorithm 2. Clearly, when  $p_t$  is small, Algorithm 2 needs to solve the QP in (19) only rarely.

---

#### Algorithm 2 SSQP-Skip

---

- 1: **Input:**  $\mathbf{x}_0 \in \mathcal{K}$ ,  $\mathbf{y}_0 = \nabla f_{i_0}(\mathbf{x}_0)$ ,  $\gamma = \bar{B}/\nu$ ,  $p > 0$ , and  $\eta_t \in (0, 1]$  and  $i_0$  is a random index.
- 2: **for**  $t = 0, 1, \dots, T - 1$
- 3:   Sample  $i_t$  randomly
- 4:   Evaluate  $\tilde{\mathbf{x}}_{t+1}$  as per (20)
- 5:   Sample  $w_t \sim \text{Bernoulli}(p_t)$  and update

$$\mathbf{x}_{t+1} = w_t \hat{\mathbf{x}}_{t+1} + (1 - w_t) \tilde{\mathbf{x}}_{t+1} \quad (21)$$

where  $\hat{\mathbf{x}}_{t+1}$  is calculated as per (19).

- 6:   Update  $\mathbf{y}_{t+1} = \mathbf{y}_t + \frac{p_t}{2\eta_t} (\mathbf{x}_{t+1} - \tilde{\mathbf{x}}_{t+1})$
  - 7: **end for**
  - 8: **Output:**  $\mathbf{x}_T$ .
- 

Having detailed the proposed SSQP-Skip algorithm, the following theorem characterizes its performance.

**Theorem 2.** Under Assumptions **A1-A4**, and  $L = \max\{\gamma L_g, L_f\}$ ,  $\kappa = L/\mu$ ,  $\eta_t = \frac{2}{\mu(t+1+\omega)}$  for  $\omega = \lfloor 4\kappa^2 \rfloor$ , and  $p_t = \sqrt{2\mu\eta_t}$ , we have the bound

$$\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}_*\|^2\right] \leq \frac{8\sigma^2}{\mu^2 T} + \frac{4\kappa^4((1+4\kappa^2)\mu^2\delta_0+4\sigma^2)}{\mu^2 T^2}, \quad (22)$$

implying an SFO complexity of  $\mathcal{O}\left(\frac{\sigma^2}{\mu^2 \epsilon} + \kappa^2 \frac{\kappa\sqrt{\delta_0} + \sigma}{\sqrt{\epsilon}}\right)$  and a QMO complexity of  $\mathcal{O}\left(\frac{\sigma}{\kappa + \mu\sqrt{\epsilon}} + \frac{\kappa\sqrt{\kappa\sqrt{\delta_0} + \sigma}}{\epsilon^{1/4}}\right)$ .

The result in Theorem 2 is the first of its kind in the context of constrained optimization, and the proof is provided

in Appendix C. The bound in Theorem 2 is even better than that obtained for the corresponding unconstrained problem in [50]. Specifically, the proof of Algorithm 2 proceeds in a similar manner and obtains a similar recursion as that in [50, Lemma C.2]. However, we utilize diminishing stepsizes to avoid the  $\log(T)$  term that appears in [50, Corollary 5.6].

#### IV. VARIANCE-REDUCED ACCELERATED SSQP ALGORITHM

In this section, we focus on the special case of  $(\mathcal{P})$  where  $f$  has a finite-sum structure and  $i_t \in \{1, \dots, n\}$  for moderately large  $n$ . In the finite-sum case, we show that variance reduction can be applied to Algorithm 1 so as to obtain an improved dependence of the SFO complexity on  $\epsilon$ . In particular, we build upon the VARAG algorithm from [31] to propose the novel accelerated variance-reduced SSQP (VARAS) algorithm. As we shall show later, the performance of the proposed algorithm is also similar to that of VARAG.

The proposed updates are summarized in Algorithm 3, and follow a similar structure as that of VARAG, except that the proximal step is replaced with a constrained minimization step similar to (14). The proposed algorithm entails several passes over the data. At the  $s$ -th epoch or pass, the algorithm needs the full gradient  $\nabla f(\tilde{\mathbf{x}}_{s-1})$ , which is used to correct the stochastic gradient of each data point. Specifically, the  $t$ -th iteration of the  $s$ -th epoch utilizes a random  $i_t \in \{1, \dots, n\}$  and  $\nabla f(\tilde{\mathbf{x}}_{s-1})$  to construct an unbiased estimate  $\tilde{\nabla}_t$  of  $\nabla f(\mathbf{y}_t)$  with a variance that decreases with  $s$ .

We remark that although there exist several variance-reduced and accelerated stochastic optimization algorithms, we specifically selected VARAG, given its good performance and a flexible structure that allows for easy modifications. Indeed, both Katyusha acceleration [51] as well as related negative momentum acceleration techniques, such as those in [52], cannot be applied here for the general convex case, as they utilize a penalty parameter within the proximal operator that is required to be small or diminishing. For instance, Katyusha uses a penalty parameter  $\alpha_s \sim \mathcal{O}(1/s)$  where  $s$  is the epoch index [51], and likewise, ASVRG uses  $\beta_s \sim \mathcal{O}(1/s)$ . Hence the minimization subproblem at each iteration would only be  $\mathcal{O}(1/s)$ -strongly convex, which would not be sufficient to counter the term arising from the application of the quadratic upper bound (41) on  $g_k$ . Also note that while the VRADA algorithm proposed in [53] achieves the best known SFO complexity in the unconstrained setting, it uses a recursively defined estimate sequence, which cannot be extended to the present setting because of the penalty term.

The following theorem, whose proof is provided in the supplementary material, establishes the oracle-complexity bounds for VARAS.

**Theorem 3.** Under Assumptions **A1-A4**, let  $L_\gamma := L_f + \gamma L_g$ ,  $\kappa := L_\gamma/\mu$ ,  $D_0 := 2B_\gamma + \frac{3L_\gamma}{2}B_x \geq 2(\mathbb{E}F(\tilde{\mathbf{x}}_0) - F(\mathbf{x}_*)) + \frac{3L_\gamma}{2}\mathbb{E}\|\mathbf{z}_0 - \mathbf{x}_*\|^2$ ,  $s_0 := \lfloor \log n \rfloor + 1$ ,  $\beta_s = \frac{1}{3\alpha_s L_\gamma}$ ,  $\omega_s = \frac{1}{2}$ , and  $T_s = 2^{s-1}$  for  $s \leq s_0$  and  $T_s = T_{s_0}$  for  $s > s_0$ . Then, we have the following oracle complexity bounds.

#### Algorithm 3 VARAS: VARIance-Reduced Accelerated SSQP

- 1: **Input:**  $\mathbf{x}_0 \in \mathcal{K} \subset \mathbb{R}^d$ ,  $\gamma = \tilde{B}/\nu$ ,  $T_s$ ,  $\alpha_s$ ,  $\omega_s$ ,  $\beta_s$ , and  $\theta_t$
- 2: Set  $\tilde{\mathbf{x}}_0 = \mathbf{z}_0 = \mathbf{x}_0$
- 3: **for**  $s = 1, 2, \dots, S$
- 4: Calculate  $\nabla f(\tilde{\mathbf{x}}_{s-1})$
- 5: Set  $\mathbf{x}_0 = \tilde{\mathbf{x}}_{s-1}$
- 6: **for**  $t = 1, 2, \dots, T_s$
- 7: Sample  $i_t$  randomly from  $\{1, \dots, n\}$
- 8: Update
 
$$\mathbf{y}_t = \frac{(1 + \mu\beta_s)(1 - \alpha_s - \omega_s)\mathbf{x}_{t-1} + \alpha_s\mathbf{z}_{t-1}}{1 + \mu\beta_s(1 - \alpha_s)} + \frac{(1 + \mu\beta_s)\omega_s}{1 + \mu\beta_s(1 - \alpha_s)}\tilde{\mathbf{x}}_{s-1} \quad (23)$$

$$\mathbf{z}_{t-1}^+ = \frac{1}{1 + \mu\beta_s}(\mathbf{z}_{t-1} + \mu\beta_s\mathbf{y}_t) \quad (24)$$

$$\tilde{\nabla}_t = \nabla f_{i_t}(\mathbf{y}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}_{s-1}) + \nabla f(\tilde{\mathbf{x}}_{s-1}) \quad (25)$$

$$\mathbf{z}_t = \arg \min_{\mathbf{u} \in \mathbb{R}^d} \alpha_s \beta_s \left( \langle \tilde{\nabla}_t, \mathbf{u} \rangle + \frac{\mu}{2} \|\mathbf{y}_t - \mathbf{u}\|^2 + h(\mathbf{u}) \right) + \frac{\alpha_s}{2} \|\mathbf{z}_{t-1} - \mathbf{u}\|^2 \quad (26)$$

$$+ \gamma \beta_s \max\{[g_k(\mathbf{y}_t) + \alpha_s \langle \nabla g_k(\mathbf{y}_t), \mathbf{u} - \mathbf{z}_{t-1}^+ \rangle]\} \quad (27)$$
- 9: **end for**
- 10: Set  $\tilde{\mathbf{x}}_s = (\sum_t \theta_t \mathbf{x}_t) / \sum_t \theta_t$
- 11: Reset  $\mathbf{z}_0 = \mathbf{z}_{T_s}$
- 12: **end for**
- 13: **Output:**  $\bar{\mathbf{x}} = \tilde{\mathbf{x}}_S$ .

- 1) When  $f_i$  are convex and  $\alpha_s = \min\{\frac{1}{2}, \frac{2}{s-s_0+4}\}$  and

$$\theta_t = \begin{cases} \frac{\beta_s}{\alpha_s}(\alpha_s + \omega_s) & 1 \leq t \leq T_s - 1 \\ \frac{\beta_s}{\alpha_s} & t = T_s \end{cases} \quad (28)$$

then the oracle complexity of Algorithm 3 is given by

$$N_{\text{SFO}} = \begin{cases} \mathcal{O}(n \log \frac{D_0}{\epsilon}) & n \geq \frac{D_0}{\epsilon} \\ \mathcal{O}\left(n \log n + \sqrt{\frac{nD_0}{\epsilon}}\right) & n < \frac{D_0}{\epsilon} \end{cases} \quad (29)$$

$$N_{\text{QMO}} = \begin{cases} \mathcal{O}\left(\frac{D_0}{\epsilon}\right) & n \geq \frac{D_0}{\epsilon} \\ \mathcal{O}\left(\sqrt{\frac{nD_0}{\epsilon}}\right) & n < \frac{D_0}{\epsilon} \end{cases} \quad (30)$$

- 2) When  $f_i$  are  $\mu$ -strongly convex, let  $\alpha_s = \min\{\frac{1}{2}, \max\{\frac{2}{s-s_0+4}, \min\{\sqrt{\frac{n}{3\kappa}}, \frac{1}{2}\}\}\}$  and  $\theta_t$  is set as in (28) if  $1 \leq s \leq s_0$  or  $s_0 < s \leq s_0 + \sqrt{\frac{12\kappa}{n}} - 4$ ,  $n < \frac{3\kappa}{4}$ . Otherwise set as

$$\theta_t = \begin{cases} \Gamma_{t-1} - (1 - \alpha_s - \omega_s)\Gamma_t & 1 \leq t \leq T_s - 1 \\ \Gamma_{t-1} & t = T_s \end{cases} \quad (31)$$

where  $\Gamma_t = (1 + \mu\beta_s)^t$ . Then, the oracle complexity is given by

$$N_{\text{SFO}} = \begin{cases} \mathcal{O}\left(n \log \frac{D_0}{\epsilon}\right) & n \geq \frac{D_0}{\epsilon} \text{ or } n \geq \frac{3\kappa}{4} \\ \mathcal{O}\left(n \log n + \sqrt{\frac{nD_0}{\epsilon}}\right) & n < \frac{D_0}{\epsilon} \leq \frac{3\kappa}{4} \\ \mathcal{O}\left(n \log n + \sqrt{n\kappa} \log \frac{4D_0}{3\kappa\epsilon}\right) & n < \frac{3\kappa}{4} \leq \frac{D_0}{\epsilon} \end{cases} \quad (32)$$

$$N_{\text{QMO}} = \begin{cases} \mathcal{O}\left(\frac{D_0}{\epsilon}\right) & n \geq \frac{D_0}{\epsilon} \\ \mathcal{O}\left(n \log \frac{D_0}{\epsilon}\right) & \frac{3\kappa}{4} < n \leq \frac{D_0}{\epsilon} \\ \mathcal{O}\left(\sqrt{\frac{nD_0}{\epsilon}}\right) & n < \frac{D_0}{\epsilon} \leq \frac{3\kappa}{4} \\ \mathcal{O}\left(n \log n + \sqrt{n\kappa} \log \frac{4D_0}{3\kappa\epsilon}\right) & n < \frac{3\kappa}{4} \leq \frac{D_0}{\epsilon} \end{cases} \quad (33)$$

The proof of Theorem 3 begins by deriving a one-step inequality based on the update in (26) and the relationships among the parameters. The resulting inequality (see the supplementary material) matches the form of [31, Lemma 6], so the remaining arguments in [31, Theorems 1–2] apply directly. Though the rates established in Theorem 3 are the fastest, the most common case is when  $\epsilon$  is small and VARAS achieves SFO complexity of  $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$  and  $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$  for the convex and strongly convex cases, respectively. These rates are clearly superior to the best known rates for constrained problems in [54]–[56]. Before concluding the theoretical results, the following remark is due.

**Remark 1.** The chosen oracle model hides the computational costs for large  $m$ , since each SFO call returns  $\{\nabla f_i(\mathbf{x}), \{g_k(\mathbf{x}), \nabla g_k(\mathbf{x})\}_{k=1}^m\}$ . Indeed, solving the QP at each iteration of SSQP may incur  $\mathcal{O}(m^3)$  floating point operations (flops), which is significantly more than the usual  $\mathcal{O}(m)$  flops incurred by similar primal–dual algorithms. While SSQP-Skip does allay this concern to an extent, it remains an open problem to see if we can design algorithms that work with only one (or a few) of the constraints at every iteration.

## V. NUMERICAL EXPERIMENTS

In this section, we analyze the performance of our proposed algorithms on two real-world problems and compare them with Adaptive Primal–Dual SGD (APriD) in [12], Generalized Online Convex Optimization (GOCO) in [13], and primal–dual stochastic subgradient (PDSS) method in [11]. Other older algorithms in Table I are not included here as they were not directly comparable. For instance, the problems in [5] and [42] are special cases of  $(\mathcal{P})$  as they consider only a single constraint, while [15] focuses on a nested finite-sum structure different from  $(\mathcal{P})$ . The work in [30] still required at least one projection. Finally, [8] employs Bregman divergence and does not provide any numerical results, making it difficult to adapt it to the Euclidean setting here.

We remark that the purpose of this section is to illustrate the effects of constraints and examine the trade-offs between the solvers. These benchmarks are not exhaustive, since we do not carry out many scaling (i.e. examining the effect of  $m$ ,  $n$ ,  $\kappa$ ) or other ablation studies. Such studies are outside the scope of the current work, and we do not expect them to yield any new insights beyond what we already know from existing literature. All experiments were run in MATLAB R2023a (Intel Core i7, 16 GB RAM), using quadprog for the intermediate QPs.

### A. Trajectory generation for an unmanned surface vehicle

Here we consider Zermelo’s navigation problem [57] in an oceanic environment where the aim is to find two-dimensional energy-optimal trajectory for an unmanned surface vehicle (USV) operating in a rectangular region  $\mathcal{X} := \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{u}\|_\infty \leq r\}$ . Let  $\mathbf{x}(t) \in \mathbb{R}^2$  denote the position of the USV at discrete time  $t \in \{1, \dots, T\}$ . Since the USV operates in a small and homogeneous area, we model the surface current velocity at coordinate  $\mathbf{y}$  as an unknown linear function  $\mathbf{v}(\mathbf{y}) = \mathbf{W}\mathbf{y} + \mathbf{z}$ . However, exact information on the ocean current at each position is unavailable; rather, several oceanographic agencies [58], [59] publish estimated measurements. As considered in [60], we seek to find an energy-efficient USV trajectory given the ensemble of ocean current estimates, denoted by  $\{\mathbf{W}_i, \mathbf{z}_i\}_{i=1}^n$ .

The energy consumption for a USV to move from  $\mathbf{x}(t-1)$  to a nearby point  $\mathbf{x}(t)$  scales cubically with the effective speed and can be modeled as  $\|\mathbf{x}(t-1) - \mathbf{x}(t) - \mathbf{v}(\mathbf{x}(t))\|^3$  [61]. Our goal is to minimize the total energy. If the maximum velocity of a USV is  $s_{\max}$  and the maximum surface current speed  $s_w \ll s_{\max}$ , then the constraint  $\|\mathbf{x}(t-1) - \mathbf{x}(t)\| \leq v_{\max} := s_{\max} - s_w$  ensures that the generated trajectories are feasible for the lower level controller. Therefore, the minimum expected-energy trajectory from the starting position  $\mathbf{p}_{\text{start}}$  to the destination position  $\mathbf{p}_{\text{dest}}$ , where  $\mathbf{x}$  collects all the coordinates across  $T$  discrete time instances, is the solution to

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^{2T}} \quad & \frac{1}{n} \sum_{i=1}^n \sum_{t=2}^T \|\mathbf{x}(t-1) - \mathbf{x}(t) - \mathbf{W}_i \mathbf{x}(t-1) - \mathbf{z}_i\|^3 \\ \text{s. t.} \quad & \mathbf{x}(1) = \mathbf{p}_{\text{start}}, \quad \mathbf{x}(T) = \mathbf{p}_{\text{dest}} \\ & \|\mathbf{x}(t-1) - \mathbf{x}(t)\|^2 \leq v_{\max}^2, \quad 2 \leq t \leq T, \end{aligned} \quad (34)$$

which has the familiar finite-sum structure with convex objective and constraints as in  $(\mathcal{P})$ .

To generate an ensemble  $\{\mathbf{W}_i, \mathbf{z}_i\}_{i=1}^n$ , we first randomly generate  $\mathbf{W} \in \mathbb{R}^{2 \times 2}$  and  $\mathbf{z} \in \mathbb{R}^2$ . Then for each  $i \in \{1, \dots, n\}$ , we generate noisy velocities  $\mathbf{v}_i(\mathbf{y}_j)$  as

$$\mathbf{v}_i(\mathbf{y}_j) = (\mathbf{I} + \text{diag}(\boldsymbol{\xi}))(\mathbf{W}\mathbf{y}_j + \mathbf{z}) \quad j = 1, 2, 3 \quad (35)$$

at the three sample positions  $\{\mathbf{y}_j\}_{j=1}^3$ , where,  $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , and solve the system of equations in (35) to get  $\{\mathbf{W}_i, \mathbf{z}_i\}$ . We consider (34) for  $n = 100$  and  $T = 40$  over a  $200 \times 200$  square region (in arbitrary units). The USV needs to travel from  $(20, 20)$  to  $(180, 180)$ , with a speed limit of 10 units per second. We initialized all the algorithms with the straight-line path joining  $\mathbf{p}_{\text{start}}$  and  $\mathbf{p}_{\text{dest}}$  with  $T$  equidistant waypoints.

The hyperparameters of each algorithm are tuned to achieve the best performance. For SSQP, the configuration  $\eta_0 = 0.009$ ,  $\eta_t = \frac{\eta_0}{\sqrt{t}}$ ,  $\gamma = 6 \times 10^5$ , and a minibatch of size 4 yielded the most favorable results. For VARAS, the best performance was observed with the setting,  $L_\gamma = 350$ ,  $\gamma = 10^6$ . We observed that APriD, GOCO, and PDSS diverged once the iterates left the feasible set. Therefore, we needed to tune their hyperparameters so as to maintain the balance between optimality gap reduction and constraint violation. The value of  $F(\mathbf{x}_*)$  was computed after running  $10^6$  iterations of SSQP, and we compare all the algorithms in terms of the relative optimality

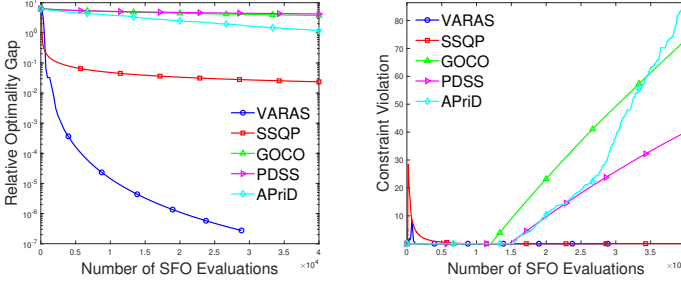


Fig. 1. Relative optimality gap and constraint violation versus the number of SFO evaluations.

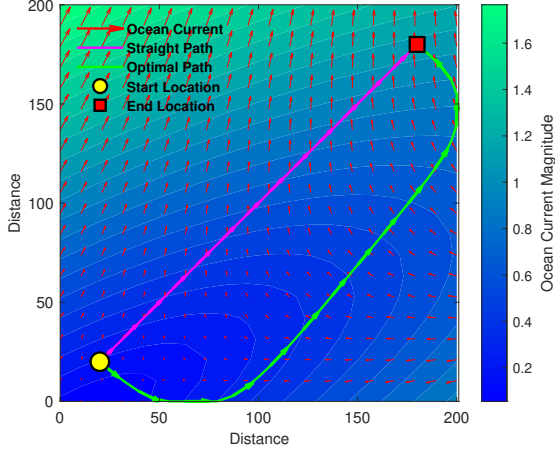


Fig. 2. Straight path and optimal path obtained by VARAS for traveling from the start to the end location in an ocean current field. The background color represents the current magnitude while the arrows represent the direction. The energy required to traverse the straight path is  $4.17 \times 10^6$  units, whereas the optimal path requires only  $0.57 \times 10^6$  units.

gap, i.e.  $\frac{F(\mathbf{x}) - F(\mathbf{x}_*)}{F(\mathbf{x}_*)}$ . Fig. 1 shows the relative optimality gap and constraint violation of different algorithms as a function of the number of SFO calls. In the initial iterations, both of our proposed algorithms, SSQP and VARAS, violate some of the constraints but eventually generate iterates within the feasible region. In contrast, other algorithms failed to return to the feasible region once they started violating the constraints, and their violations continued to increase with further iterations. We confirmed that this was not due to a poor tuning of hyperparameters but was inherent to these algorithms. Fig. 2 shows the least-energy path obtained by the VARAS algorithm.

### B. Regression with Residual Constraints

Regression is a fundamental tool in signal processing and learning. However, in many practical applications, e.g. in wireless communications [62], in addition to simply fitting a regression model to the observed data, it is desirable to ensure that, for some critical samples, the loss remains below a prescribed tolerance. We can write the constrained regression problem as

$$\begin{aligned} \min_{\theta \in \mathbb{R}^d} \quad & \frac{1}{2n} \sum_{i=1}^n \ell(y_i, b_{\theta}(\mathbf{x}_i)) \\ \text{s. t. } \quad & \ell(y_k, b_{\theta}(\mathbf{x}_k)) \leq r, \quad k \in \{1, 2, \dots, K\} \end{aligned} \quad (36)$$

where  $\ell$  denotes the loss function,  $b_{\theta}$  denotes the regression model, and  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$  are the data points whose first  $K$  tuples belong to the critical set over which the loss should be below  $r$ . While (36) is a special case of (P) whenever the objective and the constraints are convex, we consider linear regression for simplicity, i.e.,  $\ell(y_i, b_{\theta}(\mathbf{x}_i)) = (y_i - \mathbf{x}_i^T \theta)^2$ . We remark that the adaptive version of (36) has been widely studied within the framework of set-membership adaptive filtering [63], [64].

TABLE II  
PERFORMANCE COMPARISON FOR DIFFERENT THRESHOLD VALUES  
(AVERAGED OVER 50 RUNS).

		Thresholds		
		0.02	0.01	0.008
SSQP-Skip	SFO	1167	4598	7505
	QMO	189	308	377
	Time (s)	0.39	0.86	1.16
APriD	SFO	22561	27301	29395
	Time (s)	2.71	3.22	3.53
GOCO	SFO	38422	65502	84492
	Time (s)	2.93	5.18	6.67

Here, we evaluate the performance of the proposed algorithms on the Boston Housing dataset [65] which consists of 506 data points with 13 features, resulting in a  $13 \times 506$  data matrix  $\mathbf{X}$ . We normalized the data and appended a column of ones to account for the bias term, resulting in an effective feature dimension of  $d = 14$ . To generate the labels, we first sampled  $\theta_0 \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$  with  $\sigma = 1/\sqrt{d}$ , and then set  $\mathbf{y} = \mathbf{X}\theta_0 + \mathbf{n}$ , where  $\mathbf{n}$  was drawn from the standard normal distribution. Subsequently, we randomly selected  $n = 450$  data points as the dataset  $\mathcal{D}_f = \{\mathbf{x}_i, y_i\}_{i=1}^n$  and the remaining  $K = 56$  data points as the set of critical samples  $\mathcal{D}_c = \{\mathbf{x}_k, y_k\}_{k=1}^K$ . We empirically examined the optimization problem in (36) and set  $r = 1.3$ , ensuring that the feasible set is non-empty.

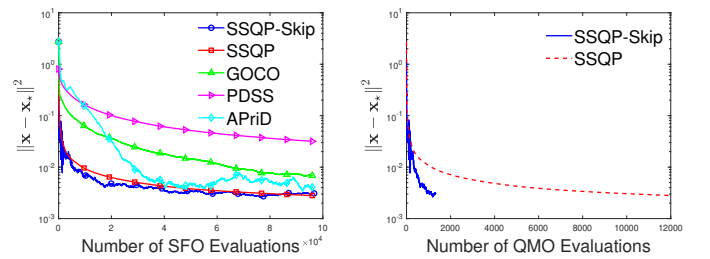


Fig. 3. Squared distance to the optimum versus the SFO and QMO calls.

The proposed algorithms SSQP and SSQP-Skip are compared against state-of-the-art APriD, GOCO, and PDSS algorithms, with all hyperparameters separately tuned to yield the best convergence rate. For SSQP, the configuration  $L = 1.1$ ,  $\mu = 0.8$ ,  $\gamma = 10^3$ , and a minibatch size of 8 yielded the best performance, whereas for SSQP-Skip the optimal setting was  $L = 1$ ,  $\mu = 0.85$ ,  $\gamma = 10^5$ , with a minibatch size of 1. We observed that in the initial  $K_s$  iterations, not skipping the QMO step in SSQP-Skip empirically improved its performance, a strategy we call *kickstart*. In our experiment, we set  $K_s = 100$  which is insignificant in comparison to the  $10^5$  iterations required by all algorithms. For GOCO,



empirically it was observed that clipping the subgradient lead to better convergence. The optimal point  $\mathbf{x}_*$  was obtained by running SSQP for a very large number of iterations. Fig. 3 shows the convergence rate in terms of the distance to the optimal point versus the number of SFO and QMO evaluations.

Since APriD, GOCO, and PDSS do not solve quadratic programs, it is natural to quantify the advantage of SSQP-Skip in terms of wall-clock time. Table II reports the SFO and QMO complexities of SSQP-Skip, APriD, and GOCO, averaged over 50 runs, required to ensure that the squared distance from the optimum is at most  $\epsilon \in \{0.02, 0.01, 0.008\}$ . We note that APriD also requires a matrix–vector multiplication in every iteration to compute the Lagrangian derivative. We did not include PDSS in this table since it required substantially more iterations to reach the same thresholds and was slower than both APriD and GOCO. As evident, SSQP-Skip is faster in terms of both SFO complexity and wall-clock times.

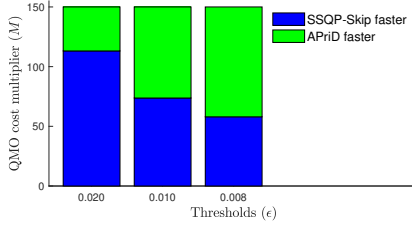


Fig. 4. SSQP-Skip is faster when  $M$  is small (blue region) and slower when  $M$  is large (green region).

Raw CPU times on modern hardware may vary widely due to differing implementations of quadratic optimization across commercial solvers and open-source libraries. To abstract away these implementation effects, we assume an SFO that returns stochastic gradient information in  $\tau$  seconds and a QP solver that completes each intermediate quadratic program in  $M\tau$  seconds, where  $M \geq 1$  captures the additional cost of solving a QP relative to an SFO-only setting. With this, the wall-clock time of SSQP-like algorithms is  $\mathcal{O}(\text{SFO} + M \times \text{QMO})$ , where SFO and QMO denote the number of SFO and QMO calls, respectively. Intuitively, if  $M$  is very large, even a few QMO calls will hurt the wall-clock time and SSQP-Skip will take longer than other algorithms. Figure 4 illustrates the critical values of  $M$  below which SSQP-Skip remains faster than APriD, based on the complexities summarized in Table II. The trend in Fig. 4 is consistent across thresholds, and both algorithms show negligible improvement below  $\epsilon \approx 0.005$ .

## VI. CONCLUSION

This paper proposes the stochastic sequential quadratic programming (SSQP) framework, where each iteration requires solving a quadratic program (QP) with linearized constraints. For the convex and strongly convex cases, SSQP achieves rates on par with those of unconstrained stochastic gradient descent. Additionally, we propose the SSQP-Skip algorithm which requires solving QPs only on a small subset of iterations, resulting in reduced wall-clock times. For the finite-sum case, we propose the accelerated variance-reduced VARAS

algorithm that also achieves near-optimal iteration complexity, improving upon existing results for constrained problems. The performance of the proposed algorithms, tested on trajectory generation and constrained regression problems, is also significantly better than the related primal–dual and other approaches in the literature.

## APPENDIX A BASIC INEQUALITIES

This section details some basic inequalities that will be repeatedly used in the proofs. Since the max function is monotonic, we have:

$$\max\{[u_k + a]_+\} - \max\{[u_k]_+\} \leq a \quad (37)$$

for any  $a \geq 0$ . Further,  $\max\{[au_k]_+\} = a \max\{[u_k]_+\}$  for any  $a \geq 0$ . Similarly, it can also be shown that

$$\max\{[u_k + y_k]_+\} \leq \max\{[u_k]_+\} + \max\{[y_k]_+\} \quad (38)$$

For a  $\mu$ -strongly convex function  $\varphi(\mathbf{x})$ , we have the quadratic lower bound

$$\varphi(\mathbf{y}) \geq \varphi(\mathbf{x}) + \langle \nabla \varphi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad (39)$$

for all  $\mathbf{x}, \mathbf{y} \in \text{dom } \varphi$ . Hence, for  $\mathbf{x}_* = \arg \min_{\mathbf{x}} \varphi(\mathbf{x})$ , we have that which implies that

$$\varphi(\mathbf{x}_*) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}_*\|^2 \leq \varphi(\mathbf{x}) \quad (40)$$

for all  $\mathbf{x} \in \text{dom } \varphi$ . If  $\varphi$  is also  $L$ -smooth, we have the quadratic upper bound as well as the co-coercivity property:

$$\varphi(\mathbf{y}) \leq \varphi(\mathbf{x}) + \langle \nabla \varphi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad (41)$$

$$\varphi(\mathbf{y}) \geq \varphi(\mathbf{x}) + \langle \nabla \varphi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L} \|\nabla \varphi(\mathbf{x}) - \nabla \varphi(\mathbf{y})\|^2. \quad (42)$$

We also list some of the common norm inequalities, which follow from the Cauchy-Schwarz inequality, Young's inequality, and the triangle inequality:

$$\langle \mathbf{u}, \mathbf{v} \rangle \leq \|\mathbf{u}\| \|\mathbf{v}\| \leq \frac{\varepsilon}{2} \|\mathbf{u}\|^2 + \frac{1}{2\varepsilon} \|\mathbf{v}\|^2 \quad (43)$$

$$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\| \quad (44)$$

for  $\varepsilon > 0$ . Combining the two for  $\varepsilon = 1$ , we obtain

$$\|\mathbf{u} + \mathbf{v}\|^2 \leq 2 \|\mathbf{u}\|^2 + 2 \|\mathbf{v}\|^2. \quad (45)$$

## APPENDIX B PROOF OF THEOREM 1

We begin by establishing a key lemma using the update equation, convexity, and the smoothness of the constraint functions  $g_k$ .

**Lemma 1.** Under Assumption A2, the update (14) implies that

$$\begin{aligned} & \langle \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_* \rangle + h(\mathbf{x}_{t+1}) + \gamma \max\{[g_k(\mathbf{x}_{t+1})]_+\} \\ & \leq h(\mathbf{x}_*) + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_*\|^2 - \frac{1}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2 \\ & \quad - \left( \frac{1}{2\eta_t} - \frac{\gamma L_g}{2} \right) \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2. \end{aligned} \quad (46)$$

*Proof:* Since  $\mathbf{x}_{t+1}$  is obtained by minimizing a  $\frac{1}{\eta_t}$ -strongly convex function in (14), we have from (40) that

$$\begin{aligned} & \langle \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_* \rangle + h(\mathbf{x}_{t+1}) + \frac{1}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ & + \gamma \max\{[g_k(\mathbf{x}_t) + \langle \nabla g_k(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle]_+\} \\ & \leq h(\mathbf{x}_*) + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_*\|^2 - \frac{1}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2 \\ & + \gamma \max\{[g_k(\mathbf{x}_t) + \langle \nabla g_k(\mathbf{x}_t), \mathbf{x}_* - \mathbf{x}_t \rangle]_+\} \end{aligned} \quad (47)$$

$$\leq h(\mathbf{x}_*) + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_*\|^2 - \frac{1}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2 \quad (48)$$

where (48) follows from the convexity of  $g_k$  and the feasibility of  $\mathbf{x}_*$ , so that

$$g_k(\mathbf{x}_t) + \langle \nabla g_k(\mathbf{x}_t), \mathbf{x}_* - \mathbf{x}_t \rangle \leq g_k(\mathbf{x}_*) \leq 0 \quad (49)$$

and hence,  $\max\{[g_k(\mathbf{x}_t) + \langle \nabla g_k(\mathbf{x}_t), \mathbf{x}_* - \mathbf{x}_t \rangle]_+\} = 0$ . Likewise, since  $g_k$  is  $L_g$ -smooth, we have from the (41) and (37)-(38) that

$$\begin{aligned} \max\{[g_k(\mathbf{x}_{t+1})]_+\} & \leq \max\{[g_k(\mathbf{x}_t) + \langle \nabla g_k(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle]_+\} \\ & + \frac{L_g}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \end{aligned} \quad (50)$$

Multiplying (50) by  $\gamma$  and substituting into (48), we obtain the required result. ■

We are now ready to establish the one-step inequality for the SSQP algorithm.

**Lemma 2.** Under Assumptions (A2) and (A3), we have for  $\eta_t \leq \frac{1}{2(L_f + \max\{\gamma L_g, L_f\})}$ , where  $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x}) + \gamma \max\{[g_k(\mathbf{x})]_+\}$ :

$$\begin{aligned} \mathbb{E}_t[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_*) & \leq \frac{1 - \mu\eta_t(1 - 4\eta_t L_f)}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_*\|^2 \\ & - \frac{1}{2\eta_t} \mathbb{E}_t[\|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2] + 2\eta_t \sigma^2 \end{aligned} \quad (51)$$

*Proof:* Since  $f$  is  $L_f$ -smooth, we have from (41) that

$$\begin{aligned} f(\mathbf{x}_{t+1}) - f(\mathbf{x}_*) & \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \\ & + \frac{L_f}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - f(\mathbf{x}_*) \end{aligned} \quad (52)$$

$$\begin{aligned} & = -(f(\mathbf{x}_*) - f(\mathbf{x}_t) - \langle \nabla f(\mathbf{x}_t), \mathbf{x}_* - \mathbf{x}_t \rangle) \\ & + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_* \rangle + \frac{L_f}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \end{aligned} \quad (53)$$

Adding (46) and rearranging, we obtain

$$\begin{aligned} F(\mathbf{x}_{t+1}) - F(\mathbf{x}_*) & \leq \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_*\|^2 - \frac{1}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2 \\ & - \frac{1 - \eta_t(L_f + \gamma L_g)}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ & + \langle \nabla f(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* \rangle \\ & + \langle \nabla f(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \\ & - (f(\mathbf{x}_*) - f(\mathbf{x}_t) - \langle \nabla f(\mathbf{x}_t), \mathbf{x}_* - \mathbf{x}_t \rangle). \end{aligned} \quad (54)$$

Since  $i_t$  is selected at random, we have that

$$\mathbb{E}_t[\langle \nabla f(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_* \rangle] = 0 \quad (55)$$

For the other term depending on  $i_t$ , we use (43) and (8) to obtain the bound

$$\begin{aligned} & \mathbb{E}_t[\langle \nabla f(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle] \\ & \stackrel{(43)}{\leq} \eta_t \mathbb{E}_t[\|\nabla f(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_t)\|^2] + \frac{1}{4\eta_t} \mathbb{E}_t[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2] \\ & \stackrel{(8)}{\leq} 4\eta_t L_f (f(\mathbf{x}_*) - f(\mathbf{x}_t) - \langle \nabla f(\mathbf{x}_t), \mathbf{x}_* - \mathbf{x}_t \rangle) + 2\eta_t \sigma^2 \\ & + \frac{1}{4\eta_t} \mathbb{E}_t[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2] \end{aligned} \quad (56)$$

Therefore, taking expectation with respect to  $i_t$  in (54) and substituting (55)-(56), we obtain

$$\begin{aligned} \mathbb{E}_t[F(\mathbf{x}_{t+1})] - F(\mathbf{x}_*) & \leq \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_*\|^2 + 2\eta_t \sigma^2 \\ & - \frac{1}{2\eta_t} \mathbb{E}_t[\|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2] - \frac{1 - 2\eta_t(L_f + \gamma L_g)}{4\eta_t} \mathbb{E}_t[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2] \\ & - (1 - 4\eta_t L_f)(f(\mathbf{x}_*) - f(\mathbf{x}_t) - \langle \nabla f(\mathbf{x}_t), \mathbf{x}_* - \mathbf{x}_t \rangle) \end{aligned} \quad (57)$$

The fourth term on the right is non-positive and can be dropped for  $\eta_t \leq \frac{1}{2(L_f + \gamma L_g)}$ . Finally, for  $\eta_t \leq \frac{1}{4L_f}$ , the last term on the right can be bounded from (39),

$$\begin{aligned} & -(1 - 4\eta_t L_f)(f(\mathbf{x}_*) - f(\mathbf{x}_t) - \langle \nabla f(\mathbf{x}_t), \mathbf{x}_* - \mathbf{x}_t \rangle) \\ & \leq -\frac{\mu(1 - 4\eta_t L_f)}{2} \|\mathbf{x}_t - \mathbf{x}_*\|^2 \end{aligned} \quad (58)$$

which upon substituting into (57), yields the required result. Further, both requirements for  $\eta_t$  are satisfied when  $\eta_t \leq \frac{1}{2(L_f + \max\{\gamma L_g, L_f\})}$ . ■

Having established the one-step inequality, we can now obtain the required oracle complexities. For the sake of brevity, recall the definition of  $\Delta_t$  from Sec. II-B and also define  $\delta_t := \mathbb{E} \|\mathbf{x}_t - \mathbf{x}_*\|^2$ . Taking full expectation in (51) and using the definitions of  $\Delta_t$  and  $\delta_t$ , we obtain:

$$\Delta_{t+1} \leq \frac{1 - \mu\eta_t(1 - 4\eta_t L_f)}{2\eta_t} \delta_t - \frac{1}{2\eta_t} \delta_{t+1} + 2\eta_t \sigma^2. \quad (59)$$

Let us consider the convex case first, where we set  $\mu = 0$ , yielding  $\Delta_{t+1} \leq \frac{1}{2\eta_t} \delta_t - \frac{1}{2\eta_t} \delta_{t+1} + 2\eta_t \sigma^2$ . Multiplying by  $\eta_t$  on both sides for  $\eta_{t+1} \leq \eta_t$  and taking sum for  $t = 0, \dots, T-1$ , we obtain:

$$\sum_{t=1}^T \eta_t \Delta_t \leq \frac{\delta_0 - \delta_T}{2} + 2\sigma^2 \sum_{t=0}^{T-1} \eta_t^2 \leq \frac{\delta_0}{2} + 2\sigma^2 \sum_{t=0}^{T-1} \eta_t^2. \quad (60)$$

Therefore, from the convexity of  $F$ , we have the following bound for the averaged iterate  $\bar{\mathbf{x}}_T := \left( \sum_{t=1}^T \eta_t \mathbf{x}_t \right) / \left( \sum_{t=1}^T \eta_t \right)$ :

$$\mathbb{E}[F(\bar{\mathbf{x}}_T)] - F(\mathbf{x}_*) \leq \frac{\sum_{t=1}^T \eta_t \Delta_t}{\sum_{t=1}^T \eta_t} \leq \frac{\delta_0 + 4\sigma^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{t=1}^T \eta_t}$$

With the stepsize rule  $\eta_t = \frac{\eta_0}{\sqrt{t+1}}$  where  $\eta_0 \leq \frac{1}{4L}$ , we obtain

$$\mathbb{E}[F(\bar{\mathbf{x}}_T)] - F(\mathbf{x}_*) \leq \frac{\delta_0 + 4\sigma^2 \eta_0^2 (1 + \log(T))}{2\eta_0 \sqrt{T}}. \quad (61)$$

Alternatively, setting  $\eta_t = \eta_0 / \sqrt{T}$  for  $\eta_0 = \min\{\frac{\sqrt{\delta_0}}{2\sigma}, \frac{1}{4L}\}$ , the bound becomes

$$\mathbb{E}[F(\bar{\mathbf{x}}_T)] - F(\mathbf{x}_*) \leq \frac{2}{\sqrt{T}} \max\{2L\delta_0, \sigma\sqrt{\delta_0}\} \quad (62)$$

Since  $\max\{[g_k(\mathbf{x}_*)]_+\} = 0$  and  $\mathbb{E}[\max\{[g_k(\bar{\mathbf{x}}_T)]_+\}] \geq 0$ , we can drop these terms from the left of (61)-(62) to yield the desired bounds on the optimality gap.

To bound the constraint violation, let  $w_T := \max_k \{g_k(\bar{\mathbf{x}}_T)\}$ . Taking expectation in (4) and rearranging, we obtain

$$\begin{aligned} f(\mathbf{x}_*) + h(\mathbf{x}_*) - \mathbb{E}[f(\bar{\mathbf{x}}_T) + h(\bar{\mathbf{x}}_T)] & \leq \|\boldsymbol{\lambda}_*\|_1 \mathbb{E}[w_T] \\ & \leq \frac{\bar{B}}{\nu} \mathbb{E}[w_T]. \end{aligned} \quad (63)$$

Adding (62) and (63) we obtain

$$\left( \gamma - \frac{\bar{B}}{\nu} \right) \mathbb{E}[w_T] \leq \frac{2}{\sqrt{T}} \max\{2L\delta_0, \sigma\sqrt{\delta_0}\}. \quad (64)$$

which yields the desired bound and hence the SFO/QMO complexity for the convex case.

For the strongly convex case, we first use (39) to write

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}_*) \geq \frac{\mu}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2 \quad (65)$$

so as to obtain

$$\delta_{t+1} \leq \frac{1-\mu\eta_t(1-4\eta_t L_f)}{1+\mu\eta_t} \delta_t + \frac{4\eta_t^2 \sigma^2}{1+\mu\eta_t} \quad (66)$$

Suppose we choose  $\eta_t \leq \frac{1}{8L} \leq \frac{1}{8L_f}$  so that  $1 - 4\eta_t L_f \geq 1/2$  and the recursion becomes

$$\delta_{t+1} \leq \frac{1-\mu\eta_t/2}{1+\mu\eta_t} \delta_t + \frac{4\eta_t^2 \sigma^2}{1+\mu\eta_t} \quad (67)$$

Thus if we set  $\eta_t = \frac{2}{\mu(t+\omega+1)}$  and choose  $\omega = \lfloor \frac{16L}{\mu} \rfloor$ , it would follow that  $\eta_t \leq \eta_0 \leq \frac{1}{8L}$  and the recursion can be written as

$$\delta_{t+1} \leq \frac{t+\omega}{t+\omega+3} \delta_t + \frac{16\sigma^2}{\mu^2(t+\omega+3)(t+\omega+1)} \quad (68)$$

Multiplying both sides by  $(t+\omega+1)(t+\omega+2)(t+\omega+3)$  and summing telescopically, we obtain

$$(T+\omega+2)(T+\omega+1)(T+\omega)\delta_T \leq \omega(\omega+1)(\omega+2)\delta_0 + \frac{8\sigma^2}{\mu^2}(T+\omega+1)(T+\omega+2). \quad (69)$$

Rearranging and bounding the terms on the right, we obtain

$$\delta_T \leq \left( \frac{\omega+2}{T+\omega+2} \right)^3 \delta_0 + \frac{8\sigma^2}{\mu^2(T+\omega)} \quad (70)$$

$$\leq \frac{(16\kappa+2)^3}{T^3} \delta_0 + \frac{8\sigma^2}{\mu^2 T} \quad (71)$$

where  $\kappa = L/\mu$ . The bound in (71) translates to an SFO complexity of  $\mathcal{O}\left(\frac{\sigma^2}{\mu^2 \epsilon} + \frac{\kappa \delta_0^{1/3}}{\epsilon^{1/3}}\right)$ .

## APPENDIX C PROOF OF THEOREM 2

Before establishing the main result, we derive some intermediate results through the following lemmas. The following result is a consequence of the update in (19) and the convexity and smoothness of  $g_k$ .

**Lemma 3.** Under Assumption A2 and for  $\mathbf{y}_* = \nabla f(\mathbf{x}_*)$ , it holds that

$$p_t \|\hat{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 \leq p_t \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 - 2\eta_t \langle \hat{\mathbf{x}}_{t+1} - \mathbf{x}_*, \mathbf{y}_t - \mathbf{y}_* \rangle - (p_t - \gamma\eta_t L_g) \|\hat{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_{t+1}\|^2 \quad (72)$$

*Proof:* We begin with establishing three point inequality for the update  $\hat{\mathbf{x}}_{t+1}$ . Since (19) involves minimizing a  $\frac{p_t}{\eta_t}$ -strongly convex function, we have that

$$\begin{aligned} & \frac{p_t}{2\eta_t} \|\tilde{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t+1}\|^2 + \langle \hat{\mathbf{x}}_{t+1} - \mathbf{x}_*, \mathbf{y}_t \rangle + h(\hat{\mathbf{x}}_{t+1}) \\ & + \gamma \max \{ [g_k(\tilde{\mathbf{x}}_{t+1}) + \langle \nabla g_k(\tilde{\mathbf{x}}_{t+1}), \hat{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_{t+1} \rangle]_+ \} \\ (40) \quad & \leq -\frac{p_t}{2\eta_t} \|\hat{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 + \frac{p_t}{2\eta_t} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 + h(\mathbf{x}_*) \\ & + \gamma \max \{ [g_k(\tilde{\mathbf{x}}_{t+1}) + \langle \nabla g_k(\tilde{\mathbf{x}}_{t+1}), \mathbf{x}_* - \tilde{\mathbf{x}}_{t+1} \rangle]_+ \}. \end{aligned} \quad (73)$$

Next, the convexity and  $L_g$ -smoothness of  $g_k$  allow us to similarly use (49) and (50), respectively, with  $\mathbf{x}_{t+1}$  in place of  $\hat{\mathbf{x}}_{t+1}$  and  $\mathbf{x}_t$  in place of  $\tilde{\mathbf{x}}_{t+1}$ , yielding

$$\begin{aligned} & \frac{p_t}{2\eta_t} \|\hat{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 \leq \frac{p_t}{2\eta_t} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 - \langle \hat{\mathbf{x}}_{t+1} - \mathbf{x}_*, \mathbf{y}_t \rangle \\ & - \left( \frac{p_t}{2\eta_t} - \frac{\gamma L_g}{2} \right) \|\hat{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_{t+1}\|^2 \\ & + h(\mathbf{x}_*) - h(\hat{\mathbf{x}}_{t+1}) - \gamma \max \{ [g_k(\hat{\mathbf{x}}_{t+1})]_+ \} \end{aligned} \quad (74)$$

Finally, the first order optimality condition of  $(\mathcal{P}_1)$  implies that

$$\begin{aligned} \mathbf{x}_* &= \arg \min_{\mathbf{u} \in \mathbb{R}^d} \frac{1}{2\eta_t} \|\mathbf{u} - \mathbf{x}_*\|^2 + \langle \mathbf{u}, \mathbf{y}_* \rangle + h(\mathbf{u}) \\ & + \gamma \max \{ [g_k(\mathbf{x}_*) + \langle \nabla g_k(\mathbf{x}_*), \mathbf{u} - \mathbf{x}_* \rangle]_+ \} \end{aligned} \quad (75)$$

for  $\mathbf{y}_* = \nabla f(\mathbf{x}_*)$ . Since (75) again involves minimization of a  $\frac{1}{\eta_t}$ -strongly convex function, we have from (40) that:

$$\begin{aligned} & h(\mathbf{x}_*) - h(\hat{\mathbf{x}}_{t+1}) + \langle \mathbf{x}_* - \hat{\mathbf{x}}_{t+1}, \mathbf{y}_* \rangle \\ & \leq \gamma \max \{ [g_k(\mathbf{x}_*) + \langle \nabla g_k(\mathbf{x}_*), \hat{\mathbf{x}}_{t+1} - \mathbf{x}_* \rangle]_+ \} \end{aligned} \quad (76)$$

$$\leq \gamma \max \{ [g_k(\hat{\mathbf{x}}_{t+1})]_+ \} \quad (77)$$

where (77) follows from the convexity of  $g_k$  and monotonicity of the  $\max\{[\cdot]_+\}$  operator. Substituting (77) in (74) and multiplying by  $2\eta_t$ , we obtain the required result. ■

The next lemma obtains a recursive inequality incorporating the effect of skipping (19). Recall that in Algorithm 2,  $w_t \sim \text{Bernoulli}(p_t)$ , and let  $\mathbb{E}_{w_t}[\cdot]$  denote the expectation with respect to  $w_t$ .

**Lemma 4.** For  $\eta_t \leq \frac{p_t}{2\gamma L_g}$  and  $\mathbf{y}_* = \nabla f(\mathbf{x}_*)$ , the updates in Algorithm 2 imply that

$$\begin{aligned} & \mathbb{E}_{w_t} [\|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2] + \frac{2\eta_t^2}{p_t^2} \mathbb{E}_{w_t} [\|\mathbf{y}_{t+1} - \mathbf{y}_*\|^2] \\ & \leq \|\mathbf{x}_t - \mathbf{x}_*\|^2 - \eta_t \langle \nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_*, \mathbf{y}_t - \mathbf{y}_* \rangle + \frac{\eta_t^2(2-p_t^2)}{p_t^2} \|\mathbf{y}_t - \mathbf{y}_*\|^2. \end{aligned} \quad (78)$$

*Proof:* From the update in (21) and the result of Lemma 3, we have that

$$\begin{aligned} \mathbb{E}_{w_t} [\|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2] &= p_t \|\hat{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 + (1-p_t) \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 \\ &\stackrel{(72)}{\leq} \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 - 2\eta_t \langle \hat{\mathbf{x}}_{t+1} - \mathbf{x}_*, \mathbf{y}_t - \mathbf{y}_* \rangle \\ &\quad - (p_t - \eta_t \gamma L_g) \|\hat{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_{t+1}\|^2. \end{aligned} \quad (79)$$

We also note from the update of  $\mathbf{y}_{t+1}$  in Algorithm 2 that

$$\mathbf{y}_{t+1} = \mathbf{y}_t + \frac{w_t p_t}{2\eta_t} (\hat{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_{t+1}) \quad (80)$$

so that

$$\begin{aligned} \mathbb{E}_{w_t} [\|\mathbf{y}_{t+1} - \mathbf{y}_*\|^2] &= p_t \left\| \mathbf{y}_t - \mathbf{y}_* + \frac{p_t}{2\eta_t} (\hat{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_{t+1}) \right\|^2 \\ &\quad + (1-p_t) \|\mathbf{y}_t - \mathbf{y}_*\|^2 \end{aligned} \quad (81)$$

$$\begin{aligned} &= \|\mathbf{y}_t - \mathbf{y}_*\|^2 + \frac{p_t^3}{4\eta_t^2} \|\hat{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_{t+1}\|^2 \\ &\quad + \frac{p_t^2}{\eta_t} \langle \hat{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_{t+1}, \mathbf{y}_t - \mathbf{y}_* \rangle \end{aligned} \quad (82)$$

Multiplying (82) by  $\frac{2\eta_t^2}{p_t^2}$  and adding with (79), we obtain

$$\begin{aligned} & \mathbb{E}_{w_t} [\|\mathbf{x}_{t+1} - \mathbf{x}_*\|^2] + \frac{2\eta_t^2}{p_t^2} \mathbb{E}_{w_t} [\|\mathbf{y}_{t+1} - \mathbf{y}_*\|^2] \\ & \leq \|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_*\|^2 - 2\eta_t \langle \tilde{\mathbf{x}}_{t+1} - \mathbf{x}_*, \mathbf{y}_t - \mathbf{y}_* \rangle \\ & \quad + \frac{2\eta_t^2}{p_t^2} \|\mathbf{y}_t - \mathbf{y}_*\|^2 - \left( \frac{p_t}{2} - \eta_t \gamma L_g \right) \|\hat{\mathbf{x}}_{t+1} - \tilde{\mathbf{x}}_{t+1}\|^2 \end{aligned} \quad (83)$$

where the last term can be dropped if  $p_t > 2\eta_t\gamma L_g$ .

Next, we have from (20) that

$$\begin{aligned}\|\tilde{\mathbf{x}}_{t+1} - \mathbf{x}_\star\|^2 &= \|\mathbf{x}_t - \mathbf{x}_\star - \eta_t \nabla f_{i_t}(\mathbf{x}_t) + \eta_t \mathbf{y}_t\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}_\star - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star) + \eta_t (\mathbf{y}_t - \mathbf{y}_\star)\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}_\star - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star)\|^2 + \eta_t^2 \|\mathbf{y}_t - \mathbf{y}_\star\|^2 \\ &\quad + 2\eta_t \langle \mathbf{x}_t - \mathbf{x}_\star - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star), \mathbf{y}_t - \mathbf{y}_\star \rangle\end{aligned}\quad (84)$$

and similarly

$$\begin{aligned}\langle \tilde{\mathbf{x}}_{t+1} - \mathbf{x}_\star, \mathbf{y}_t - \mathbf{y}_\star \rangle &= \langle \mathbf{x}_t - \mathbf{x}_\star - \eta_t \nabla f_{i_t}(\mathbf{x}_t) + \eta_t \mathbf{y}_t, \mathbf{y}_t - \mathbf{y}_\star \rangle \\ &= \langle \mathbf{x}_t - \mathbf{x}_\star - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star) + \eta_t (\mathbf{y}_t - \mathbf{y}_\star), \mathbf{y}_t - \mathbf{y}_\star \rangle \\ &= \eta_t \|\mathbf{y}_t - \mathbf{y}_\star\|^2 + \langle \mathbf{x}_t - \mathbf{x}_\star - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star), \mathbf{y}_t - \mathbf{y}_\star \rangle.\end{aligned}\quad (85)$$

Substituting (84)-(85) in (83), we obtain the required result.  $\blacksquare$

Having derived the key recursive inequality, we now proceed with proving the main result.

*Proof of Theorem 2.* From the update in (20), we obtain

$$\begin{aligned}\|\mathbf{x}_t - \mathbf{x}_\star - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star)\|^2 &= \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\ &\quad + \eta_t^2 \|\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star\|^2 - 2\eta_t \langle \mathbf{x}_t - \mathbf{x}_\star, \nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star \rangle.\end{aligned}\quad (86)$$

Taking expectation with respect to the random variable  $i_t$ , we obtain

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{x}_t - \mathbf{x}_\star - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star)\|^2] &= \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\ &\quad + \eta_t^2 \mathbb{E}_t[\|\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star\|^2] - 2\eta_t \langle \mathbf{x}_t - \mathbf{x}_\star, \nabla f(\mathbf{x}_t) - \mathbf{y}_\star \rangle.\end{aligned}\quad (87)$$

Recalling that  $\mathbf{y}_\star = \nabla f(\mathbf{x}_\star)$  and using (8), the second term on the right can be bounded as

$$\mathbb{E}_t[\|\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star\|^2] \leq 4L_f D_f(\mathbf{x}_\star, \mathbf{x}_t) + 2\sigma^2. \quad (88)$$

We further have that

$$\langle \mathbf{x}_t - \mathbf{x}_\star, \nabla f(\mathbf{x}_t) - \mathbf{y}_\star \rangle = D_f(\mathbf{x}_\star, \mathbf{x}_t) + D_f(\mathbf{x}_t, \mathbf{x}_\star). \quad (89)$$

Substituting (88)-(89) into (87), we obtain

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{x}_t - \mathbf{x}_\star - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star)\|^2] &\leq \|\mathbf{x}_t - \mathbf{x}_\star\|^2 \\ &\quad - 2\eta_t(1 - 2\eta_t L_f) D_f(\mathbf{x}_\star, \mathbf{x}_t) - 2\eta_t D_f(\mathbf{x}_t, \mathbf{x}_\star) + 2\eta_t^2 \sigma^2.\end{aligned}\quad (90)$$

Here, we can drop the non-positive second term on the right since  $\eta_t \leq \frac{p_t}{2L} \leq \frac{1}{2L_f}$  and use (39) to obtain

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{x}_t - \mathbf{x}_\star - \eta_t (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{y}_\star)\|^2] &\leq (1 - \mu\eta_t) \|\mathbf{x}_t - \mathbf{x}_\star\|^2 + 2\eta_t^2 \sigma^2\end{aligned}\quad (91)$$

Taking full expectation in (78) and substituting (91), we obtain

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2] &+ \frac{2\eta_t^2}{p_t^2} \mathbb{E}[\|\mathbf{y}_{t+1} - \mathbf{y}_\star\|^2] \\ &\leq (1 - \mu\eta_t) \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}_\star\|^2] + 2\eta_t^2 \sigma^2 \\ &\quad + \frac{\eta_t^2(2-p_t^2)}{p_t^2} \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_\star\|^2].\end{aligned}\quad (92)$$

Let us denote  $\Psi_{t+1} := \mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}_\star\|^2 + \frac{\eta_t}{\mu} \|\mathbf{y}_{t+1} - \mathbf{y}_\star\|^2]$ . From the initialization and Assumption A3, we have that  $\mathbb{E}\|\mathbf{y}_0 - \mathbf{y}_\star\|^2 \leq 2L_f^2 \|\mathbf{x}_0 - \mathbf{x}_\star\|^2 + 4\sigma^2$ , implying that  $\Psi_0 \leq$

$\delta_0 + \frac{\eta-1}{\mu} \mathbb{E}\|\mathbf{y}_0 - \mathbf{y}_\star\|^2 \leq (1 + 4\kappa^2) \delta_0 + \frac{4\sigma^2}{\mu^2}$  for  $\eta_{-1} \leq \frac{2}{\mu}$ . If we set  $p_t = \sqrt{2\mu\eta_t}$ , we would obtain the one-step inequality:

$$\Psi_{t+1} \leq (1 - \mu\eta_t) \Psi_t + 2\eta_t^2 \sigma^2 \quad (93)$$

for  $\eta_t \leq \eta_{t-1}$ . Recall that Lemma 4 also requires that  $\eta_t \leq p_t/2L$  or equivalently  $\eta_t \leq \frac{\mu}{2L^2}$ . Therefore, if we set  $\eta_t = \frac{2}{\mu(t+\omega+1)}$  for  $t \geq -1$  with  $\omega = \lceil \frac{4L^2}{\mu^2} \rceil$ , and proceed as in (68)-(71), we obtain the bound

$$\Psi_T \leq \frac{8\sigma^2}{\mu^2 T} + \frac{\omega^2 \Psi_0}{T^2} \leq \frac{8\sigma^2}{\mu^2 T} + \frac{4\kappa^4((1+4\kappa^2)\mu^2 \delta_0 + 4\sigma^2)}{\mu^2 T^2} \quad (94)$$

The obtained bounds hence translate to an SFO complexity of  $\mathcal{O}\left(\frac{\sigma^2}{\mu^2 \epsilon} + \kappa^2 \frac{\kappa \sqrt{\delta_0 + \sigma}}{\sqrt{\epsilon}}\right)$ . However, the average number of calls to the QMO are bounded as  $\sum_{t=1}^{T-1} p_t \leq 4\sqrt{T+\omega}$  or  $\mathcal{O}\left(\frac{\sigma}{\kappa + \mu\sqrt{\epsilon}} + \frac{\kappa \sqrt{\kappa \sqrt{\delta_0 + \sigma}}}{\epsilon^{1/4}}\right)$ .  $\square$

## REFERENCES

- [1] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," in *Proc. of the Conf. on Uncertainty in Artificial Intelligence*, 2010, pp. 733–742.
- [2] A. F. Martins, N. A. Smith, M. Figueiredo, and P. Aguiar, "Structured sparsity in structured prediction," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 2011, pp. 1500–1511.
- [3] S. Melacci, M. Maggini, and M. Gori, "Semi-supervised learning with constraints for multi-view object recognition," in *Proc. of the Intl. Conf. on Artificial Neural Networks*. Springer, 2009, pp. 653–662.
- [4] I. Necoara and N. K. Singh, "Stochastic subgradient for composite convex optimization with functional constraints," *Journal of Machine Learning Research*, vol. 23, no. 265, pp. 1–35, 2022.
- [5] G. Lan and Z. Zhou, "Algorithms for stochastic optimization with function or expectation constraints," *Computational Optimization and Applications*, vol. 76, no. 2, pp. 461–498, 2020.
- [6] A. Nedić and I. Necoara, "Random minibatch subgradient algorithms for convex problems with functional constraints," *Applied Mathematics and Optimization*, vol. 80, no. 3, pp. 801–833, 2019.
- [7] K. Basu and P. Nandy, "Optimal convergence for stochastic optimization with multiple expectation constraints," *arXiv preprint arXiv:1906.03401*, 2019.
- [8] A. Bayandina, P. Dvurechensky, A. Gasnikov, F. Stonyakin, and A. Titov, "Mirror descent and convex optimization problems with non-smooth inequality constraints," in *Large-Scale and Distributed Optimization*. Springer, 2018, pp. 181–213.
- [9] Y. Xu, "Primal-dual stochastic gradient method for convex programs with many functional constraints," *SIAM Journal on Optimization*, vol. 30, no. 2, pp. 1664–1692, 2020.
- [10] E. Yazdandoost Hamedani, A. Jalilzadeh, and N. Serhat Aybat, "A randomized block-coordinate primal-dual method for large-scale stochastic saddle point problems," *arXiv e-prints*, pp. arXiv–1907, 2019.
- [11] A. N. Madavan and S. Bose, "A stochastic primal-dual method for optimization with conditional value at risk constraints," *Journal of Optimization Theory and Applications*, vol. 190, no. 2, pp. 428–460, 2021.
- [12] Y. Yan and Y. Xu, "Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs," *Mathematical Programming Computation*, vol. 14, no. 2, pp. 319–363, 2022.
- [13] J. Yuan and A. Lampserski, "Online convex optimization for cumulative constraints," in *Proc. of the Intl. Conf. on Neural Information Processing Systems*, 2018, p. 6140–6149.
- [14] H. Yu and M. J. Neely, "Online convex optimization with stochastic constraints," in *Proc. of the Intl. Conf. on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.
- [15] Q. Lin, R. Ma, and T. Yang, "Level-set methods for finite-sum constrained convex optimization," in *Intl. Conf. on Machine Learning*. PMLR, 2018, pp. 3112–3121.
- [16] S.-P. Han and O. L. Mangasarian, "Exact penalty functions in nonlinear programming," *Mathematical programming*, vol. 17, no. 1, pp. 251–269, 1979.
- [17] G.-H. Lin and M. Fukushima, "Some exact penalty results for nonlinear programs and mathematical programs with equilibrium constraints," *Journal of Optimization Theory and Applications*, vol. 118, no. 1, pp. 67–80, 2003.

- [18] J. Zhang and L. Xiao, "Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization," *Mathematical Programming*, pp. 1–43, 2021.
- [19] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1997.
- [20] P. E. Gill and E. Wong, "Sequential quadratic programming methods," in *Mixed integer nonlinear programming*. Springer, 2012, pp. 147–224.
- [21] F. E. Curtis, D. P. Robinson, and B. Zhou, "Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints," *SIAM Journal on Optimization*, vol. 34, no. 4, pp. 3592–3622, 2024.
- [22] N. Doikov and Y. Nesterov, "High-order optimization methods for fully composite problems," *SIAM Journal on Optimization*, vol. 32, no. 3, pp. 2402–2427, 2022.
- [23] D. Davis and D. Drusvyatskiy, "Stochastic model-based minimization of weakly convex functions," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 207–239, 2019.
- [24] X. Wang, S. Ma, and Y.-x. Yuan, "Penalty methods with stochastic approximation for stochastic nonlinear programming," *Mathematics of computation*, vol. 86, no. 306, pp. 1793–1820, 2017.
- [25] X. Xiao, "Penalized stochastic gradient methods for stochastic convex optimization with expectation constraints," *Optimization-online*, 2019.
- [26] S. T. Thomdapu and K. Rajawat, "Optimal design of queuing systems via compositional stochastic programming," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8460–8474, 2019.
- [27] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.
- [28] L. Zhang, T. Yang, R. Jin, and X. He, " $\mathcal{O}(\log t)$  projections for stochastic optimization of smooth and strongly convex functions," in *Intl Conf. on Machine Learning*. PMLR, 2013, pp. 1121–1129.
- [29] J. Chen, T. Yang, Q. Lin, L. Zhang, and Y. Chang, "Optimal stochastic strongly convex optimization with a logarithmic number of projections," in *Thirty-Second Conf. on Uncertainty in Artificial Intelligence*. AUAI Press, 2016, pp. 122–131.
- [30] M. Mahdavi, T. Yang, R. Jin, S. Zhu, and J. Yi, "Stochastic gradient descent with only one projection," in *Proc. of the Intl. Conf. on Neural Information Processing Systems*, 2012, pp. 494–502.
- [31] G. Lan, Z. Li, and Y. Zhou, "A unified variance-reduced accelerated gradient method for convex optimization," in *Proc. of the Intl. Conf. on Neural Information Processing Systems*, 2019, pp. 10 462–10 472.
- [32] A. Jalilzadeh, "Primal-dual incremental gradient method for nonsmooth and convex optimization problems," *Optimization Letters*, vol. 15, no. 8, pp. 2541–2554, 2021.
- [33] O. Fercoq, A. Alacaoglu, I. Necoara, and V. Cevher, "Almost surely constrained convex optimization," in *International Conf. on Machine Learning*. PMLR, 2019, pp. 1910–1919.
- [34] A. Kundu, F. Bach, and C. Bhattacharya, "Convex optimization over intersection of simple sets: improved convergence rate guarantees via an exact penalty approach," in *International Conf. on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 958–967.
- [35] D. Boob, Q. Deng, and G. Lan, "Stochastic first-order methods for convex and nonconvex functional constrained optimization," *Mathematical Programming*, vol. 197, no. 1, pp. 215–279, 2023.
- [36] Z. Akhtar, A. Singh Bedi, and K. Rajawat, "Conservative stochastic optimization with expectation constraints," *IEEE Trans. Signal Process.*, vol. 69, pp. 3190–3205, 2021.
- [37] S. T. Thomdapu and K. Rajawat, "Optimizing QOS for erasure-coded wireless data centers," in *IEEE Intl. Conf. on Commun.*, 2021, pp. 1–6.
- [38] S. T. Thomdapu, H. Vardhan, and K. Rajawat, "Stochastic compositional gradient descent under compositional constraints," *IEEE Trans. Signal Process.*, vol. 71, pp. 1115–1127, 2023.
- [39] I. Necoara and A. Nedić, "Minibatch stochastic subgradient-based projection algorithms for feasibility problems with convex inequalities," *Computational Optimization and Applications*, vol. 80, no. 1, pp. 121–152, 2021.
- [40] X. Wei, H. Yu, Q. Ling, and M. J. Neely, "Solving non-smooth constrained programs with lower complexity than  $\mathcal{O}(1/\epsilon)$ : a primal-dual homotopy smoothing approach," in *Proc. of the Intl. Conf. on Neural Information Processing Systems*, 2018, pp. 3999–4009.
- [41] T. Yang, Q. Lin, and L. Zhang, "A richer theory of convex constrained optimization with reduced projections and improved rates," in *International Conf. on Machine Learning*. PMLR, 2017, pp. 3901–3910.
- [42] H. Guo, H. Wei, X. Liu, and L. Ying, "Online convex optimization with hard constraints: towards the best of two worlds and beyond," in *Proc. of the Intl. Conf. on Neural Information Processing Systems*, 2022, pp. 36 426–36 439.
- [43] A. Sinha and R. Vaze, "Optimal algorithms for online convex optimization with adversarial constraints," in *Proc. of the Intl. Conf. on Neural Information Processing Systems*, 2024, pp. 41 274–41 302.
- [44] B. M. Idrees, S. D. Sharma, and K. Rajawat, "Decentralized stochastic successive convex approximation for composite non-convex problems with non-linear functional constraints," in *IEEE ICASSP*, 2025.
- [45] J. C. Duchi and F. Ruan, "Stochastic methods for composite and weakly convex optimization problems," *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 3229–3259, 2018.
- [46] D. Drusvyatskiy and C. Paquette, "Efficiency of minimizing compositions of convex functions and smooth maps," *Mathematical Programming*, vol. 178, no. 1, pp. 503–558, 2019.
- [47] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," in *Proc. of the Conf. Fairness, Accountability, and Transparency in Machine Learning*, 2017.
- [48] A. Khaled, O. Sebbouh, N. Loizou, R. M. Gower, and P. Richtárik, "Unified analysis of stochastic gradient methods for composite convex and smooth optimization," *Journal of Optimization Theory and Applications*, vol. 199, no. 2, pp. 499–540, 2023.
- [49] E. Gorbunov, F. Hanzely, and P. Richtárik, "A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent," in *International Conf. on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 680–690.
- [50] K. Mishchenko, G. Malinovsky, S. U. Stich, and P. Richtarik, "ProxSkip: Yes! local gradient steps provably lead to communication acceleration! finally!" in *Proc. of the Intl. Conf. on Machine Learning*, 2022, pp. 15 750–15 769.
- [51] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 8194–8244, 2017.
- [52] F. Shang, L. Jiao, K. Zhou, J. Cheng, Y. Ren, and Y. Jin, "Asvrg: Accelerated proximal SVRG," in *Asian Conf. on Machine Learning*. PMLR, 2018, pp. 815–830.
- [53] C. Song, Y. Jinag, and Y. Ma, "Variance reduction via accelerated dual averaging for finite-sum optimization," in *Proc. of the Intl. Conf. on Neural Information Processing Systems*, 2020, pp. 833–844.
- [54] Y. Xu, "Iteration complexity of inexact augmented lagrangian methods for constrained convex programming," *Mathematical Programming*, vol. 185, no. 1, pp. 199–244, 2021.
- [55] Q. Lin, S. Nadarajah, and N. Soheili, "A level-set method for convex optimization with a feasible solution path," *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 3290–3311, 2018.
- [56] H. Yu and M. J. Neely, "A simple parallel algorithm with an  $\mathcal{O}(1/t)$  convergence rate for general convex programs," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 759–783, 2017.
- [57] E. Zermelo, "Über das navigationsproblem bei ruhender oder veränderlicher windverteilung," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 11, no. 2, pp. 114–124, 1931.
- [58] Mercator Ocean International. (2025) Mercator ocean – ocean forecasters. [Online; accessed 8-Sept-2025]. [Online]. Available: <https://www.mercator-ocean.eu/>
- [59] Copernicus Marine Service. (2025) Copernicus marine environment monitoring service. [Online; accessed 8-Sept-2025]. [Online]. Available: <https://marine.copernicus.eu/>
- [60] C. Yoo, J. J. Heon Lee, S. Anstee, and R. Fitch, "Path planning in uncertain ocean currents using ensemble forecasts," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2021, pp. 8323–8329.
- [61] D. Jones and G. A. Hollinger, "Planning energy-efficient trajectories in strong disturbances," *IEEE Robot. Autom. Lett.*, vol. 2, no. 4, pp. 2080–2087, 2017.
- [62] H. Song, P. Shi, C.-C. Lim, W.-A. Zhang, and L. Yu, "Set-membership estimation for complex networks subject to linear and nonlinear bounded attacks," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 31, no. 1, pp. 163–173, 2019.
- [63] K. Zhu, Z. Wang, H. Dong, and G. Wei, "Set-membership filtering for two-dimensional systems with dynamic event-triggered mechanism," *Automatica*, vol. 143, p. 110416, 2022.
- [64] A. Flores and R. C. de Lamare, "Set-membership adaptive kernel nlms algorithms: Design and analysis," *Signal Processing*, vol. 154, 2019.
- [65] (2025) Libsvm regression data sets. LIBSVM Data Repository, [Online; accessed 8-Sept-2025]. [Online]. Available: <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html>

## APPENDIX D: SUPPLEMENTARY MATERIAL

## PROOF OF THEOREM 3

For the sake of brevity, let us denote

$$\delta_t^z := \mathbb{E} \|\mathbf{z}_t - \mathbf{x}_\star\|^2, \quad \tilde{\Delta}_s = \mathbb{E} [F(\tilde{\mathbf{x}}_s)] - F(\mathbf{x}_\star). \quad (95)$$

We begin with deriving some preliminary results. Using the definitions of  $\mathbf{x}_t$ ,  $\mathbf{y}_t$ ,  $\mathbf{z}_t$ , and  $\mathbf{z}_{t-1}^+$  in Algorithm 3, we see that

$$\mathbf{x}_t - \mathbf{y}_t = \alpha_s (\mathbf{z}_t - \mathbf{z}_{t-1}^+). \quad (96)$$

$$\mathbf{y}_t - \alpha_s \mathbf{z}_{t-1}^+ = (1 - \alpha_s - \omega_s) \mathbf{x}_{t-1} + \omega_s \tilde{\mathbf{x}}_{s-1}. \quad (97)$$

The key to proving the required result is the following one-step inequality, which looks similar to the result in [31, Lemma 6] but requires a different proof from that in the proximal case.

*Lemma 5.* If the parameters  $\alpha_s$ ,  $\omega_s$  and  $\beta_s$  satisfy

$$\alpha_s + \omega_s \leq 1 \quad (98)$$

$$1 + \mu\beta_s - \alpha_s\beta_s L_\gamma > 0 \quad (99)$$

$$\omega_s - \frac{\alpha_s\beta_s L_f}{1 + \mu\beta_s - \alpha_s\beta_s L_\gamma} \geq 0 \quad (100)$$

then it holds that

$$\begin{aligned} & \frac{\beta_s}{\alpha_s} \Delta_t + (1 + \mu\beta_s) \frac{1}{2} \delta_t^z \\ & \leq \frac{\beta_s}{\alpha_s} (1 - \alpha_s - \omega_s) \Delta_{t-1} + \frac{\beta_s \omega_s}{\alpha_s} \tilde{\Delta}_{s-1} + \frac{1}{2} \delta_{t-1}^z. \end{aligned} \quad (101)$$

*Proof:* We begin with using the equality in (96) and the  $L_g$ -smoothness of  $g_k$  to obtain

$$\begin{aligned} & g_k(\mathbf{y}_t) + \alpha_s \langle \nabla g_k(\mathbf{y}_t), \mathbf{z}_t - \mathbf{z}_{t-1}^+ \rangle \\ & \stackrel{(96)}{=} g_k(\mathbf{y}_t) + \langle \nabla g_k(\mathbf{y}_t), \mathbf{x}_t - \mathbf{y}_t \rangle \stackrel{(A2)}{\geq} g_k(\mathbf{y}_t) - \frac{L_g}{2} \|\mathbf{x}_t - \mathbf{y}_t\|^2 \\ & \stackrel{(96)}{=} g_k(\mathbf{y}_t) - \frac{\alpha_s^2 L_g}{2} \|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2 \end{aligned} \quad (102)$$

Further, from Jensen's inequality, and from (24), we obtain

$$\|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2 \leq \frac{1}{1 + \mu\beta_s} \|\mathbf{z}_t - \mathbf{z}_{t-1}\|^2 + \frac{\mu\beta_s}{1 + \mu\beta_s} \|\mathbf{z}_t - \mathbf{y}_t\|^2. \quad (103)$$

Next, since  $\mathbf{z}_t$ -update in (26) involves minimizing an  $\alpha_s(1 + \mu\beta_s)$ -strongly convex function, we have that

$$\begin{aligned} & \frac{\alpha_s\beta_s\mu}{2} \|\mathbf{y}_t - \mathbf{x}_\star\|^2 + \frac{\alpha_s}{2} \|\mathbf{z}_{t-1} - \mathbf{x}_\star\|^2 - \frac{(1 + \mu\beta_s)\alpha_s}{2} \|\mathbf{z}_t - \mathbf{x}_\star\|^2 \\ & + \gamma\beta_s \max\{[g_k(\mathbf{y}_t) + \alpha_s \langle \nabla g_k(\mathbf{y}_t), \mathbf{x}_\star - \mathbf{z}_{t-1}^+ \rangle]_+\} \\ & \stackrel{(40)}{\geq} \alpha_s\beta_s \langle \tilde{\nabla}_t, \mathbf{z}_t - \mathbf{x}_\star \rangle + \frac{\alpha_s}{2} \|\mathbf{z}_{t-1} - \mathbf{z}_t\|^2 \\ & + \gamma\beta_s \max\{[g_k(\mathbf{y}_t) + \alpha_s \langle \nabla g_k(\mathbf{y}_t), \mathbf{z}_t - \mathbf{z}_{t-1}^+ \rangle]_+\} \\ & + \frac{\alpha_s\beta_s\mu}{2} \|\mathbf{y}_t - \mathbf{z}_t\|^2 + \alpha_s\beta_s h(\mathbf{z}_t) - \alpha_s\beta_s h(\mathbf{x}_\star) \\ & \stackrel{(102)}{\geq} \alpha_s\beta_s \langle \tilde{\nabla}_t, \mathbf{z}_t - \mathbf{x}_\star \rangle + \frac{\alpha_s}{2} \|\mathbf{z}_{t-1} - \mathbf{z}_t\|^2 \\ & + \gamma\beta_s \max\{[g_k(\mathbf{x}_t)]_+\} - \frac{\gamma\alpha_s^2\beta_s L_g}{2} \|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2 \\ & + \frac{\alpha_s\beta_s\mu}{2} \|\mathbf{y}_t - \mathbf{z}_t\|^2 + \alpha_s\beta_s h(\mathbf{z}_t) - \alpha_s\beta_s h(\mathbf{x}_\star). \end{aligned} \quad (104)$$

$$\quad (105)$$

Substituting (103) into (105) and re-arranging, we obtain

$$\begin{aligned} & \frac{\alpha_s\beta_s\mu}{2} \|\mathbf{y}_t - \mathbf{x}_\star\|^2 + \frac{\alpha_s}{2} \|\mathbf{z}_{t-1} - \mathbf{x}_\star\|^2 \\ & + \gamma\beta_s \max\{[g_k(\mathbf{y}_t) + \alpha_s \langle \nabla g_k(\mathbf{y}_t), \mathbf{x}_\star - \mathbf{z}_{t-1}^+ \rangle]_+\} \\ & - (1 + \mu\beta_s) \frac{\alpha_s}{2} \|\mathbf{z}_t - \mathbf{x}_\star\|^2 - \alpha_s\beta_s h(\mathbf{z}_t) + \alpha_s\beta_s h(\mathbf{x}_\star) \\ & \geq \alpha_s\beta_s \langle \tilde{\nabla}_t, \mathbf{z}_t - \mathbf{x}_\star \rangle + \gamma\beta_s \max\{[g_k(\mathbf{x}_t)]_+\} \\ & + \frac{(1 + \mu\beta_s)\alpha_s - \alpha_s^2\beta_s\gamma L_g}{2} \|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2 \end{aligned} \quad (106)$$

Further, from the convexity of  $g_k$  and (98), we have that

$$\begin{aligned} & g_k(\mathbf{y}_t) + \alpha_s \langle \nabla g_k(\mathbf{y}_t), \mathbf{x}_\star - \mathbf{z}_{t-1} \rangle \\ & \leq g_k(\mathbf{y}_t + \alpha_s(\mathbf{x}_\star - \mathbf{z}_{t-1})) \\ & \stackrel{(97)}{=} g_k((1 - \alpha_s - \omega_s)\mathbf{x}_{t-1} + \alpha_s\mathbf{x}_\star + \omega_s\tilde{\mathbf{x}}_{s-1}) \\ & \leq (1 - \alpha_s - \omega_s)g_k(\mathbf{x}_{t-1}) + \alpha_s g_k(\mathbf{x}_\star) + \omega_s g_k(\tilde{\mathbf{x}}_{s-1}) \\ & \leq (1 - \alpha_s - \omega_s)g_k(\mathbf{x}_{t-1}) + \omega_s g_k(\tilde{\mathbf{x}}_{s-1}) \end{aligned} \quad (107)$$

where the last inequality uses the fact that  $g_k(\mathbf{x}_\star) \leq 0$ . Therefore, from the monotonicity of the  $\max[\cdot]_+$  operator, we obtain

$$\begin{aligned} & \max\{[g_k(\mathbf{y}_t) + \alpha_s \langle \nabla g_k(\mathbf{y}_t), \mathbf{x}_\star - \mathbf{z}_{t-1} \rangle]_+\} \\ & \stackrel{(107)}{\leq} \max\{[(1 - \alpha_s - \omega_s)g_k(\mathbf{x}_{t-1}) + \omega_s g_k(\tilde{\mathbf{x}}_{s-1})]_+\} \\ & \stackrel{(38)}{\leq} (1 - \alpha_s - \omega_s) \max\{[g_k(\mathbf{x}_{t-1})]_+\} \\ & \quad + \omega_s \max\{[g_k(\tilde{\mathbf{x}}_{s-1})]_+\}. \end{aligned} \quad (108)$$

Substituting (109) into (106), we obtain

$$\begin{aligned} & \alpha_s\beta_s \langle \tilde{\nabla}_t, \mathbf{z}_t - \mathbf{x}_\star \rangle + \frac{(1 + \mu\beta_s)\alpha_s - \alpha_s^2\beta_s\gamma L_g}{2} \|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2 \\ & + \gamma\beta_s \max\{[g_k(\mathbf{x}_t)]_+\} + (1 + \mu\beta_s) \frac{\alpha_s}{2} \|\mathbf{z}_t - \mathbf{x}_\star\|^2 \\ & + \alpha_s\beta_s h(\mathbf{z}_t) - \alpha_s\beta_s h(\mathbf{x}_\star) \\ & \leq \frac{\alpha_s\beta_s\mu}{2} \|\mathbf{y}_t - \mathbf{x}_\star\|^2 + \frac{\alpha_s}{2} \|\mathbf{z}_{t-1} - \mathbf{x}_\star\|^2 \\ & + \gamma\beta_s(1 - \alpha_s - \omega_s) \max\{[g_k(\mathbf{x}_{t-1})]_+\} \\ & + \gamma\beta_s\omega_s \max\{[g_k(\tilde{\mathbf{x}}_{s-1})]_+\} \end{aligned} \quad (109)$$

Since  $f$  is  $L_f$ -smooth and convex, we have that

$$\begin{aligned} & f(\mathbf{x}_t) \stackrel{(41)}{\leq} f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_t - \mathbf{y}_t \rangle + \frac{L_f}{2} \|\mathbf{x}_t - \mathbf{y}_t\|^2 \\ & \stackrel{(96),(27)}{=} f(\mathbf{y}_t) + (1 - \alpha_s - \omega_s) \langle \nabla f(\mathbf{y}_t), \mathbf{x}_{t-1} - \mathbf{y}_t \rangle \\ & + \alpha_s \langle \nabla f(\mathbf{y}_t), \mathbf{z}_t - \mathbf{y}_t \rangle + \omega_s \langle \nabla f(\mathbf{y}_t), \tilde{\mathbf{x}}_{s-1} - \mathbf{y}_t \rangle \\ & + \frac{\alpha_s^2 L_f}{2} \|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2 \end{aligned} \quad (110)$$

$$\begin{aligned} & \stackrel{(A2)}{\leq} (1 - \alpha_s - \omega_s) f(\mathbf{x}_{t-1}) \\ & + \alpha_s (f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_\star - \mathbf{y}_t \rangle + \langle \nabla f(\mathbf{y}_t), \mathbf{z}_t - \mathbf{x}_\star \rangle) \\ & + \omega_s (f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \tilde{\mathbf{x}}_{s-1} - \mathbf{y}_t \rangle) + \frac{\alpha_s^2 L_f}{2} \|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2 \end{aligned} \quad (111)$$

Adding (112) and (110), we obtain

$$\begin{aligned}
f(\mathbf{x}_t) + \gamma \max\{[g_k(\mathbf{x}_t)]_+\} &\leq -\alpha_s \langle \tilde{\nabla}_t, \mathbf{z}_t - \mathbf{x}_\star \rangle \\
&\quad - \frac{(1 + \mu\beta_s)\alpha_s - \beta_s\alpha_s^2(L_f + \gamma L_g)}{2\beta_s} \|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2 \\
&\quad + \frac{\alpha_s}{2\beta_s} \|\mathbf{z}_{t-1} - \mathbf{x}_\star\|^2 - (1 + \mu\beta_s) \frac{\alpha_s}{2\beta_s} \|\mathbf{z}_t - \mathbf{x}_\star\|^2 \\
&\quad + \omega_s(f(\tilde{\mathbf{x}}_{s-1}) + \gamma \max\{[g_k(\tilde{\mathbf{x}}_{s-1})]_+\}) \\
&\quad + \omega_s(f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \tilde{\mathbf{x}}_{s-1} - \mathbf{y}_t \rangle) - f(\tilde{\mathbf{x}}_{s-1}) \\
&\quad + \alpha_s \left( f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_\star - \mathbf{y}_t \rangle + \frac{\mu}{2} \|\mathbf{y}_t - \mathbf{x}_\star\|^2 \right) \\
&\quad + (1 - \alpha_s - \omega_s) \left( f(\mathbf{x}_{t-1}) + \gamma \max\{[g_k(\mathbf{x}_{t-1})]_+\} \right) \\
&\quad + \alpha_s \langle \nabla f(\mathbf{y}_t), \mathbf{z}_t - \mathbf{x}_\star \rangle - \alpha_s h(\mathbf{z}_t) + \alpha_s h(\mathbf{x}_\star). \quad (113)
\end{aligned}$$

Moreover, by convexity of  $h$ , we have that

$$-\alpha_s h(\mathbf{z}_t) \leq -h(\mathbf{x}_t) + (1 - \alpha_s - \omega_s)h(\mathbf{x}_{t-1}) + \omega_s h(\tilde{\mathbf{x}}_{s-1}). \quad (114)$$

Substituting (114) in (113) and using the facts that  $f$  is  $\mu$ -strongly convex and  $\max\{[g_k(\mathbf{x}_\star)]_+\} = 0$ , we obtain

$$\begin{aligned}
F(\mathbf{x}_t) &\leq \alpha_s \langle \nabla f(\mathbf{y}_t) - \tilde{\nabla}_t, \mathbf{z}_t - \mathbf{x}_\star \rangle \\
&\quad - \frac{(1 + \mu\beta_s)\alpha_s - \beta_s\alpha_s^2(L_f + \gamma L_g)}{2\beta_s} \|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2 \\
&\quad - (1 + \mu\beta_s) \frac{\alpha_s}{2\beta_s} \|\mathbf{z}_t - \mathbf{x}_\star\|^2 + (1 - \alpha_s - \omega_s)F(\mathbf{x}_{t-1}) \\
&\quad + \alpha_s F(\mathbf{x}_\star) + \omega_s F(\tilde{\mathbf{x}}_{s-1}) + \frac{\alpha_s}{2\beta_s} \|\mathbf{z}_{t-1} - \mathbf{x}_\star\|^2 \\
&\quad - \omega_s D_f(\tilde{\mathbf{x}}_{s-1}, \mathbf{y}_t) \quad (115)
\end{aligned}$$

Next, we take expectation with respect to  $i_t$  and consider the different terms in (115) separately. First note that since  $\mathbf{y}_t$  and  $\mathbf{z}_{t-1}^+$  are independent of  $i_t$  and hence  $\mathbb{E}_t[\tilde{\nabla}] = \nabla f(\mathbf{y}_t)$ , the first term in (115) can be bounded as

$$\begin{aligned}
&\alpha_s \mathbb{E}_t[\langle \nabla f(\mathbf{y}_t) - \tilde{\nabla}_t, \mathbf{z}_t - \mathbf{x}_\star \rangle] \\
&= \alpha_s \mathbb{E}_t[\langle \nabla f(\mathbf{y}_t) - \tilde{\nabla}_t, \mathbf{z}_t - \mathbf{z}_{t-1}^+ \rangle] \\
&\quad + \alpha_s \langle \nabla f(\mathbf{y}_t) - \mathbb{E}_t[\tilde{\nabla}_t], \mathbf{z}_{t-1}^+ - \mathbf{x}_\star \rangle \\
&\stackrel{(43)}{\leq} \frac{\alpha_s \beta_s}{2(1 + \mu\beta_s - \alpha_s \beta_s L_\gamma)} \mathbb{E}_t[\|\nabla f(\mathbf{y}_t) - \tilde{\nabla}_t\|^2] \\
&\quad + \frac{(1 + \mu\beta_s)\alpha_s - \alpha_s^2 \beta_s L_\gamma}{2\beta_s} \mathbb{E}_t[\|\mathbf{z}_t - \mathbf{z}_{t-1}^+\|^2]. \quad (116)
\end{aligned}$$

where recall that  $L_\gamma := L_f + \gamma L_g$  and from (99). Here, since  $\mathbb{E}_t[\nabla f_{i_t}(\tilde{\mathbf{x}}_{s-1})] = \nabla f(\tilde{\mathbf{x}}_{s-1})$ , the variance of  $\tilde{\nabla}$  can be bounded as

$$\begin{aligned}
&\mathbb{E}_t[\|\tilde{\nabla}_t - \nabla f(\mathbf{y}_t)\|^2] \leq \mathbb{E}_t[\|\nabla f_{i_t}(\mathbf{y}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}_{s-1})\|^2] \\
&\stackrel{(42)}{\leq} 2L_f(\mathbb{E}_t[f_{i_t}(\tilde{\mathbf{x}}_{s-1}) - f_{i_t}(\mathbf{y}_t)] - \langle \nabla f_{i_t}(\mathbf{y}_t), \tilde{\mathbf{x}}_{s-1} - \mathbf{y}_t \rangle) \\
&= 2L_f D_f(\tilde{\mathbf{x}}_{s-1}, \mathbf{y}_t) \quad (117)
\end{aligned}$$

Substituting (117) into (116), and adding with (115) after taking expectation with respect to  $i_t$ , we obtain

$$\begin{aligned}
&\mathbb{E}_t[F(\mathbf{x}_t) + (1 + \mu\beta_s) \frac{\alpha_s}{2\beta_s} \|\mathbf{z}_t - \mathbf{x}_\star\|^2] \\
&\leq (1 - \alpha_s - \omega_s)F(\mathbf{x}_{t-1}) + \frac{\alpha_s}{2\beta_s} \|\mathbf{z}_{t-1} - \mathbf{x}_\star\|^2 \\
&\quad + \omega_s F(\tilde{\mathbf{x}}_{s-1}) + \alpha_s F(\mathbf{x}_\star) \\
&\quad - (\omega_s - \frac{\alpha_s \beta_s L_f}{1 + \mu\beta_s - \alpha_s \beta_s L_\gamma}) D_f(\tilde{\mathbf{x}}_{s-1}, \mathbf{y}_t) \quad (118)
\end{aligned}$$

where observe that the last term is non-positive from (100) and can be dropped. Taking full expectation and re-arranging, we obtain the required result in (101) ■

It is remarked that the conditions required in (98)-(100) are satisfied by the choice of parameters in the statement of Theorem 3. The statement of Lemma 5 will subsequently be used for each case in Theorem 3.

#### A. Proof of Theorem 3(1)

For general convex function  $f$ , setting  $\mu = 0$  in (101),

$$\frac{\beta_s}{\alpha_s} \Delta_t + \frac{1}{2} \delta_t^z \leq \frac{\beta_s}{\alpha_s} (1 - \alpha_s - \omega_s) \Delta_{t-1} + \frac{\beta_s \omega_s}{\alpha_s} \tilde{\Delta}_{s-1} + \frac{1}{2} \delta_{t-1}^z$$

Substituting  $\theta_t$  as specified in (28) at line 10 of Alg. 3 and summing the recursive expression for  $t = 1, \dots, T_s$ ,

$$\begin{aligned}
\tilde{\Delta}_s \sum_{t=1}^{T_s} \theta_t &\leq \sum_{t=1}^{T_s} \theta_t \Delta_t \leq \left[ \frac{\beta_s}{\alpha_s} (1 - \alpha_s) + (T_s - 1) \frac{\beta_s \omega_s}{\alpha_s} \right] \tilde{\Delta}_{s-1} \\
&\quad + \frac{1}{2} (\delta_{T_s-1}^z - \delta_{T_s}^z) \quad (119)
\end{aligned}$$

where, we have utilized the fact that  $\mathbf{x}_0 = \tilde{\mathbf{x}}_{s-1}$ ,  $\tilde{\mathbf{x}}_s = \sum_t \theta_t \mathbf{x}_t / \sum_t \theta_t$ , and the convexity of  $F$ . Let us denote

$$\mathcal{L}_s \stackrel{(28)}{:=} \sum_{t=1}^{T_s} \theta_t = \frac{\beta_s}{\alpha_s} + (T_s - 1) \frac{\beta_s (\alpha_s + \omega_s)}{\alpha_s} \quad (120)$$

$$\mathcal{R}_s := \frac{\beta_s}{\alpha_s} (1 - \alpha_s) + (T_s - 1) \frac{\beta_s \omega_s}{\alpha_s}. \quad (121)$$

so that  $\mathcal{L}_s \tilde{\Delta}_s = \mathcal{R}_s \tilde{\Delta}_{s-1} + \frac{1}{2} (\delta_{T_s-1}^z - \delta_{T_s}^z)$ . Summing over  $j = 1, \dots, s$  and rewriting,

$$\mathcal{L}_s \Delta_s + \sum_{j=1}^{s-1} (\mathcal{L}_j - \mathcal{R}_{j+1}) \tilde{\Delta}_j \leq \mathcal{R}_1 \tilde{\Delta}_0 + \frac{1}{2} \delta_0^z - \frac{1}{2} \delta_{T_s}^z \quad (122)$$

where the last term can be dropped.

Next, we can show that  $\nu_s := \mathcal{L}_s - \mathcal{R}_{s+1} \geq 0$  for  $s \geq 1$ . For  $1 \leq s < s_0$ , we have  $\alpha_s = \alpha_{s+1} = \omega_s = 1/2$ ,  $\beta_s = \beta_{s+1}$ , and  $T_{s+1} = 2T_s$ , so that  $\nu_s = 0$ . For  $s \geq s_0$ ,

$$\begin{aligned}
\nu_s &= \frac{\beta_s}{\alpha_s} - \frac{\beta_{s+1}}{\alpha_{s+1}} + \beta_{s+1} + (T_{s_0} - 1) \left[ \frac{\beta_s (\alpha_s + \omega_s)}{\alpha_s} - \frac{\beta_{s+1} \omega_{s+1}}{\alpha_{s+1}} \right] \\
&= \frac{1}{24L_\gamma} (2 + (T_{s_0} - 1) (2(s - s_0 + 4) - 1)) \geq 0, \quad (123)
\end{aligned}$$

Setting  $\bar{\mathbf{x}}_s := \sum_{j=1}^{s-1} \nu_j \tilde{\mathbf{x}}_j / \sum_{k=1}^{s-1} \nu_k$  and using the convexity of  $F$ , we obtain

$$\begin{aligned}
&\mathcal{L}_s \tilde{\Delta}_s + (\mathbb{E}[F(\bar{\mathbf{x}}_s)] - F(\mathbf{x}_\star)) \sum_{j=1}^{s-1} \nu_j \leq \mathcal{R}_1 \tilde{\Delta}_0 + \frac{1}{2} \delta_{T_0}^z \\
&\Rightarrow \mathcal{L}_s \tilde{\Delta}_s \leq \mathcal{R}_1 \tilde{\Delta}_0 + \frac{1}{2} \delta_{T_0}^z \quad (124)
\end{aligned}$$

where from  $(\mathcal{P}_1)$ , we have used  $F(\bar{\mathbf{x}}_s) \geq F(\mathbf{x}_*)$ . Here,  $\mathcal{R}_1 = 2/3L_\gamma$ . In the case when  $1 \leq s \leq s_0$ , we have that  $\mathcal{L}_s = \frac{2^{s+1}}{3L_\gamma}$  so that  $\tilde{\Delta}_s \leq 2^{-(s+1)}D_0$ . For  $s > s_0$ , we have

$$\begin{aligned} \mathcal{L}_s &= \frac{1}{3L_\gamma \alpha_s^2} \left[ 1 + (T_{s_0} - 1) \left( \alpha_s + \frac{1}{2} \right) \right] \\ &= \frac{(s-s_0+4)(T_{s_0}-1)}{6L_\gamma} + \frac{(s-s_0+4)^2(T_{s_0}+1)}{24L_\gamma} \geq \frac{(s-s_0+4)^2 n}{48L_\gamma} \end{aligned}$$

where the last inequality follows from  $T_{s_0} = 2^{\lfloor \log_2 n \rfloor} \geq n/2$ . Hence we obtain

$$\tilde{\Delta}_s \leq \frac{16D_0}{(s-s_0+4)^2 n}. \quad (125)$$

First consider the case when  $n \geq D_0/\epsilon$  for a given  $\epsilon$ . The algorithm cannot run for more than  $s_0$  epochs, which can easily be checked as  $2^{-(S_l-1)}D_0 \leq \epsilon$  which implies, that the total number of epochs the algorithm runs is given by

$$S_l = \min \left\{ \log \frac{D_0}{\epsilon}, s_0 \right\} = \log \frac{D_0}{\epsilon} \quad (126)$$

since  $s_0 = \lfloor \log n \rfloor + 1 \geq \log \frac{D_0}{\epsilon}$ . The SFO and QMO complexities are then given by

$$N_{\text{QMO}} = \sum_{s=1}^{S_l} T_s = \mathcal{O} \left( \min \left\{ \frac{D_0}{\epsilon}, n \right\} \right) = \mathcal{O} \left( \frac{D_0}{\epsilon} \right) \quad (127)$$

$$N_{\text{SFO}} = nS_l + \sum_{s=1}^{S_l} T_s = \mathcal{O} \left( n \log \frac{D_0}{\epsilon} \right) \quad (128)$$

where we have used the fact that  $n \geq \frac{D_0}{\epsilon}$ . Next we consider the case when  $n < D_0/\epsilon$  and it is possible to evaluate true gradient for more than  $s_0$  epochs as

$$S_h = \left\lceil \sqrt{\frac{16D_0}{n\epsilon}} + s_0 - 4 \right\rceil \quad (129)$$

The total number of gradient evaluations of  $f_i$  is

$$N_{\text{SFO}} = ns_0 + \sum_{s=1}^{s_0} T_s + (S_h - s_0)(n + T_{s_0}) \quad (130)$$

$$\begin{aligned} &= ns_0 + 2^{s_0} - 1 + (n + T_{s_0}) S_h - s_0 2^{s_0-1} - ns_0 \\ &\leq (n + T_{s_0}) S_h \leq (n + 2^{s_0-1}) \left( \sqrt{\frac{16D_0}{n\epsilon}} + s_0 \right) \\ &\leq (n + n) \left( \sqrt{\frac{16D_0}{n\epsilon}} + \log n \right) = \mathcal{O} \left( n \log n + \sqrt{\frac{nD_0}{\epsilon}} \right) \end{aligned}$$

$$N_{\text{QMO}} = \sum_{s=1}^{s_0} T_s + T_{s_0} (S_h - s_0) \quad (131)$$

$$= (2^{s_0} - 1) + (2^{s_0-1}) \left( \sqrt{\frac{16nD_0}{\epsilon}} - 4 \right) = \mathcal{O} \left( \sqrt{\frac{nD_0}{\epsilon}} \right)$$

Combining, we obtain the desired bound in (29).

### B. Proof of Theorem 3(2)

For this result, we separately consider different cases based on values of  $s$  and  $n$ .

1) *Case  $s \leq s_0$* : In this case  $\alpha_s = \omega_s = \frac{1}{2}$ ,  $\beta_s = \frac{2}{3L_\gamma}$  and  $T_s = 2^{s-1}$ , so we can write (101) as

$$\frac{\beta_s}{\alpha_s} \Delta_t + (1 + \mu\beta_s) \frac{1}{2} \delta_t^z \leq \frac{\beta_s \omega_s}{\alpha_s} \tilde{\Delta}_{s-1} + \frac{1}{2} \delta_{t-1}^z. \quad (132)$$

Summing over  $t = 1, \dots, T_s$ , we obtain

$$\frac{\beta_s}{\alpha_s} \sum_{t=1}^{T_s} \Delta_t + \frac{1}{2} \delta_{T_s}^z + \frac{\mu\beta_s}{2} \sum_{t=1}^{T_s} \delta_t^z \leq \frac{\beta_s T_s}{2\alpha_s} \tilde{\Delta}_{s-1} + \frac{1}{2} \delta_0^z. \quad (133)$$

Using the definitions of  $\tilde{\mathbf{x}}_s$  and  $\theta_t$ , we obtain

$$\begin{aligned} \frac{4T_s}{3L_\gamma} \tilde{\Delta}_s + \frac{1}{2} \delta_{T_s}^z &\leq \frac{4T_s}{6L_\gamma} \tilde{\Delta}_{s-1} + \frac{1}{2} \delta_{T_{s-1}}^z \\ &= \frac{4T_{s-1}}{3L_\gamma} \tilde{\Delta}_{s-1} + \frac{1}{2} \delta_{T_{s-1}}^z \end{aligned} \quad (134)$$

Applying the inequality recursively over  $s$ ,

$$\frac{4T_s}{3L_\gamma} \tilde{\Delta}_s + \frac{1}{2} \delta_{T_s}^z \leq \frac{4}{3L_\gamma} \tilde{\Delta}_0 + \frac{1}{2} \delta_{T_0}^z \quad (135)$$

By substituting  $T_s = 2^{s-1}$ , we conclude  $\tilde{\Delta}_s \leq 2^{-(s+1)}D_0$ . Now by (128) and (127) we get the SFO and QMO complexities.

2) *Case  $s > s_0$  and  $n \geq \frac{3\kappa}{4}$* : In this case  $\alpha_s = \omega_s = \frac{1}{2}$ ,  $\beta_s = \frac{2}{3L_\gamma}$ , and  $T_s = T_{s_0} = 2^{s_0-1}$  so (101) yields

$$\frac{4}{3L_\gamma} \Delta_t + \left( 1 + \frac{2}{3\kappa} \right) \frac{1}{2} \delta_t^z \leq \frac{2}{3L_\gamma} \tilde{\Delta}_{s-1} + \frac{1}{2} \delta_{t-1}^z \quad (136)$$

Multiplying both sides by  $\theta_t = \Gamma_{t-1} = \left( 1 + \frac{2}{3\kappa} \right)^{t-1}$ ,

$$\frac{4\Gamma_{t-1}}{3L_\gamma} \Delta_t + \frac{\Gamma_t}{2} \delta_t^z \leq \frac{2\Gamma_{t-1}}{3L_\gamma} \tilde{\Delta}_{s-1} + \frac{\Gamma_{t-1}}{2} \delta_{t-1}^z. \quad (137)$$

Summing over  $t = 1, \dots, T_s$ , we obtain

$$\frac{4}{3L_\gamma} \sum_{t=1}^{T_s} \theta_t \Delta_t + \frac{\Gamma_{T_s}}{2} \delta_{T_s}^z \leq \frac{2}{3L_\gamma} \tilde{\Delta}_{s-1} \sum_{t=1}^{T_s} \theta_t + \frac{1}{2} \delta_{T_{s-1}}^z$$

For  $s \geq s_0$ , it holds that  $n \geq 2^{s_0-1} = T_{s_0} = 2^{\lfloor \log n \rfloor} \geq n/2$ , and hence it follows

$$\Gamma_{T_s} = \left( 1 + \frac{2}{3\kappa} \right)^{T_s} \geq 1 + \frac{2\mu T_{s_0}}{3L_\gamma} \geq 1 + \frac{T_{s_0}}{2n} \geq \frac{5}{4} \quad (138)$$

Denoting  $\Theta_s := \sum_{t=1}^{T_s} \theta_t \geq T_s$  and applying these inequalities we obtain

$$\frac{5}{4} \times \frac{2}{3L_\gamma} \tilde{\Delta}_s + \frac{5}{4\Theta_s} \times \frac{1}{2} \delta_{T_s}^z \leq \frac{2}{3L_\gamma} \tilde{\Delta}_{s-1} + \frac{1}{2\Theta_s} \delta_{T_{s-1}}^z \quad (139)$$

which upon continuing recursively for  $s \geq s_0$ , yields

$$\begin{aligned} \frac{2}{3L_\gamma} \tilde{\Delta}_s + \frac{1}{2\Theta_s} \delta_{T_s}^z &\leq \left( \frac{4}{5} \right)^{s-s_0} \left[ \frac{2}{3L_\gamma} \tilde{\Delta}_{s_0} + \frac{1}{2\Theta_{s_0}} \delta_{T_{s_0}}^z \right] \\ &\stackrel{(135)}{\leq} \left( \frac{4}{5} \right)^{s-s_0} \left[ \frac{4}{3L_\gamma T_{s_0}} \tilde{\Delta}_0 + \frac{1}{2T_{s_0}} \delta_{T_0}^z \right] \end{aligned} \quad (140)$$

Substituting  $T_{s_0} = 2^{s_0-1}$  and  $D_0$ , we obtain the final result for this case as  $\tilde{\Delta}_s \leq \left( \frac{4}{5} \right)^{s-s_0} \frac{D_0}{2^{s_0}} \leq \left( \frac{4}{5} \right)^s D_0$ . Observing that VARAS runs for  $S = \mathcal{O}(\log \frac{D_0}{\epsilon})$  epochs, we bound SFO and QMO evaluations as,

$$N_{\text{SFO}} = nS + \sum_{s=1}^S T_s \leq 2nS = \mathcal{O} \left( n \log \frac{D_0}{\epsilon} \right). \quad (141)$$

$$N_{\text{QMO}} = \sum_{s=1}^S T_s = \mathcal{O} \left( n + (S - s_0)n \right) = \mathcal{O} \left( n \log \frac{D_0}{\epsilon} \right) \quad (142)$$



3) *Case*  $s_0 < s \leq s_0 + \sqrt{\frac{12\kappa}{n}} - 4$  and  $n < \frac{3\kappa}{4}$ : In this case  $\alpha_s = \frac{2}{s-s_0+4}$ ,  $\omega_s = \frac{1}{2}$ ,  $\beta_s = \frac{s-s_0+4}{6L_\gamma}$ , and  $T_s = T_{s_0} = 2^{s_0-1}$ . Observe that the parameter setting is same as in the smooth convex case with  $\mu = 0$ . Hence the same result holds for positive  $\mu$  values too which is

$$\mathcal{L}_s \tilde{\Delta}_s + \frac{1}{2} \delta_{T_s}^z \leq \mathcal{R}_{s_0+1} \tilde{\Delta}_{s_0} + \frac{1}{2} \delta_{T_{s_0}}^z \leq \frac{D_0}{3L_\gamma}. \quad (143)$$

where the last inequality follows since  $\mathcal{L}_{s_0} \geq \frac{2T_{s_0}}{3L_\gamma}$ . From the analysis in previous subsection,  $\mathcal{L}_s \geq (s - s_0 + 4)^2 \frac{n}{48L_\gamma}$  and hence  $\tilde{\Delta}_s \leq \frac{16D_0}{(s-s_0+4)^2 n}$ . Following (130) and (131), we bound gradient and QP evaluations.

4) *Case*  $s > s_0 + \sqrt{\frac{12\kappa}{n}} - 4$  and  $n < \frac{3\kappa}{4}$ : In this case,  $\alpha_s = \sqrt{\frac{n}{3\kappa}}$ ,  $\omega_s = \frac{1}{2}$ ,  $\beta_s = \frac{1}{\sqrt{3nL_\gamma\mu}}$ , and  $T_s = T_{s_0} = 2^{s_0-1}$ . By multiplying with  $\Gamma_{t-1}$  on both sides of (101), we obtain

$$\begin{aligned} \frac{\beta_s}{\alpha_s} \Gamma_{t-1} \Delta_t + \Gamma_t \frac{1}{2} \delta_t^z &\leq \frac{\beta_s \omega_s}{\alpha_s} \Gamma_{t-1} \tilde{\Delta}_{s-1} \\ &+ \frac{\beta_s}{\alpha_s} (1 - \alpha_s - \omega_s) \Gamma_{t-1} \Delta_{t-1} + \Gamma_{t-1} \frac{1}{2} \delta_{t-1}^z \end{aligned} \quad (144)$$

Summing over  $t = 1, \dots, T_s$ , we obtain

$$\begin{aligned} \frac{\beta_s}{\alpha_s} \sum_{t=1}^{T_s} \theta_t \Delta_t + \frac{\beta_s}{\alpha_s} \sum_{t=1}^{T_s-1} (1 - \alpha_s - \omega_s) \Gamma_t \Delta_t \\ \leq \frac{\beta_s}{\alpha_s} \sum_{t=1}^{T_s} (1 - \alpha_s - \omega_s) \Gamma_{t-1} \Delta_{t-1} + \frac{\beta_s \omega_s}{\alpha_s} \sum_{t=1}^{T_s} \Gamma_{t-1} \tilde{\Delta}_{s-1} \\ + \frac{1}{2} \delta_{T_{s-1}}^z - \frac{\Gamma_{T_s}}{2} \delta_{T_s}^z \end{aligned} \quad (145)$$

By canceling the terms, and simplifying, we get

$$\begin{aligned} \frac{\beta_s}{\alpha_s} \left( \sum_{t=1}^{T_{s_0}} \theta_t \right) \tilde{\Delta}_s + \frac{\Gamma_{T_{s_0}}}{2} \delta_{T_s}^z \\ \leq \frac{\beta_s}{\alpha_s} \left[ 1 - \alpha_s - \omega_s + \omega_s \sum_{t=1}^{T_{s_0}} \Gamma_{t-1} \right] \tilde{\Delta}_{s-1} + \frac{1}{2} \delta_{T_{s-1}}^z \end{aligned} \quad (146)$$

As in [31, Lemma 11], we can establish that  $\sum_{t=1}^{T_{s_0}} \theta_t \geq \Gamma_{T_{s_0}} \Omega_{s_0}$  where

$$\Omega_{s_0} := \frac{\beta_s}{\alpha_s} \left[ 1 - \alpha_s - \omega_s + \omega_s \sum_{t=1}^{T_{s_0}} \Gamma_{t-1} \right] \geq \frac{(\bar{s}_0 - s_0 + 4)^2 T_{s_0}}{24L_\gamma}.$$

where  $\bar{s}_0 = s_0 + \sqrt{\frac{12\kappa}{n}} - 4$ , so that (146) can be written as

$$\begin{aligned} \Omega_{s_0} \tilde{\Delta}_s + \frac{1}{2} \delta_{T_s}^z &\leq \frac{1}{\Gamma_{T_{s_0}}} \left[ \Omega_{s_0} \tilde{\Delta}_{s-1} + \frac{1}{2} \delta_{T_{s-1}}^z \right] \\ &\leq \frac{1}{\Gamma_{T_{s_0}}} \left[ \Omega_{s_0} \tilde{\Delta}_{\bar{s}_0} + \frac{1}{2} \delta_{T_{\bar{s}_0}}^z \right] \end{aligned} \quad (147)$$

From the analysis in previous subsection, it holds  $\mathcal{L}_{\bar{s}_0} \geq \frac{(\bar{s}_0 - s_0 + 4)^2 T_{s_0}}{24L_\gamma} = \frac{T_{s_0}}{2n\mu}$ . Denoting  $\mathcal{C}_s = (1 + \mu\beta_s)^{-T_{s_0}(s-\bar{s}_0)}$ , we conclude this case with

$$\begin{aligned} \tilde{\Delta}_s &\leq \mathcal{C}_s \tilde{\Delta}_{\bar{s}_0} + \frac{12L_\gamma \mathcal{C}_s}{(\bar{s}_0 - s_0 + 4)^2 T_{s_0}} \delta_{T_{\bar{s}_0}}^z \\ &\leq \frac{24L_\gamma \mathcal{C}_s}{(\bar{s}_0 - s_0 + 4)^2 T_{s_0}} \left( \mathcal{L}_{\bar{s}_0} \tilde{\Delta}_{\bar{s}_0} + \frac{1}{2} \delta_{T_{\bar{s}_0}}^z \right) \\ &\leq \frac{24L_\gamma \mathcal{C}_s}{(\bar{s}_0 - s_0 + 4)^2 T_{s_0}} \frac{D_0}{3L_\gamma} \leq (1 + \sqrt{3n\kappa})^{\frac{-n(s-\bar{s}_0)}{2}} \frac{D_0}{3\kappa/4}. \end{aligned} \quad (148)$$

We note that, here the number of epochs is bounded by  $S = \bar{s}_0 + 2\sqrt{\frac{3L_\gamma}{n\mu}} \log \frac{4D_0}{3\kappa\epsilon}$ . Thus the number of gradient evaluations is bounded by

$$\begin{aligned} N_{\text{SFO}} &= \sum_{s=1}^S (n + T_s) \\ &= \sum_{s=1}^{s_0} (n + T_s) + \sum_{s=s_0+1}^{\bar{s}_0} (n + T_{s_0}) + (n + T_{s_0})(S - \bar{s}_0) \\ &\leq 2n \log n + 2n \left( \sqrt{\frac{12\kappa}{n}} - 4 \right) + 4n \sqrt{\frac{12\kappa}{n}} \log \frac{4D_0}{3\kappa\epsilon} \\ &= \mathcal{O} \left( n \log n + \sqrt{n\kappa} \log \frac{4D_0}{3\kappa\epsilon} \right) \end{aligned} \quad (149)$$

and QP evaluations are bounded by

$$\begin{aligned} N_{\text{QMO}} &= \sum_{s=1}^S T_s = \sum_{s=1}^{s_0} T_s + \sum_{s=s_0+1}^{\bar{s}_0} T_{s_0} + T_{s_0}(S - \bar{s}_0) \\ &= \mathcal{O} \left( n + n \left( \sqrt{\frac{12\kappa}{n}} - 4 \right) + 2n \sqrt{\frac{12\kappa}{n}} \log \frac{4D_0}{3\kappa\epsilon} \right) \\ &= \mathcal{O} \left( n \log n + \sqrt{n\kappa} \log \frac{4D_0}{3\kappa\epsilon} \right) \end{aligned} \quad (150)$$