Variance Reduction and Low Sample Complexity in Stochastic Optimization via Proximal Point Method

Jiaming Liang JIAMING.LIANG@ROCHESTER.EDU
Goergen Institute for Data Science and Artificial Intelligence, and Department of Computer Science, University of Rochester, Rochester, NY 14620

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

High-probability guarantees in stochastic optimization are often obtained only under strong noise assumptions such as sub-Gaussian tails. We show that such guarantees can also be achieved under the weaker assumption of bounded variance by developing a stochastic proximal point method. This method combines a proximal subproblem solver, which inherently reduces variance, with a probability booster that amplifies per-iteration reliability into high-confidence results. The analysis demonstrates convergence with low sample complexity, without restrictive noise assumptions or reliance on mini-batching.

Keywords: stochastic optimization, high probability, sample complexity, iteration complexity, proximal point method, variance reduction

1. Introduction

A key challenge in stochastic optimization is obtaining high-confidence guarantees, which are crucial in practice but typically require restrictive assumptions such as sub-Gaussian noise. In this paper, we address this challenge by introducing a stochastic proximal point method (SPPM) for solving the stochastic convex composite optimization problem

$$\phi_* := \min_{x \in \mathbb{R}^d} \{\phi(x) := f(x) + h(x)\}, \tag{1}$$

where $f(x) = \mathbb{E}_\xi[F(x,\xi)]$ is the expectation of a stochastic function $F(x,\xi)$. We analyze the sample complexity of SPPM under the following assumptions: h is closed, convex, and admits an efficiently computable proximal mapping; f is closed and strongly convex; and the stochastic gradient oracle of f is unbiased with bounded variance.

Under the above assumptions, standard results on stochastic approximation (SA) methods (Nemirovski and Yudin, 1983; Polyak and Juditsky, 1992; Ghadimi and Lan, 2013; Liang et al., 2023) provide non-asymptotic convergence guarantees in expectation. Specifically, for some tolerance $\varepsilon > 0$, one can obtain an ε-solution $x \in \operatorname{dom} h$ satisfying $\mathbb{E}[\phi(x)] - \phi_* \leq \varepsilon$ within $\mathcal{O}(1/\varepsilon)$ queries to the stochastic gradient oracle of f. High-probability guarantees of the form $\mathbb{P}(\phi(x) - \phi_* \leq \varepsilon) \geq 1 - p$ for given $\varepsilon > 0$ and $p \in (0,1)$ can also be derived using Markov's inequality. However, this yields a sample complexity of $\mathcal{O}(1/(\varepsilon p))$, which can be unsatisfactory when p is small. Without modifying the algorithm, prior works (Nemirovski et al., 2009; Juditsky and Nesterov, 2014; Lan, 2012; Ghadimi and Lan, 2012, 2013) improve the dependence of p from $\mathcal{O}(1/p)$ to $\mathcal{O}(\log(1/p))$ by assuming a stronger condition on the stochastic gradient noise, namely a sub-Gaussian distribution. This assumption, however, is more restrictive than the bounded variance condition considered in this paper.

Recent papers Nazin et al. (2019); Hsu and Sabato (2016); Davis et al. (2021) employ gradient clipping or classical probability tools, such as robust distance estimation (RDE) by Nemirovski and Yudin (1983), to reduce the dependence of sample complexity on p without imposing restrictive assumptions on stochastic gradient noise. In particular, Davis et al. (2021) proposes an interesting framework that incorporates an arbitrary SA method and RDE into the proximal point method (PPM). For strongly convex and smooth problems, the condition number of the proximal subproblem improves with smaller prox stepsizes, leading to sample complexity of with only logarithmic dependence on $1/p$. While conceptually appealing, the approach in Davis et al. (2021) raises several practical questions. First, the method is described in terms of a generic minimization oracle within PPM, leaving open which specific SA method might work best in practice. Second, the algorithm relies on a sequence of decreasing prox stepsizes that depend on the strong convexity parameter μ, which is often unknown or difficult to estimate. Designing adaptive or universal methods that achieve optimal iteration complexity without prior knowledge of μ remains an active line of research, even in deterministic settings (Aujol et al., 2024; Nesterov, 2013; Lan et al., 2023). Finally, although their analysis also improves the dependence of sample complexity on the condition number compared with Hsu and Sabato (2016), this improvement similarly relies on access to μ. Thus, while the framework in Davis et al. (2021) offers valuable theoretical insights, its practical applicability may depend on further advances in adaptive methods.

SPPM in a nutshell. Our algorithm, SPPM, builds on PPM with a *constant* prox stepsize $\lambda > 0$. At iteration k, given the current prox–center $\bar{z}_{k-1}$, we consider the proximal subproblem

$$\hat{z}_k := \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \phi(x) + \frac{1}{2\lambda} \|x - \bar{z}_{k-1}\|^2 \right\}. \tag{2}$$

SPPM does not require solving (2) exactly. Instead, it queries a *proximal subproblem solver* (PSS) n times to obtain statistically independent approximate solutions, and then applies a *probability booster* (PB) that selects one candidate. The booster is designed so that, with high probability, the selected point lies within a prescribed accuracy of the exact proximal point $\hat{z}_k$. This selected point then becomes the next prox–center, i.e., $\bar{z}_k \leftarrow \text{PB}(\{\text{PSS outputs}\})$, and the process repeats.

Our analysis establishes high-probability guarantees for a PPM scheme with a *fixed* λ. The main difficulty is that the stochastic errors incurred in solving each proximal subproblem do not naturally diminish over time; instead, they can accumulate across iterations. This requires a careful argument to show that the error remains controlled and does not overwhelm the contraction effect of the proximal operator. By contrast, the high-confidence framework of Davis et al. (2021) (proxBoost) relies on a *geometrically decaying* prox stepsize rule, which progressively suppresses noise across stages and simplifies the concentration analysis. However, geometric decay has drawbacks: (i) it introduces stage lengths and decay factors that must be scheduled or tuned in advance, often depending on the problem horizon; and (ii) once the stepsize becomes very small, progress can stagnate due to overly conservative steps. Our results demonstrate that one can obtain high-probability convergence *without* resorting to geometric decaying stepsizes, by combining a constant-λ PPM with a probability booster that carefully controls the stochastic error at each iteration.

Contributions. Our contributions build on two crucial subroutines, namely PSS and PB. PSS returns inexact proximal points whose distribution has *reduced variance* relative to raw stochastic gradient steps, which reflects an intrinsic variance-reduction effect of proximal regularization. PB further amplifies this benefit by aggregating n independent PSS calls and selecting a statistically

reliable candidate, yielding a point that is sufficiently close to $\hat{z}_k$ with probability at least $1-p$ while requiring only $O(\log(1/p))$ extra samples. Combining these ideas, we propose SPPM, a stochastic proximal point scheme with a constant prox stepsize λ (with a lower bound), together with the implementable PSS and PB oracles. Under only a bounded-variance noise assumption, we establish a high-probability convergence guarantee with failure probability at most p and a sample complexity that scales $O(\log(1/p))$; see the main results of the paper, namely Theorem 1 (high-probability guarantee) and Theorem 2 (low sample complexity).

2. Stochastic Proximal Point Method and Main Results

This section describes the assumptions made on problem (1), presents the main algorithm SPPM to solve (1), and gives an overview of the main results of the paper.

Let Ξ denote the support of random vector ξ and assume that the following conditions on (1) hold:

(A1) for almost every $\xi \in \Xi$, there exist a stochastic function oracle $F(\cdot,\xi) : \operatorname{dom} h \to \mathbb{R}$ and a stochastic gradient oracle $s(\cdot,\xi) : \operatorname{dom} h \to \mathbb{R}^d$ satisfying

$$f(x) = \mathbb{E}[F(x,\xi)], \quad \nabla f(x) = \mathbb{E}[s(x,\xi)] \in \partial f(x), \quad \forall x \in \operatorname{dom} h;$$

(A2) for every $x \in \operatorname{dom} h$, we have $\mathbb{E}[\|s(x,\xi) - \nabla f(x)\|^2] \le \sigma^2$;

(A3) for every $x, y \in \operatorname{dom} h$,

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|;$$

(A4) f is μ-strongly convex, h is convex, and $\operatorname{dom} h \subset \operatorname{dom} f$;

(A5) $\operatorname{dom} h$ is bounded with diameter $D > 0$.

It is well-known that Assumptions (A3) and (A4) imply that for every $x, y \in \operatorname{dom} h$,

$$\frac{\mu}{2}\|y-x\|^2 \le f(y) - f(x) - \langle \nabla f(x), y-x\rangle \le \frac{L}{2}\|y-x\|^2. \tag{3}$$

Now, we formally present SPPM in Algorithm 1. SPPM relies on two key oracles, namely PSS and PB, which are given and analyzed in Sections 3 and 5, respectively. Step 1 of Algorithm 1 repeatedly calls PSS for n times to generate independent pairs $\{z_k^j, w_k^j\}_{j=1}^n$ and each pair satisfies a guarantee of low probability. Among the n pairs output by PSS, Step 2 calls PB to select one candidate so that the same guarantee holds at a high confidence level.

Algorithm 1 Stochastic Proximal Point Method, SPPM$(\bar{z}_0, \alpha, \lambda, n, q, I, K)$

Input: Initial point $\bar{z}_0 \in \operatorname{dom} h$, scalars $\alpha \in (0,1)$ and $\lambda > 0$, and integers $n, q, I, K \ge 1$.
for $k = 1, \ldots, K$ **do**
  Step 1. Generate independent pairs $(z_k^1, w_k^1), \ldots, (z_k^n, w_k^n)$ by calling n times the oracle PSS$(\bar{z}_{k-1}, \alpha, \lambda, I)$;
  Step 2. Generate $(\bar{z}_k, \bar{w}_k)$ by calling the oracle PB$(\{(z_k^j, w_k^j)\}_{j=1}^n, \bar{z}_{k-1}, q, \lambda)$.
end for

Under Assumptions (A1)–(A5) on (1) and mild conditions on the input of Algorithm 1, Theorem 1 establishes a high-probability guarantee for obtaining an ε-solution of (1). In addition, Theorem 2 provides a bound on the sample complexity of stochastic gradients that is logarithmic in the confidence parameter. A further contribution of this work is the observation that the oracle PSS (Algorithm 2) offers a new form of variance reduction without requiring mini-batches of samples (see the discussion at the end of Section 3). This effect arises naturally from employing PPM within stochastic optimization, and highlights the intrinsic variance-reduction benefits of our approach.

3. Proximal Subproblem Solver

This section introduces and analyzes the first key oracle used in Algorithm 1, namely the proximal subproblem solver (PSS). A detailed description of PSS is provided in Algorithm 2.

Algorithm 2 Proximal Subproblem Solver, PSS$(x_0, \alpha, \lambda, I)$

Input: Initial point $x_0 \in \operatorname{dom} h$, scalars $\alpha \in (0,1)$ and $\lambda > 0$, and integer $I \geq 1$.
for $i = 1, \ldots, I+1$ **do**
Step 1. Take an independent sample ξ_{i-1} of r.v. ξ and set $s_{i-1} = s(x_{i-1}, \xi_{i-1})$;
Step 2. Compute

$$x_i = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ h(x) + \langle S_i, x \rangle + \frac{1}{2\lambda}\|x - x_0\|^2 \right\}, \tag{4}$$

$$y_i = \begin{cases} x_i, & \text{if } i = 1, \\ \alpha y_{i-1} + (1-\alpha)x_i, & \text{otherwise}, \end{cases} \tag{5}$$

where

$$S_i = \begin{cases} s_0, & \text{if } i = 1, \\ \alpha S_{i-1} + (1-\alpha)s_{i-1}, & \text{otherwise}. \end{cases} \tag{6}$$

end for
Output: x_{I+1} and y_{I+1}.

Our goal in this section is to study the following proximal subproblem

$$\hat{x} := \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \phi^\lambda(x) := \phi(x) + \frac{1}{2\lambda}\|x - x_0\|^2 \right\} \tag{7}$$

and analyze the solution pair (x_{I+1}, y_{I+1}) returned by Algorithm 2. The central result is Proposition 1, which establishes the convergence rate of the expected primal gap of (7). More importantly, it demonstrates that Algorithm 2 significantly reduces the variance term within the bound of the expected primal gap.

To streamline our presentation, we introduce the following definitions:

$$\Phi(\cdot, \xi) = F(\cdot, \xi) + h(\cdot), \quad \ell(\cdot; x, \xi) = f(x) + \langle s(x, \xi), \cdot - x \rangle + h(\cdot), \tag{8}$$

$$u_i := \begin{cases} \Phi(x_1, \xi_1) + \frac{1}{2\lambda}\|x_1 - x_0\|^2, & \text{if } i = 1, \\ \alpha u_{i-1} + (1-\alpha)\left[\phi(x_i) + \frac{1}{2\lambda}\|x_i - x_0\|^2\right], & \text{otherwise}, \end{cases} \tag{9}$$

$$\mathcal{L}_i(\cdot) := \begin{cases} F(x_0, \xi_0) + \langle s_0, \cdot - x_0 \rangle + h(\cdot), & \text{if } i = 1, \\ \alpha \mathcal{L}_{i-1}(\cdot) + (1-\alpha)\ell(\cdot; x_{i-1}, \xi_{i-1}), & \text{otherwise}, \end{cases} \tag{10}$$

and

$$\mathcal{L}_i^\lambda(\cdot) := \mathcal{L}_i(\cdot) + \frac{1}{2\lambda}\|\cdot - x_0\|^2, \quad t_i := u_i - \mathcal{L}_i^\lambda(x_i). \tag{11}$$

We are ready to state the main result of this section. The following proposition demonstrates that the bound $\mathbb{E}[t_{I+1}]$ on the expected primal gap of (7) decreases as I increases.

Proposition 1 *Assuming*

$$\alpha \geq \frac{I/2 + \lambda L}{1 + I/2 + \lambda L}, \tag{12}$$

then we have

$$\mathbb{E}[t_{I+1}] \leq \alpha^I \left(\sigma D + \frac{LD^2}{2}\right) + \frac{\lambda \sigma^2}{I}. \tag{13}$$

The full proof of Proposition 1 is deferred to Appendix D.4. In the rest of this section, we develop several intermediate technical results that will serve as key building blocks for the proof of Proposition 1. We begin by observing some relations. It is easy to see from (4), (6), and (11) that

$$x_i = \operatorname*{argmin}_{x \in \mathbb{R}^d} \mathcal{L}_i^\lambda(x). \tag{14}$$

This observation and the fact that $\mathcal{L}_i^\lambda$ is $(1/\lambda)$-strongly convex imply that for every $x \in \operatorname{dom} h$,

$$\mathcal{L}_i^\lambda(x) \geq \mathcal{L}_i^\lambda(x_i) + \frac{1}{2\lambda}\|x - x_i\|^2. \tag{15}$$

The result outlined below provides some basic relations that are frequently used in our analysis. We provide its proof in Appendix D.1.

Lemma 1 *For every $i \geq 1$, we have*

$$\mathbb{E}[\phi^\lambda(y_i)] \leq \mathbb{E}[u_i], \tag{16}$$

$$\mathbb{E}[\mathcal{L}_i(x)] \leq \phi(x), \quad \forall x \in \operatorname{dom} h, \tag{17}$$

$$\phi(x_i) - \ell(x_i; x_{i-1}, \xi_{i-1}) \leq \|\nabla f(x_{i-1}) - s_{i-1}\|\|x_i - x_{i-1}\| + \frac{L}{2}\|x_i - x_{i-1}\|^2. \tag{18}$$

The next technical result shows that the expectation $\mathbb{E}[t_i]$ provides an upper bound on the primal gap of (7). The proof is deferred to Appendix D.2.

Lemma 2 *For every $i \geq 1$, define*

$$r_i := \frac{\lambda \|\nabla f(x_i) - s_i\|^2}{I}. \tag{19}$$

Then, the following statements hold:

(a) for every $i \geq 1$, $\mathbb{E}[r_i] \leq \lambda\sigma^2/I$;

(b) $\mathbb{E}[t_1] \leq \sigma D + LD^2/2$ *where* σ *and* D *are as in Assumptions (A2) and (A5), respectively;*

(c) for every $i \geq 1$, $\mathbb{E}[t_i] \geq \mathbb{E}[\phi^\lambda(y_i) - \phi^\lambda(\hat{x})]$.

The next lemma establishes a relation between t_i and r_i, defined in (11) and (19), respectively. This relation plays a key role in the proof of Proposition 1, where it is used to show that t_{I+1} is small in expectation. The proof of the lemma is deferred to Appendix D.3.

Lemma 3 *Assuming* (12) *holds, then for every* $i \geq 2$, *we have*

$$t_i \leq \alpha^{i-1} t_1 + (1-\alpha) \sum_{j=1}^{i-1} \alpha^{i-j-1} r_j, \tag{20}$$

where t_i *and* r_i *are as in* (11) *and* (19), *respectively.*

We conclude this section by offering some insight into Proposition 1. First, in (13), the terms $\lambda\sigma^2/I$ and $\alpha^I \left(\sigma D + LD^2/2\right)$ can be interpreted as variance and bias components, respectively. Second, the tradeoff between these components is governed by the choice of the prox stepsize λ, since the bias term depends on λ through α as defined in (12). Third, increasing the number of iterations I in Algorithm 2 reduces both the bias and variance terms. Unlike standard stochastic gradient methods, inequality (13) indicates a variance reduction by a factor of I, not through sample averaging but by performing multiple iterations to solve the proximal subproblem (7). This distinctive variance-reduction effect is natural, as Algorithm 2 effectively uses $I+1$ stochastic gradient samples.

4. Analysis of Proximal Point Method

This section analyzes Step 1 of Algorithm 1 and prepares the technical results necessary for the analysis of the PB oracle in Step 2 of Algorithm 1. The main result in this section is Proposition 2, which gives a probability guarantee of the suboptimality of the proximal subproblem (2).

Before starting the analysis, we need to properly translate the notation from an inner viewpoint to an outer (proximal point) perspective, i.e., using SPPM (Algorithm 1) iteration index k instead of PSS (Algorithm 2) iteration index i. Consider the j-th call to PSS in Step 1 of Algorithm 1, and let x_0^j and (x_{I+1}^j, y_{I+1}^j) denote the initial point and output for PSS, respectively. We know that for every $j \in \{1, \ldots, n\}$,

$$\bar{z}_{k-1} = x_0^j, \quad z_k^j = x_{I+1}^j, \quad w_k^j = y_{I+1}^j. \tag{21}$$

For simplicity, we denote z_k^j and w_k^j by z_k and w_k, respectively, ignoring the query index j. Therefore, in view of (21), the notational convention we adopt in this section is

$$\bar{z}_{k-1} = x_0^j, \quad z_k = x_{I+1}^j, \quad w_k = y_{I+1}^j. \tag{22}$$

Although we omit the index j for simplicity, we note that all the results in this section hold for each $j \in \{1, \ldots, n\}$, i.e., any call to PSS in Step 1 of Algorithm 1.

We are now ready to state the main result of Section 4, which provides a probability guarantee of the suboptimality of the proximal subproblem (2).

Proposition 2 *For every $k \geq 1$, we have*

$$\mathbb{P}\left(\phi(w_k) + \frac{1}{2\lambda}\|w_k - \bar{z}_{k-1}\|^2 - \phi(\hat{z}_k) - \frac{1}{2\lambda}\|\hat{z}_k - \bar{z}_{k-1}\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\hat{z}_k - z_k\|^2 \leq 8\varepsilon_k\right) \geq \frac{3}{4}, \tag{23}$$

where $\hat{z}_k$ is as in (2).

Proposition 2 follows directly from Markov's inequality together with Lemma 7, which we establish at the end of this section. To this end, we first develop several intermediate technical results that serve as key ingredients towards the proof of Lemma 7.

We begin by using the notation in (22) to restate key results from Section 3 in the proximal point framework. The following lemma serves as the starting point for our proximal point analysis, and its proof is deferred to Appendix E.1.

Lemma 4 *For every $k \geq 1$, let*

$$\Gamma_k = \mathcal{L}_{I+1}^j, \quad \varepsilon_k = \alpha^I\left(\sigma D + \frac{LD^2}{2}\right) + \frac{\lambda\sigma^2}{I}, \tag{24}$$

where $\mathcal{L}_{I+1}^j$ means $\mathcal{L}_{I+1}$ considered in the analysis of the j-th call to PSS in Step 1 of Algorithm 1 and j is an arbitrary index in $\{1,\ldots,n\}$. Then, the following relations hold

$$\mathbb{E}[\Gamma_k(x)] \leq \phi(x), \quad \forall x \in \operatorname{dom} h, \tag{25}$$

$$z_k = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\Gamma_k(x) + \frac{1}{2\lambda}\|x - \bar{z}_{k-1}\|^2\right\}, \tag{26}$$

$$\mathbb{E}\left[\phi(w_k) + \frac{1}{2\lambda}\|w_k - \bar{z}_{k-1}\|^2 - \Gamma_k(z_k) - \frac{1}{2\lambda}\|z_k - \bar{z}_{k-1}\|^2\right] \leq \varepsilon_k. \tag{27}$$

The following lemma is a technical result that translates the stationarity condition of (26) into conditions resembling the convexity inequality. Its proof is given in Appendix E.2.

Lemma 5 *Define*

$$v_k := \frac{\bar{z}_{k-1} - z_k}{\lambda}, \quad \eta_k := \phi(w_k) - \Gamma_k(z_k) - \langle v_k, w_k - z_k\rangle. \tag{28}$$

Then, we have

$$\phi(x) \geq \mathbb{E}[\phi(w_k) + \langle v_k, x - w_k\rangle - \eta_k], \quad \forall x \in \operatorname{dom} h, \tag{29}$$

$$\mathbb{E}[\eta_k] \leq \varepsilon_k - \frac{1}{2\lambda}\mathbb{E}\left[\|w_k - z_k\|^2\right]. \tag{30}$$

The lemma presented next is crucial for re-establishing the μ-strong convexity of ϕ. To achieve this, we introduce an auxiliary function q_k and outline its key properties in the subsequent discussion. We provide the proof in Appendix E.3.

Lemma 6 *Define*

$$q_k(x) := \phi(w_k) + \langle v_k, x - w_k\rangle + \frac{\mu}{4}\|x - w_k\|^2 - 2\eta_k. \tag{31}$$

Then, the following statements hold:

a) $\mathbb{E}[q_k(x)] \leq \phi(x)$ *for every* $x \in \operatorname{dom} h$*;*

b) $\mathbb{E}[q_k(z_k)] \geq \mathbb{E}[\Gamma_k(z_k)] + \left(\frac{1}{2\lambda} + \frac{\mu}{4}\right)\mathbb{E}[\|z_k - w_k\|^2] - \varepsilon_k$.

Combining the technical results above, the next lemma shows that the suboptimality of (2) is under control in expectation. Its proof is deferred to Appendix E.4.

Lemma 7 *For every* $k \geq 1$ *and* $x \in \operatorname{dom} h$*, we have*

$$\mathbb{E}\left[\phi(w_k) + \frac{1}{2\lambda}\|w_k - \bar{z}_{k-1}\|^2 - \phi(x) - \frac{1}{2\lambda}\|x - \bar{z}_{k-1}\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|x - z_k\|^2\right] \leq 2\varepsilon_k.$$

5. Probability Booster

This section introduces and analyzes the second key oracle used in Algorithm 1, namely PB. It is designed to select, with high probability, a sufficiently accurate candidate from several approximate proximal solutions, thus ensuring the reliability of the overall scheme. A detailed description is provided in Algorithm 3. PB itself relies on two auxiliary subroutines: second tertile selection (STS) and robust gradient estimation (RGE), whose presentations are deferred to Appendices B and C for clarity. The design of Algorithm 3 is motivated by Algorithm 9 of Davis et al. (2021).

Algorithm 3 Probability Booster, $\text{PB}(\{(z^j, w^j)\}_{j=1}^n, \bar{z}, q, \lambda)$

Input: Independent pairs $(z^1, w^1), \ldots, (z^n, w^n)$ generated by PSS with initial point $\bar{z}$, integer $q \geq 1$, and scalar $\lambda > 0$.
Step 1. Call oracle $\mathcal{J}_1 = \text{STS}(\{w^j\}_{j=1}^n, d_2(\cdot,\cdot))$;
Step 2. Call oracle $\mathcal{J}_2 = \text{STS}(\{z^j\}_{j=1}^n, d_2(\cdot,\cdot))$;
Step 3. Fix an arbitrary $j_0 \in \mathcal{J}_1 \cap \mathcal{J}_2$ and set $\tilde{w} := w^{j_0}$. Call oracle $\bar{s}(\tilde{w}) = \text{RGE}(\tilde{w}, n, q)$;
Step 4. Define the metric $d_h(x,y) := |h(x) - h(y) + \langle \bar{s}(\tilde{w}) + (\tilde{w} - \bar{z})/\lambda, x - y\rangle|$. Call oracle $\mathcal{J}_3 = \text{STS}(\{w^j\}_{j=1}^n, d_h(\cdot,\cdot))$.
Output: A pair (z^j, w^j) for an arbitrary $j \in \mathcal{J}_1 \cap \mathcal{J}_2 \cap \mathcal{J}_3$.

Note that the subroutine STS, used in Steps 1, 2, and 4, requires a metric as input. Here, $d_2(x,y) = \|x - y\|_2$ denotes the Euclidean distance and is therefore a valid metric, while Lemma 8 establishes that $d_h(\cdot,\cdot)$ is also a metric.

Recall that Step 1 of Algorithm 1 generates n pairs $\{(z_k^j, w_k^j)\}_{j=1}^n$ by invoking the oracle PSS (Algorithm 2). Proposition 2 provides a low-probability guarantee on the suboptimality of each pair (z_k^j, w_k^j), as stated in (23). The purpose of Algorithm 3 is to boost this low-probability bound into a high-probability guarantee by selecting one of the n pairs such that an inequality analogous to (23) holds with probability close to one.

To simplify our notation, we exclude the iteration index k and focus on the following proximal subproblem. Given the prox-center $\bar{z} \in \operatorname{dom} h$ and the prox stepsize $\lambda > 0$, define

$$\phi^\lambda(\cdot) := \phi(\cdot) + \frac{1}{2\lambda}\|\cdot - \bar{z}\|^2, \quad f^\lambda(\cdot) := f(\cdot) + \frac{1}{2\lambda}\|\cdot - \bar{z}\|^2, \tag{32}$$

and consider

$$\hat{z} := \operatorname*{argmin}_{x \in \mathbb{R}^d} \phi^\lambda(x). \tag{33}$$

The optimality condition of the proximal subproblem (33) gives

$$-\nabla f^{\lambda}(\hat{z}) = -\left(\nabla f(\hat{z}) + \frac{\hat{z}-\bar{z}}{\lambda}\right) \in \partial h(\hat{z}),$$

where $\partial h(\hat{z})$ denotes the subdifferential set of h at $\hat{z}$. This implies that for every $x \in \operatorname{dom} h$,

$$D_h(x,\hat{z}) := h(x) - h(\hat{z}) + \langle \nabla f^{\lambda}(\hat{z}), x-\hat{z}\rangle \geq 0, \tag{34}$$

where D_h is the Bregman divergence for h. Let $\bar{s}^{\lambda}(\tilde{w}) := \bar{s}(\tilde{w}) + (\tilde{w}-\bar{z})/\lambda$. Then, it follows from the definition of $d_h(\cdot,\cdot)$ in Step 4 of Algorithm 3 that

$$d_h(w^j,\hat{z}) = |h(w^j) - h(\hat{z}) + \langle \bar{s}^{\lambda}(\tilde{w}), w^j - \hat{z}\rangle|. \tag{35}$$

For every $j = 1,\ldots,n$, define the event B_j as follows

$$B_j := \left\{ \phi^{\lambda}(w^j) - \phi^{\lambda}(\hat{z}) + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\hat{z}-z^j\|^2 \leq \tau \right\}, \tag{36}$$

where τ is considered as (in view of (23))

$$\tau = 8\varepsilon_k \overset{(24)}{=} 8\left[\alpha^I\left(\sigma D + \frac{LD^2}{2}\right) + \frac{\lambda\sigma^2}{I}\right]. \tag{37}$$

Then it follows from Proposition 2 that a low probability guarantee holds for each B_j, that is,

$$\mathbb{P}(B_j \text{ occurs}) \geq \frac{3}{4}. \tag{38}$$

The following result establishes that the output of Algorithm 3 satisfies a high-probability version of the inequality in (36), up to a factor of the condition number. For clarity, the proof is deferred to Appendix F.

Proposition 3 *Assume that (12) holds. If the input q of Algorithm 3 satisfies*

$$q \geq \frac{18(1+\lambda\mu)\sigma^2}{\lambda L^2 \tau}, \tag{39}$$

where τ is as in (37), then with probability at least $1 - 2\exp(-n/72)$, the pair (z^j, w^j) returned by PB satisfies

$$\phi^{\lambda}(w^j) - \phi^{\lambda}(\hat{z}) + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\hat{z}-z^j\|^2 \leq 12\tau + 57\kappa\tau, \tag{40}$$

where $\hat{z}$ is as in (33) and condition number $\kappa = L/\mu > 1$.

6. High-Probability Result and Low Sample Complexity

This section provides the full analysis of Algorithm 1 and establishes its two main guarantees: the high-probability convergence result (Theorem 1) and the sample complexity bound (Theorem 2), which constitute the central contributions of this paper.

The following proposition establishes a high-probability guarantee for a single iteration of Algorithm 1. It yields a key recursive relation that serves as the foundation for proving the subsequent theorems.

Proposition 4 *Assuming that* (12) *and* (39) *hold and considering a full iteration of Algorithm 1, then for every* $k \geq 1$ *and with probability at least* $1 - 2\exp(-n/72)$*, we have*

$$\phi(\bar{w}_k) - \phi_* - \frac{1}{2\lambda}\|x_* - \bar{z}_{k-1}\|^2 + \frac{1+\lambda\mu}{\lambda(4+\lambda\mu)}\|x_* - \bar{z}_k\|^2 \leq 12\tau + 57\kappa\tau, \tag{41}$$

where x_* *is the solution to* (1) *and* τ *is as in* (37)*.*

Proof: We restate Proposition 3 using the notation in Algorithm 1 as: if (12) and (39) hold, then with probability at least $1 - 2\exp(-n/72)$, the pair $(\bar{z}_k, \bar{w}_k)$ generated in Step 2 of Algorithm 1 satisfies

$$\phi(\bar{w}_k) + \frac{1}{2\lambda}\|\bar{w}_k - \bar{z}_{k-1}\|^2 - \phi(\hat{z}_k) - \frac{1}{2\lambda}\|\hat{z}_k - \bar{z}_{k-1}\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\hat{z}_k - \bar{z}_k\|^2 \overset{(40)}{\leq} \delta, \tag{42}$$

where $\hat{z}_k$ is as in (2) and $\delta = 12\tau + 57\kappa\tau$. Using the fact that the proximal subproblem in (2) is $(\mu + 1/\lambda)$-strongly convex, we have

$$\phi_* + \frac{1}{2\lambda}\|x_* - \bar{z}_{k-1}\|^2 \geq \phi(\hat{z}_k) + \frac{1}{2\lambda}\|\hat{z}_k - \bar{z}_{k-1}\|^2 + \frac{1+\lambda\mu}{2\lambda}\|x_* - \hat{z}_k\|^2.$$

This inequality and (42) then imply that with probability at least $1 - 2\exp(-n/72)$,

$$\phi(\bar{w}_k) - \left(\phi_* + \frac{1}{2\lambda}\|x_* - \bar{z}_{k-1}\|^2\right) + \frac{1+\lambda\mu}{2\lambda}\|x_* - \hat{z}_k\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\bar{z}_k - \hat{z}_k\|^2 \leq \delta. \tag{43}$$

Furthermore, it follows from the triangle inequality and the Cauchy-Schwarz inequality that

$$\|x_* - \bar{z}_k\|^2 \leq [2 + (2+\lambda\mu)]\left(\frac{1}{2}\|x_* - \hat{z}_k\|^2 + \frac{1}{2+\lambda\mu}\|\bar{z}_k - \hat{z}_k\|^2\right),$$

and hence that

$$\frac{1+\lambda\mu}{\lambda(4+\lambda\mu)}\|x_* - \bar{z}_k\|^2 \leq \frac{1+\lambda\mu}{2\lambda}\|x_* - \hat{z}_k\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\bar{z}_k - \hat{z}_k\|^2.$$

Combining the above inequality and (43), we conclude that (41) holds with probability at least $1 - 2\exp(-n/72)$. ∎

We are now prepared to state the main theorem, which establishes the high-probability guarantee of Algorithm 1.

Theorem 1 *Assume that* $\lambda\mu \geq 3$*,* (12)*, and* (39) *hold. For given* $\varepsilon > 0$ *and* $p \in (0,1)$*, consider the solution sequence* $\{\bar{w}_k\}_{k=1}^K$ *generated by Algorithm 1, if the input triple* (K, I, n) *satisfies*

$$K = \mathcal{O}\left(\log\frac{1}{\varepsilon}\right), \quad I = \mathcal{O}\left(\max\left\{\kappa\log\frac{\kappa}{\varepsilon}, \frac{\kappa\sigma^2}{\mu\varepsilon}\right\}\right), \quad n = \mathcal{O}\left(\log\frac{1}{p}\right), \tag{44}$$

where $\kappa = L/\mu > 1$ *is the condition number, then we have*

$$\mathbb{P}\left(\min_{1\leq k\leq K}\phi(\bar{w}_k) - \phi_* \leq \varepsilon\right) \geq 1 - p. \tag{45}$$

Proof: Define

$$a_k = \phi(\bar{w}_k) - \phi_*, \quad b_k = \frac{1}{2\lambda}\|x_* - \bar{z}_k\|^2, \quad \theta = \frac{2+2\lambda\mu}{4+\lambda\mu}, \quad \delta = 12\tau + 57\kappa\tau. \tag{46}$$

Then, using Proposition 4, we have with probability at least $1 - 2\exp(-n/72)$,

$$a_k \overset{(41)}{\leq} b_{k-1} - \theta b_k + \delta. \tag{47}$$

Multiplying (47) by θ^{k-1} and summing the resulting inequality from $k = 1$ to K, we have

$$\sum_{k=1}^{K} \theta^{k-1} a_k \leq \sum_{k=1}^{K} \theta^{k-1}(b_{k-1} - \theta b_k + \delta) = b_0 - \theta^K b_K + \sum_{k=1}^{K} \theta^{k-1}\delta, \tag{48}$$

with probability at least

$$\left(1 - 2\exp\left(-\frac{n}{72}\right)\right)^K \geq 1 - 2K\exp\left(-\frac{n}{72}\right). \tag{49}$$

Dividing (48) by $\sum_{k=1}^{K} \theta^{k-1}$ and using (46), we have

$$\min_{1\leq k\leq K} \phi(\bar{w}_k) - \phi_* = \min_{1\leq k\leq K} a_k \leq \frac{b_0}{\sum_{k=1}^{K}\theta^{k-1}} + \delta \leq \frac{\|\bar{z}_0 - x_*\|^2}{2\lambda\sum_{k=1}^{K}\theta^{k-1}} + 12\tau + 57\kappa\tau, \tag{50}$$

with probability as in (49). It follows from the assumption that $\lambda\mu \geq 3$ and the definition of θ in (46) that $\theta \geq 8/7$, which together with (50), implies that

$$\min_{1\leq k\leq K} \phi(\bar{w}_k) - \phi_* \leq \frac{\|\bar{z}_0 - x_*\|^2}{14\lambda\left[\left(\frac{8}{7}\right)^K - 1\right]} + 12\tau + 57\kappa\tau, \tag{51}$$

with probability as in (49). It is clear that if the first condition in (44) holds, i.e.,

$$K = \mathcal{O}\left(\log\frac{1}{\varepsilon}\right), \tag{52}$$

then the first term on the right-hand side of (51) satisfies

$$\frac{\|\bar{z}_0 - x_*\|^2}{14\lambda\left[\left(\frac{8}{7}\right)^K - 1\right]} = \mathcal{O}(\varepsilon). \tag{53}$$

Since the last term $57\kappa\tau$ in (51) dominates the second term 12τ, in view of (37), we only need to derive a bound on I such that

$$\alpha^I\left(\sigma D + \frac{LD^2}{2}\right) + \frac{\lambda\sigma^2}{I} \leq \frac{\varepsilon}{\kappa}. \tag{54}$$

Using the above inequality and the fact that $\alpha \leq e^{\alpha-1}$, we know the iteration count of Algorithm 2 is

$$I = \mathcal{O}\left(\max\left\{\frac{1}{1-\alpha}\log\frac{\kappa}{\varepsilon}, \frac{\lambda\kappa\sigma^2}{\varepsilon}\right\}\right).$$

It thus follows from (12) and $\lambda\mu \geq 3$ that

$$I = \mathcal{O}\left(\max\left\{\kappa \log \frac{\kappa}{\varepsilon}, \frac{\kappa\sigma^2}{\mu\varepsilon}\right\}\right), \tag{55}$$

which is the second condition in (44). Now, assuming (52) and (55) hold, then we know (53) and (54) should also hold. Putting (51), (53), and (54) together, we have

$$\mathbb{P}\left(\min_{1\leq k\leq K} \phi(\bar{w}_k) - \phi_* \leq \varepsilon\right) \geq 1 - 2K\exp\left(-\frac{n}{72}\right) \overset{(52)}{\approx} 1 - \exp\left(-\frac{n}{72}\right).$$

Finally, the above inequality and the last condition in (44) imply that (45) holds. ■

We now arrive at the final main result, which establishes the low sample complexity of Algorithm 1.

Theorem 2 *For given $\varepsilon > 0$ and $p \in (0,1)$, to find a solution $\bar{w} \in \operatorname{dom} h$ by Algorithm 1 such that*

$$\mathbb{P}\left(\phi(\bar{w}) - \phi_* \leq \varepsilon\right) \geq 1 - p,$$

we need $\mathcal{O}\left(\log\frac{1}{p}\log\frac{1}{\varepsilon}\right)$ calls to Algorithm 2 and $\mathcal{O}(\log\frac{1}{\varepsilon})$ calls to Algorithm 3. Moreover, the sample complexity of stochastic gradients in Algorithm 1 is

$$\mathcal{O}\left(\max\left\{\kappa \log \frac{\kappa}{\varepsilon}, \frac{\kappa\sigma^2}{\mu\varepsilon}\right\}\log\frac{1}{p}\log\frac{1}{\varepsilon}\right). \tag{56}$$

Proof: It is clear that Algorithm 1 requires nK PSS oracles (Algorithm 2) and K PB oracles (Algorithm 3). Using Theorem 1, we know that the numbers of calls to PSS and PB oracles are $\mathcal{O}\left(\log\frac{1}{p}\log\frac{1}{\varepsilon}\right)$ and $\mathcal{O}(\log\frac{1}{\varepsilon})$, respectively. From Algorithm 2, we know that each PSS oracle has $I+1$ iterations and each iteration takes one stochastic gradient sample. Therefore, PSS takes $nK(I+1)$ samples in total. From Algorithm 3, we know that each PB oracle queries an RGE subroutine and each RGE takes nq stochastic gradient samples. Therefore, PB takes nKq samples in total. Using (37), (39), and the fact that $\lambda\mu = \Omega(1)$, we have

$$q \overset{(39)}{=} \mathcal{O}\left(\frac{\sigma^2}{\lambda L^2 \tau}\right) \overset{(37)}{=} \mathcal{O}\left(\frac{I}{\lambda^2 L^2}\right) = \mathcal{O}\left(\frac{I}{\kappa^2}\right).$$

Therefore, the total sample complexity is $\mathcal{O}(nKI)$, which is (56) in view of (44). ■

7. Conclusions

In this paper, we studied the stochastic convex composite optimization problem (1) with the goal of establishing high-probability guarantees, namely $\mathbb{P}(\phi(x) - \phi_* \leq \varepsilon) \geq 1 - p$ for given $\varepsilon > 0$ and $p \in (0,1)$. Existing approaches that achieve $\mathcal{O}(\log(1/p))$ dependence on p in the sample complexity typically rely on restrictive noise assumptions such as sub-Gaussian tails. Our contribution is to show that such guarantees are attainable under the much milder bounded-variance assumption by designing SPPM, a stochastic proximal point scheme with a constant prox stepsize. The key ingredients are the subroutine PSS, which both approximates the proximal subproblem (2) and

reduces variance, and the subroutine PB, which ensures that each selected iterate is sufficiently accurate with high probability. Together these yield the main results of the paper, namely Theorem 1 (high-probability guarantee) and Theorem 2 (low sample complexity).

We close by discussing several directions for future work. First, while our analysis shows that SPPM achieves logarithmic dependence on $1/p$, its overall complexity still carries an overhead proportional to the condition number κ. A natural question is whether an accelerated variant of SPPM could reduce this dependence to $\sqrt{\kappa}$, in analogy with accelerated gradient methods. Although acceleration is often believed to amplify noise, it is worth investigating whether a restarted acceleration scheme, interpreted through the proximal point framework (see Liang (2025)) in the same way that we have analyzed PSS here, could instead lead to improved variance reduction. Second, our current theorems require certain assumptions on problem parameters to state the guarantees (such as $\lambda\mu \geq 3$ in Theorem 1). An important extension would be to weaken or remove these assumptions in order to design methods that adapt automatically to the problem structure. Finally, it would be interesting to explore variants of SPPM with a variable prox stepsize λ_k, which may further expand the flexibility of the framework and sharpen the theoretical guarantees.

Acknowledgments

This work was partially supported by GIDS-AI seed funding and AFOSR grant FA9550-25-1-0182.

References

J.-F. Aujol, L. Calatroni, C. Dossal, H. Labarrière, and A. Rondepierre. Parameter-free FISTA by adaptive restart and backtracking. *SIAM Journal on Optimization*, 34(4):3259–3285, 2024.

D. Davis, D. Drusvyatskiy, L. Xiao, and J. Zhang. From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22(49), 2021.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22 (4):1469–1492, 2012.

S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(1):543–582, 2016.

A. Juditsky and Y. Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

G. Lan, Y. Ouyang, and Z. Zhang. Optimal and parameter-free gradient minimization methods for smooth optimization. *Available at arXiv:2310.12139*, 2023.

J. Liang. Unifying restart accelerated gradient and proximal bundle methods. *Available at arXiv:2501.04165*, 2025.

J. Liang, V. Guigues, and R. D. C. Monteiro. A single cut proximal bundle method for stochastic convex composite optimization. *Mathematical Programming*, 2023.

A. V. Nazin, A. S. Nemirovsky, A. B. Tsybakov, and A. B. Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80:1607–1627, 2019.

A. Nemirovski and D. Yudin. Problem complexity and method efficiency in optimization. 1983.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Appendix A. Technical Results

This section collects several useful technical results used throughout our analysis.

Lemma 8 *Recall $d_h(\cdot,\cdot)$ defined in Step 4 of Algorithm 3. Then $d_h(\cdot,\cdot)$ is a metric.*

Proof: Recall the notation $\bar{s}^\lambda(\tilde{w}) := \bar{s}(\tilde{w}) + (\tilde{w} - \bar{z})/\lambda$ used in Section 5. Thus, it follows from the definition of $d_h(x,y)$ in Step 4 of Algorithm 3 that

$$d_h(x,y) = |h(x) - h(y) + \langle \bar{s}^\lambda(\tilde{w}), x - y\rangle|.$$

Since $d_h(x,y) \geq 0$ and $d_h(x,x) = 0$, nonnegativity holds. It is clear that $d_h(x,y) = d_h(y,x)$. Hence, symmetry also holds. Finally, we also have the triangle inequality $d_h(x,y) + d_h(y,z) \geq d_h(x,z)$, since

$$\begin{aligned} d_h(x,z) &= |h(x) - h(y) + h(y) - h(z) + \langle \bar{s}^\lambda(\tilde{w}), x - y + y - z\rangle| \\ &\leq |h(x) - h(y) + \langle \bar{s}^\lambda(\tilde{w}), x - y\rangle| + |h(y) - h(z) + \langle \bar{s}^\lambda(\tilde{w}), y - z\rangle| \\ &= d_h(x,y) + d_h(y,z). \end{aligned}$$

■

Lemma 9 *Recall $D_h(\cdot,\hat{z})$ and $d_h(\cdot,\hat{z})$ given in (34) and (35), respectively. Then, for every $x \in \operatorname{dom} h$, we have*

$$d_h(x,\hat{z}) \leq D_h(x,\hat{z}) + \left(\|\bar{s}(\tilde{w}) - \nabla f(\hat{z})\| + \frac{1}{\lambda}\|\tilde{w} - \hat{z}\|\right)\|x - \hat{z}\|, \tag{57}$$

$$D_h(x,\hat{z}) \leq d_h(x,\hat{z}) + \left(\|\bar{s}(\tilde{w}) - \nabla f(\hat{z})\| + \frac{1}{\lambda}\|\tilde{w} - \hat{z}\|\right)\|x - \hat{z}\|, \tag{58}$$

where $\hat{z}$ is as in (33), and $\bar{s}(\tilde{w})$ and $\tilde{w}$ are defined in Step 3 of Algorithm 3.

Proof: Following the notation in Section 5, we note that

$$\bar{s}^\lambda(\tilde{w}) = \bar{s}(\tilde{w}) + \frac{\tilde{w} - \bar{z}}{\lambda}, \quad \nabla f^\lambda(\hat{z}) = \nabla f(\hat{z}) + \frac{\hat{z} - \bar{z}}{\lambda}.$$

Using (35), the above relations, and the triangle inequality, we have for every $x \in \operatorname{dom} h$,

$$\begin{aligned} d_h(x,\hat{z}) &\overset{(35)}{=} |h(x) - h(\hat{z}) + \langle \nabla f^\lambda(\hat{z}), x - \hat{z}\rangle + \langle \bar{s}^\lambda(\tilde{w}) - \nabla f^\lambda(\hat{z}), x - \hat{z}\rangle| \\ &\leq |h(x) - h(\hat{z}) + \langle \nabla f^\lambda(\hat{z}), x - \hat{z}\rangle| + |\langle \bar{s}(\tilde{w}) - \nabla f(\hat{z}), x - \hat{z}\rangle| + \frac{1}{\lambda}|\langle \tilde{w} - \hat{z}, x - \hat{z}\rangle|. \end{aligned}$$

It follows from the definition of D_h in (34) and the Cauchy-Schwarz inequality that (57) holds. The proof of (58) follows similarly. ■

Lemma 10 *Recall $D_h(\cdot,\hat{z})$ and $\phi^\lambda(\cdot)$ given in (34) and (32), respectively. Then, for every $x \in \operatorname{dom} h$, we have*

$$\frac{1 + \lambda\mu}{2\lambda}\|x - \hat{z}\|^2 + D_h(x,\hat{z}) \leq \phi^\lambda(x) - \phi^\lambda(\hat{z}) \leq \frac{1 + \lambda L}{2\lambda}\|x - \hat{z}\|^2 + D_h(x,\hat{z}). \tag{59}$$

Proof: It follows from (3) and the definition of f^λ in (32) that for every $x \in \operatorname{dom} h$,

$$\frac{1+\lambda\mu}{2\lambda}\|x-\hat{z}\|^2 \leq f^\lambda(x) - f^\lambda(\hat{z}) - \langle \nabla f^\lambda(\hat{z}), x-\hat{z}\rangle \leq \frac{1+\lambda L}{2\lambda}\|x-\hat{z}\|^2.$$

Adding $D_h(x,\hat{z})$ to the above inequality and using the fact that $\phi^\lambda(\cdot) = f^\lambda(\cdot) + h(\cdot)$ completes the proof. ■

Lemma 11 *Consider a sequence of independent events $\{A_j\}_{j=1}^n$ where each event occurs with probability at least $3/4$. Define index set $\mathcal{J} := \{j \in \{1,\ldots,n\} : A_j \text{ occurs}\}$. Then, we have*

$$\mathbb{P}\left(|\mathcal{J}| > \frac{2n}{3}\right) \geq 1 - \exp\left(-\frac{n}{72}\right). \tag{60}$$

Proof: Recall that Hoeffding's inequality states that: let $X_1,\ldots,X_n$ be independent random variables such that $a_j \leq X_j \leq b_j$ and $S_n = X_1 + \ldots + X_n$, then for all $t > 0$,

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] \leq -t\right) \leq \exp\left(\frac{-2t^2}{\sum_{j=1}^n (b_j - a_j)^2}\right). \tag{61}$$

Define the random indicator variable X_j associated with A_j as follows

$$X_j := \begin{cases} 1, & \text{if } A_j \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, X_j's are independent. For every $j = 1,\ldots,n$, we have $a_j = 0$, $b_j = 1$, and

$$\mathbb{E}[X_j] = \mathbb{P}(X_j = 1) = \mathbb{P}(A_j \text{ occurs}) \geq \frac{3}{4}. \tag{62}$$

It follows Hoeffding's inequality with $t = n/12$ that

$$\mathbb{P}\left(S_n \geq \frac{2n}{3}\right) \overset{(62)}{\geq} \mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq -\frac{n}{12}\right) \overset{(61)}{\geq} 1 - \exp\left(-\frac{n}{12}\right).$$

Therefore, (60) immediately follows. ■

Appendix B. Second Tertile Selection

This section presents and analyzes the subroutine STS used in Algorithm 3. Algorithm 4 below is motivated by Algorithm 8 of Davis et al. (2021).

Algorithm 4 Second Tertile Selection, STS$(Z, d(\cdot,\cdot))$

Input: A set of points $Z = \{z^1,\ldots,z^n\} \subset \operatorname{dom} h$ and a metric $d(\cdot,\cdot)$ on $\operatorname{dom} h$.
for $j = 1,\ldots,n$ **do**
  Step 1. Compute $\rho_j = \min\left\{\rho > 0 : \left|B_{\rho,d}\left(z^j\right) \cap Z\right| > 2n/3\right\}$;
end for
Step 2. Compute the second tertile $\bar{\rho}$ of $(\rho_1,\ldots,\rho_n)$.
Output: $\mathcal{J} = \{j \in \{1,\ldots,n\} : \rho_j \leq \bar{\rho}\}$.

Here, $B_{\rho,d}(z)$ denotes a d-metric ball centered at z with radius ρ, that is, $B_{\rho,d}(z) = \{x : d(x,z) \leq \rho\}$. Note that STS takes as input a collection of points $\{z^j\}_{j=1}^n$ and a metric $d(\cdot,\cdot)$. In particular, we use $d_2(\cdot,\cdot)$ and $d_h(\cdot,\cdot)$ in Algorithm 3. The output $\mathcal{J}$ of STS is an index set with a cardinality of at least $2n/3$.

The following lemma outlines the key property of Algorithm 4, specifically its ability to maintain proximity for most input points in Z to any point. Typically, this lemma is employed in conjunction with Lemma 11 to boost the probability from low to high confidence, impacting the overall quality by at most a constant factor.

Lemma 12 *Let $d(\cdot,\cdot)$ be a metric on $\operatorname{dom} h$. Consider a collection of points $Z = \{z^1,\ldots,z^n\}$ and a point $\tilde{z} \in \operatorname{dom} h$ satisfying*

$$|B_{\epsilon,d}(\tilde{z}) \cap Z| > \frac{2n}{3} \tag{63}$$

for some $\epsilon > 0$. Then, output index set $\mathcal{J}$ of Algorithm 4) satisfies

$$d(z^j,\tilde{z}) \leq 3\epsilon, \quad \forall j \in \mathcal{J}. \tag{64}$$

Proof: Consider two arbitrary points $z^i, z^j \in B_{\epsilon,d}(\tilde{z})$. Since $d(\cdot,\cdot)$ is a metric, the triangle inequality holds, and hence

$$d(z^i,z^j) \leq d(z^i,\tilde{z}) + d(\tilde{z},z^j) \leq 2\epsilon.$$

Thus, for any $z^j \in B_{\epsilon,d}(\tilde{z})$ fixed, we have $|B_{2\epsilon,d}(z^j) \cap Z| > \frac{2n}{3}$ and consequently $\rho_j \leq 2\epsilon$ (see Step 1 of Algorithm 4). Note that there are at least $2n/3$ such z^j. This observation further implies that the second tertile $\bar{\rho} \leq 2\epsilon$ (see Step 2 of Algorithm 4).

Now, we consider an arbitrary index $j \in \mathcal{J}$. Since both $B_{\epsilon,d}(\tilde{z})$ and $B_{\rho_j,d}(z^j)$ contain at least $2n/3$ points, by the pigeonhole principle, there must exist a point z at the intersection $B_{\epsilon,d}(\tilde{z}) \cap B_{\rho_j,d}(z^j)$. Using the triangle inequality, we conclude that

$$d(\tilde{z},z^j) \leq d(\tilde{z},z) + d(z,z^j) \leq \epsilon + \rho_j \leq \epsilon + 2\epsilon = 3\epsilon.$$

Since the above inequality holds for any $j \in \mathcal{J}$, (64) immediately follows. ∎

Appendix C. Robust Gradient Estimation

This section presents and analyzes the subroutine RGE used in Algorithm 3. Algorithm 5 below is motivated by Davis et al. (2021).

Algorithm 5 Robust Gradient Estimation, RGE(x,n,q)

Input: A point $x \in \operatorname{dom} h$ and integers $n, q \geq 1$.
for $j = 1,\ldots,n$ **do**
  Step 1. Generate q independent stochastic gradients $s(x,\xi_j^1),\ldots,s(x,\xi_j^q)$ and compute $\bar{s}_j(x) = \frac{1}{q}\sum_{i=1}^q s(x,\xi_j^i)$;
end for
Step 2. Call oracle $\mathcal{J} := \mathrm{STS}(S(x), d_2(\cdot,\cdot))$ where $S(x) = \{\bar{s}_1(x),\ldots,\bar{s}_n(x)\}$.
Output: $\bar{s}_{j^*}(x)$ for an arbitrary index $j^* \in \mathcal{J}$.

In contrast to PSS, i.e., Algorithm 2, RGE achieves variance reduction by using a batch of q independent stochastic gradients. The following lemma presents a concentration inequality of the stochastic gradient estimate.

Lemma 13 *In Algorithm 5, if $q \geq 4\sigma^2/\delta^2$ where σ is as in Assumption (A2) and $\delta > 0$ is some scalar, then for any $x \in \operatorname{dom} h$ and $n \geq 1$, the output $\bar{s}_{j^*}(x)$ satisfies*

$$\mathbb{P}\left(\|\bar{s}_{j^*}(x) - \mathbb{E}[\bar{s}_{j^*}(x)]\| \leq 3\delta\right) \geq 1 - \exp\left(-\frac{n}{72}\right). \tag{65}$$

Proof: We first note that $\nabla f(x) = \mathbb{E}[\bar{s}_{j^*}(x)]$ due to Assumption (A1), and hence use $\nabla f(x)$ throughout the proof for simplicity. Using Assumptions (A1) and (A2), we know that the estimate $\bar{s}_j(x)$ has a variance reduction by a factor of q, that is,

$$\mathbb{E}\left[\|\bar{s}_j(x) - \nabla f(x)\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{q}\sum_{i=1}^{q} s(x, \xi_j^i) - \nabla f(x)\right\|^2\right] = \frac{1}{q^2}\sum_{i=1}^{q}\mathbb{E}\left[\left\|s(x, \xi_j^i) - \nabla f(x)\right\|^2\right] \leq \frac{\sigma^2}{q}.$$

It follows from the assumption that $q \geq 4\sigma^2/\delta^2$ and Markov's inequality that

$$\mathbb{P}\left(\|\bar{s}_j(x) - \nabla f(x)\|^2 \geq \delta^2\right) \leq \frac{\sigma^2}{q\delta^2} \leq \frac{1}{4}.$$

Hence, for every $j = 1, \ldots, n$, we have

$$\mathbb{P}\left(\|\bar{s}_j(x) - \nabla f(x)\| \leq \delta\right) \geq \frac{3}{4}.$$

Using Lemma 11 with $A_j = \{\|\bar{s}_j(x) - \nabla f(x)\| \leq \delta\}$, we have

$$\mathbb{P}\left(|B_{\delta,d_2}(\nabla f(x)) \cap S(x)| > \frac{2n}{3}\right) \geq 1 - \exp\left(-\frac{n}{72}\right).$$

This means condition (63) in Lemma 12 holds with probability at least $1 - \exp\left(-\frac{n}{72}\right)$. Since we call the oracle $\mathcal{J} = \mathrm{STS}(S(x), d_2(\cdot,\cdot))$ in Step 2, it follows from Lemma 12 that for every $j \in \mathcal{J}$,

$$\mathbb{P}\left(\|\bar{s}_j(x) - \nabla f(x)\| \leq 3\delta\right) \geq 1 - \exp\left(-\frac{n}{72}\right).$$

Therefore, (65) holds for any $j^* \in \mathcal{J}$. ∎

Appendix D. Deferred Proofs from Section 3

D.1. Proof of Lemma 1

Proof: We first prove (16) by induction. It is easy to verify that (16) holds for $i = 1$ using (9) and assumption (A1). Assume that (16) holds for some $i \geq 1$. Then, using (5), (9), the induction hypothesis, and the convexity of ϕ^λ, we conclude that

$$\mathbb{E}[u_{i+1}] \overset{(9),(16)}{\geq} \alpha\mathbb{E}[\phi^\lambda(y_i)] + (1-\alpha)\mathbb{E}[\phi^\lambda(x_{i+1})] \geq \mathbb{E}[\phi^\lambda(\alpha y_i + (1-\alpha)x_{i+1})] \overset{(5)}{=} \mathbb{E}[\phi^\lambda(y_{i+1})].$$

Now, we prove (17) again by induction. It is easy to verify that (17) holds for $i = 1$ using (10), assumption (A1), and the convexity of f. Assume that (17) holds for some $i \geq 1$. Then, using (10), the induction hypothesis, and the convexity of ϕ^λ, we conclude that

$$\mathbb{E}[\mathcal{L}_{i+1}(x)] \overset{(10)}{=} \alpha\mathbb{E}[\mathcal{L}_i(x)] + (1-\alpha)\mathbb{E}[\ell(x; x_i, \xi_i)] \overset{(17)}{\leq} \alpha\phi(x) + (1-\alpha)\phi(x) = \phi(x).$$

Finally, we prove (18). Using the definitions of ϕ and $\ell(\cdot; x, \xi)$ in (1) and (8), respectively, we have

$$\begin{aligned}\phi(x_i) - \ell(x_i; x_{i-1}, \xi_{i-1}) &= f(x_i) - f(x_{i-1}) - \langle s_{i-1}, x_i - x_{i-1}\rangle \\ &\leq \langle \nabla f(x_{i-1}) - s_{i-1}, x_i - x_{i-1}\rangle + \frac{L}{2}\|x_i - x_{i-1}\|^2,\end{aligned}$$

where the inequality is due to the second inequality in (3). Now, (18) directly follows from the above inequality and the Cauchy-Schwarz inequality. ∎

D.2. Proof of Lemma 2

Proof: (a) This statement follows directly from (19), the fact that $s_i = s(x_i, \xi_i)$, and assumption (A2).

(b) Let

$$e_i := \Phi(x_i, \xi_i) - \phi(x_i) = F(x_i, \xi_i) - f(x_i). \tag{66}$$

Using the definitions of t_i and $\mathcal{L}_i^\lambda$ in (11), and u_i in (9), we have

$$\begin{aligned}t_1 &\overset{(11)}{=} u_1 - \mathcal{L}_1^\lambda(x_1) \\ &\overset{(9),(11)}{=} \Phi(x_1, \xi_1) - [F(x_0, \xi_0) + \langle s_0, x_1 - x_0\rangle + h(x_1)] \\ &\overset{(8),(66)}{=} e_1 - e_0 + \phi(x_1) - \ell(x_1; x_0, \xi_0) \\ &\overset{(18)}{\leq} e_1 - e_0 + \|\nabla f(x_0) - s_0\|\|x_1 - x_0\| + \frac{L}{2}\|x_1 - x_0\|^2,\end{aligned} \tag{67}$$

where the inequality is due to (18). Thus, the above inequality and assumption (A5) imply that

$$t_1 \leq e_1 - e_0 + \|\nabla f(x_0) - s_0\|D + \frac{L}{2}D^2. \tag{68}$$

It follows from (66), and assumptions (A1) and (A2) that

$$\mathbb{E}[e_1] = 0, \quad \mathbb{E}[e_0] = 0, \quad \mathbb{E}[\|\nabla f(x_0) - s_0\|^2] \leq \sigma^2.$$

Hence, the statement follows by taking expectation of (68) and using the above three relations.

(c) It follows from (14) and (17) that

$$\mathbb{E}[\mathcal{L}_i^\lambda(x_i)] \overset{(14)}{\leq} \mathbb{E}[\mathcal{L}_i^\lambda(\hat{x})] \overset{(17)}{\leq} \mathbb{E}[\phi^\lambda(\hat{x})].$$

Using the above inequality and (16), we have

$$\mathbb{E}[\phi^\lambda(y_i) - \phi^\lambda(\hat{x})] \leq \mathbb{E}[u_i - \mathcal{L}_i^\lambda(x_i)].$$

Hence, the last statement follows from the definition of t_i in (17). ∎

D.3. Proof of Lemma 3

Proof: It suffices to prove that for every $i \geq 2$,

$$t_i \leq \alpha t_{i-1} + (1-\alpha) r_{i-1}, \tag{69}$$

since it is clear that (20) follows immediately from (69) and an induction argument. Let $i \geq 2$ be given. It follows from the definitions of $\mathcal{L}_i$ and $\mathcal{L}_i^\lambda$ in (10) and (11), respectively, that

$$\begin{aligned}
\mathcal{L}_i^\lambda(x_i) - (1-\alpha)\ell(x_i; x_{i-1}, \xi_{i-1}) &= \alpha \mathcal{L}_{i-1}(x_i) + \frac{1}{2\lambda}\|x_i - x_0\|^2 \\
&= \alpha \mathcal{L}_{i-1}^\lambda(x_i) + \frac{1-\alpha}{2\lambda}\|x_i - x_0\|^2 \\
&\overset{(15)}{\geq} \alpha \left[\mathcal{L}_{i-1}^\lambda(x_{i-1}) + \frac{1}{2\lambda}\|x_i - x_{i-1}\|^2\right] + \frac{1-\alpha}{2\lambda}\|x_i - x_0\|^2,
\end{aligned}$$

where the inequality is due to (15). Rearranging the terms in the above inequality and using (11), (18), (19), and (12), we have

$$\begin{aligned}
&\mathcal{L}_i^\lambda(x_i) - \alpha \mathcal{L}_{i-1}^\lambda(x_{i-1}) \geq (1-\alpha)\left[\ell(x_i; x_{i-1}, \xi_{i-1}) + \frac{1}{2\lambda}\|x_i - x_0\|^2 + \frac{\alpha}{2\lambda(1-\alpha)}\|x_i - x_{i-1}\|^2\right] \\
&\overset{(11),(18)}{\geq} (1-\alpha)\phi^\lambda(x_i) + (1-\alpha)\left[\frac{\alpha}{2\lambda(1-\alpha)}\|x_i - x_{i-1}\|^2 - \|\nabla f(x_{i-1}) - s_{i-1}\|\|x_i - x_{i-1}\| - \frac{L}{2}\|x_i - x_{i-1}\|^2\right] \\
&\overset{(12)}{\geq} (1-\alpha)\phi^\lambda(x_i) + (1-\alpha)\left[\frac{I}{4\lambda}\|x_i - x_{i-1}\|^2 - \|\nabla f(x_{i-1}) - s_{i-1}\|\|x_i - x_{i-1}\|\right] \\
&\geq (1-\alpha)\phi^\lambda(x_i) - (1-\alpha)\frac{\lambda\|\nabla f(x_{i-1}) - s_{i-1}\|^2}{I} \overset{(19)}{=} (1-\alpha)\phi^\lambda(x_i) - (1-\alpha) r_{i-1},
\end{aligned}$$

where the last inequality is by the AM-GM inequality. Rearranging the above inequality and using the definition of t_i in (11), identity (9), and the fact that $i \geq 2$, we then conclude that

$$\begin{aligned}
\mathcal{L}_i^\lambda(x_i) + (1-\alpha) r_{i-1} &\geq (1-\alpha)\phi^\lambda(x_i) + \alpha \mathcal{L}_{i-1}^\lambda(x_{i-1}) \\
&\overset{(11)}{=} (1-\alpha)\phi^\lambda(x_i) + \alpha(u_{i-1} - t_{i-1}) \overset{(9)}{=} u_i - \alpha t_{i-1},
\end{aligned}$$

which, in view of the definition of t_i in (11) again, implies (69). ∎

D.4. Proof of Proposition 1

Proof: It follows from Lemma 3 that

$$t_{I+1} \leq \alpha^I t_1 + (1-\alpha)\sum_{j=1}^{I} \alpha^{I-j} r_j.$$

Taking the expectation of the above inequality and using Lemma 2, we have

$$\mathbb{E}[t_{I+1}] \leq \alpha^I \mathbb{E}[t_1] + (1-\alpha)\sum_{j=1}^{I} \alpha^{I-j}\mathbb{E}[r_j] \leq \alpha^I\left(\sigma D + \frac{LD^2}{2}\right) + (1-\alpha)\frac{\lambda\sigma^2}{I}\sum_{j=1}^{I}\alpha^{I-j}.$$

Therefore, (13) immediately holds. ∎

Appendix E. Deferred Proofs from Section 4

E.1. Proof of Lemma 4

Proof: Consider the j-th call to PSS (i.e., Algorithm 2). It clearly follows from (17) and the definition of Γ_k in (24) that (25) holds. Now, we prove (26). It follows from (14) with $i = I + 1$ that

$$x_{I+1}^j = \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \mathcal{L}_{I+1}^j(x) + \frac{1}{2\lambda}\|x - x_0^j\|^2 \right\}.$$

In view of (22) and the definition of Γ_k in (24), the above identity immediately implies (26). Finally, we prove (27). Using (13) and the definitions of t_i and ε_k in (11) and (24), resoectively, we have

$$\varepsilon_k \overset{(13),(24)}{\geq} \mathbb{E}\left[t_{I+1}^j\right] \overset{(11)}{=} \mathbb{E}\left[u_{I+1}^j - (\mathcal{L}_{I+1}^j)^\lambda(x_{I+1}^j)\right].$$

It thus follows from (16) and the definitions of ϕ^λ and $\mathcal{L}_i^\lambda$ in (7) and (11), respectively, that

$$\begin{aligned}\varepsilon_k &\overset{(16)}{\geq} \mathbb{E}\left[\phi^\lambda(y_{I+1}^j) - (\mathcal{L}_{I+1}^j)^\lambda(x_{I+1}^j)\right] \\ &\overset{(7),(11)}{=} \mathbb{E}\left[\phi(y_{I+1}^j) + \frac{1}{2\lambda}\|y_{I+1}^j - x_0^j\|^2 - \mathcal{L}_{I+1}^j(x_{I+1}^j) - \frac{1}{2\lambda}\|x_{I+1}^j - x_0^j\|^2\right].\end{aligned}$$

In view of $(\bar{z}_{k-1}, z_k, w_k)$ and Γ_k given in (22) and (24), respectively, the above inequality immediately implies (27). ∎

E.2. Proof of Lemma 5

Proof: It follows from the definition of v_k in (28) and the optimality condition of (26) that $v_k \in \partial\Gamma_k(z_k)$, i.e., for every $x \in \operatorname{dom} h$,

$$\Gamma_k(x) \geq \Gamma_k(z_k) + \langle v_k, x - z_k\rangle \overset{(28)}{=} \phi(w_k) + \langle v_k, x - w_k\rangle - \eta_k,$$

where the identity is due to the definition of η_k in (28). Thus, (29) follows by taking expectation of the above inequality and using (25). Now, we prove (30). Using (27) and the definition of η_k in (28), we have

$$\begin{aligned}\mathbb{E}[\eta_k] &= \mathbb{E}[\phi(w_k) - \Gamma_k(z_k) - \langle v_k, w_k - z_k\rangle] \\ &\leq \varepsilon_k + \mathbb{E}\left[-\frac{1}{2\lambda}\|w_k - \bar{z}_{k-1}\|^2 + \frac{1}{2\lambda}\|z_k - \bar{z}_{k-1}\|^2 - \langle v_k, w_k - z_k\rangle\right] \\ &= \varepsilon_k - \mathbb{E}\left[\frac{1}{2\lambda}\|w_k - z_k\|^2\right],\end{aligned}$$

where the last identity is due to the definition of v_k in (28). ∎

E.3. Proof of Lemma 6

Proof: a) Assumption (A5) implies that $\phi(x) - \mathbb{E}[\langle v_k, x\rangle]$ has a unique global minimum $\bar{y}$. Thus, for every $x \in \operatorname{dom} h$, we have

$$\phi(x) \geq \phi(\bar{y}) + \mathbb{E}[\langle v_k, x - \bar{y}\rangle] + \frac{\mu}{2}\|x - \bar{y}\|^2. \tag{70}$$

It follows from (29) with $x = \bar{y}$ that

$$\phi(\bar{y}) \geq \mathbb{E}[\phi(w_k) + \langle v_k, \bar{y} - w_k \rangle - \eta_k] \tag{71}$$

Combining (70) and (71), we conclude that for every $x \in \operatorname{dom} h$,

$$\begin{aligned}
\phi(x) &\geq \mathbb{E}[\phi(w_k) + \langle v_k, x - w_k \rangle - \eta_k] + \frac{\mu}{2}\|x - \bar{y}\|^2 \\
&= \mathbb{E}[\phi(w_k) + \langle v_k, x - w_k \rangle] - \mathbb{E}\left[\eta_k + \frac{\mu}{2}\|\bar{y} - w_k\|^2\right] + \frac{\mu}{2}(\mathbb{E}[\|\bar{y} - w_k\|^2] + \|x - \bar{y}\|^2) \\
&\geq \mathbb{E}[\phi(w_k) + \langle v_k, x - w_k \rangle - \eta_k'] + \frac{\mu}{4}\mathbb{E}[\|x - w_k\|^2]
\end{aligned} \tag{72}$$

where

$$\eta_k' := \eta_k + \frac{\mu}{2}\|\bar{y} - w_k\|^2. \tag{73}$$

Using (70) with $x = w_k$ and taking expectation of the resulting inequality, we have

$$\mathbb{E}[\phi(w_k)] \geq \phi(\bar{y}) + \mathbb{E}[\langle v_k, w_k - \bar{y} \rangle] + \frac{\mu}{2}\mathbb{E}[\|w_k - \bar{y}\|^2].$$

It follows from the above inequality and (71) that

$$\frac{\mu}{2}\mathbb{E}[\|w_k - \bar{y}\|^2] \leq \mathbb{E}[\eta_k],$$

which together with (73) implies that

$$\mathbb{E}[\eta_k'] \leq 2\mathbb{E}[\eta_k].$$

Statement (a) now follows from (72), the above inequality, and the definition of q_k in (31).

b) Taking $x = z_k$ in (31) and using the definition of η_k in (28), we have

$$q_k(z_k) = \phi(w_k) + \langle v_k, z_k - w_k \rangle + \frac{\mu}{4}\|z_k - w_k\|^2 - 2\eta_k = \Gamma_k(z_k) + \frac{\mu}{4}\|z_k - w_k\|^2 - \eta_k.$$

The statement now follows from taking expectation of the above inequality and using (30). ∎

E.4. Proof of Lemma 7

Proof: Using the definition of q_k in (31), it can be verified that

$$\begin{aligned}
&\mathbb{E}[q_k(x)] + \frac{1}{2\lambda}\mathbb{E}[\|x - \bar{z}_{k-1}\|^2] - \left(\frac{1}{2\lambda} + \frac{\mu}{4}\right)\mathbb{E}[\|x - z_k\|^2] \\
=&\mathbb{E}[q_k(z_k)] + \frac{1}{2\lambda}\mathbb{E}[\|z_k - \bar{z}_{k-1}\|^2] + \frac{\mu}{2}\mathbb{E}[\langle z_k - w_k, x - z_k \rangle].
\end{aligned}$$

It thus follows from Lemma 6(b) that

$$\begin{aligned}
&\mathbb{E}[q_k(x)] + \frac{1}{2\lambda}\mathbb{E}[\|x - \bar{z}_{k-1}\|^2] - \left(\frac{1}{2\lambda} + \frac{\mu}{4}\right)\mathbb{E}[\|x - z_k\|^2] \\
\geq&\mathbb{E}[\Gamma_k(z_k)] + \frac{1}{2\lambda}\mathbb{E}[\|z_k - \bar{z}_{k-1}\|^2] - \varepsilon_k + \left(\frac{1}{2\lambda} + \frac{\mu}{4}\right)\mathbb{E}[\|z_k - w_k\|^2] + \frac{\mu}{2}\mathbb{E}[\langle z_k - w_k, x - z_k \rangle].
\end{aligned} \tag{74}$$

Note that by the AM-GM inequality and the Cauchy-Schwarz inequality, we have

$$\left(\frac{1}{2\lambda}+\frac{\mu}{4}\right)\|z_k-w_k\|^2+\frac{\lambda\mu^2}{4(2+\lambda\mu)}\|x-z_k\|^2\geq-\frac{\mu}{2}\langle z_k-w_k,x-z_k\rangle,$$

and hence

$$\left(\frac{1}{2\lambda}+\frac{\mu}{4}\right)\mathbb{E}[\|z_k-w_k\|^2]+\frac{\mu}{2}\mathbb{E}[\langle z_k-w_k,x-z_k\rangle]\geq-\frac{\lambda\mu^2}{4(2+\lambda\mu)}\mathbb{E}[\|x-z_k\|^2].$$

Plugging the above inequality into (74) and using (27), we obtain

$$\begin{aligned}&\mathbb{E}[q_k(x)]+\frac{1}{2\lambda}\mathbb{E}[\|x-\bar{z}_{k-1}\|^2]-\left(\frac{1}{2\lambda}+\frac{\mu}{4}\right)\mathbb{E}[\|x-z_k\|^2]\\ \geq&\mathbb{E}[\phi(w_k)]+\frac{1}{2\lambda}\mathbb{E}[\|w_k-\bar{z}_{k-1}\|^2]-2\varepsilon_k-\frac{\lambda\mu^2}{4(2+\lambda\mu)}\mathbb{E}[\|x-z_k\|^2].\end{aligned}$$

The lemma now follows from the above inequality, Lemma 6(a), and rearranging the terms. ∎

Appendix F. Deferred Proof of Proposition 3

Proof: Consider events $\{B_j\}_{j=1}^n$ defined in (36) and define index set $\mathcal{J}_0$ and event E_1 as follows

$$\mathcal{J}_0:=\{j\in\{1,\dots,n\}:B_j\text{ occurs}\},\quad E_1:=\left\{|\mathcal{J}_0|>\frac{2n}{3}\right\}.$$

It follows from (38) and Lemma 11 that

$$\mathbb{P}(E_1)\geq1-\exp\left(-\frac{n}{72}\right).\tag{75}$$

It clearly follows from the first inequality in (59) that for every $j=1,\dots,n$,

$$\frac{1+\lambda\mu}{2\lambda}\|w^j-\hat{z}\|^2+D_h(w^j,\hat{z})+\frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|z^j-\hat{z}\|^2\leq\phi^\lambda(w^j)-\phi^\lambda(\hat{z})+\frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|z^j-\hat{z}\|^2.$$

The above inequality, (34), and (36) imply that

$$\|w^j-\hat{z}\|\leq\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}},\quad D_h(w^j,\hat{z})\leq\tau,\quad\|z^j-\hat{z}\|\leq\sqrt{\frac{(2+\lambda\mu)\lambda\tau}{1+\lambda\mu}},\quad\forall j\in\mathcal{J}_0.\tag{76}$$

Now, we are ready to analyze Steps 1 and 2 of Algorithm 3. Assuming that the event E_1 occurs, then the three inequalities in (76) hold for more than $2/3$ indices $j\in\{1,\dots,n\}$. Now, condition (63) of Lemma 12 is satisfied. Applying Lemma 12 to Step 1 of Algorithm 3 and using the first inequality in (76), we have

$$\|w^j-\hat{z}\|\overset{(76),(64)}{\leq}3\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}},\quad\forall j\in\mathcal{J}_1.\tag{77}$$

Applying Lemma 12 to Step 2 of Algorithm 3 and using the third inequality in (76), we also have

$$\|z^j - \hat{z}\| \overset{(76),(64)}{\leq} 3\sqrt{\frac{(2+\lambda\mu)\lambda\tau}{1+\lambda\mu}}, \quad \forall j \in \mathcal{J}_2. \tag{78}$$

Next, we analyze Step 3 of Algorithm 3. Condition (39) satisfies the assumption $q \geq 4\sigma^2/\delta^2$ in Lemma 13 with

$$\delta = \frac{L}{3}\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}.$$

Hence, applying Lemma 13 to Step 3 of Algorithm 3 and noting that $\nabla f(\tilde{w}) = \mathbb{E}[\bar{s}(\tilde{w})]$, we have

$$\mathbb{P}\left(E_2 \,|\, E_1\right) \overset{(65)}{\geq} 1 - \exp\left(-\frac{n}{72}\right), \tag{79}$$

where event E_2 is defined as

$$E_2 := \left\{ \|\bar{s}(\tilde{w}) - \nabla f(\tilde{w})\| \leq L\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}} \right\}. \tag{80}$$

From now on, we suppose that $E_1 \cap E_2$ occurs. Since $\tilde{w} = w^{j_0}$ with $j_0 \in \mathcal{J}_1$, using the triangle inequality, (77) with $j = j_0$, (80), and the fact that f is L-smooth, we have

$$\|\tilde{w} - \hat{z}\| \overset{(77)}{\leq} 3\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}, \tag{81}$$

and

$$\begin{aligned}\|\bar{s}(\tilde{w}) - \nabla f(\hat{z})\| &\leq \|\bar{s}(\tilde{w}) - \nabla f(\tilde{w})\| + \|\nabla f(\tilde{w}) - \nabla f(\hat{z})\| \\ &\overset{(80)}{\leq} L\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}} + L\|\tilde{w} - \hat{z}\| \overset{(81)}{\leq} 4L\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}.\end{aligned} \tag{82}$$

Using Lemma 9, (81), and (82), we have

$$d_h(w^j, \hat{z}) \overset{(57)}{\leq} D_h(w^j, \hat{z}) + 4\left(L + \frac{1}{\lambda}\right)\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}\|w^j - \hat{z}\|, \quad \forall j \in \{1, \ldots, n\}.$$

Hence, it follows from (76) that

$$d_h(w^j, \hat{z}) \leq \tau + \frac{8(1+\lambda L)\tau}{1+\lambda\mu}, \quad \forall j \in \mathcal{J}_0. \tag{83}$$

Now, we are ready to analyze Step 4 of Algorithm 3. Note that (83) holds for more than $2/3$ indices $j \in \{1, \ldots, n\}$, which means condition (63) of Lemma 12 is satisfied. Applying Lemma 12 to Step 4 of Algorithm 3 and using (83), we have

$$d_h(w^j, \hat{z}) \overset{(83),(64)}{\leq} 3\tau + \frac{24(1+\lambda L)\tau}{1+\lambda\mu}, \quad \forall j \in \mathcal{J}_3. \tag{84}$$

To this end, we are in a position to prove (40). Using Lemma 9, (81), and (82), we have

$$D_h(w^j,\hat{z}) \overset{(58)}{\leq} d_h(w^j,\hat{z}) + 4\left(L+\frac{1}{\lambda}\right)\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}\|w^j-\hat{z}\|, \quad \forall j\in\{1,\dots,n\}. \tag{85}$$

It clearly follows from Algorithm 4 that

$$|\mathcal{J}_1|>\frac{2n}{3}, \quad |\mathcal{J}_2|>\frac{2n}{3}, \quad |\mathcal{J}_3|>\frac{2n}{3}.$$

By the pigeonhole principle, the intersection $\mathcal{J}_1\cap\mathcal{J}_2\cap\mathcal{J}_3$ must be nonempty. Hence, it follows from (85), (84), and (77) that

$$D_h(w^j,\hat{z}) \leq 3\tau + \frac{48(1+\lambda L)\tau}{1+\lambda\mu}, \quad \forall j\in\mathcal{J}_1\cap\mathcal{J}_3.$$

Consider an arbitrary index $j\in\mathcal{J}_1\cap\mathcal{J}_2\cap\mathcal{J}_3$. Using the second inequality in (59), the above inequality, (77), and (78), we conclude

$$\begin{aligned}\phi^\lambda(w^j)-\phi^\lambda(\hat{z})+\frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|z^j-\hat{z}\|^2 &\overset{(59)}{\leq} \frac{1+\lambda L}{2\lambda}\|w^j-\hat{z}\|^2 + D_h(w^j,\hat{z}) + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|z^j-\hat{z}\|^2\\ &\leq \frac{1+\lambda L}{1+\lambda\mu}9\tau + \left(3\tau+\frac{48(1+\lambda L)\tau}{1+\lambda\mu}\right)+9\tau\\ &\leq 9\kappa\tau+3\tau+48\kappa\tau+9\tau,\end{aligned}$$

where the last inequality is due to the fact that $(a+b)/(c+d)\leq\max\{a/c,b/d\}$ for $a,b,c,d>0$. Therefore, (40) follows. Finally, combining (75) and (79), we complete the proof

$$\mathbb{P}(E_1\cap E_2)=\mathbb{P}(E_2|E_1)\mathbb{P}(E_1)\geq\left(1-\exp\left(-\frac{n}{72}\right)\right)^2\geq 1-2\exp\left(-\frac{n}{72}\right).$$

■