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

Jiaming Liang *

February 14, 2024

Abstract

This paper proposes a stochastic proximal point method to solve a stochastic convex composite optimization problem. High probability results in stochastic optimization typically hinge on restrictive assumptions on the stochastic gradient noise, for example, sub-Gaussian distributions. Assuming only weak conditions such as bounded variance of the stochastic gradient, this paper establishes a low sample complexity to obtain a high probability guarantee on the convergence of the proposed method. Additionally, a notable aspect of this work is the development of a subroutine to solve the proximal subproblem, which also serves as a novel technique for variance reduction.

Key words. stochastic convex composite optimization, high probability, sample complexity, iteration complexity, proximal point method, variance reduction

AMS subject classifications. 49M37, 65K05, 68Q25, 90C25, 90C30, 90C60

1 Introduction

The main goal of this paper is to present a stochastic proximal point method (SPPM) to solve a stochastic convex composite optimization problem

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

where $f(x) = \mathbb{E}_\xi[F(x,\xi)]$, and study the sample complexity of the proposed SPPM. We assume that h is convex, f is strongly convex, and that the stochastic gradient oracle of f is unbiased and has a bounded variance.

Under the above assumptions, standard results [4, 9, 11, 13] about stochastic approximation (SA) methods provide non-asymptotic convergence guarantees in expectation, that is, for some tolerance $\varepsilon > 0$, one can find an ε-solution $x \in \operatorname{dom} h$ satisfying $\mathbb{E}[\phi(x)] - \phi_* \le \varepsilon$ within $\mathcal{O}(1/\varepsilon)$ queries to the stochastic gradient oracle of f. Pobability results such as $\mathbb{P}(\phi(x) - \phi_* \le \varepsilon) \ge 1 - p$ for given $\varepsilon > 0$ and $p \in (0,1)$ are also available using Markov's inequality. The resulting sample complexity for the stochastic gradient oracle of f is $\mathcal{O}(1/(\varepsilon p))$, which is undesirable, as p can be close enough to 0. Without changing the algorithm, previous works [3, 4, 6, 7, 10] provide a solution to improve the sample complexity in terms of p from $\mathcal{O}(1/p)$ to $\mathcal{O}(\log(1/p))$ by assuming

*Goergen Institute for Data Science and Department of Computer Science, University of Rochester, Rochester, NY 14620 (email: jiaming.liang@rochester.edu).

a stronger condition on the stochastic gradient noise, namely a sub-Gaussian distribution, which is more restrictive than the bounded variance as we assume in this paper.

Recent papers [2, 5] leverage classical probability tools, such as robust distance estimation (RDE) by [11], to reduce the dependence of sample complexity on p without making restrictive assumptions on stochastic gradient noise. In particular, [2] proposes a novel and interesting approach by incorporating an arbitrary existing SA method and RDE into the proximal point method (PPM). For a strongly convex and smooth problem, the condition number for its proximal subproblem becomes better with a smaller prox stepsize. As a result, the sample complexity of their proposed methods has only a logarithmic dependence on $1/p$. However, the methods proposed in [2] are arguably more conceptual than practical. First, [2] does not specify which SA method to use within PPM but assumes some minimization oracle. Moreover, it is unclear which SA method may have the best performance with PPM and RDE. Second, their proposed methods use a sequence of decreasing prox stepsizes, which require the strong convexity parameter μ. However, the parameter μ is usually unknown or difficult to estimate. Even in the deterministic setting, it is still an active research question to design adaptive/universal methods [1, 8, 12] to establish optimal iteration complexity for strongly convex problems. Third, another improvement in [2] over [5] is the dependence of sample complexity on the condition number, while it is also contingent upon the knowledge of μ. Consequently, such an improvement might not be achievable in the absence of a reliable estimate of μ.

Our approach SPPM is built upon the constant prox stepsize PPM that does not need μ or other problem parameters. By the nature of PPM, SPPM is an iterative method. At the k-th iteration of SPPM, given the prox-center $\bar{z}_{k-1}$, we want to solve the following proximal subproblem

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

where λ is the prox stepsize. SPPM calls a proximal subproblem solver (PSS) oracle to approximately solve (2) for n times, and then calls another probability booster (PB) oracle to choose one of the n approximate solutions. It can be shown that this chosen solution is close enough to the exact solution $\hat{z}_k$ with high probability. The approximate solution obtained in the k-th iteration will be the prox-center $\bar{z}_k$ for the $(k+1)$-th iteration.

The contributions of this paper are summarized in the following. First, we propose an implementable and parameter-free SA-type method called SPPM to solve (1). Second, without making more restrictive noise assumptions than bounded variance, we prove a convergence guarantee of SPPM with probability at least $1-p$ and low sample complexity that has only a logarithmic dependence on $1/p$. Third, our proposed subroutine PSS approximately solves (2) and reduces the variance of its approximate solution at the same time. This brings us a novel variance reduction technique due to the usage of PMM.

Organization of the paper. Subsection 1.1 presents basic notation and definitions used throughout the paper. Section 2 formally describes problem (1) and presents SPPM. Section 3 provides the oracle PSS and its analysis. Section 4 gives a proximal point analysis of SPPM, which serves as the backbone of the analysis of this paper. Section 5 presents and analyzes the oracle PB. Section 6 provides the main results of the paper, namely a high probability guarantee and a low sample complexity bound. Section 7 gives some concluding remarks and potential directions for future investigation. Appendix A describes several useful technical results. Finally, Appendices B and C provide the additional oracles needed in the paper and their analysis.

1.1 Basic notation and definitions

Let $\mathbb{R}$ denote the set of real numbers. Let $\mathbb{R}^d$ denote the standard d-dimensional Euclidean space equipped with inner product and norm denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively. For a given function $\psi : \mathbb{R}^d \to (-\infty, +\infty]$, let $\operatorname{dom} \psi := \{x \in \mathbb{R}^d : \psi(x) < \infty\}$ denote the effective domain of ψ. We say ψ is proper if $\operatorname{dom} \psi \neq \emptyset$. A proper function $\psi : \mathbb{R}^d \to (-\infty, +\infty]$ is μ-strongly convex for some $\mu > 0$ if

$$\psi(tx + (1-t)y) \le t\psi(x) + (1-t)\psi(y) - \frac{\mu}{2}\|x - y\|^2$$

for every $t \in [0, 1]$ and $x, y \in \operatorname{dom} \psi$. The *subdifferential* of ψ at $x \in \operatorname{dom} \psi$ is denoted by

$$\partial\psi(x) := \left\{ s \in \mathbb{R}^d : \psi(y) \ge \psi(x) + \langle s, y - x \rangle, \forall y \in \mathbb{R}^d \right\}. \tag{3}$$

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 $x, y \in \operatorname{dom} h$,

$$\frac{\mu}{2}\|y - x\|^2 \le f(y) - f(x) - \langle \nabla f(x), y - x \rangle \le \frac{L}{2}\|y - x\|^2. \tag{4}$$

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 of them so that the same guarantee holds with a high confidence level.

Algorithm 1 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. Call the oracle PSS$(\bar{z}_{k-1}, \alpha, \lambda, I)$ for n times and generate independent pairs $(z_k^1, w_k^1), \ldots, (z_k^n, w_k^n)$;
  Step 2. Call the oracle PB$(\{(z_k^j, w_k^j)\}_{j=1}^n, \bar{z}_{k-1}, q, \lambda)$ to generate $(\bar{z}_k, \bar{w}_k)$.
end for

Under Assumptions (A1)-(A5) on (1) and mild requirements on the input of Algorithm 1, we provide a high probability guarantee to obtain an ε-solution of (1) in Theorem 6.2. We further establish a low sample complexity bound on stochastic gradients in Theorem 6.3. Another interesting finding of this paper is that the oracle PSS (i.e., Algorithm 2) provides a novel variance reduction technique without using a batch of samples. (See the discussion at the end of Section 3.) This advantage inherently arises from the use of PPM in solving stochastic optimization problems (1).

3 Proximal subproblem solver

This section presents and analyzes the first key oracle, namely PSS, used in Algorithm 1. Algorithm 2 below gives a detailed description of PSS.

Algorithm 2 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

$$x_i = \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ h(x) + \langle S_i, x\rangle + \frac{1}{2\lambda}\|x - x_0\|^2 \right\}, \tag{5}$$

$$y_i = \begin{cases} x_i, & \text{if } i = 1, \\ \alpha y_{i-1} + (1-\alpha)x_i, & \text{otherwise,} \end{cases} \tag{6}$$

where

$$S_i = \begin{cases} s_0, & \text{if } i = 1, \\ \alpha S_{i-1} + (1-\alpha)s_{i-1}, & \text{otherwise.} \end{cases} \tag{7}$$

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

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

$$\hat{x} := \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \phi^\lambda(x) := \phi(x) + \frac{1}{2\lambda}\|x - x_0\|^2 \right\} \tag{8}$$

and analyze the solution pair (x_{I+1}, y_{I+1}) returned by Algorithm 2. The main result of this section is Proposition 3.4, which establishes the convergence rate of the expected primal gap of (8). 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:

$$\Phi(\cdot, \xi) = F(\cdot, \xi) + h(\cdot), \quad \ell(\cdot; x, \xi) = f(x) + \langle s(x, \xi), \cdot - x\rangle + h(\cdot), \tag{9}$$

$$u_i := \begin{cases} \Phi(x_1, \xi_1) + \frac{1}{2\lambda}\|x_1 - x_0\|^2, & \text{if } i = 1, \\ \alpha u_{i-1} + (1-\alpha)\left[\phi(x_i) + \frac{1}{2\lambda}\|x_i - x_0\|^2\right], & \text{otherwise,} \end{cases} \tag{10}$$

$$\mathcal{L}_i(\cdot) := \begin{cases} F(x_0, \xi_0) + \langle s_0, \cdot - x_0\rangle + h(\cdot), & \text{if } i = 1, \\ \alpha\mathcal{L}_{i-1}(\cdot) + (1-\alpha)\ell(\cdot; x_{i-1}, \xi_{i-1}), & \text{otherwise,} \end{cases} \tag{11}$$

and

$$\mathcal{L}_i^\lambda(\cdot) := \mathcal{L}_i(\cdot) + \frac{1}{2\lambda}\|\cdot - x_0\|^2. \tag{12}$$

We make some simple observations. It is easy to see from (5), (7), and (12) that

$$x_i = \underset{x\in\mathbb{R}^d}{\operatorname{argmin}}\, \mathcal{L}_i^\lambda(x). \tag{13}$$

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

$$\mathcal{L}_i^\lambda(x) \ge \mathcal{L}_i^\lambda(x_i) + \frac{1}{2\lambda}\|x - x_i\|^2. \tag{14}$$

The result outlined below provides some basic relations that are frequently used in our analysis.

Lemma 3.1. *For every $i \ge 1$, we have*

$$\mathbb{E}[\phi^\lambda(y_i)] \le \mathbb{E}[u_i], \tag{15}$$

$$\mathbb{E}[\mathcal{L}_i(x)] \le \phi(x), \quad \forall x \in \operatorname{dom} h, \tag{16}$$

$$\phi(x_i) - \ell(x_i; x_{i-1}, \xi_{i-1}) \le \|\nabla f(x_{i-1}) - s_{i-1}\|\|x_i - x_{i-1}\| + \frac{L}{2}\|x_i - x_{i-1}\|^2. \tag{17}$$

Proof: We first prove (15) by induction. It is easy to verify that (15) holds for $i = 1$ using (10) and assumption (A1). Assume that (15) holds for some $i \ge 1$. Then, using (6), (10), the induction hypothesis, and the convexity of ϕ^λ, we conclude that

$$\mathbb{E}[u_{i+1}] \overset{(10),(15)}{\ge} \alpha\mathbb{E}[\phi^\lambda(y_i)] + (1-\alpha)\mathbb{E}[\phi^\lambda(x_{i+1})] \ge \mathbb{E}[\phi^\lambda(\alpha y_i + (1-\alpha)x_{i+1})] \overset{(6)}{=} \mathbb{E}[\phi^\lambda(y_{i+1})].$$

Now, we prove (16) again by induction. It is easy to verify that (16) holds for $i = 1$ using (11), assumption (A1), and the convexity of f. Assume that (16) holds for some $i \ge 1$. Then, using (11), the induction hypothesis, and the convexity of ϕ^λ, we conclude that

$$\mathbb{E}[\mathcal{L}_{i+1}(x)] \overset{(11)}{=} \alpha\mathbb{E}[\mathcal{L}_i(x)] + (1-\alpha)\mathbb{E}[\ell(x; x_i, \xi_i)] \overset{(16)}{\le} \alpha\phi(x) + (1-\alpha)\phi(x) = \phi(x).$$

Finally, we prove (17). Using the definitions of ϕ and $\ell(\cdot; x, \xi)$ in (1) and (9), respectively, we have

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

where the inequality is due to the second inequality in (4). Now, (17) directly follows from the above inequality and the Cauchy-Schwarz inequality. ∎

The technical result below introduces a key quantity t_i, whose expectation $\mathbb{E}[t_i]$ is an upper bound on the primal gap of (8).

Lemma 3.2. *For every $i \ge 1$, define*

$$t_i := u_i - \mathcal{L}_i^\lambda(x_i), \quad r_i := \frac{\lambda\|\nabla f(x_i) - s_i\|^2}{I}, \tag{18}$$

where u_i is as in (10) and $\mathcal{L}_i^\lambda$ is as in (12). 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 $i \geq 1$, $\mathbb{E}[t_i] \geq \mathbb{E}[\phi^\lambda(y_i) - \phi^\lambda(\hat{x})]$.

Proof: (a) This statement follows directly from (18), the fact that $s_i = s(x_i, \xi_i)$, and assumption (A2).

(b) Let

$$e_i := \Phi(x_i, \xi_i) - \phi(x_i) = F(x_i, \xi_i) - f(x_i). \tag{19}$$

Using the definitions of t_i, u_i, and $\mathcal{L}_i^\lambda$ in (18), (10), and (12), respectively, we have

$$\begin{aligned} t_1 &\overset{(18)}{=} u_1 - \mathcal{L}_1^\lambda(x_1) \\ &\overset{(10),(12)}{=} \Phi(x_1, \xi_1) - [F(x_0, \xi_0) + \langle s_0, x_1 - x_0 \rangle + h(x_1)] \\ &\overset{(9),(19)}{=} e_1 - e_0 + \phi(x_1) - \ell(x_1; x_0, \xi_0) \\ &\overset{(17)}{\leq} e_1 - e_0 + \|\nabla f(x_0) - s_0\| \|x_1 - x_0\| + \frac{L}{2}\|x_1 - x_0\|^2, \end{aligned} \tag{20}$$

where the inequality is due to (17). Thus, the above inequality and assumption (A5) imply that

$$t_1 \leq e_1 - e_0 + \|\nabla f(x_0) - s_0\| D + \frac{L}{2} D^2. \tag{21}$$

It follows from (19), 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 (21) and using the above three relations.

(c) It follows from (13) and (16) that

$$\mathbb{E}[\mathcal{L}_i^\lambda(x_i)] \overset{(13)}{\leq} \mathbb{E}[\mathcal{L}_i^\lambda(\hat{x})] \overset{(16)}{\leq} \mathbb{E}[\phi^\lambda(\hat{x})].$$

Using the above inequality and (15), 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 (16). ∎

The next lemma provides a useful relation for t_i and r_i defined in (18). This relation will then be used in Proposition 3.4 to show that t_{I+1} is small in expectation.

Lemma 3.3. *Assuming*

$$\alpha \geq \frac{I/2 + \lambda L}{1 + I/2 + \lambda L}, \tag{22}$$

then, for every $i \geq 2$, we have

$$t_i \leq \alpha^{i-1} t_1 + (1 - \alpha) \sum_{j=1}^{i-1} \alpha^{i-j-1} r_j, \tag{23}$$

where t_i and r_i are as in (18).

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

$$t_i \leq \alpha t_{i-1} + (1-\alpha) r_{i-1}, \tag{24}$$

since it is clear that (23) follows immediately from (24) 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 (11) and (12), respectively, that

$$\begin{aligned}
\mathcal{L}_i^\lambda(x_i) - (1-\alpha)\ell(x_i; x_{i-1}, \xi_{i-1}) &= \alpha \mathcal{L}_{i-1}(x_i) + \frac{1}{2\lambda}\|x_i - x_0\|^2 \\
&= \alpha \mathcal{L}_{i-1}^\lambda(x_i) + \frac{1-\alpha}{2\lambda}\|x_i - x_0\|^2 \\
&\overset{(14)}{\geq} \alpha \left[\mathcal{L}_{i-1}^\lambda(x_{i-1}) + \frac{1}{2\lambda}\|x_i - x_{i-1}\|^2 \right] + \frac{1-\alpha}{2\lambda}\|x_i - x_0\|^2,
\end{aligned}$$

where the inequality is due to (14). Rearranging the terms in the above inequality and using (12), (17), (18), and (22), we have

$$\begin{aligned}
&\mathcal{L}_i^\lambda(x_i) - \alpha \mathcal{L}_{i-1}^\lambda(x_{i-1}) \geq (1-\alpha) \left[\ell(x_i; x_{i-1}, \xi_{i-1}) + \frac{1}{2\lambda}\|x_i - x_0\|^2 + \frac{\alpha}{2\lambda(1-\alpha)}\|x_i - x_{i-1}\|^2 \right] \\
&\overset{(12),(17)}{\geq} (1-\alpha)\phi^\lambda(x_i) + (1-\alpha) \left[\frac{\alpha}{2\lambda(1-\alpha)}\|x_i - x_{i-1}\|^2 - \|\nabla f(x_{i-1}) - s_{i-1}\|\|x_i - x_{i-1}\| - \frac{L}{2}\|x_i - x_{i-1}\|^2 \right] \\
&\overset{(22)}{\geq} (1-\alpha)\phi^\lambda(x_i) + (1-\alpha) \left[\frac{I}{4\lambda}\|x_i - x_{i-1}\|^2 - \|\nabla f(x_{i-1}) - s_{i-1}\|\|x_i - x_{i-1}\| \right] \\
&\geq (1-\alpha)\phi^\lambda(x_i) - (1-\alpha)\frac{\lambda\|\nabla f(x_{i-1}) - s_{i-1}\|^2}{I} \overset{(18)}{=} (1-\alpha)\phi^\lambda(x_i) - (1-\alpha) r_{i-1},
\end{aligned}$$

where the last inequality by the AM-GM inequality. Rearranging the above inequality and using the definition of t_i in (18), identity (10), and the fact that $i \geq 2$, we then conclude that

$$\begin{aligned}
\mathcal{L}_i^\lambda(x_i) + (1-\alpha) r_{i-1} &\geq (1-\alpha)\phi^\lambda(x_i) + \alpha \mathcal{L}_{i-1}^\lambda(x_{i-1}) \\
&\overset{(18)}{=} (1-\alpha)\phi^\lambda(x_i) + \alpha(u_{i-1} - t_{i-1}) \overset{(10)}{=} u_i - \alpha t_{i-1},
\end{aligned}$$

which, in view of the definition of t_i in (18) again, implies (24). ■

Finally, 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 (8) decreases as I grows.

Proposition 3.4. *Assuming* (22) *holds, then we have*

$$\mathbb{E}[t_{I+1}] \leq \alpha^I \left(\sigma D + \frac{LD^2}{2} \right) + \frac{\lambda \sigma^2}{I}. \tag{25}$$

Proof: It follows from Lemma 3.3 that

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

Taking expectation of the above inequality and using Lemma 3.2, 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, (25) immediately holds. ∎

We conclude this section by providing some insight into Proposition 3.4. First, the terms $\lambda\sigma^2/I$ and $\alpha^I\left(\sigma D + LD^2/2\right)$ in (25) can be roughly regarded as the variance and bias components, respectively. Second, the tradeoff between bias and variance is managed by selecting the prox stepsize λ. The bias term is affected by λ via α as depicted in (22). Third, increasing the number of iterations I in Algorithm 2 leads to a reduction in both bias and variance terms. Unlike typical stochastic gradient methods, inequality (25) suggests a variance reduction by a factor of I. Notably, this reduction is not achieved using the well-known sample average technique. Instead, it occurs by executing multiple steps to solve the proximal subproblem (8). This distinctive approach to variance reduction by a factor of I is logical, considering that Algorithm 2 ultimately utilizes $I+1$ 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 4.5, 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 $j \in \{1,\ldots,n\}$,

$$\bar z_{k-1} = x_0^j, \quad z_k^j = x_{I+1}^j, \quad w_k^j = y_{I+1}^j. \tag{26}$$

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 (26), the notational convention we adopt in this section is

$$\bar z_{k-1} = x_0^j, \quad z_k = x_{I+1}^j, \quad w_k = y_{I+1}^j. \tag{27}$$

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 first use the notation in (27) to translate some important results in Section 3 into the proximal point perspective. The following result is the starting point for the proximal point analysis.

Lemma 4.1. *For every $k \ge 1$, let*

$$\Gamma_k = \mathcal{L}_{I+1}^j, \quad \varepsilon_k = \alpha^I\left(\sigma D + \frac{LD^2}{2}\right) + \frac{\lambda\sigma^2}{I}, \tag{28}$$

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

$$\mathbb{E}[\Gamma_k(x)] \le \phi(x), \quad \forall x \in \operatorname{dom} h, \tag{29}$$

$$z_k = \operatorname*{argmin}_{x\in\mathbb{R}^d}\left\{\Gamma_k(x) + \frac{1}{2\lambda}\|x - \bar z_{k-1}\|^2\right\}, \tag{30}$$

$$\mathbb{E}\left[\phi(w_k) + \frac{1}{2\lambda}\|w_k - \bar z_{k-1}\|^2 - \Gamma_k(z_k) - \frac{1}{2\lambda}\|z_k - \bar z_{k-1}\|^2\right] \le \varepsilon_k. \tag{31}$$

Proof: Consider the j-th call to PSS (i.e., Algorithm 2). It clearly follows from (16) and the definition of Γ_k in (28) that (29) holds. Now, we prove (30). It follows from (13) with $i = I+1$ that

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

In view of (27) and the definition of Γ_k in (28), the above identity immediately implies (30). Finally, we prove (31). Using (25) and the definitions of t_i and ε_k in (18) and (28), resoectively, we have

$$\varepsilon_k \overset{(25),(28)}{\geq} \mathbb{E}\left[t_{I+1}^j\right] \overset{(18)}{=} \mathbb{E}\left[u_{I+1}^j - (\mathcal{L}_{I+1}^j)^\lambda(x_{I+1}^j)\right].$$

It thus follows from (15) and the definitions of ϕ^λ and $\mathcal{L}_i^\lambda$ in (8) and (12), respectively, that

$$\begin{aligned}\varepsilon_k &\overset{(15)}{\geq} \mathbb{E}\left[\phi^\lambda(y_{I+1}^j) - (\mathcal{L}_{I+1}^j)^\lambda(x_{I+1}^j)\right] \\ &\overset{(8),(12)}{=} \mathbb{E}\left[\phi(y_{I+1}^j) + \frac{1}{2\lambda}\|y_{I+1}^j - x_0^j\|^2 - \mathcal{L}_{I+1}^j(x_{I+1}^j) - \frac{1}{2\lambda}\|x_{I+1}^j - x_0^j\|^2\right].\end{aligned}$$

In view of $(\bar z_{k-1}, z_k, w_k)$ and Γ_k given in (27) and (28), respectively, the above inequality immediately implies (31). ■

The following lemma is a technical result that generalizes the stationarity condition of (30) from the proximal point perspective in the stochastic setting.

Lemma 4.2. *Define*

$$v_k := \frac{\bar z_{k-1} - z_k}{\lambda}, \quad \eta_k := \phi(w_k) - \Gamma_k(z_k) - \langle v_k, w_k - z_k\rangle. \tag{32}$$

Then, we have

$$\phi(x) \geq \mathbb{E}[\phi(w_k) + \langle v_k, x - w_k\rangle - \eta_k], \quad \forall x \in \operatorname{dom} h, \tag{33}$$

$$\mathbb{E}[\eta_k] \leq \varepsilon_k - \frac{1}{2\lambda}\mathbb{E}\left[\|w_k - z_k\|^2\right]. \tag{34}$$

Proof: It follows from the definition of v_k in (32) and the optimality condition of (30) that $v_k \in \partial\Gamma_k(z_k)$, i.e., for every $x \in \operatorname{dom} h$,

$$\Gamma_k(x) \geq \Gamma_k(z_k) + \langle v_k, x - z_k\rangle \overset{(32)}{=} \phi(w_k) + \langle v_k, x - w_k\rangle - \eta_k,$$

where the identity is due to the definition of η_k in (32). Thus, (33) follows by taking expectation of the above inequality and using (29). Now, we prove (34). Using (31) and the definition of η_k in (32), 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 (32). ■

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.

Lemma 4.3. *Define*

$$q_k(x) := \phi(w_k) + \langle v_k, x - w_k \rangle + \frac{\mu}{4}\|x - w_k\|^2 - 2\eta_k. \tag{35}$$

Then, the following statements hold:

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

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

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

$$\phi(x) \ge \phi(\bar{y}) + \mathbb{E}[\langle v_k, x - \bar{y} \rangle] + \frac{\mu}{2}\|x - \bar{y}\|^2. \tag{36}$$

It follows from (33) with $x = \bar{y}$ that

$$\phi(\bar{y}) \ge \mathbb{E}[\phi(w_k) + \langle v_k, \bar{y} - w_k \rangle - \eta_k] \tag{37}$$

Combining (36) and (37), we conclude that for every $x \in \operatorname{dom} h$,

$$\begin{aligned}
\phi(x) &\ge \mathbb{E}[\phi(w_k) + \langle v_k, x - w_k \rangle - \eta_k] + \frac{\mu}{2}\|x - \bar{y}\|^2 \\
&= \mathbb{E}[\phi(w_k) + \langle v_k, x - w_k \rangle] - \mathbb{E}\left[\eta_k + \frac{\mu}{2}\|\bar{y} - w_k\|^2\right] + \frac{\mu}{2}(\mathbb{E}[\|\bar{y} - w_k\|^2] + \|x - \bar{y}\|^2) \\
&\ge \mathbb{E}[\phi(w_k) + \langle v_k, x - w_k \rangle - \eta_k'] + \frac{\mu}{4}\mathbb{E}[\|x - w_k\|^2] \qquad (38)
\end{aligned}$$

where

$$\eta_k' := \eta_k + \frac{\mu}{2}\|\bar{y} - w_k\|^2. \tag{39}$$

Using (36) with $x = w_k$ and taking expectation of the resulting inequality, we have

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

It follows from the above inequality and (37) that

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

which together with (39) implies that

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

Statement (a) now follows from (38), the above inequality, and the definition of q_k in (35).

b) Taking $x = z_k$ in (35) and using the definition of η_k in (32), 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 (34). ■

Combining the technical results above, the next lemma shows that the suboptimality of (2) is under control in expectation.

Lemma 4.4. *For every $k \geq 1$ and $x \in \mathrm{dom}\, h$, we have*

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

Proof: Using the definition of q_k in (35), 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 4.3(b) that

$$\begin{aligned}&\mathbb{E}[q_k(x)] + \frac{1}{2\lambda}\mathbb{E}[\|x - \bar{z}_{k-1}\|^2] - \left(\frac{1}{2\lambda} + \frac{\mu}{4}\right)\mathbb{E}[\|x - z_k\|^2] \\ \geq&\mathbb{E}[\Gamma_k(z_k)] + \frac{1}{2\lambda}\mathbb{E}[\|z_k - \bar{z}_{k-1}\|^2] - \varepsilon_k + \left(\frac{1}{2\lambda} + \frac{\mu}{4}\right)\mathbb{E}[\|z_k - w_k\|^2] + \frac{\mu}{2}\mathbb{E}[\langle z_k - w_k, x - z_k\rangle]. \quad (40)\end{aligned}$$

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 (40) and using (31), 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 4.3(a), and rearranging the terms. ∎

Finally, we are ready to present a probability guarantee of the suboptimality of the proximal subproblem (2).

Proposition 4.5. *For every $k \geq 1$, we have*

$$\mathbb{P}\left(\phi(w_k) + \frac{1}{2\lambda}\|w_k - \bar{z}_{k-1}\|^2 - \phi(\hat{z}_k) - \frac{1}{2\lambda}\|\hat{z}_k - \bar{z}_{k-1}\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\hat{z}_k - z_k\|^2 \leq 8\varepsilon_k\right) \geq \frac{3}{4}, \quad (41)$$

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

Proof: Using Lemma 4.4 and Markov's inequality, we have every $x \in \mathrm{dom}\, h$,

$$\mathbb{P}\left(\phi(w_k) + \frac{1}{2\lambda}\|w_k - \bar{z}_{k-1}\|^2 - \phi(x) - \frac{1}{2\lambda}\|x - \bar{z}_{k-1}\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|x - z_k\|^2 \geq 8\varepsilon_k\right) \leq \frac{1}{4}.$$

The proposition now immediately follows by taking $x = \hat{z}_k$ in the above inequality. ∎

5 Probability booster

This section presents and analyzes the second key oracle, namely PB, used in Algorithm 1. Algorithm 3 below gives a detailed description of PB. We note that Algorithm 3 also relies on two other subroutines, namely second tertile selection (STS) and robust gradient estimation (RGE). For the sake of simplicity, we defer their presentation and discussion to Appendices B and C.

Algorithm 3 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 $\mathcal{J}_1 = \mathrm{STS}(\{w^j\}_{j=1}^n, d_2(\cdot,\cdot))$;
Step 2. Call oracle $\mathcal{J}_2 = \mathrm{STS}(\{z^j\}_{j=1}^n, d_2(\cdot,\cdot))$;
Step 3. Fix an arbitrary $j_0 \in \mathcal{J}_1 \cap \mathcal{J}_2$ and set $\tilde{w} := w^{j_0}$. Call oracle $\bar{s}(\tilde{w}) = \mathrm{RGE}(\tilde{w}, n, q)$;
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 $\mathcal{J}_3 = \mathrm{STS}(\{w^j\}_{j=1}^n, d_h(\cdot,\cdot))$.
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 as input a metric. Here, $d_2(x, y) = \|x - y\|_2$ denotes the Euclidean distance and is therefore a metric. We prove $d_h(\cdot,\cdot)$ is a metric in Lemma A.1.

Recall that Step 1 of Algorithm 1 generates n pairs $\{(z_k^j, w_k^j)\}_{j=1}^n$ by calling the oracle PSS (i.e., Algorithm 2) and Proposition 4.5 gives a low probability guarantee on the suboptimality (see (41)) holds for each pair (z_k^j, w_k^j). Our goal in Algorithm 3 is to select one of those n pairs such that a similar inequality as in (41) holds with a high probability.

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 $\lambda > 0$, we consider

$$\hat{z} := \operatorname*{argmin}_{x \in \mathbb{R}^d} \phi^\lambda(x), \tag{42}$$

where

$$\phi^\lambda(\cdot) := \phi(\cdot) + \frac{1}{2\lambda}\|\cdot - \bar{z}\|^2, \quad f^\lambda(\cdot) := f(\cdot) + \frac{1}{2\lambda}\|\cdot - \bar{z}\|^2. \tag{43}$$

The optimality condition of the proximal subproblem (42) 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 $x \in \operatorname{dom} h$,

$$D_h(x, \hat{z}) := h(x) - h(\hat{z}) + \langle \nabla f^\lambda(\hat{z}), x - \hat{z}\rangle \geq 0, \tag{44}$$

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

$$d_h(w^j, \hat{z}) = |h(w^j) - h(\hat{z}) + \langle \bar{s}^\lambda(\tilde{w}), w^j - \hat{z}\rangle|. \tag{45}$$

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

$$B_j := \left\{\phi^\lambda(w^j) - \phi^\lambda(\hat{z}) + \frac{1 + \lambda\mu}{\lambda(2 + \lambda\mu)}\|\hat{z} - z^j\|^2 \leq \tau\right\}, \tag{46}$$

where τ is considered as (in view of (41))

$$\tau = 8\varepsilon_k \overset{(28)}{=} 8\left[\alpha^I\left(\sigma D + \frac{LD^2}{2}\right) + \frac{\lambda\sigma^2}{I}\right]. \tag{47}$$

Then it follows from Proposition 4.5 that a low probability guarantee holds for each B_j, that is,

$$\mathbb{P}(B_j \text{ occurs}) \geq \frac{3}{4}. \tag{48}$$

The following result shows that the output of Algorithm 3 satisfies a high probability guarantee for an inequality similar to the one in (46), differing at most by a factor of the condition number.

Proposition 5.1. *Assume that* (22) *holds. If the input q of Algorithm 3 satisfies*

$$q \geq \frac{18(1+\lambda\mu)\sigma^2}{\lambda L^2 \tau}, \tag{49}$$

where τ is as in (47)*, then with probability at least $1 - 2\exp(-n/72)$, the pair (z^j, w^j) returned by PB satisfies*

$$\phi^\lambda(w^j) - \phi^\lambda(\hat{z}) + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\hat{z} - z^j\|^2 \leq 12\tau + 57\kappa\tau, \tag{50}$$

where $\hat{z}$ is as in (42) *and condition number $\kappa = L/\mu > 1$.*

Proof: Consider events $\{B_j\}_{j=1}^n$ defined in (46) and define index set $\mathcal{J}_0$ and event E_1 as follows

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

It follows from (48) and Lemma A.4 that

$$\mathbb{P}(E_1) \geq 1 - \exp\left(-\frac{n}{72}\right). \tag{51}$$

It clearly follows from the first inequality in (80) that for every $j = 1,\ldots,n$,

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

The above inequality, (44), and (46) imply that

$$\|w^j - \hat{z}\| \leq \sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}, \quad D_h(w^j,\hat{z}) \leq \tau, \quad \|z^j - \hat{z}\| \leq \sqrt{\frac{(2+\lambda\mu)\lambda\tau}{1+\lambda\mu}}, \quad \forall j \in \mathcal{J}_0. \tag{52}$$

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 (52) hold for more than $2/3$ indices $j \in \{1,\ldots,n\}$. Now, condition (84) of Lemma B.1 is satisfied. Applying Lemma B.1 to Step 1 of Algorithm 3 and using the first inequality in (52), we have

$$\|w^j - \hat{z}\| \overset{(52),(85)}{\leq} 3\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}, \quad \forall j \in \mathcal{J}_1. \tag{53}$$

Applying Lemma B.1 to Step 2 of Algorithm 3 and using the third inequality in (52), we also have

$$\|z^j - \hat{z}\| \overset{(52),(85)}{\leq} 3\sqrt{\frac{(2+\lambda\mu)\lambda\tau}{1+\lambda\mu}}, \quad \forall j \in \mathcal{J}_2. \tag{54}$$

Next, we analyze Step 3 of Algorithm 3. Condition (49) satisfies the assumption $q \geq 4\sigma^2/\delta^2$ in Lemma C.1 with

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

Hence, applying Lemma C.1 to Step 3 of Algorithm 3 and noting that $\nabla f(\tilde{w}) = \mathbb{E}[\bar{s}(\tilde{w})]$, we have

$$\mathbb{P}\left(E_2 \,|\, E_1\right) \overset{(86)}{\geq} 1 - \exp\left(-\frac{n}{72}\right), \tag{55}$$

where event E_2 is defined as

$$E_2 := \left\{ \|\bar{s}(\tilde{w}) - \nabla f(\tilde{w})\| \leq L\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}} \right\}. \tag{56}$$

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, (53) with $j = j_0$, (56), and the fact that f is L-smooth, we have

$$\|\tilde{w} - \hat{z}\| \overset{(53)}{\leq} 3\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}, \tag{57}$$

and

$$\begin{aligned} \|\bar{s}(\tilde{w}) - \nabla f(\hat{z})\| &\leq \|\bar{s}(\tilde{w}) - \nabla f(\tilde{w})\| + \|\nabla f(\tilde{w}) - \nabla f(\hat{z})\| \\ &\overset{(56)}{\leq} L\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}} + L\|\tilde{w} - \hat{z}\| \overset{(57)}{\leq} 4L\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}. \end{aligned} \tag{58}$$

Using Lemma A.2, (57), and (58), we have

$$d_h(w^j, \hat{z}) \overset{(78)}{\leq} D_h(w^j, \hat{z}) + 4\left(L + \frac{1}{\lambda}\right)\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}\|w^j - \hat{z}\|, \quad \forall j \in \{1, \ldots, n\}.$$

Hence, it follows from (52) that

$$d_h(w^j, \hat{z}) \leq \tau + \frac{8(1+\lambda L)\tau}{1+\lambda\mu}, \quad \forall j \in \mathcal{J}_0. \tag{59}$$

Now, we are ready to analyze Step 4 of Algorithm 3. Note that (59) holds for more than 2/3 indices $j \in \{1, \ldots, n\}$, which means condition (84) of Lemma B.1 is satisfied. Applying Lemma B.1 to Step 4 of Algorithm 3 and using (59), we have

$$d_h(w^j, \hat{z}) \overset{(59),(85)}{\leq} 3\tau + \frac{24(1+\lambda L)\tau}{1+\lambda\mu}, \quad \forall j \in \mathcal{J}_3. \tag{60}$$

To this end, we are in a position to prove (50)*.* Using Lemma A.2, (57), and (58), we have

$$D_h(w^j, \hat{z}) \overset{(79)}{\leq} d_h(w^j, \hat{z}) + 4\left(L + \frac{1}{\lambda}\right)\sqrt{\frac{2\lambda\tau}{1+\lambda\mu}}\|w^j - \hat{z}\|, \quad \forall j \in \{1, \ldots, n\}. \tag{61}$$

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 (61), (60), and (53) 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 (80), the above inequality, (53), and (54), we conclude

$$\begin{aligned}
\phi^\lambda(w^j) - \phi^\lambda(\hat{z}) + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|z^j - \hat{z}\|^2 &\overset{(80)}{\leq} \frac{1+\lambda L}{2\lambda}\|w^j - \hat{z}\|^2 + D_h(w^j, \hat{z}) + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|z^j - \hat{z}\|^2 \\
&\leq \frac{1+\lambda L}{1+\lambda\mu}9\tau + \left(3\tau + \frac{48(1+\lambda L)\tau}{1+\lambda\mu}\right) + 9\tau \\
&\leq 9\kappa\tau + 3\tau + 48\kappa\tau + 9\tau,
\end{aligned}$$

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, (50) follows. Finally, combining (51) and (55), we complete the proof

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

■

6 High probability result and low sample complexity

This section is devoted to the complete analysis of Algorithm 1 to establish both the high probability result (Theorem 6.2) and the sample complexity bound (Theorem 6.3) of Algorithm 1, which are the main results of the paper.

The following proposition states a high probability result for running one iteration of Algorithm 1. It provides a crucial recursive formula in proving the subsequent theorems.

Proposition 6.1. *Assuming that* (22) *and* (49) *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*

$$\phi(\bar{w}_k) - \phi_* - \frac{1}{2\lambda}\|x_* - \bar{z}_{k-1}\|^2 + \frac{1+\lambda\mu}{\lambda(4+\lambda\mu)}\|x_* - \bar{z}_k\|^2 \leq 12\tau + 57\kappa\tau, \tag{62}$$

where x_* is the solution to (1) *and τ is as in* (47)*.*

Proof: We restate Proposition 5.1 using the notation in Algorithm 1 as: if (22) and (49) 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

$$\phi(\bar{w}_k) + \frac{1}{2\lambda}\|\bar{w}_k - \bar{z}_{k-1}\|^2 - \phi(\hat{z}_k) - \frac{1}{2\lambda}\|\hat{z}_k - \bar{z}_{k-1}\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\hat{z}_k - \bar{z}_k\|^2 \overset{(50)}{\leq} \delta, \tag{63}$$

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 (63) then imply that with probability at least $1 - 2\exp(-n/72)$,

$$\phi(\bar{w}_k) - \left(\phi_* + \frac{1}{2\lambda}\|x_* - \bar{z}_{k-1}\|^2\right) + \frac{1+\lambda\mu}{2\lambda}\|x_* - \hat{z}_k\|^2 + \frac{1+\lambda\mu}{\lambda(2+\lambda\mu)}\|\bar{z}_k - \hat{z}_k\|^2 \leq \delta. \tag{64}$$

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 (64), we conclude that (62) holds with probability at least $1 - 2\exp(-n/72)$. ■

We are now ready to present the main result on the high probability guarantee of Algorithm 1.

Theorem 6.2. *Assume that $\lambda\mu \geq 3$, (22), and (49) 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*

$$K = \mathcal{O}\left(\log\frac{1}{\varepsilon}\right), \quad I = \mathcal{O}\left(\max\left\{\kappa\log\frac{\kappa}{\varepsilon}, \frac{\kappa\sigma^2}{\mu\varepsilon}\right\}\right), \quad n = \mathcal{O}\left(\log\frac{1}{p}\right), \tag{65}$$

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

$$\mathbb{P}\left(\min_{1\leq k\leq K}\phi(\bar{w}_k) - \phi_* \leq \varepsilon\right) \geq 1 - p. \tag{66}$$

Proof: Define

$$a_k = \phi(\bar{w}_k) - \phi_*, \quad b_k = \frac{1}{2\lambda}\|x_* - \bar{z}_k\|^2, \quad \theta = \frac{2+2\lambda\mu}{4+\lambda\mu}, \quad \delta = 12\tau + 57\kappa\tau. \tag{67}$$

Then, using Proposition 6.1, we have with probability at least $1 - 2\exp(-n/72)$,

$$a_k \overset{(62)}{\leq} b_{k-1} - \theta b_k + \delta. \tag{68}$$

Multiplying (68) by θ^{k-1} and summing the resulting inequality from $k = 1$ to K, we have

$$\sum_{k=1}^{K} \theta^{k-1} a_k \le \sum_{k=1}^{K} \theta^{k-1}(b_{k-1} - \theta b_k + \delta) = b_0 - \theta^K b_K + \sum_{k=1}^{K} \theta^{k-1}\delta, \tag{69}$$

with probability at least

$$\left(1 - 2\exp\left(-\frac{n}{72}\right)\right)^K \ge 1 - 2K\exp\left(-\frac{n}{72}\right). \tag{70}$$

Dividing (69) by $\sum_{k=1}^{K} \theta^{k-1}$ and using (67), we have

$$\min_{1\le k\le K} \phi(\bar{w}_k) - \phi_* = \min_{1\le k\le K} a_k \le \frac{b_0}{\sum_{k=1}^{K} \theta^{k-1}} + \delta \le \frac{\|\bar{z}_0 - x_*\|^2}{2\lambda \sum_{k=1}^{K} \theta^{k-1}} + 12\tau + 57\kappa\tau, \tag{71}$$

with probability as in (70). It follows from the assumption that $\lambda\mu \ge 3$ and the definition of θ in (67) that $\theta \ge 8/7$, which together with (71), implies that

$$\min_{1\le k\le K} \phi(\bar{w}_k) - \phi_* \le \frac{\|\bar{z}_0 - x_*\|^2}{14\lambda\left[\left(\frac{8}{7}\right)^K - 1\right]} + 12\tau + 57\kappa\tau, \tag{72}$$

with probability as in (70). It is clear that if the first condition in (65) holds, i.e.,

$$K = \mathcal{O}\left(\log\frac{1}{\varepsilon}\right), \tag{73}$$

then the first term on the right-hand side of (72) satisfies

$$\frac{\|\bar{z}_0 - x_*\|^2}{14\lambda\left[\left(\frac{8}{7}\right)^K - 1\right]} = \mathcal{O}(\varepsilon). \tag{74}$$

Since the last term $57\kappa\tau$ in (72) dominates the second term 12τ, in view of (47), we only need to derive a bound on I such that

$$\alpha^I\left(\sigma D + \frac{LD^2}{2}\right) + \frac{\lambda\sigma^2}{I} \le \frac{\varepsilon}{\kappa}. \tag{75}$$

Using the above inequality and the fact that $\alpha \le e^{\alpha-1}$, 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 (22) and $\lambda\mu \ge 3$ that

$$I = \mathcal{O}\left(\max\left\{\kappa\log\frac{\kappa}{\varepsilon}, \frac{\kappa\sigma^2}{\mu\varepsilon}\right\}\right), \tag{76}$$

which is the second condition in (65). Now, assuming (73) and (76) hold, then we know (74) and (75) should also hold. Putting (72), (74), and (75) together, we have

$$\mathbb{P}\left(\min_{1\le k\le K} \phi(\bar{w}_k) - \phi_* \le \varepsilon\right) \ge 1 - 2K\exp\left(-\frac{n}{72}\right) \overset{(73)}{\approx} 1 - \exp\left(-\frac{n}{72}\right).$$

Finally, the above inequality and the last condition in (65) imply that (66) holds. ■

We are now ready to present the main result on the low sample complexity of Algorithm 1.

Theorem 6.3. *For given $\varepsilon > 0$ and $p \in (0,1)$, to find a solution $\bar{w} \in \operatorname{dom} h$ by Algorithm 1 such that*

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

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

$$\mathcal{O}\left(\max\left\{\kappa \log \frac{\kappa}{\varepsilon}, \frac{\kappa\sigma^2}{\mu\varepsilon}\right\} \log \frac{1}{p} \log \frac{1}{\varepsilon}\right). \tag{77}$$

Proof: It is clear that Algorithm 1 requires nK PSS oracles (Algorithm 2) and K PB oracles (Algorithm 3). Using Theorem 6.2, 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 (47), (49), and the fact that $\lambda\mu = \mathcal{O}(1)$, we have

$$q \overset{(49)}{=} \mathcal{O}\left(\frac{\sigma^2}{\lambda L^2 \tau}\right) \overset{(47)}{=} \mathcal{O}\left(\frac{I}{\lambda^2 L^2}\right) = \mathcal{O}\left(\frac{I}{\kappa^2}\right).$$

Therefore, the total sample complexity is $\mathcal{O}(nKI)$, which is (77) in view of (65). ∎

7 Conclusions

In this paper, we study the stochastic convex composite optimization problem (1) and aim to establish high probability guarantees in a small primal gap, namely $\mathbb{P}(\phi(x) - \phi_* \le \varepsilon) \ge 1 - p$ for given $\varepsilon > 0$ and $p \in (0,1)$. To establish $\mathcal{O}(\log(1/p))$ dependence on p in the sample complexity, typical results in the literature often need stringent assumptions such as sub-Gaussian noise distributions. Our contribution is the introduction of an SA-type method based on PPM, named SPPM, which ensures a high probability guarantee with low sample complexity, assuming only bounded variance in the stochastic gradient. Furthermore, we introduce a subroutine PSS within SPPM, serving not only to solve the proximal subproblem (2) but also as a novel variance reduction technique.

We conclude by exploring potential extensions of our analysis presented in this paper. First, it is worth noting that while our sample complexity demonstrates a logarithmic dependence on $1/p$, SPPM incurs an overhead cost proportional to the condition number κ. A feasible approach to reduce this dependence on κ is to develop an accelerated variant of SPPM, which is expected to reduce overhead cost to $\sqrt{\kappa}$, paralleling the efficiency improvement typically seen in accelerated gradient methods. Traditionally, there is a belief that acceleration methods amplify the noise level. However, we aim to investigate whether a restarted acceleration method, viewed through the lens of PPM as we have done with PSS in this paper, could in fact reduce the variance. Second, while SPPM does not require any problem parameters (such as μ and L) as input, Theorems 6.2 and 6.3 do presuppose certain assumptions on its input. Our ultimate objective is to develop a realistic parameter-free method by removing or at least weakening those assumptions. Finally, in our pursuit of a parameter-free method, we are also interested in devising a variant of SPPM that incorporates a variable prox stepsize λ_k. This adaptation aims to fully exploit the potential and versatility of PPM.

References

[1] J.-F. Aujol, L. Calatroni, C. Dossal, H. Labarrière, and A. Rondepierre. Parameter-free FISTA by adaptive restart and backtracking. *Available at arXiv:2307.14323*, 2023.

[2] 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.

[3] 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.

[4] 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.

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

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

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

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

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

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

[11] A. Nemirovskij and D. Yudin. Problem complexity and method efficiency in optimization. 1983.

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

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

A Technical results

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

Lemma A.1. *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 A.2. *Recall $D_h(\cdot, \hat{z})$ and $d_h(\cdot, \hat{z})$ given in (44) and (45), respectively. Then, for every $x \in \operatorname{dom} h$, we have*

$$d_h(x, \hat{z}) \leq D_h(x, \hat{z}) + \left(\|\bar{s}(\tilde{w}) - \nabla f(\hat{z})\| + \frac{1}{\lambda}\|\tilde{w} - \hat{z}\| \right) \|x - \hat{z}\|, \tag{78}$$

$$D_h(x, \hat{z}) \leq d_h(x, \hat{z}) + \left(\|\bar{s}(\tilde{w}) - \nabla f(\hat{z})\| + \frac{1}{\lambda}\|\tilde{w} - \hat{z}\| \right) \|x - \hat{z}\|, \tag{79}$$

where $\hat{z}$ is as in (42), 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 (45), the above relations, and the triangle inequality, we have for every $x \in \operatorname{dom} h$,

$$\begin{aligned} d_h(x, \hat{z}) &\overset{(45)}{=} |h(x) - h(\hat{z}) + \langle \nabla f^\lambda(\hat{z}), x - \hat{z}\rangle + \langle \bar{s}^\lambda(\tilde{w}) - \nabla f^\lambda(\hat{z}), x - \hat{z}\rangle| \\ &\leq |h(x) - h(\hat{z}) + \langle \nabla f^\lambda(\hat{z}), x - \hat{z}\rangle| + |\langle \bar{s}(\tilde{w}) - \nabla f(\hat{z}), x - \hat{z}\rangle| + \frac{1}{\lambda}|\langle \tilde{w} - \hat{z}, x - \hat{z}\rangle|. \end{aligned}$$

It follows from the definition of D_h in (44) and the Cauchy-Schwarz inequality that (78) holds. The proof of (79) follows similarly. ■

Lemma A.3. *Recall $D_h(\cdot, \hat{z})$ and $\phi^\lambda(\cdot)$ given in (44) and (43), respectively. Then, for every $x \in \operatorname{dom} h$, we have*

$$\frac{1 + \lambda\mu}{2\lambda}\|x - \hat{z}\|^2 + D_h(x, \hat{z}) \leq \phi^\lambda(x) - \phi^\lambda(\hat{z}) \leq \frac{1 + \lambda L}{2\lambda}\|x - \hat{z}\|^2 + D_h(x, \hat{z}). \tag{80}$$

Proof: It follows from (4) and the definition of f^λ in (43) 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 A.4. *Consider a sequence of pairwise 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*

$$\mathbb{P}\left(|\mathcal{J}| > \frac{2n}{3}\right) \geq 1 - \exp\left(-\frac{n}{72}\right). \tag{81}$$

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 $t > 0$,

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] \leq -t\right) \leq \exp\left(\frac{-2t^2}{\sum_{j=1}^n (b_j - a_j)^2}\right). \tag{82}$$

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 pairwise independent. For every $j = 1, \ldots, n$, we have $a_j = 0$, $b_j = 1$, and

$$\mathbb{E}[X_j] = \mathbb{P}(X_j = 1) = \mathbb{P}(A_j \text{ occurs}) \geq \frac{3}{4}. \tag{83}$$

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

$$\mathbb{P}\left(S_n \geq \frac{2n}{3}\right) \overset{(83)}{\geq} \mathbb{P}\left(S_n - \mathbb{E}[S_n] \geq -\frac{n}{12}\right) \overset{(82)}{\geq} 1 - \exp\left(-\frac{n}{12}\right).$$

Therefore, (81) immediately follows. ∎

B Second tertile selection

This section presents and analyzes the subroutine STS used in Algorithm 3.

Algorithm 4 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 A.4 to boost the probability from low to high confidence, impacting the overall quality by at most a constant factor.

Lemma B.1. *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*

$$|B_{\epsilon,d}(\tilde{z}) \cap Z| > \frac{2n}{3} \tag{84}$$

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

$$d(z^j, \tilde{z}) \le 3\epsilon, \quad \forall j \in \mathcal{J}. \tag{85}$$

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

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

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 $\rho_j \le 2\epsilon$ (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 $\bar{\rho} \le 2\epsilon$ (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

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

Since the above inequality holds for any $j \in \mathcal{J}$, (85) immediately follows. ■

C Robust gradient estimation

This section presents and analyzes the subroutine RGE used in Algorithm 3.

Algorithm 5 RGE(x, n, q)

Input: A point $x \in \operatorname{dom} h$ and integers $n, q \ge 1$.
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 $j^* \in \mathcal{J}$.

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

$$\mathbb{P}\left(\|\bar{s}_{j^*}(x) - \mathbb{E}[\bar{s}_{j^*}(x)]\| \le 3\delta\right) \ge 1 - \exp\left(-\frac{n}{72}\right). \tag{86}$$

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,

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

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 A.4 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 (84) in Lemma B.1 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 B.1 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, (86) holds for any $j^* \in \mathcal{J}$. ∎