Some Unified Theory for Variance Reduced Prox-Linear Methods

Yue Wu[*] Benjamin Grimmer[†]

Abstract

This work considers the nonconvex, nonsmooth problem of minimizing a composite objective of the form $f(g(x))+h(x)$ where the inner mapping g is a smooth finite summation or expectation amenable to variance reduction. In such settings, prox-linear methods can enjoy variance-reduced speed-ups despite the existence of nonsmoothness. We provide a unified convergence theory applicable to a wide range of common variance-reduced vector and Jacobian constructions. Our theory (i) only requires operator norm bounds on Jacobians (whereas prior works used potentially much larger Frobenius norms), (ii) provides state-of-the-art high probability guarantees, and (iii) allows inexactness in proximal computations.

1 Introduction

In this work, we consider nonsmooth, nonconvex problems

$$\min_{x\in\mathbb{R}^n} \Phi(x) := f(g(x)) + h(x) \tag{1.1}$$

where $f\colon \mathbb{R}^m \to \mathbb{R}$ and $h\colon \mathbb{R}^n \to \mathbb{R}$ are convex functions and $g\colon \mathbb{R}^n \to \mathbb{R}^m$ is a differentiable mapping. Note that although f is convex and g is smooth, their composition may be neither convex nor smooth. This "convex-composite" optimization model is surprisingly versatile. As two classic example applications,

- **Nonlinear Programming.** Consider minimizing an objective function $h(x)$ subject to functional constraints $g^{(l)}(x) \leq 0$ for $l = 1\dots m$. Then letting $g(x) = [g^{(1)}(x), \dots, g^{(m)}(x)]$ and $f(z)$ be either an indicator function for the nonpositive orthant or an exact penalty $f(z) = \sum_{i=1}^m C\max\{z_i, 0\}$ for sufficiently large C, any such nonlinear program can be cast in the form (1.1). Of particular interest here are settings where each constraint $g^{(l)}(x)$ takes the form of a summation $\frac{1}{N}\sum_{j=1}^N g_j^{(l)}(x)$ as occurs across machine learning tasks.
- **Nonlinear Equation Solving/Regression.** Consider a solving system of equations $0 = g(x) := \mathbb{E}_{\xi\sim D} g_\xi(x)$, given only oracles for sampling from D and first-order queries about individual samples $g_\xi(x)$. If one measures solution quality in some norm $f(z) = \|z\|$, minimizing solution error takes the form (1.1). Any additional regularization can be additional modeled by $h(x)$, for example, setting $h(x) = \|x\|_1$.

We focus on reducing the number of first-order queries needed to elements of the finite summations g_i or expectations g_ξ as occur above. Our approach is based on leveraging two well-studied tools in first-order optimization, discussed briefly below: *variance reduction* and *prox-linear methods*. This

[*] Johns Hopkins University, Department of Applied Mathematics and Statistics, ywu166@jhu.edu
[†] Johns Hopkins University, Department of Applied Mathematics and Statistics, grimmer@jhu.edu

combination was recently considered by Zhang and Xiao [1] and Tran-Dinh et al. [2], motivating our work.

Variance Reduction. Throughout, we assume $g\colon \mathbb{R}^n \to \mathbb{R}^m$ is either a finite sum

$$g(x) = \frac{1}{N}\sum_{j=1}^{N} g_j(x) \tag{1.2}$$

or, more generally, an expectation

$$g(x) = \mathbb{E}_{\xi\sim D}[g_\xi(x)], \tag{1.3}$$

and that oracles for evaluating components $g_\xi(\cdot)$ and their Jacobian's $g'_\xi(\cdot)$ are given. Given samples $\xi \sim D$, these oracle evaluations provide unbiased estimates of $g(\cdot)$ and $g'(\cdot)$. Variance reduction techniques enable the construction of lower variance estimators where a high accuracy (large batch) estimate only needs to be computed every τ iterations. For the most classic style of update, due to [1, 3], every τ steps would use estimators of the form

$$\begin{cases} \widetilde{g}_0 = \frac{1}{A}\sum_{\xi\in\mathcal{A}_0} g_\xi(x_0) \\ \widetilde{g}_i = \frac{1}{a}\sum_{\xi\in\mathcal{A}_i}\Big(g_\xi(x_i) - g_\xi(x_0)\Big) + \widetilde{g}_0 \qquad \forall i = 1,\dots,\tau-1 \end{cases} \tag{1.4}$$

where the batches $\mathcal{A}_i$ can be much smaller than $\mathcal{A}_0$. When g is given by a finite summation (1.2), $\widetilde{g}_0$ could be computed exactly. At the cost of additional Jacobian evaluations $g'_\xi(\cdot)$, further refined schemes have been considered [1, 4]

$$\widetilde{g}_i = \frac{1}{a}\sum_{\xi\in\mathcal{A}_i}\Big(g_\xi(x_i) - g_\xi(x_0) - g'_\xi(x_0)(x_i - x_0)\Big) + \widetilde{g}_0 + \widetilde{J}_0(x_i - x_0) \tag{1.5}$$

where $\widetilde{J}_0$ is an unbiased estimate of $g'(x_0)$. Methods specifically targeting root-finding were recently given by [5] and generalizing to allow relative smoothness by [6]. See the survey [7] for more historical context.

Prox-linear Methods. Note a fundamental difficulty in (1.1) is that the composition of a convex function f with a smooth function g may be nonconvex. In contrast, the composition of a convex function with a linear function always remains convex. This motivates replacing $g(\cdot)$ by its linearization $g(x_k) + g'(x_k)(\cdot - x_k)$. Repeatedly minimizing this relaxed convex problem, with an added proximal term, is known as the "prox-linear method" [8–13]

$$x_+ = \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{f\left(g(x) + g'(x)(y-x)\right) + h(y) + \frac{M}{2}\|y-x\|_2^2\right\} \tag{1.6}$$

given some proximal parameter $M > 0$. If the above argmin is only computed approximately, perhaps via some first-order method for convex optimization using (sub)gradients of f, we refer to this as an "inexact prox-linear method".

This prox-linear step provides a generalized notion of stationarity for composite nonsmooth, nonconvex problems. Denote the generalized gradient at some x by

$$\mathcal{G}_M(x) := M(x - x_+) \in \partial\left(f(g(x) + g'(x)(\cdot - x)) + h\right)(x_+) \tag{1.7}$$

where x_+ is defined as the exact prox-linear step (1.6). The optimality condition defining x_+ in (1.6) ensures $\mathcal{G}_M(x) \in \partial\left(f(g(x) + g'(x)(\cdot - x)) + h\right)(x_+)$. By the sum and chain rules of subdifferential calculus, there exists $\lambda \in \partial f(g(x) + g'(x)(x_+ - x))$ and $\zeta \in \partial h(x_+)$ such that $\mathcal{G}_M(x) = g'(x)\lambda + \zeta$. Hence if $\|\mathcal{G}_M(x)\| \le \epsilon$, then together λ, $g'(x)$, and ζ provide a small subgradient certifying stationarity where each of these differential objects is taken at points near x. See the survey [14] for more historical context on prox-linear methods and similar approximate notions of stationarity.

Our Contributions. We analyze variance-reduced, prox-linear methods, iterating

$$\begin{cases} \widetilde{g}_k, \widetilde{J}_k & \leftarrow \text{Variance Reduced Estimates of } g(x_k), g'(x_k) \\ x_{k+1} & \leftarrow \text{Approximate Minimizer of } f\left(\widetilde{g}_k + \widetilde{J}_k(x - x_k)\right) + h(x) + \frac{M}{2}\|x - x_k\|_2^2. \end{cases} \tag{1.8}$$

We provide a unified theory for the oracle complexity with respect to evaluations of the vector $g_\xi(\cdot)$ and its Jacobian $g'_\xi(\cdot)$, for a range of variance-reduced approaches to constructing $\widetilde{g}_k$ and $\widetilde{J}_k$. Our main theorem (Theorem 3.1) offers three main advances:

- **Operator Norm Assumptions.** Our theory only relies on uniform bounds on the variation of Jacobians in operator norm of the form

$$\|g'_\xi(x) - g'_\xi(y)\|_{\mathsf{op}} \leq L_g\|x - y\|_2, \qquad \text{and} \qquad \|g'_\xi(x) - g'(x)\|_{\mathsf{op}} \leq \sigma_{g'}. \tag{1.9}$$

  Prior works have instead used the Frobenius norm (see related work discussion below). As a result, the "constants" in prior works may be up to a dimension-dependent factor of $\sqrt{\min\{n, m\}}$ times larger than those considered here.

- **A Pareto Frontier of State-of-the-Art Guarantees.** Our theory provides guarantees that various prox-linear methods produce a (ϵ, Δ)-h.p. stationary point, meaning with probability $1 - \Delta$, some x_k has $\|\mathcal{G}_M(x_k)\|_2^2 \leq \epsilon$. Depending on the relative cost of evaluating g_ξ and g'_ξ evaluations in (1.3) or the relative size of $1/\epsilon$ and N in (1.2), the best-known method varies. See the many state-of-the-art corollaries in Section 3.1.

- **Accounting of Inexact Proximal Computations.** Our theory allows for inexact prox-linear steps. Section 3.2 provides guarantees including the cost of subroutines. For example, guarantees follow for doubly stochastic problems where f is also defined as an expectation, requiring inexact minimization.

Outline. The remainder of this section discusses related work. Section 2 provides preliminaries and introduces the general algorithm considered. Section 3 states our unified convergence theorem and applies it to produce state-of-the-art guarantees for several variance-reduction schemes. Finally, Section 4 provides our technical analysis.

1.1 Related Work

The setting (1.1) was recently addressed by two works [1, 2]. A key insight was their identification that prox-linear methods can benefit from variance reduction despite the existence of nonsmoothness. Although both of these prior works are motivated by bounds on operator norms of Jacobians, their proof techniques relied on uniformly bounding Jacobian matrices in the Frobenius norm[1]. Our

[1] Both prior works [1, 2] rely on matrix generalizations of mean-squared error bounding lemmas typical to the analysis of methods with stochastic gradient vectors (see [15, Lemma 1] and [16, Lemma 2] for the essential vector arguments being generalized). At their core, such lemmas rely on a classic bias-variance decomposition: given a space $\mathcal{E}$ with inner product $\langle \cdot, \cdot \rangle$, a random variable $X_\xi \in \mathcal{E}$ and some fixed $Y \in \mathcal{E}$, one has

$$\mathbb{E}_\xi\|X_\xi - Y\|_{\langle \cdot,\cdot \rangle}^2 = \|\mathbb{E}_\xi[X_\xi] - Y\|_{\langle \cdot,\cdot \rangle}^2 + \mathbb{E}_\xi\|X_\xi - \mathbb{E}_{\xi'}[X_{\xi'}]\|_{\langle \cdot,\cdot \rangle}^2$$

where $\|\cdot\|_{\langle \cdot,\cdot \rangle}$ denotes the norm associated with the given inner product. In the space of matrices, one could apply this reasoning with the trace inner product to relate Frobenius norms. However, such relationships do not hold for norms without an associated inner product (e.g., matrix operator norms), and so prior works, even if not denoted, require the potentially larger Frobenius norm.

analysis relies only on operator norm bounds, closing this theoretical gap and offering improvements by dimension-dependent factors. To make formal comparisons, denote our "constants" from (1.9) as $(L_{g,\mathtt{op}}, \sigma_{g',\mathtt{op}})$ and their parallels using the Frobenius norm as by $(L_{g,\mathtt{Frob}}, \sigma_{g',\mathtt{Frob}})$. Note $L_{g,\mathtt{op}} \leq L_{g,\mathtt{Frob}}$ and $\sigma_{g',\mathtt{op}} \leq \sigma_{g',\mathtt{Frob}}$.

When g is given by a finite summation (1.2), Corollaries 3.13-3.17 show stationary points can be reached with high probability using at most $O(N + N^{4/5}\frac{L_{g,\mathtt{op}}}{\epsilon})$ evaluations of g_j and g_j', improving prior expectation guarantees of $O(N + N^{4/5}\frac{L_{g,\mathtt{Frob}}}{\epsilon})$.

When g is given by an expectation (1.3), prior works assuming stronger Frobenius norm bounds proved $O(\sigma^2_{g',\mathtt{Frob}}/\epsilon^{3/2})$ evaluations of $g_\xi'(x)$ suffice to reach expected stationarity. Corollaries 3.6-3.10 of our unified, operator norm-based, variance-reduced theory achieve high probability stationarity guarantees of $O(\sigma^2_{g',\mathtt{op}}/\epsilon^{5/3})$. For example, this yields an improvement whenever $\frac{\sigma_{g',\mathtt{Frob}}}{\sigma_{g',\mathtt{op}}} \geq 1/\epsilon^{1/12}$.

2 Preliminaries

First, we briefly summarize our basic notations. Let $O(\cdot)$ and $\Theta(\cdot)$ denote their standard asymptotic notations, both w.r.t $\epsilon \to 0$ and $N \to \infty$. In addition, we use $\widetilde{\Theta}$ instead of Θ to omit the multiplicative logarithmic terms in ϵ. For any distribution D, we denote its support by $\mathrm{supp}(D)$. Throughout, $\|\cdot\|_2$ is the 2-norm on Euclidean space and $\|\cdot\|_{\mathrm{op}}$ is the spectral norm of a matrix. We use several notions of Lipschitz continuity: A vector-valued function $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is l-Lipschitz if $\|\varphi(x) - \varphi(y)\|_2 \leq l\|x - y\|_2$ for any $x, y \in \mathbb{R}^n$, a matrix-valued function $\varphi : \mathbb{R}^n \to \mathbb{R}^{m_1 \times m_2}$ is L-Lipschitz if $\|\varphi(x) - \varphi(y)\|_{\mathrm{op}} \leq L\|x - y\|_2$ for any $x, y \in \mathbb{R}^n$. For a convex function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, a vector $v \in \mathbb{R}^n$ is a subgradient of φ at $x_0 \in \mathbb{R}^n$ if $\varphi(x) \geq \varphi(x_0) + \langle v, x - x_0 \rangle$ for all $x \in \mathbb{R}^n$. The subdifferential of φ at x_0, defined as the set of all subgradients of φ at x_0, is denoted by $\partial\varphi(x_0)$. For $M \geq 0$, a function $\varphi(x)$ is M-strongly convex if $\varphi(x) - \frac{M}{2}\|x\|_2^2$ is convex.

Throughout, we assume the following conditions hold for f, g, h defining (1.1):

1. The function $f : \mathbb{R}^m \to \mathbb{R}$ is convex and l_f-Lipschitz.
2. For any $\xi \in \mathrm{supp}(D)$, the function $g_\xi : \mathbb{R}^n \to \mathbb{R}^m$ is $l_{g,\xi}$-Lipschitz, and its Jacobian $g_\xi' : \mathbb{R}^n \to \mathbb{R}^{m \times n}$ is $L_{g,\xi}$-Lipschitz. In addition, $l_g := \sup_{\xi \in \mathrm{supp}(D)} l_{g,\xi}$ and $L_g := \sup_{\xi \in \mathrm{supp}(D)} L_{g,\xi}$ are both finite.
3. The function $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, convex, and proper.

The second condition above is under the general expectation setting of (1.3). If g is in the special finite average form of (1.2), then it only requires the existence of finitely many Lipschitz constants $\{l_{g,j}, L_{g,j} : j = 1, ..., N\}$. The Lipschitz conditions of f, g_ξ, and g_ξ' give the following pair of facts.

Proposition 2.1. *The function g is l_g-Lipschitz and g' is L_g-Lipschitz.*

Proposition 2.2. *For any $x, y \in \mathbb{R}^n$,*

$$\Big|f(g(x)) - f\Big(g(y) + g'(y)(x - y)\Big)\Big| \leq \frac{l_f L_g}{2}\|x - y\|_2^2.$$

2.1 A General Variance Reduced Prox-Linear Method

Algorithm 1 presents the general method our unified theory covers. This method proceeds via two nested loops. As inputs, we require a total number of outer iterations to be run K and a number of iterations for each inner loop $\tau_0, ..., \tau_{K-1}$. A typical variance-reduced method may compute an exact or high-accuracy estimate of $g(x)$ and $g'(x)$ once per outer loop while using cheaper estimates at each inner loop.

As useful notations, let $\Sigma_\tau = \sum_{k=0}^{K-1} \tau_k$ denote the total number of iterations. For $K \in \mathbb{N}_+$ and $\boldsymbol{\tau} \in \mathbb{N}_+^K$, we define the index set $\mathcal{I}(K, \boldsymbol{\tau}) = \{(k,i) \in \mathbb{N}^2 : 0 \leq k \leq K-1, 0 \leq i \leq \tau_k - 1\}$. So $\mathcal{I}(K, \boldsymbol{\tau})$ corresponds to all the inner iterations in Algorithm 1. Algorithm 1 then proceeds following the general pattern of (1.8) with the (k,i)-th iteration consists of an estimation step and an optimization step, using a predefined `estimator` and `solver` as described below.

```
Algorithm 1: Generalized Variance Reduced, Inexact Prox-Linear Method
Input: Initialization $x_0^0$, $M > 0$, Iteration bounds $K$,$\boldsymbol{\tau} = (\tau_0, ..., \tau_{K-1})$, an estimation
       method estimator($x, i; \theta$), a solver solver($s, \bar{\epsilon}, \bar{\delta}$).
for $k = 0, ..., K-1$ do
    for $i = 0, ..., \tau_k - 1$ do
        Compute $\widetilde{g}_i^k$ and $\widetilde{J}_i^k$ using the predefined method, $(\widetilde{g}_i^k, \widetilde{J}_i^k) \leftarrow$ estimator($x_i^k, i; \theta$).
        Minimize $s_i^k(x) := f(\widetilde{g}_i^k + \widetilde{J}_i^k(x - x_i^k)) + h(x) + \frac{M}{2}\|x - x_i^k\|_2^2$ by the known solver, and
          get an inexact solution $x_{i+1}^k \leftarrow$ solver($s_i^k, \bar{\epsilon}, \bar{\delta}$).
    end for
    Set $x_0^{k+1} = x_{\tau_k}^k$.
end for
```

2.1.1 Estimation Step At each step (k,i), Algorithm 1 requires an estimator `estimator`$(x, i; \theta)$, treated for now as a black-box, which produces stochastic estimates of $g(x_i^k)$ and $g'(x_i^k)$, denoted $\widetilde{g}_i^k$ and $\widetilde{J}_i^k$. As examples, see the several estimators (Est$_0$)–(Est$_4$) in Section 3.1.

As indicated by our notation, the estimator `estimator`$(x, i; \theta)$ is allowed to depend on i but not k. For example, the most classic variance reduction [3] computes an exact (or high accuracy) estimates of $g(x_0^k)$ and $g'(x_0^k)$ when $i = 0$ and then leverage these past estimates to cheaply estimate $g(x_i^k)$ and $g'(x_i^k)$ when $i > 0$. This process is repeated at every outer iteration k. In particular, the estimators considered here will have a "memory" of the most recent x_0^k and potentially the component evaluations g_ξ and g'_ξ previously computed there. All additional parameters of `estimator` are captured by θ, taken from some space Θ. For example, if `estimator` is some mini-batch method, then θ contains the batch sizes used at each iteration.

For our guarantees to apply, we require abstract high probability bounds on the estimation errors $\|\widetilde{g}_i^k - g(x_i^k)\|_2$ and $\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}}$ that grow at most quadratically and linearly in $\|x_i^k - x_0^k\|_2$. This is natural since as $\|x_i^k - x_0^k\|_2$ grows, any variance reduction scheme leveraging a memory of x_0^k ought to incur larger errors. Any additional constraints on the selection of the parameters θ are captured by $\mathcal{C}(K, \boldsymbol{\tau}, \Delta)$.

Assumption 2.3 (Abstract bounds for estimation errors)**.** *For a fixed* `estimator`*, there exist five non-negative functions of* $(K, \boldsymbol{\tau}, \theta, \Delta)$*, denoted as* $\gamma_0, \gamma_1, \gamma_2, \lambda_0, \lambda_1$*, such that for any* $K \in \mathbb{N}_+$*,* $\boldsymbol{\tau} \in \mathbb{N}_+^K$ *and* $\Delta \in (0,1)$*, there exists a set* $\mathcal{C}(K, \boldsymbol{\tau}, \Delta) \subseteq \Theta$ *such that for any* $\theta \in \mathcal{C}$*, with probability at least* $1 - \Delta$*, the following two inequalities simultaneously hold for all* $(k,i) \in \mathcal{I}(K, \boldsymbol{\tau})$*:*

$$\|\widetilde{g}_i^k - g(x_i^k)\|_2 \leq \gamma_0(K, \boldsymbol{\tau}, \theta, \Delta) + \gamma_1(K, \boldsymbol{\tau}, \theta, \Delta)\|x_i^k - x_0^k\|_2 + \gamma_2(K, \boldsymbol{\tau}, \theta, \Delta)\|x_i^k - x_0^k\|_2^2,$$
$$\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}} \leq \lambda_0(K, \boldsymbol{\tau}, \theta, \Delta) + \lambda_1(K, \boldsymbol{\tau}, \theta, \Delta)\|x_i^k - x_0^k\|_2.$$

The functions $\{\gamma_l\}_{l=0}^2$ and $\{\lambda_l\}_{l=0}^1$ may also depend on some quantities like m, n and the Lipschitz constants l_f, l_g, L_g. Since these are all fixed constants, we omit them and only keep the algorithmic parameters $(K, \boldsymbol{\tau}, \theta, \Delta)$ in the arguments of the functions. Again Section 3.1 provides specific examples of `estimator` with explicit forms for the set $\mathcal{C}(K, \boldsymbol{\tau}, \Delta)$ and functions $\{\gamma_l\}_{l=0}^2$, $\{\lambda_l\}_{l=0}^1$.

2.1.2 Optimization Step In the optimization step, we need to (inexactly) solve the subproblem $\min s_i^k(x)$. Formally, we assume access to a known solver, $\texttt{solver}(s, \bar{\epsilon}, \bar{\delta})$, that returns an inexact solution x_{sol}. Algorithm 1 uses `solver` in black-box fashion, only requiring the following assumption:

Assumption 2.4. *For $\bar{\epsilon}, \bar{\delta} > 0$ and the problem $\min_x s(x)$, with probability at least $1-\bar{\delta}$, $\texttt{solver}(s, \bar{\epsilon}, \bar{\delta})$ returns an $\bar{\epsilon}$-optimal solution x_{sol}, i.e., $s(x_{sol}) \leq \inf_x s(x) + \bar{\epsilon}$.*

As possible instantiations of `solver`, we consider four example subroutines and provide bounds on the resulting total oracle complexities with respect to f in Section 3.2.

3 Main Results

Given any estimator and solver satisfying Assumptions 2.3 and 2.4, our main result provides a general set of conditions for algorithmic parameters which guarantees the production of an ϵ-stationary point with high probability.

Theorem 3.1. *Suppose Assumption 2.3 holds for* `estimator`*, and Assumption 2.4 holds for* `solver`*. Assume $\Phi^* := \inf_x \Phi(x) > -\infty$. Fix an $M > 5 l_f L_g$. For any $\Delta \in (0,1)$ and $\epsilon > 0$, with probability at least $1 - \Delta$, Algorithm 1's iterates satisfy:*

$$\frac{1}{\Sigma_\tau} \sum_{k=0}^{K-1} \sum_{i=0}^{\tau_k - 1} \|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon,$$

provided the parameters $K, \boldsymbol{\tau}, \theta, \bar{\epsilon}, \bar{\delta}$ satisfy[2]

$$\theta \in \mathcal{C}(K, \boldsymbol{\tau}, \Delta/2), \tag{3.1}$$

$$\bar{\delta} \leq \Delta/(2\Sigma_\tau), \tag{3.2}$$

$$\bar{\epsilon} \leq \epsilon/(5 \cdot 30M), \tag{3.3}$$

$$\Sigma_\tau \geq 5 \cdot 30M(\Phi(x_0^0) - \Phi^*)/\epsilon, \tag{3.4}$$

$$\gamma_0(K, \boldsymbol{\tau}, \theta, \Delta/2) \leq \epsilon/(5 \cdot 125 l_f M), \tag{3.5}$$

$$\lambda_0^2(K, \boldsymbol{\tau}, \theta, \Delta/2) \leq L_g \epsilon/(5 \cdot 95 l_f M), \tag{3.6}$$

$$(1 + \tau_{max})^2 \gamma_1^2(K, \boldsymbol{\tau}, \theta, \Delta/2) \leq L_g \epsilon/(5 \cdot 525 l_f M), \tag{3.7}$$

$$(1 + \tau_{max})^2 \gamma_2(K, \boldsymbol{\tau}, \theta, \Delta/2) \leq 3L_g/50, \tag{3.8}$$

$$(1 + \tau_{max})^2 \lambda_1^2(K, \boldsymbol{\tau}, \theta, \Delta/2) \leq 3L_g^2/38. \tag{3.9}$$

3.1 Convergence Rate Corollaries for a Range of VR Schemes

Next we apply this result to several estimation schemes, providing optimized algorithmic parameters (e.g., batch sizes, loop durations τ_k). These applications all amount to simple applications of concentration inequalities to establish a lemma ensuring Assumption 2.3 and then calculations based on Theorem 3.1 to provide optimized parameter selections and guarantees. Such sample calculations

[2] The $\tau_{\max}$ here denotes $\max\{\tau_0, ..., \tau_{K-1}\}$.

deriving Corollaries 3.6-3.7 are given in Section 4.2. As all remaining derivations of corollaries are effectively identical, they are omitted. An interested reader can find them in Appendix A.2.

The Mini-Batch Method. We first discuss a simple mini-batch method as a warm-up example. At the (k,i)-th iteration, we generate an index set $\mathcal{A}_i^k$ of size A and another index set $\mathcal{B}_i^k$ of size B, both by sampling from distribution D. Then we construct $\widetilde{g}_i^k$ and $\widetilde{J}_i^k$ using the sample mean over the index sets, parameterized by $\theta = (A,B) \in \mathbb{N}_+^2$. We can explicitly express this estimator for use in Algorithm 1 as

$$\texttt{estimator}_0(x_i^k, i; \theta) : \begin{cases} \widetilde{g}_i^k = \frac{1}{A}\sum_{\xi\in\mathcal{A}_i^k} g_\xi(x_i^k) \\ \widetilde{J}_i^k = \frac{1}{B}\sum_{\xi\in\mathcal{B}_i^k} g'_\xi(x_i^k). \end{cases} \tag{Est$_0$}$$

The construction above is for the expectation setting in (1.3). For the special finite average case in (1.2), it reduces to sampling with replacement from $\{1,...,N\}$, then $\widetilde{g}_i^k = \frac{1}{A}\sum_{j\in\mathcal{A}_i^k} g_j(x_i^k)$ and $\widetilde{J}_i^k = \frac{1}{B}\sum_{j\in\mathcal{B}_i^k} g'_j(x_i^k)$. To control the estimation error of $\widetilde{g}_i^k$ and $\widetilde{J}_i^k$, we need the following assumption.

Assumption 3.2. *There exist constants σ_g and $\sigma_{g'}$ such that for any $\xi \in \operatorname{supp}(D)$ and any $x \in \mathbb{R}^n$, $\|g_\xi(x) - g(x)\|_2 \le \sigma_g$ and $\|g'_\xi(x) - g'(x)\|_{\mathrm{op}} \le \sigma_{g'}$.*

Such uniform bounds suffice to ensure Assumption 2.3 holds for the above mini-batching estimator. The following lemma provides explicit values for the associated set $\mathcal{C}(K,\boldsymbol{\tau},\Delta)$, and functions $\{\gamma_l\}_{l=0}^2$, $\{\lambda_l\}_{l=0}^1$. This lemma follows as a consequence of standard concentration inequalities.

Lemma 3.3. *Suppose Assumption 3.2 holds. If $\texttt{estimator}(x_i^k, i; \theta)$ is defined by (Est$_0$), where $\theta = (A,B)$ and $\Theta = \mathbb{N}_+^2$, then Assumption 2.3 holds with the following choices of $\mathcal{C}(K,\boldsymbol{\tau},\Delta)$, $\{\gamma_l\}_{l=0}^2$ and $\{\lambda_l\}_{l=0}^1$:*

$$\mathcal{C}(K,\boldsymbol{\tau},\Delta) = \left\{(A,B)\in\mathbb{N}_+^2 : A \ge \frac{4}{9}\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right), \text{and } B \ge \frac{4}{9}\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)\right\},$$

$$\gamma_0(K,\boldsymbol{\tau},\theta,\Delta) = \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right)}, \quad \lambda_0(K,\boldsymbol{\tau},\theta,\Delta) = \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)},$$

$$\gamma_1 = \gamma_2 = \lambda_1 = 0.$$

Substituting the results of Lemma 3.3 into conditions (3.1)–(3.9), Theorem 3.1 provides constraints on the parameters $K,\boldsymbol{\tau},\theta$ which guarantee minibatching produces a stationary point with high probability. Furthermore, since $(K,\boldsymbol{\tau})$ controls the number of iterations in Algorithm 1 and $\theta = (A,B)$ determines the batch sizes at each evaluation, one can optimize their selection over this feasible region. Directly doing so, the following corollary provides such optimized choices.

Corollary 3.4. *Consider any $\Delta \in (0,1)$, $M > 5l_fL_g$, and any sufficiently small $\epsilon > 0$. Suppose Assumption 2.4 holds for $\texttt{solver}$, Assumption 3.2 holds for g, $\inf_x \Phi(x) > -\infty$, and $\texttt{estimator}$ is defined by (Est$_0$). Set $\Sigma_\tau = \lceil C_\Sigma\cdot\epsilon^{-1}\rceil$, $A = \lceil C_A\cdot\epsilon^{-2}\cdot\log(\frac{4(m+1)\Sigma_\tau}{\Delta})\rceil$, $B = \lceil C_B\cdot\epsilon^{-1}\cdot\log(\frac{4(m+n)\Sigma_\tau}{\Delta})\rceil$, $\overline{\delta} \le \Delta/(2\Sigma_\tau)$, $\overline{\epsilon} \le \epsilon/(5\cdot 30M)$, where C_Σ, C_A, C_B are some constants, then with probability at least $1-\Delta$: (i) Algorithm 1's iterates satisfy:*

$$\frac{1}{\Sigma_\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau_k-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \le \epsilon,$$

and (ii) the oracle complexities for evaluations and Jacobians of inner components $g_\xi(\cdot)$ respectively are at most

$$\widetilde{\Theta}\left(\epsilon^{-3}\log(1/\Delta)\right) \quad \textit{and} \quad \widetilde{\Theta}\left(\epsilon^{-2}\log(1/\Delta)\right).$$

Note the two complexities $\widetilde{\Theta}\left(\epsilon^{-3}\log(1/\Delta)\right)$ and $\widetilde{\Theta}\left(\epsilon^{-2}\log(1/\Delta)\right)$ match the high probability guarantees in [2]. Up to logarithm terms, our high probability results also agree with the expectation results of [1]. In both cases, our theory improves prior Frobenius norm bounds to matrix operator norms.

3.1.1 Expectation Case Methods Given g is defined as an expectation (1.3), we consider two different variance reduced schemes below, following the forms of (1.4) and (1.5). Our unified theorem's guarantees for the first scheme requires fewer evaluations of g'_ξ while the second requires fewer evaluations of g_ξ. As a result, both methods may be state-of-the-art depending on the relative cost of these two operations.

Application of Standard Variance Reduction for Expectations. First, we consider the variance-reduced estimator, defined in two cases, $i = 0$ and $i > 0$, as

$$\texttt{estimator}_1(x_i^k, i; \theta) : \begin{cases} \widetilde{g}_0^k = \frac{1}{A}\sum_{\xi\in\mathcal{A}_0^k} g_\xi(x_0^k) \\ \widetilde{J}_0^k = \frac{1}{B}\sum_{\xi\in\mathcal{B}_0^k} g'_\xi(x_0^k) \\ \widetilde{g}_i^k = \frac{1}{a}\sum_{\xi\in\mathcal{A}_i^k}\Big(g_\xi(x_i^k) - g_\xi(x_0^k)\Big) + \widetilde{g}_0^k \\ \widetilde{J}_i^k = \frac{1}{b}\sum_{\xi\in\mathcal{B}_i^k}\Big(g'_\xi(x_i^k) - g'_\xi(x_0^k)\Big) + \widetilde{J}_0^k. \end{cases} \tag{Est$_1$}$$

This estimator simply applies the classic variance reduced update (1.4) independently to estimate both g_ξ and g'_ξ: At the (k,i)-th iteration, we generate index sets $\mathcal{A}_i^k$ and $\mathcal{B}_i^k$ by sampling from distribution D. At the start of each epoch, namely $i = 0$, the batch sizes are set to be $|\mathcal{A}_0^k| = A$ and $|\mathcal{B}_0^k| = B$. We still use the sample mean to construct $\widetilde{g}_0^k$ and $\widetilde{J}_0^k$, same as the mini-batch method. In the case $i > 0$, we set $|\mathcal{A}_i^k| = a$ and $|\mathcal{B}_i^k| = b$, with $a < A$ and $b < B$. It is also worth noting that, unlike the mini-batch method, $\texttt{estimator}_1$ is history-dependent, since the construction of $\widetilde{g}_i^k$ and $\widetilde{J}_i^k$ involves the past iterate x_0^k.

In (Est$_1$), the parameter of $\texttt{estimator}_1$ is $\theta = (A, B, a, b) \in \mathbb{N}_+^4$, which captures the batch sizes. The set $\mathcal{C}(K, \boldsymbol{\tau}, \Delta)$ and functions $\{\gamma_l\}_{l=0}^2$, $\{\lambda_l\}_{l=0}^1$ are given below.

Lemma 3.5. *Suppose Assumption 3.2 holds. If* $\texttt{estimator}(x_i^k, i; \theta)$ *is defined by (Est$_1$), where* $\theta = (A, B, a, b)$ *and* $\Theta = \mathbb{N}_+^4$*, then Assumption 2.3 holds with*

$$\mathcal{C}(K, \boldsymbol{\tau}, \Delta) = \left\{(A, B, a, b) \in \mathbb{N}_+^4 : A, a \geq \frac{4}{9}\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right), \textit{and } B, b \geq \frac{4}{9}\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)\right\},$$

$$\gamma_0(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right)}, \quad \gamma_1(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{4l_g}{\sqrt{a}}\sqrt{\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right)}, \quad \gamma_2 = 0,$$

$$\lambda_0(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)}, \quad \lambda_1(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{4L_g}{\sqrt{b}}\sqrt{\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)}.$$

Using Lemma 3.5 enables us to select the parameters $K, \boldsymbol{\tau}, \theta$ in Theorem 3.1 and analyze the oracle complexities there yielding the following pair of results. Proofs of this lemma and both resulting corollaries are given in Section 4.2.

Corollary 3.6 (Algorithmic guarantee)**.** *Consider any* $\Delta \in (0, 1)$*,* $M > 5l_f L_g$*, integer* $\tau > 0$*, and any sufficiently small* $\epsilon > 0$*. Suppose Assumption 2.4 holds for* $\texttt{solver}$*, Assumption 3.2 holds for function* g*,* $\inf_x \Phi(x) > -\infty$*,* $\texttt{estimator}$ *is defined by (Est$_1$), and* $\boldsymbol{\tau}$ *is restricted to the form* $\tau_0 = \cdots = \tau_{K-1} = \tau$*. Set* $K = \lceil \frac{C_\Sigma\cdot\epsilon^{-1}}{\tau} \rceil$, $A = \lceil C_A \cdot \epsilon^{-2} \cdot \log(\frac{4(m+1)K\tau}{\Delta}) \rceil$, $B = \lceil C_B \cdot \epsilon^{-1} \cdot \log(\frac{4(m+n)K\tau}{\Delta}) \rceil$, $a = \lceil C_a \cdot (1+\tau)^2 \cdot \epsilon^{-1} \cdot \log(\frac{4(m+1)K\tau}{\Delta}) \rceil$, $b = \lceil C_b \cdot (1+\tau)^2 \cdot \log(\frac{4(m+n)K\tau}{\Delta}) \rceil$, $\overline{\delta} = \Delta/(2K\tau)$,

$\bar{\epsilon} = \epsilon/(5 \cdot 30M)$, where $C_\Sigma, C_A, C_B, C_a, C_b$ are some constants, then with probability at least $1 - \Delta$: (i) Algorithm 1's iterates satisfy:

$$\frac{1}{K\tau} \sum_{k=0}^{K-1} \sum_{i=0}^{\tau-1} \|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon,$$

and (ii) provided $\tau = O(\epsilon^{-1})$, the oracle complexities for evaluations and Jacobians of inner components $g_\xi(\cdot)$ respectively are at most

$$\widetilde{\Theta}\left((\epsilon^{-3}\tau^{-1} + \epsilon^{-2}\tau^2)\log(1/\Delta)\right) \tag{3.10}$$

$$\text{and} \quad \widetilde{\Theta}\left((\epsilon^{-2}\tau^{-1} + \epsilon^{-1}\tau^2)\log(1/\Delta)\right). \tag{3.11}$$

The oracle complexity upper bounds for evaluations and Jacobians in Corollary 3.6 can be optimized through careful selection of the epoch length τ.

Corollary 3.7 (Optimized complexity bounds)**.** *The minimal asymptotic rates, with respect to τ, of (3.10) and (3.11) are $\widetilde{\Theta}(\epsilon^{-8/3}\log(1/\Delta))$ and $\widetilde{\Theta}(\epsilon^{-5/3}\log(1/\Delta))$ respectively, which are simultaneously achieved by setting $\tau = \Theta(\epsilon^{-1/3})$.*

Comparing the two rates in Corollary 3.7 with those in Corollary 3.4 suggests the variance reduced estimator $\texttt{estimator}_1$ dominates the mini-batch method $\texttt{estimator}_0$ in the sense that, $\texttt{estimator}_1$ achieves $\frac{1}{K\tau}\sum_{k=1}^{K}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ in high probability with less evaluations of $g_\xi(\cdot)$ and $g'_\xi(\cdot)$.

Application of Modified Variance Reduction for Expectations. Next, we consider a modified variance-reduced estimator that leverages Jacobian evaluations to provide a better estimate of the value of $g_\xi(x)$. Again we consider $i = 0$ and $i > 0$ with $\texttt{estimator}_2(x_i^k, i; \theta)$ defined as

$$\begin{cases} \widetilde{g}_0^k = \frac{1}{A}\sum_{\xi\in\mathcal{A}_0^k} g_\xi(x_0^k) \\ \widetilde{J}_0^k = \frac{1}{B}\sum_{\xi\in\mathcal{B}_0^k} g'_\xi(x_0^k) \\ \widetilde{g}_i^k = \frac{1}{a}\sum_{\xi\in\mathcal{A}_i^k}\left(g_\xi(x_i^k) - g_\xi(x_0^k) - g'_\xi(x_0^k)(x_i^k - x_0^k)\right) + \widetilde{g}_0^k + \widetilde{J}_0^k(x_i^k - x_0^k) \\ \widetilde{J}_i^k = \frac{1}{b}\sum_{\xi\in\mathcal{B}_i^k}\left(g'_\xi(x_i^k) - g'_\xi(x_0^k)\right) + \widetilde{J}_0^k. \end{cases} \tag{Est$_2$}$$

Here (Est$_2$) applies a standard variance reduction update (1.4) to estimate the Jacobian and a first-order corrected update (1.5) to estimate the value of g itself. Again this estimator is parameterized by the four batch sizes $\theta = (A, B, a, b) \in \mathbb{N}_+^4$. The following lemma characterizes this modified scheme in terms of Assumption 2.3.

Lemma 3.8. *Suppose Assumption 3.2 holds. If $\texttt{estimator}(x_i^k, i; \theta)$ is defined by (Est$_2$), where $\theta = (A, B, a, b)$ and $\Theta = \mathbb{N}_+^4$, then Assumption 2.3 holds with*

$$\mathcal{C}(K, \boldsymbol{\tau}, \Delta) = \left\{(A, B, a, b) \in \mathbb{N}_+^4 : A, a \geq \frac{4}{9}\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right), \text{and } B, b \geq \frac{4}{9}\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)\right\},$$

$$\gamma_0(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right)}, \quad \gamma_1(K, \boldsymbol{\tau}, \theta, \Delta) = \lambda_0(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)},$$

$$\gamma_2(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{2L_g}{\sqrt{a}}\sqrt{\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right)}, \quad \lambda_1(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{4L_g}{\sqrt{b}}\sqrt{\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)}.$$

From this, we have the following two corollaries analyzing $\texttt{estimator}_2$.

Corollary 3.9 (Algorithmic guarantee)**.** *Consider any $\Delta \in (0,1)$, $M > 5l_f L_g$, integer $\tau > 0$, and any sufficiently small $\epsilon > 0$. Suppose Assumption 2.4 holds for* `solver`*, Assumption 3.2 holds for function g, $\inf_x \Phi(x) > -\infty$,* `estimator` *is defined by (Est$_2$), and $\boldsymbol{\tau}$ is restricted to the form $\tau_0 = \cdots = \tau_{K-1} = \tau$. Set $K = \lceil \frac{C_\Sigma \cdot \epsilon^{-1}}{\tau} \rceil$, $A = \lceil C_A \cdot \epsilon^{-2} \cdot \log(\frac{4(m+1)K\tau}{\Delta}) \rceil$, $B = \lceil C_B \cdot (1+\tau)^2 \cdot \epsilon^{-1} \cdot \log(\frac{4(m+n)K\tau}{\Delta}) \rceil$, $a = \lceil C_a \cdot (1+\tau)^4 \cdot \log(\frac{4(m+1)K\tau}{\Delta}) \rceil$, $b = \lceil C_b \cdot (1+\tau)^2 \cdot \log(\frac{4(m+n)K\tau}{\Delta}) \rceil$, $\overline{\delta} = \Delta/(2K\tau)$, $\overline{\epsilon} = \epsilon/(5 \cdot 30M)$, where $C_\Sigma, C_A, C_B, C_a, C_b$ are some constants, then with probability at least $1-\Delta$: (i) Algorithm 1's iterates satisfy:*

$$\frac{1}{K\tau} \sum_{k=0}^{K-1} \sum_{i=0}^{\tau-1} \|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon,$$

and (ii) provided $\tau = O(\epsilon^{-1})$, the oracle complexities for evaluations and Jacobians of inner components $g_\xi(\cdot)$ respectively are at most

$$\widetilde{\Theta}\left((\epsilon^{-3}\tau^{-1} + \epsilon^{-1}\tau^4) \log(1/\Delta) \right) \tag{3.12}$$

$$\text{and} \quad \widetilde{\Theta}\left((\epsilon^{-2}\tau + \epsilon^{-1}\tau^4) \log(1/\Delta) \right). \tag{3.13}$$

Corollary 3.10 (Optimized complexity bounds)**.** *(i) The minimal asymptotic evaluation complexity, with respect to τ, of (3.12) is $\widetilde{\Theta}(\epsilon^{-13/5} \log(1/\Delta))$, achieved by setting $\tau = \Theta(\epsilon^{-2/5})$. In this case, (3.13) is also $\widetilde{\Theta}(\epsilon^{-13/5} \log(1/\Delta))$. (ii) The minimal asymptotic Jacobian complexity, with respect to τ, of (3.13) is $\widetilde{\Theta}(\epsilon^{-2} \log(1/\Delta))$, achieved at $\tau = \Theta(1)$. In this case, (3.12) is $\widetilde{\Theta}(\epsilon^{-3} \log(1/\Delta))$.*

Remark 3.11. *We can compare the asymptotic rates in Corollary 3.10 with those in Corollary 3.7. Note that $\frac{13}{5} < \frac{8}{3}$ and $\frac{5}{3} < 2$. So Corollary 3.10(i) suggests the optimized asymptotic evaluation complexity bound of* `estimator`$_2$ *is lower than that of* `estimator`$_1$*. Corollary 3.10(ii) implies that the asymptotic Jacobian complexity bound of* `estimator`$_2$ *is always higher than the optimized bound of* `estimator`$_1$*. Hence neither method's guarantee uniformly dominates the other. Depending on the relative cost between evaluations and Jacobians, the best method varies.*

3.1.2 Finite Average Case Methods

Now we focus on the finite average setting in (1.2) and consider the natural extensions of the above estimators. Again, we find neither one of these two estimator's guarantees dominates the other. In this case, which method's guarantees are stronger depends on the relative size of $1/\epsilon$ and the number of summands N.

Application of Standard Variance Reduction for Finite Averages. First, we consider a variant of `estimator`$_1$, defined in two cases, $i = 0$ and $i > 0$, as

$$\texttt{estimator}_3(x_i^k, i; \theta) : \begin{cases} \widetilde{g}_0^k = \frac{1}{N}\sum_{j=1}^N g_j(x_0^k) = g(x_0^k) \\ \widetilde{J}_0^k = \frac{1}{N}\sum_{j=1}^N g_j'(x_0^k) = g'(x_0^k) \\ \widetilde{g}_i^k = \frac{1}{a}\sum_{j \in \mathcal{A}_i^k} \left(g_j(x_i^k) - g_j(x_0^k) \right) + \widetilde{g}_0^k \\ \widetilde{J}_i^k = \frac{1}{b}\sum_{j \in \mathcal{B}_i^k} \left(g_j'(x_i^k) - g_j'(x_0^k) \right) + \widetilde{J}_0^k. \end{cases} \tag{Est$_3$}$$

For each $k = 1, ..., K$, at the start of the k-th epoch, this estimator now constructs $\widetilde{g}_0^k$ and $\widetilde{J}_0^k$ exactly. At the iterations with $i > 0$, we generate an index set $\mathcal{A}_i^k$ of size a and another index set $\mathcal{B}_i^k$ of size b, both by sampling with replacement from $\{1, \ldots, N\}$. Note since $g_j(x_0^k)$ and $g_j'(x_0^k)$ are all evaluated for all $j \in \{1, \ldots, N\}$, one can store these in memory for use later in the construction of $\widetilde{g}_i^k$ and $\widetilde{J}_i^k$. So at the (k,i)-th iteration (for $i > 0$), the terms in $\{g_j(x_0^k) : j \in \mathcal{A}_i^k\}$ and $\{g_j'(x_0^k) : j \in \mathcal{B}_i^k\}$ can be simply called from the past data. Only the terms in $\{g_j(x_i^k) : j \in \mathcal{A}_i^k\}$ and $\{g_j'(x_i^k) : j \in \mathcal{B}_i^k\}$, namely those involves x_i^k, are needed to be evaluated. This estimator is parameterized by $\theta = (a, b) \in \mathbb{N}_+^2$, describing both batch sizes utilized. The set $\mathcal{C}(K, \boldsymbol{\tau}, \Delta)$, and functions $\{\gamma_l\}_{l=0}^2$, $\{\lambda_l\}_{l=0}^1$ are given below.

Lemma 3.12. *If* $\texttt{estimator}(x_i^k, i; \theta)$ *is defined by (Est₃), where* $\theta = (a, b)$ *and* $\Theta = \mathbb{N}_+^2$*, then Assumption 2.3 holds with the following*

$$\mathcal{C}(K, \boldsymbol{\tau}, \Delta) = \left\{(a, b) \in \mathbb{N}_+^2 : a \geq \frac{4}{9}\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right), b \geq \frac{4}{9}\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)\right\},$$

$$\gamma_0 = \gamma_2 = \lambda_0 = 0, \quad \gamma_1(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{4l_g}{\sqrt{a}}\sqrt{\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right)},$$

$$\lambda_1(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{4L_g}{\sqrt{b}}\sqrt{\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)}.$$

Just as done before, Lemma 3.12 and Theorem 3.1 provide recommendations for $K, \boldsymbol{\tau}, \theta$ and enable analysis of resulting oracle complexities.

Corollary 3.13 (Algorithmic guarantee)**.** *Consider any* $\Delta \in (0, 1)$*,* $M > 5l_f L_g$*, integer* $\tau > 0$*, and any sufficiently small* $\epsilon > 0$*. Suppose Assumption 2.4 holds for* $\texttt{solver}$*,* $\inf_x \Phi(x) > -\infty$*,* $\texttt{estimator}$ *is defined by (Est₃), and* $\boldsymbol{\tau}$ *is restricted to the form* $\tau_0 = \cdots = \tau_{K-1} = \tau$*. Set* $K = \lceil \frac{C_\Sigma \cdot \epsilon^{-1}}{\tau} \rceil$*,* $a = \lceil C_a \cdot (1+\tau)^2 \cdot \epsilon^{-1} \cdot \log(\frac{4(m+1)K\tau}{\Delta}) \rceil$*,* $b = \lceil C_b \cdot (1+\tau)^2 \cdot \log(\frac{4(m+n)K\tau}{\Delta}) \rceil$*,* $\bar{\delta} = \Delta/(2K\tau)$*,* $\bar{\epsilon} = \epsilon/(5 \cdot 30M)$*, where* C_Σ, C_a, C_b *are some constants, then with probability at least* $1 - \Delta$*: (i) Algorithm* 1*'s iterates satisfy:*

$$\frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon,$$

and (ii) the oracle complexities for evaluations and Jacobians of inner components $g_\xi(\cdot)$ *respectively are at most*

$$\widetilde{\Theta}\left(N + \epsilon^{-1}\tau^3 + N\epsilon^{-1}\tau^{-1} + \epsilon^{-2}\tau^2\right) \tag{3.14}$$

$$\text{and} \quad \widetilde{\Theta}\left(N + \tau^3 + N\epsilon^{-1}\tau^{-1} + \epsilon^{-1}\tau^2\right). \tag{3.15}$$

Corollary 3.14 (Optimized complexity bounds)**.** *(i) The minimal asymptotic evaluation complexity, with respect to* τ*, of (3.14) is* $\widetilde{\Theta}\left(N + \epsilon^{-2} + N^{2/3}\epsilon^{-4/3}\right)$*, achieved by setting* $\tau = \Theta\left(\max\{1, N^{1/3}\epsilon^{1/3}\}\right)$*. In this case, (3.15) will become* $\widetilde{\Theta}\left(\min\{N\epsilon^{-1}, N + N^{2/3}\epsilon^{-4/3}\}\right)$*. (ii) The minimal asymptotic Jacobian complexity, with respect to* τ*, of (3.15) is* $\widetilde{\Theta}\left(N + N^{2/3}\epsilon^{-1}\right)$*, achieved by setting* $\tau = \Theta\left(N^{1/3}\right)$*. In this case, (3.14) becomes* $\widetilde{\Theta}\left(N\epsilon^{-1} + N^{2/3}\epsilon^{-2}\right)$*.*

Application of Modified Variance Reduction for Finite Averages. Similarly, we can also incorporate exact evaluation into $\texttt{estimator}_2$. The modified estimator $\texttt{estimator}_4(x_i^k, i; \theta)$ is defined in two cases, $i = 0$ and $i > 0$, as

$$\begin{cases} \widetilde{g}_0^k = \frac{1}{N}\sum_{j=1}^N g_j(x_0^k) = g(x_0^k) \\ \widetilde{J}_0^k = \frac{1}{N}\sum_{j=1}^N g_j'(x_0^k) = g'(x_0^k) \\ \widetilde{g}_i^k = \frac{1}{a}\sum_{j \in \mathcal{A}_i^k}\left(g_j(x_i^k) - g_j(x_0^k) - g_j'(x_0^k)(x_i^k - x_0^k)\right) + \widetilde{g}_0^k + \widetilde{J}_0^k(x_i^k - x_0^k) \\ \widetilde{J}_i^k = \frac{1}{b}\sum_{j \in \mathcal{B}_i^k}\left(g_j'(x_i^k) - g_j'(x_0^k)\right) + \widetilde{J}_0^k. \end{cases} \tag{Est$_4$}$$

This estimator is the natural generalization of (Est$_2$) to utilize exact computations at the start of each epoch. For `estimator`$_4$, we have the following sequence of results, including choices of $\mathcal{C}(K, \boldsymbol{\tau}, \Delta)$, $\{\gamma_l\}_{l=0}^2$, $\{\lambda_l\}_{l=0}^1$, and algorithmic analysis.

Lemma 3.15. *If* `estimator`$(x_i^k, i; \theta)$ *is defined by (Est$_4$), where $\theta = (a, b)$ and $\Theta = \mathbb{N}_+^2$, then Assumption 2.3 holds with the following*

$$\mathcal{C}(K, \boldsymbol{\tau}, \Delta) = \left\{(a, b) \in \mathbb{N}_+^2 : a \geq \frac{4}{9}\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right), b \geq \frac{4}{9}\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)\right\},$$
$$\gamma_0 = \gamma_1 = \lambda_0 = 0, \quad \gamma_2(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{2L_g}{\sqrt{a}}\sqrt{\log\left(\frac{2(m+1)\Sigma_\tau}{\Delta}\right)},$$
$$\lambda_1(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{4L_g}{\sqrt{b}}\sqrt{\log\left(\frac{2(m+n)\Sigma_\tau}{\Delta}\right)}.$$

Corollary 3.16 (Algorithmic guarantee)**.** *Consider any $\Delta \in (0, 1)$, $M > 5 l_f L_g$, integer $\tau > 0$, and any sufficiently small $\epsilon > 0$. Suppose Assumption 2.4 holds for* `solver`*, $\inf_x \Phi(x) > -\infty$,* `estimator` *is defined by (Est$_4$), and $\boldsymbol{\tau}$ is restricted to the form $\tau_0 = \cdots = \tau_{K-1} = \tau$. Set $K = \lceil \frac{C_\Sigma \cdot \epsilon^{-1}}{\tau} \rceil$, $a = \lceil C_a \cdot (1+\tau)^4 \cdot \log(\frac{4(m+1)K\tau}{\Delta}) \rceil$, $b = \lceil C_b \cdot (1+\tau)^2 \cdot \log(\frac{4(m+n)K\tau}{\Delta}) \rceil$, $\bar{\delta} = \Delta/(2K\tau)$, $\bar{\epsilon} = \epsilon/(5 \cdot 30M)$, where C_Σ, C_a, C_b are some constants, then with probability at least $1 - \Delta$: (i) Algorithm 1's iterates satisfy:*

$$\frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon,$$

and (ii) the oracle complexities for evaluations and Jacobians of inner components $g_\xi(\cdot)$ respectively are at most

$$\widetilde{\Theta}\left(N + \tau^5 + N\epsilon^{-1}\tau^{-1} + \epsilon^{-1}\tau^4\right) \tag{3.16}$$
$$\text{and} \quad \widetilde{\Theta}\left(N + \tau^3 + N\epsilon^{-1}\tau^{-1} + \epsilon^{-1}\tau^2\right). \tag{3.17}$$

Corollary 3.17 (Optimized complexity bounds)**.** *(i) The minimal asymptotic evaluation complexity, with respect to τ, of (3.16) is $\widetilde{\Theta}\left(N + N^{4/5}\epsilon^{-1}\right)$, achieved by setting $\tau = \Theta\left(N^{1/5}\right)$. In this case, (3.17) is also $\widetilde{\Theta}\left(N + N^{4/5}\epsilon^{-1}\right)$. (ii) The minimal asymptotic Jacobian complexity, with respect to τ, of (3.17) is $\widetilde{\Theta}\left(N + N^{2/3}\epsilon^{-1}\right)$, achieved by setting $\tau = \Theta\left(N^{1/3}\right)$. In this case, (3.16) becomes $\widetilde{\Theta}\left(N^{5/3} + N^{4/3}\epsilon^{-1}\right)$.*

Remark 3.18. *Similar to Remark 3.11, we can compare the asymptotic rates in Corollary 3.14 and Corollary 3.17. The Jacobian complexity parts are the same in these two Corollaries, which is not a surprise, because (3.15) and (3.17) have the same form. We focus on comparing the oracle complexity for evaluations in the two Corollaries. The optimized asymptotic evaluation complexity bound for* `estimator`$_3$ *and* `estimator`$_4$ *are $\widetilde{\Theta}(N + N^{2/3}\epsilon^{-4/3} + \epsilon^{-2})$ and $\widetilde{\Theta}(N + N^{4/5}\epsilon^{-1})$ respectively.*

There are two parameters N and ϵ here. Suppose $N = \Theta(\epsilon^{-p})$. Then the previous two asymptotic rates become $\widetilde{\Theta}(\epsilon^{-p_3})$ and $\widetilde{\Theta}(\epsilon^{-p_4})$, where $p_3 = \max\{p, \frac{2}{3}p + \frac{4}{3}, 2\}$, $p_4 = \max\{p, \frac{4}{5}p + 1\}$. Note that

$$\begin{cases} p_3 > p_4, & \textit{if } p < \frac{5}{2} \\ p_3 < p_4, & \textit{if } \frac{5}{2} < p < 5 \\ p_3 = p_4, & \textit{if } p = \frac{5}{2} \textit{ or } p \geq 5. \end{cases}$$

So the optimized asymptotic evaluation complexity bound of $\texttt{estimator}_3$ *is strictly lower if* $\frac{5}{2} < p < 5$, *and* $\texttt{estimator}_4$ *has the lower one if* $p < \frac{5}{2}$. *This dichotomy suggests neither method's guarantee uniformly dominates the other. The best method varies depending on the relative rate between* N *and the accuracy* ϵ.

3.1.3 Methods with Randomized Epoch Durations As a last application, we showcase an example application of Theorem 3.1 with random epoch durations τ_k, as were considered by prior variance-reduced works like [3, 17]. Consider the following general scheme to determine K and $\boldsymbol{\tau}$ given some S_τ: Sample $\tau_0, \tau_1, ...$ independently from some distribution D_τ belonging to a parametric distribution family $\{D_\tau(\cdot; \tau_+, \theta_\tau) : (\tau_+, \theta_\tau) \in \mathbb{N}_+ \times \Theta_\tau\}$. Then generate K and $\boldsymbol{\tau}$ as

$$\begin{cases} K & \leftarrow \inf\{N : \sum_{k=1}^{N} \tau_k \geq S_\tau\} \\ \boldsymbol{\tau} & \leftarrow (\tau_1, ..., \tau_K). \end{cases} \tag{3.18}$$

In (3.18), if S_τ is much larger than each τ_k, then $\sum_{k=0}^{K-1} \tau_k$ will be approximately equal to S_τ. To make this relationship rigorous, we assume the following pair of conditions on the parametric family of generating distribution where the integer parameter τ_+ provides a bound on the size of each τ_k and control via C_τ of its expected value.

Assumption 3.19. *(i) The support of* $D_\tau(\cdot; \tau_+, \theta_\tau)$ *is a subset of* $\{1, ..., \tau_+\}$ *for any* $(\tau_+, \theta_\tau) \in \mathbb{N}_+ \times \Theta_\tau$. *(ii) There exist a constant* C_τ, *such that* $C_\tau \mathbb{E}_{\tau \sim D_\tau(\cdot; \tau_+, \theta_\tau)}[\tau] \geq \tau_+$ *for any* $(\tau_+, \theta_\tau) \in \mathbb{N}_+ \times \Theta_\tau$.

The intuition behind Assumption 3.19(ii) is trying to connect this scheme of varying τ_k with our previous theory of fixed τ_k. Consider a degenerated distribution $\widetilde{D}_{\tau_+}$ where $\tau \equiv \tau_+$, then Assumption 3.19(ii) controls the expectation ratio between $\widetilde{D}_{\tau_+}$ and $D_\tau(\cdot; \tau_+, \theta_\tau)$ by a constant upper bound C_τ. This intuitively suggests that if we replace $D_\tau(\cdot; \tau_+, \theta_\tau)$ by $\widetilde{D}_{\tau_+}$ in (3.18) while keeping S_τ unchanged, the returned K will increase at most by some constant factor. Indeed, we can prove this holds with high probability, which leads to the following result.

Corollary 3.20 (Algorithmic guarantee for the scheme of varying τ). *Consider any* $\Delta \in (0, 1)$, $M > 5l_f L_g$ *and any sufficiently small* $\epsilon > 0$. *Suppose* $\texttt{estimator}$ *is defined by (Est*$_1$*),* $(K, \boldsymbol{\tau})$ *is generated by* (3.18), *Assumption 2.4 holds for* $\texttt{solver}$, *Assumption 3.2 holds for function* g, *Assumption 3.19 holds for the generating distribution, and* $\inf_x \Phi(x) > -\infty$. *Set* $\tau_+ = \lceil \epsilon^{-1/3} \rceil$, $S_\tau = \lceil C_\Sigma \cdot \epsilon^{-1} \rceil$, $A = \lceil C_A \cdot \epsilon^{-2} \cdot \log(\frac{5(m+1)S_\tau}{\Delta}) \rceil$, $B = \lceil C_B \cdot \epsilon^{-1} \cdot \log(\frac{5(m+n)S_\tau}{\Delta}) \rceil$, $a = \lceil C_a \cdot (1+\tau_+)^2 \cdot \epsilon^{-1} \cdot \log(\frac{5(m+1)S_\tau}{\Delta}) \rceil$, $b = \lceil C_b \cdot (1 + \tau_+)^2 \cdot \log(\frac{5(m+n)S_\tau}{\Delta}) \rceil$, $\overline{\delta} = \frac{\Delta}{2(S_\tau + \tau_+)}$, $\overline{\epsilon} = \epsilon/(5 \cdot 30M)$, *where* $C_\Sigma, C_A, C_B, C_a, C_b$ *are some constants, then with probability at least* $1 - \Delta - \exp(-C_p \epsilon^{-2/3})$*:*[3] *(i) Algorithm* 1*'s iterates satisfy:*

$$\frac{1}{\Sigma_\tau} \sum_{k=0}^{K-1} \sum_{i=0}^{\tau_k - 1} \|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon,$$

and (ii) the oracle complexity for evaluations and Jacobians of inner components $g_\xi(\cdot)$ *are at most* $\widetilde{\Theta}(\epsilon^{-8/3} \log(1/\Delta))$ *and* $\widetilde{\Theta}(\epsilon^{-5/3} \log(1/\Delta))$ *respectively.*

Note the two complexities here, $\widetilde{\Theta}(\epsilon^{-8/3} \log(1/\Delta))$ and $\widetilde{\Theta}(\epsilon^{-5/3} \log(1/\Delta))$, match the bounds in Corollary 3.7. Note the probability bound of $1 - \Delta - \exp(-C_p \epsilon^{-2/3})$ slightly differs from the $1 - \Delta$ in Corollary 3.6. So Corollary 3.20 recovers the oracle complexities of fixed epoch duration setting, despite an exponentially small setback in probability guarantee. We used $\texttt{estimator}_1$ as an example to illustrate the idea of randomizing epoch duration. Similar scheme can also be applied to all of the other estimators discussed in previous sections.

[3] Here C_p is also a constant.

3.2 On the Computational Costs of Solver Subroutines

To provide a complete accounting for the computational cost of a variance reduced method, one ought to additionally consider the cost of (inexactly) computing proximal steps, i.e., evaluating `solver`. A prox-linear step is required at every iteration of Algorithm 1. Hence by Theorem 3.1, $\Sigma_\tau = O(1/\epsilon)$ (inexact) solves are needed.

For example, if f is sufficiently simple, one may be able to exactly minimize s_i^k, setting $\texttt{solver}(s, \bar{\epsilon}, \bar{\delta}) = \operatorname{argmin} s_i^k$. For example, the subproblem for nonlinear regression problems with $f(z) = \|z\|_2^2$ is least squares minimization, which can be solved exactly as a linear system. Alternatively, if $f(z) = \max_{j=1\ldots m} z_j$, then (1.6) is a quadratic program of dimension m. Hence, the total cost of Algorithm 1's proximal solves is $O(1/\epsilon)$ (inexact) linear system or quadratic program solves, respectively. As a second example, if f has uniformly L_f-Lipschitz gradient, a linearly convergent (accelerated) gradient method can be applied to each strongly convex proximal subproblem s_i^k. The resulting total number of gradient oracle calls to f is then $O\left(\frac{1}{\epsilon}\log(1/\epsilon)\right)$.

As a more interesting example, consider a doubly stochastic composite problem

$$\min_x \mathbb{E}_\zeta f_\zeta(\mathbb{E}_\xi g_\xi(x)) + h(x).$$

Given only samples of ζ and ξ, one cannot directly construct unbiased estimators of subgradients of $\mathbb{E}_\zeta f_\zeta(\mathbb{E}_\xi g_\xi(x))$, preventing the application of many direct stochastic first-order methods. See [18,19]. Regardless, if each f_ζ is uniformly l_f-Lipschitz, a stochastic proximal subgradient method can be applied to minimize the subproblem s_i^k. After $O\left(1/\bar{\epsilon}\right)$ steps, an $\bar{\epsilon}$-minimizer can be guaranteed [20]. Hence the total number of subgradient oracle calls needed to f at most $O\left(1/\epsilon^2\right)$. Noting we measure stationarity by the gradient norm *squared*, this agrees with the subgradient method's nonsmooth, nonconvex $O(1/\epsilon^4)$ rate [21] when unbiased subgradients are available.

4 Analysis

Recall that the objective function is $\Phi(x) = f(g(x)) + h(x)$. The standard prox-linear method may consider a linearized proximal subproblem with the following objective function at each iteration:

$$l_i^k(x) := f\Big(g(x_i^k) + g'(x_i^k)(x - x_i^k)\Big) + h(x) + \frac{M}{2}\|x - x_i^k\|_2^2.$$

In our algorithm, we replace $g(x_i^k)$ and $g'(x_i^k)$ with stochastic estimates $\widetilde{g}_i^k$ and $\widetilde{J}_i^k$, resulting in the following stochastic linearized objective function at each iteration:

$$s_i^k(x) := f\Big(\widetilde{g}_i^k + \widetilde{J}_i^k(x - x_i^k)\Big) + h(x) + \frac{M}{2}\|x - x_i^k\|_2^2.$$

Since $l_i^k(x)$ and $s_i^k(x)$ are M-strongly convex, they have unique minimizers, denoted

$$\widehat{x}_{i+1}^k := \arg\min l_i^k(x), \qquad \text{and} \qquad \widetilde{x}_{i+1}^k := \arg\min s_i^k(x).$$

Noting $l_i^k(x)$ is the objective function of prox-linear step in (1.6), by the definition in (1.7), our measure of stationarity at the iterate x_i^k is $\|\mathcal{G}_M(x_i^k)\|_2 = M\|x_i^k - \widehat{x}_{i+1}^k\|_2$.

4.1 Proofs for our Main Unified Convergence Theorem

In this part, we will prove a sequence of lemmas leading to the unified theory in Theorem 4.7. Our main result, Theorem 3.1, is a consequence of Theorem 4.7. Here we give an overview of our analysis.

We first prove three lemmas only depending on the basic setting of prox-linear methods, without the specific assumptions for `estimator` or `solver`. Lemma 4.1 is a useful result that upper bounds the prox-linear error. The upper bound involves the estimation error terms $\|\widetilde{g}_i^k - g(x_i^k)\|_2$ and $\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}}$. Lemma 4.2 provides a one-step property for x_i^k, $\widehat{x}_{i+1}^k$ and $\widetilde{x}_{i+1}^k$. In particular, it upper bounds the distance $\|\widehat{x}_{i+1}^k - x_i^k\|_2$ by $\|\widetilde{x}_{i+1}^k - x_i^k\|_2$ and the estimation error terms. Lemma 4.3 provides a descent property for the objective function Φ, though only between $\Phi(\widetilde{x}_{i+1}^k)$ and $\Phi(x_i^k)$.

To apply this inductively, we require a descent between $\Phi(x_{i+1}^k)$ and $\Phi(x_i^k)$. Assumption 2.4 for `solver` enables us to relate x_{i+1}^k to $\widetilde{x}_{i+1}^k$ with high probability. Lemma 4.4 uses this to give such a descent property. Lemma 4.5 ultimate combines our results to give an upper bound for $\|\widehat{x}_{i+1}^k - x_i^k\|_2$, which is proportional to $\|\mathcal{G}_M(x_i^k)\|_2$. Assumption 2.3 for `estimator` then allows us to uniformly bound error terms with high probability, formalized in Lemma 4.6. Applying a careful induction with the upper bound in Lemma 4.6 to cancel accumulated terms $\|x_i^k - x_0^k\|_2$ and $\|x_{i+1}^k - x_i^k\|_2$ suffices to give our ultimate result in Theorem 4.7, an upper bound for $\frac{1}{\Sigma_\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau_k-1}\|\mathcal{G}_M(x_i^k)\|_2^2$.

Lemma 4.1. *For any $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$, the following holds for any x:*

$$\begin{aligned}&\left|f\left(\widetilde{g}_i^k + \widetilde{J}_i^k(x - x_i^k)\right) - f\left(g(x_i^k) + g'(x_i^k)(x - x_i^k)\right)\right| \\ &\le l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + \frac{l_f}{2L_g}\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}}^2 + \frac{l_f L_g}{2}\|x - x_i^k\|_2^2.\end{aligned}$$

Proof of Lemma 4.1. Applying in order the Lipschitz continuity of f, triangle inequality, operator norm definition, and bounding $a \cdot b$ by $\frac{1}{2L_g}a^2 + \frac{L_g}{2}b^2$ yields

$$\begin{aligned}&\left|f\left(\widetilde{g}_i^k + \widetilde{J}_i^k(x - x_i^k)\right) - f\left(g(x_i^k) + g'(x_i^k)(x - x_i^k)\right)\right| \\ &\le l_f\|\widetilde{g}_i^k + \widetilde{J}_i^k(x - x_i^k) - g(x_i^k) - g'(x_i^k)(x - x_i^k)\|_2 \\ &\le l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + l_f\|\widetilde{J}_i^k(x - x_i^k) - g'(x_i^k)(x - x_i^k)\|_2 \\ &\le l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + l_f\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}} \cdot \|x - x_i^k\|_2 \\ &\le l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + \frac{l_f}{2L_g}\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}}^2 + \frac{l_f L_g}{2}\|x - x_i^k\|_2^2.\end{aligned}$$

□

Lemma 4.2. *For any $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$,*

$$\left(\frac{M}{2} - \frac{l_f L_g}{2}\right)\|\widehat{x}_{i+1}^k - x_i^k\|_2^2 \le 2l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + \frac{l_f}{L_g}\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}}^2 + \left(M + \frac{l_f L_g}{2}\right)\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2.$$

Proof of Lemma 4.2. Recall that $\widehat{x}_{i+1}^k$ and $\widetilde{x}_{i+1}^k$ are the minimizers of the M-strongly convex functions $l_i^k(x)$ and $s_i^k(x)$ respectively. So $l_i^k(\widetilde{x}_{i+1}^k) \ge l_i^k(\widehat{x}_{i+1}^k) + \frac{M}{2}\|\widetilde{x}_{i+1}^k - \widehat{x}_{i+1}^k\|_2^2$ and $s_i^k(\widehat{x}_{i+1}^k) \ge s_i^k(\widetilde{x}_{i+1}^k) + \frac{M}{2}\|\widehat{x}_{i+1}^k - \widetilde{x}_{i+1}^k\|_2^2$, i.e.

$$\begin{aligned}&f\left(g(x_i^k) + g'(x_i^k)(\widetilde{x}_{i+1}^k - x_i^k)\right) + h(\widetilde{x}_{i+1}^k) + \frac{M}{2}\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2 \\ &\ge f\left(g(x_i^k) + g'(x_i^k)(\widehat{x}_{i+1}^k - x_i^k)\right) + h(\widehat{x}_{i+1}^k) + \frac{M}{2}\|\widehat{x}_{i+1}^k - x_i^k\|_2^2 + \frac{M}{2}\|\widetilde{x}_{i+1}^k - \widehat{x}_{i+1}^k\|_2^2\end{aligned}$$

and

$$\begin{aligned}&f\left(\widetilde{g}_i^k + \widetilde{J}_i^k(\widehat{x}_{i+1}^k - x_i^k)\right) + h(\widehat{x}_{i+1}^k) + \frac{M}{2}\|\widehat{x}_{i+1}^k - x_i^k\|_2^2 \\ &\ge f\left(\widetilde{g}_i^k + \widetilde{J}_i^k(\widetilde{x}_{i+1}^k - x_i^k)\right) + h(\widetilde{x}_{i+1}^k) + \frac{M}{2}\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2 + \frac{M}{2}\|\widehat{x}_{i+1}^k - \widetilde{x}_{i+1}^k\|_2^2.\end{aligned}$$

Summing the two inequalities above, we get

$$\begin{aligned} M\|\widehat{x}_{i+1}^k - \widetilde{x}_{i+1}^k\|_2^2 \leq\ & f(g(x_i^k) + g'(x_i^k)(\widetilde{x}_{i+1}^k - x_i^k)) - f(\widetilde{g}_i^k + \widetilde{J}_i^k(\widetilde{x}_{i+1}^k - x_i^k)) \\ & + f(\widetilde{g}_i^k + \widetilde{J}_i^k(\widehat{x}_{i+1}^k - x_i^k)) - f(g(x_i^k) + g'(x_i^k)(\widehat{x}_{i+1}^k - x_i^k)). \end{aligned} \tag{4.1}$$

Let $x = \widetilde{x}_{i+1}^k$ and $\widehat{x}_{i+1}^k$ in Lemma 4.1 respectively, we have

$$\begin{aligned} & f\Big(g(x_i^k) + g'(x_i^k)(\widetilde{x}_{i+1}^k - x_i^k)\Big) - f\Big(\widetilde{g}_i^k + \widetilde{J}_i^k(\widetilde{x}_{i+1}^k - x_i^k)\Big) \\ & \leq l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + \frac{l_f}{2L_g}\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}}^2 + \frac{l_f L_g}{2}\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2, \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} & f\Big(\widetilde{g}_i^k + \widetilde{J}_i^k(\widehat{x}_{i+1}^k - x_i^k)\Big) - f\Big(g(x_i^k) + g'(x_i^k)(\widehat{x}_{i+1}^k - x_i^k)\Big) \\ & \leq l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + \frac{l_f}{2L_g}\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}}^2 + \frac{l_f L_g}{2}\|\widehat{x}_{i+1}^k - x_i^k\|_2^2. \end{aligned} \tag{4.3}$$

Combining (4.1), (4.2), and (4.3) yields

$$M\|\widehat{x}_{i+1}^k - \widetilde{x}_{i+1}^k\|_2^2 \leq 2l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + \frac{l_f}{L_g}\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}}^2 + \frac{l_f L_g}{2}\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2 + \frac{l_f L_g}{2}\|\widehat{x}_{i+1}^k - x_i^k\|_2^2.$$

Then noting that $\|\widehat{x}_{i+1}^k - \widetilde{x}_{i+1}^k\|_2^2 \geq -\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2 + \frac{1}{2}\|\widehat{x}_{i+1}^k - x_i^k\|_2^2$ gives the claim. □

Lemma 4.3. *For any* $(k, i) \in \mathcal{I}(K, \boldsymbol{\tau})$,

$$\Phi(\widetilde{x}_{i+1}^k) \leq \Phi(x_i^k) - (M - l_f L_g)\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2 + 2l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + \frac{l_f}{2L_g}\|\widetilde{J}_i^k - g'(x_i^k)\|_{\text{op}}^2.$$

Proof of Lemma 4.3. By strong convexity, $s_i^k(x_i^k) \geq s_i^k(\widetilde{x}_{i+1}^k) + \frac{M}{2}\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2$, i.e.

$$s_i^k(x_i^k) \geq f(\widetilde{g}_i^k + \widetilde{J}_i^k(\widetilde{x}_{i+1}^k - x_i^k)) + h(\widetilde{x}_{i+1}^k) + M\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2. \tag{4.4}$$

By Lipschitz continuity of f,

$$s_i^k(x_i^k) - \Phi(x_i^k) = f(\widetilde{g}_i^k) + h(x_i^k) - f(g(x_i^k)) - h(x_i^k) \leq l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2. \tag{4.5}$$

From (4.4) and (4.5), we have

$$\Phi(x_i^k) + l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 \geq f(\widetilde{g}_i^k + \widetilde{J}_i^k(\widetilde{x}_{i+1}^k - x_i^k)) + h(\widetilde{x}_{i+1}^k) + M\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2. \tag{4.6}$$

Let $x = \widetilde{x}_{i+1}^k$ and $y = x_i^k$ in Proposition 2.2,

$$\Phi(\widetilde{x}_{i+1}^k) - h(\widetilde{x}_{i+1}^k) = f(g(\widetilde{x}_{i+1}^k)) \leq f\Big(g(x_i^k) + g'(x_i^k)(\widetilde{x}_{i+1}^k - x_i^k)\Big) + \frac{l_f L_g}{2}\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2. \tag{4.7}$$

Combining (4.6) and (4.7), we have

$$\begin{aligned} \Phi(\widetilde{x}_{i+1}^k) & \leq h(\widetilde{x}_{i+1}^k) + f\Big(g(x_i^k) + g'(x_i^k)(\widetilde{x}_{i+1}^k - x_i^k)\Big) + \frac{l_f L_g}{2}\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2 \\ & \leq \Phi(x_i^k) + l_f\|\widetilde{g}_i^k - g(x_i^k)\|_2 + \left(\frac{l_f L_g}{2} - M\right)\|\widetilde{x}_{i+1}^k - x_i^k\|_2^2 \\ & \quad + f\Big(g(x_i^k) + g'(x_i^k)(\widetilde{x}_{i+1}^k - x_i^k)\Big) - f\Big(\widetilde{g}_i^k + \widetilde{J}_i^k(\widetilde{x}_{i+1}^k - x_i^k)\Big). \end{aligned} \tag{4.8}$$

By Lemma 4.1,

$$\begin{aligned}&f\Big(g(x_i^k)+g'(x_i^k)(\widetilde{x}_{i+1}^k-x_i^k)\Big)-f\Big(\widetilde{g}_i^k+\widetilde{J}_i^k(\widetilde{x}_{i+1}^k-x_i^k)\Big)\\&\le l_f\|g(x_i^k)-\widetilde{g}_i^k\|_2+\frac{l_f}{2L_g}\|g'(x_i^k)-\widetilde{J}_i^k\|_{\mathrm{op}}^2+\frac{l_fL_g}{2}\|\widetilde{x}_{i+1}^k-x_i^k\|_2^2.\end{aligned}\tag{4.9}$$

Finally, combining (4.8) and (4.9) gives the claim. □

Lemma 4.4. *Suppose Assumption 2.4 holds for* `solver`*, then for an arbitrary* $(k,i)\in\mathcal{I}(K,\boldsymbol{\tau})$*, the following holds with probability at least* $1-\overline{\delta}$*:*

$$\begin{aligned}\Phi(x_{i+1}^k)-\Phi(x_i^k)\le\bar{\epsilon}-\left(\frac{M}{2}-l_fL_g\right)\|x_{i+1}^k-x_i^k\|_2^2-\left(\frac{M}{2}-2l_fL_g\right)\|\widetilde{x}_{i+1}^k-x_i^k\|_2^2\\+4l_f\|\widetilde{g}_i^k-g(x_i^k)\|_2+\frac{3l_f}{2L_g}\|\widetilde{J}_i^k-g'(x_i^k)\|_{\mathrm{op}}^2.\end{aligned}$$

With probability at least $1-\overline{\delta}\Sigma_\tau$*, the inequality above holds for all* $(k,i)\in\mathcal{I}(K,\boldsymbol{\tau})$*.*

Proof of Lemma 4.4. Fix an arbitrary $(k,i)\in\mathcal{I}(K,\boldsymbol{\tau})$. We can split $\Phi(x_{i+1}^k)-\Phi(x_i^k)$ into the sum of three parts:

$$\begin{aligned}\Phi(x_{i+1}^k)-\Phi(x_i^k)=\Big[s_i^k(x_{i+1}^k)-s_i^k(\widetilde{x}_{i+1}^k)\Big]+\Big[(\Phi-s_i^k)(x_{i+1}^k)-(\Phi-s_i^k)(\widetilde{x}_{i+1}^k)\Big]\\+\Big[\Phi(\widetilde{x}_{i+1}^k)-\Phi(x_i^k)\Big].\end{aligned}\tag{4.10}$$

By Lemma 4.3,

$$\Phi(\widetilde{x}_{i+1}^k)-\Phi(x_i^k)\le(l_fL_g-M)\|\widetilde{x}_{i+1}^k-x_i^k\|_2^2+2l_f\|\widetilde{g}_i^k-g(x_i^k)\|_2+\frac{l_f}{2L_g}\|\widetilde{J}_i^k-g'(x_i^k)\|_{\mathrm{op}}^2.\tag{4.11}$$

By Assumption 2.4, there exists a subset $\mathcal{E}_{k,i}$ of the whole probability space, such that $\mathbb{P}(\mathcal{E}_{k,i})\ge1-\overline{\delta}$, and the following inequality holds on $\mathcal{E}_{k,i}$:

$$s_i^k(x_{i+1}^k)-s_i^k(\widetilde{x}_{i+1}^k)\le\bar{\epsilon}.\tag{4.12}$$

Then it remains to deal with $(\Phi-s_i^k)(x_{i+1}^k)-(\Phi-s_i^k)(\widetilde{x}_{i+1}^k)$. Note that $(\Phi-s_i^k)(x)=f(g(x))-f(\widetilde{g}_i^k+\widetilde{J}_i^k(x-x_i^k))-\frac{M}{2}\|x-x_i^k\|_2^2$, so

$$\begin{aligned}&(\Phi-s_i^k)(x_{i+1}^k)-(\Phi-s_i^k)(\widetilde{x}_{i+1}^k)\\&=f(g(x_{i+1}^k))-f\Big(\widetilde{g}_i^k+\widetilde{J}_i^k(x_{i+1}^k-x_i^k)\Big)-\frac{M}{2}\|x_{i+1}^k-x_i^k\|_2^2\\&\quad-f(g(\widetilde{x}_{i+1}^k))+f\Big(\widetilde{g}_i^k+\widetilde{J}_i^k(\widetilde{x}_{i+1}^k-x_i^k)\Big)+\frac{M}{2}\|\widetilde{x}_{i+1}^k-x_i^k\|_2^2\\&\le\Big|f(g(x_{i+1}^k))-f\Big(\widetilde{g}_i^k+\widetilde{J}_i^k(x_{i+1}^k-x_i^k)\Big)\Big|-\frac{M}{2}\|x_{i+1}^k-x_i^k\|_2^2\\&\quad+\Big|f(g(\widetilde{x}_{i+1}^k))+f\Big(\widetilde{g}_i^k+\widetilde{J}_i^k(\widetilde{x}_{i+1}^k-x_i^k)\Big)\Big|+\frac{M}{2}\|\widetilde{x}_{i+1}^k-x_i^k\|_2^2.\end{aligned}\tag{4.13}$$

Let $y=x_i^k$ in Proposition 2.2, and combine it with Lemma 4.1, we get the following inequality for any x:

$$\begin{aligned}|f(g(x))-f(\widetilde{g}_i^k+\widetilde{J}_i^k(x-x_i^k))|&\le\Big|f(g(x))-f\Big(g(x_i^k)+g'(x_i^k)(x-x_i^k)\Big)\Big|\\&\quad+\Big|f\Big(g(x_i^k)+g'(x_i^k)(x-x_i^k)\Big)-f\Big(\widetilde{g}_i^k+\widetilde{J}_i^k(x-x_i^k)\Big)\Big|\\&\le l_f\|\widetilde{g}_i^k-g(x_i^k)\|_2+\frac{l_f}{2L_g}\|\widetilde{J}_i^k-g'(x_i^k)\|_{\mathrm{op}}^2+l_fL_g\|x-x_i^k\|_2^2.\end{aligned}\tag{4.14}$$

Let $x = x^k_{i+1}$ and $\widetilde{x}^k_{i+1}$ in (4.14) respectively, and plug into (4.13),

$$\begin{aligned}
&(\Phi - s^k_i)(x^k_{i+1}) - (\Phi - s^k_i)(\widetilde{x}^k_{i+1}) \\
&\le 2l_f\|\widetilde{g}^k_i - g(x^k_i)\|_2 + \frac{l_f}{L_g}\|\widetilde{J}^k_i - g'(x^k_i)\|^2_{\mathrm{op}} + \left(l_f L_g - \frac{M}{2}\right)\|x^k_{i+1} - x^k_i\|^2_2 + \left(l_f L_g + \frac{M}{2}\right)\|\widetilde{x}^k_{i+1} - x^k_i\|^2_2.
\end{aligned} \tag{4.15}$$

Finally, on the set $\mathcal{E}_{k,i}$, we can use (4.11), (4.12), and (4.15) to upper bound the three parts on the right side of (4.10) :

$$\begin{aligned}
&\Phi(x^k_{i+1}) - \Phi(x^k_i) \\
&= s^k_i(x^k_{i+1}) - s^k_i(\widetilde{x}^k_{i+1}) + (\Phi - s^k_i)(x^k_{i+1}) - (\Phi - s^k_i)(\widetilde{x}^k_{i+1}) + \Phi(\widetilde{x}^k_{i+1}) - \Phi(x^k_i) \\
&\le \bar{\epsilon} + \left(l_f L_g - \frac{M}{2}\right)\|x^k_{i+1} - x^k_i\|^2_2 + \left(2l_f L_g - \frac{M}{2}\right)\|\widetilde{x}^k_{i+1} - x^k_i\|^2_2 \\
&+ 4l_f\|\widetilde{g}^k_i - g(x^k_i)\|_2 + \frac{3l_f}{2L_g}\|\widetilde{J}^k_i - g'(x^k_i)\|^2_{\mathrm{op}}.
\end{aligned}$$

The inequality above holds for all $(k,i) \in \mathcal{I}(K, \boldsymbol{\tau})$ on the set $\cap_{(k,i)\in\mathcal{I}(K,\boldsymbol{\tau})}\mathcal{E}_{k,i}$, which has probability at least $1 - \bar{\delta}\Sigma_\tau$ by a simple union bound. □

Lemma 4.5. *Suppose Assumption 2.4 holds for* `solver`*. If $M > 5l_f L_g$, then with probability at least $1 - \bar{\delta}\Sigma_\tau$, the following holds for all $(k,i) \in \mathcal{I}(K, \boldsymbol{\tau})$:*

$$\begin{aligned}
\frac{2M}{5}\|\widehat{x}^k_{i+1} - x^k_i\|^2_2 \le 12\bar{\epsilon} &+ 12\left(\Phi(x^k_i) - \Phi(x^k_{i+1})\right) + 50l_f\|\widetilde{g}^k_i - g(x^k_i)\|_2 \\
&+ 19\frac{l_f}{L_g}\|\widetilde{J}^k_i - g'(x^k_i)\|^2_{\mathrm{op}} - 12\left(\frac{M}{2} - l_f L_g\right)\|x^k_{i+1} - x^k_i\|^2_2.
\end{aligned}$$

Proof of Lemma 4.5. $M > 5l_f L_g$ implies $\frac{M}{2} - \frac{l_f L_g}{2} > \frac{2M}{5}$. Then from Lemma 4.2, we have

$$\frac{2M}{5}\|\widehat{x}^k_{i+1} - x^k_i\|^2_2 \le 2l_f\|\widetilde{g}^k_i - g(x^k_i)\|_2 + \frac{l_f}{L_g}\|\widetilde{J}^k_i - g'(x^k_i)\|^2_{\mathrm{op}} + \left(M + \frac{l_f L_g}{2}\right)\|\widetilde{x}^k_{i+1} - x^k_i\|^2_2. \tag{4.16}$$

By Lemma 4.4, there exists a subset $\mathcal{E}$ of the whole probability space, such that $\mathbb{P}(\mathcal{E}) \ge 1 - \bar{\delta}\Sigma_\tau$, and the following inequality holds for all $(k,i) \in \mathcal{I}(K, \boldsymbol{\tau})$ on $\mathcal{E}$:

$$\begin{aligned}
&\left(\frac{M}{2} - l_f L_g\right)\|x^k_{i+1} - x^k_i\|^2_2 + \left(\frac{M}{2} - 2l_f L_g\right)\|\widetilde{x}^k_{i+1} - x^k_i\|^2_2 \\
&\le \bar{\epsilon} + \Phi(x^k_i) - \Phi(x^k_{i+1}) + 4l_f\|\widetilde{g}^k_i - g(x^k_i)\|_2 + \frac{3l_f}{2L_g}\|\widetilde{J}^k_i - g'(x^k_i)\|^2_{\mathrm{op}}.
\end{aligned} \tag{4.17}$$

Note that $12(\frac{M}{2} - 2l_f L_g) > M + \frac{l_f L_g}{2}$, we can multiply (4.17) by 12 and combine with (4.16) to get the claim for all $(k,i) \in \mathcal{I}(K, \boldsymbol{\tau})$ on $\mathcal{E}$, with probability at least $1 - \bar{\delta}\Sigma_\tau$. □

Lemma 4.6. *Suppose Assumption 2.3 holds for* `estimator`*, and Assumption 2.4 holds for* `solver`*. Fix an $M > 5l_f L_g$. Then for any $K \in \mathbb{N}_+$, $\boldsymbol{\tau} \in \mathbb{N}^K_+$, $\Delta \in (0,1)$, $\theta \in \mathcal{C}(K, \boldsymbol{\tau}, \Delta)$, and an arbitrary set of positive reals $\{\alpha_{(k,i)} > 0 : (k,i) \in \mathcal{I}(K, \boldsymbol{\tau})\}$, with probability at least $1 - \bar{\delta}\Sigma_\tau - \Delta$, the following* *holds for all* $(k,i)\in\mathcal{I}(K,\boldsymbol{\tau})$*:*

$$\begin{aligned}\frac{2M}{5}\|\widehat{x}_{i+1}^k-x_i^k\|_2^2 \le 12\overline{\epsilon}+12\left(\Phi(x_i^k)-\Phi(x_{i+1}^k)\right)-12\left(\frac{M}{2}-l_fL_g\right)\|x_{i+1}^k-x_i^k\|_2^2\\ +\left(50l_f\gamma_0(K,\boldsymbol{\tau},\theta,\Delta)+25l_f\alpha_{(k,i)}\gamma_1(K,\boldsymbol{\tau},\theta,\Delta)+38\frac{l_f}{L_g}\lambda_0^2(K,\boldsymbol{\tau},\theta,\Delta)\right)\\ +\left(50l_f\gamma_2(K,\boldsymbol{\tau},\theta,\Delta)+\frac{25l_f\gamma_1(K,\boldsymbol{\tau},\theta,\Delta)}{\alpha_{(k,i)}}+38\frac{l_f}{L_g}\lambda_1^2(K,\boldsymbol{\tau},\theta,\Delta)\right)\|x_i^k-x_0^k\|_2^2.\end{aligned}$$

Proof of Lemma 4.6. As a shorthand, we use γ_l, λ_l for $\gamma_l(\cdot,\cdot,\cdot,\cdot)$, $\lambda_l(\cdot,\cdot,\cdot,\cdot)$ as the arguments are clear from context. Note that $M>5l_fL_g$, then by Lemma 4.5, there exists a subset $\mathcal{E}_1$ of the whole probability space, such that $\mathbb{P}(\mathcal{E}_1)\ge 1-\overline{\delta}\Sigma_\tau$, and the following inequality holds for all $(k,i)\in\mathcal{I}(K,\boldsymbol{\tau})$ on $\mathcal{E}_1$:

$$\begin{aligned}\frac{2M}{5}\|\widehat{x}_{i+1}^k-x_i^k\|_2^2 \le 12\overline{\epsilon}+12\left(\Phi(x_i^k)-\Phi(x_{i+1}^k)\right)+50l_f\|\widetilde{g}_i^k-g(x_i^k)\|_2\\ +19\frac{l_f}{L_g}\|\widetilde{J}_i^k-g'(x_i^k)\|_{\text{op}}^2-12\left(\frac{M}{2}-l_fL_g\right)\|x_{i+1}^k-x_i^k\|_2^2.\end{aligned}\tag{4.18}$$

By Assumption 2.3, there exists a subset $\mathcal{E}_2$ of the whole probability space, such that $\mathbb{P}(\mathcal{E}_2)\ge 1-\Delta$, and the following two inequalities hold for all $(k,i)\in\mathcal{I}(K,\boldsymbol{\tau})$ on $\mathcal{E}_2$:

$$\begin{aligned}&\|\widetilde{g}_i^k-g(x_i^k)\|_2\le\gamma_0+\gamma_1\|x_i^k-x_0^k\|_2+\gamma_2\|x_i^k-x_0^k\|_2^2,\\ &\|\widetilde{J}_i^k-g'(x_i^k)\|_{\text{op}}\le\lambda_0+\lambda_1\|x_i^k-x_0^k\|_2.\end{aligned}$$

Then the next two inequalities also hold for all $(k,i)\in\mathcal{I}(K,\boldsymbol{\tau})$ on $\mathcal{E}_2$:

$$\|\widetilde{g}_i^k-g(x_i^k)\|_2\le(\gamma_0+\frac{\alpha_{(k,i)}\gamma_1}{2})+(\gamma_2+\frac{\gamma_1}{2\alpha_{(k,i)}})\|x_i^k-x_0^k\|_2^2,\tag{4.19}$$

$$\|\widetilde{J}_i^k-g'(x_i^k)\|_{\text{op}}^2\le 2\lambda_0^2+2\lambda_1^2\|x_i^k-x_0^k\|_2^2,\tag{4.20}$$

where the $\alpha_{(k,i)}$ in (4.19) can be arbitrary positive real number. Use (4.19) and (4.20) to upper bound $\|\widetilde{g}_i^k-g(x_i^k)\|_2$ and $\|\widetilde{J}_i^k-g'(x_i^k)\|_{\text{op}}^2$ in (4.18) on the set $\mathcal{E}_1\cap\mathcal{E}_2$:

$$\begin{aligned}\frac{2M}{5}\|\widehat{x}_{i+1}^k-x_i^k\|_2^2 \le 12\overline{\epsilon}+12\left(\Phi(x_i^k)-\Phi(x_{i+1}^k)\right)-12\left(\frac{M}{2}-l_fL_g\right)\|x_{i+1}^k-x_i^k\|_2^2\\ +(50l_f\gamma_0+25l_f\alpha_{(k,i)}\gamma_1+38\frac{l_f}{L_g}\lambda_0^2)+(50l_f\gamma_2+\frac{25l_f\gamma_1}{\alpha_{(k,i)}}+38\frac{l_f}{L_g}\lambda_1^2)\|x_i^k-x_0^k\|_2^2,\end{aligned}$$

for all $(k,i)\in\mathcal{I}(K,\boldsymbol{\tau})$ on $\mathcal{E}_1\cap\mathcal{E}_2$, which has probability at least $1-\overline{\delta}\Sigma_\tau-\Delta$. □

Theorem 4.7. *Suppose Assumption 2.3 holds for* `estimator`*, and Assumption 2.4 holds for* `solver`*. Fix an* $M>5l_fL_g$*. If some* $K\in\mathbb{N}_+$*,* $\boldsymbol{\tau}\in\mathbb{N}_+^K$*,* $\Delta\in(0,1)$ *and* $\theta\in\mathcal{C}(K,\boldsymbol{\tau},\Delta)$ *satisfy*

$$50(1+\tau_{max})^2\gamma_2(K,\boldsymbol{\tau},\theta,\Delta)\le 3L_g\quad\text{and}\quad 38(1+\tau_{max})^2\lambda_1^2(K,\boldsymbol{\tau},\theta,\Delta)\le 3L_g^2\tag{4.21}$$

where $\tau_{max}=\max\{\tau_0,...,\tau_{K-1}\}$*, then the following holds with probability at least* $1-\overline{\delta}\Sigma_\tau-\Delta$*:*

$$\begin{aligned}&\frac{1}{\Sigma_\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau_k-1}\|\mathcal{G}_M(x_i^k)\|_2^2\\ &\le 30M\overline{\epsilon}+30M\frac{\Phi(x_0^0)-\Phi(x_0^K)}{\Sigma_\tau}+125Ml_f\gamma_0(K,\boldsymbol{\tau},\theta,\Delta)\\ &\quad+525M\frac{l_f}{L_g}(1+\tau_{max})^2\gamma_1^2(K,\boldsymbol{\tau},\theta,\Delta)+95M\frac{l_f}{L_g}\lambda_0^2(K,\boldsymbol{\tau},\theta,\Delta).\end{aligned}\tag{4.22}$$

Proof of Theorem 4.7. As a shorthand, we use γ_l, λ_l for $\gamma_l(\cdot,\cdot,\cdot,\cdot)$, $\lambda_l(\cdot,\cdot,\cdot,\cdot)$ as the arguments are clear from context. For each $k = 0, ..., K-1$, we define a sequence of numbers $c_{(k,0)}, ..., c_{(k,\tau_k)}$ in the following backward recursion way: First let $c_{(k,\tau_k)} = 0$. For $i = \tau_k - 1, ..., 0$, we define

$$c_{(k,i)} = (1 + \frac{1}{\tau_k})c_{(k,i+1)} + d_k, \quad \text{where} \quad d_k := 50 l_f \gamma_2 + \frac{3 l_f L_g}{(1+\tau_k)^2} + 38\frac{l_f}{L_g}\lambda_1^2.$$

Then $c_{(k,i)} + \tau_k d_k = (1+\frac{1}{\tau_k})(c_{(k,i+1)} + \tau_k d_k)$, from which we get $c_{(k,i)} = \tau_k d_k (1+\frac{1}{\tau_k})^{\tau_k - i} - \tau_k d_k$. By the fact $d_k \geq 0$ and $c_{(k,\tau_k)} = 0$, we also have $0 = c_{(k,\tau_k)} \leq c_{(k,\tau_k-1)} \leq \cdots \leq c_{(k,0)}$.

By letting $\alpha_{(k,i)} = \overline{\alpha}_k := \frac{25\gamma_1}{3L_g}(1+\tau_k)^2$ in Lemma 4.6, there exists a subset $\mathcal{E}$ of the whole probability space, such that $\mathbb{P}(\mathcal{E}) \geq 1 - \overline{\delta}\Sigma_\tau - \Delta$, and the inequality below holds for all $(k,i) \in \mathcal{I}(K, \boldsymbol{\tau})$ on $\mathcal{E}$:

$$\begin{aligned} \frac{2M}{5}\|\widehat{x}_{i+1}^k - x_i^k\|_2^2 \leq 12\overline{\epsilon} + 12\left(\Phi(x_i^k) - \Phi(x_{i+1}^k)\right) - 12(\frac{M}{2} - l_f L_g)\|x_{i+1}^k - x_i^k\|_2^2 \\ + (50 l_f\gamma_0 + 25 l_f \overline{\alpha}_k \gamma_1 + 38\frac{l_f}{L_g}\lambda_0^2) + d_k\|x_i^k - x_0^k\|_2^2. \end{aligned} \tag{4.23}$$

Using the fact that $12(\frac{M}{2} - l_f L_g) \geq 18 l_f L_g \geq (1+\tau_k)c_{(k,1)} \geq (1+\tau_k)c_{(k,i+1)}$ for all $i = 0, ..., \tau_k - 1$,[4] we have

$$\begin{aligned} -12(\frac{M}{2} - l_f L_g)\|x_{i+1}^k - x_i^k\|_2^2 &\leq -(1+\tau_k)c_{(k,i+1)}\|x_{i+1}^k - x_i^k\|_2^2 \\ &\leq -c_{(k,i+1)}\|x_{i+1}^k - x_0^k\|_2^2 + (1+\frac{1}{\tau_k})c_{(k,i+1)}\|x_i^k - x_0^k\|_2^2 \end{aligned} \tag{4.24}$$

for all $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$. Combining (4.23) and (4.24) leads to

$$\begin{aligned} \frac{2M}{5}\|\widehat{x}_{i+1}^k - x_i^k\|_2^2 = 12\overline{\epsilon} + 12\left(\Phi(x_i^k) - \Phi(x_{i+1}^k)\right) - c_{(k,i+1)}\|x_{i+1}^k - x_0^k\|_2^2 \\ + c_{(k,i)}\|x_i^k - x_0^k\|_2^2 + (50 l_f\gamma_0 + 25 l_f\overline{\alpha}_k\gamma_1 + 38\frac{l_f}{L_g}\lambda_0^2) \end{aligned}$$

for all $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ on $\mathcal{E}$. Then we can fix an arbitrary k, let i range over $0, ..., \tau_k - 1$ and take the sum. So the following holds for all $k = 0, ..., K-1$ on $\mathcal{E}$:

$$\begin{aligned} &\frac{2M}{5}\sum_{i=0}^{\tau_k - 1}\|\widehat{x}_{i+1}^k - x_i^k\|_2^2 \\ \leq\ & 12\tau_k\overline{\epsilon} + 12\left(\Phi(x_0^k) - \Phi(x_{\tau_k}^k)\right) - c_{(k,\tau_k)}\|x_{\tau_k}^k - x_0^k\|_2^2 \\ & + c_{(k,0)}\|x_0^k - x_0^k\|_2^2 + \tau_k(50 l_f\gamma_0 + 25 l_f\overline{\alpha}_k\gamma_1 + 38\frac{l_f}{L_g}\lambda_0^2) \\ =\ & 12\tau_k\overline{\epsilon} + 12\left(\Phi(x_0^k) - \Phi(x_0^{k+1})\right) + \tau_k(50 l_f\gamma_0 + 25 l_f\overline{\alpha}_k\gamma_1 + 38\frac{l_f}{L_g}\lambda_0^2), \end{aligned}$$

[4] It remains to check $18 l_f L_g \geq (1+\tau_k)c_{(k,1)}$ in this claim. By the recursion formula, $c_{(k,1)} = \tau_k d_k(1+\frac{1}{\tau_k})^{\tau_k - 1} - \tau_k d_k < e\tau_k d_k - \tau_k d_k < 2\tau_k d_k$. So $(1+\tau_k)c_{(k,1)} < 2(1+\tau_k)^2 d_k \leq 18 l_f L_g$, where the last step is from (4.21):

$$(1+\tau_k)^2 d_k = (1+\tau_k)^2\left(50 l_f\gamma_2 + \frac{3 l_f L_g}{(1+\tau_k)^2} + 38\frac{l_f}{L_g}\lambda_1^2\right) \leq 3 l_f L_g + 3 l_f L_g + 3 l_f L_g.$$

where the last step is because $c_{(k,\tau_k)} = 0$, and $x_0^{k+1} = x_{\tau_k}^k$ (see Algorithm 1). Finally, we sum over $k = 0, ..., K-1$, and use the fact that $\overline{\alpha}_k = \frac{25\gamma_1}{3L_g}(1+\tau_k)^2 \le \frac{25\gamma_1}{3L_g}(1+\tau_{\max})^2$: On $\mathcal{E}$, we have

$$\frac{2M}{5}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau_k-1}\|\widehat{x}_{i+1}^k - x_i^k\|_2^2 \le 12\Sigma_\tau \cdot \overline{\epsilon} + 12\left(\Phi(x_0^0) - \Phi(x_0^K)\right)$$
$$+ \Sigma_\tau\left(50 l_f\gamma_0 + 210\frac{l_f}{L_g}(1+\tau_{\max})^2\gamma_1^2 + 38\frac{l_f}{L_g}\lambda_0^2\right).$$

Multiplying $5M/(2\Sigma_\tau)$ on both sides completes the proof. □

Proof of Theorem 3.1. Replace Δ by $\Delta/2$ in Theorem 4.7, then Theorem 4.7 requires (3.1), and (4.21) becomes (3.8), (3.9). Under conditions (3.3)–(3.7), the 5 terms on the right hand side of (4.22) are all at most $\epsilon/5$. In addition, suppose (3.2) holds, i.e., $\Sigma_\tau\overline{\delta} \le \Delta/2$, then the probability bound in Theorem 4.7 becomes $1-\Delta$. Therefore, under conditions (3.1)–(3.9), Theorem 4.7 gives $\frac{1}{\Sigma_\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau_k-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \le \epsilon$ with probability at least $1-\Delta$. □

4.2 Sample Derivations of Corollaries 3.6 and 3.7

We provide the direct calculations of the claimed guarantees for $\texttt{estimator}_1$ discussed in Section 3.1. To do this, we first introduce a needed concentration inequality (Section 4.2.1) and technical bounds related to establishing $\texttt{estimator}_1$ satisfies Assumption 2.3 (Section 4.2.2). From these calculations, Corollaries 3.6 and 3.7 both follow (Section 4.2.3). The derivations of the remaining corollaries in Section 3.1 are deferred to Appendix A.2.

4.2.1 Concentration Inequality

Lemma 4.8 (Matrix Bernstein)**.** *Let $X_1, ..., X_n$ be independent random matrices of common dimension $d_1 \times d_2$. Assume $\mathbb{E}[X_k] = 0$ and $\|X_k\|_{\text{op}} \le L$ for each $k = 1, ..., n$ where L is some constant. If $n \ge \frac{4}{9}\log\left(\frac{d_1+d_2}{\delta}\right)$ for some $\delta \in (0,1)$, then*

$$\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_{\text{op}} \le \frac{2L}{\sqrt{n}}\sqrt{\log\left(\frac{d_1+d_2}{\delta}\right)}$$

holds with probability at least $1-\delta$.

Proof of Lemma 4.8. Using a classic Matrix Bernstein bound (see [22]), one has

$$\mathbb{P}\left(\left\|\sum_{k=1}^n X_k\right\|_{\text{op}} \ge t\right) \le (d_1+d_2)\cdot\exp\left(\frac{-t^2/2}{V + Lt/3}\right), \tag{4.25}$$

where $V := \max\left\{\left\|\sum_{k=1}^n \mathbb{E}\left[X_k X_k^T\right]\right\|_{\text{op}}, \left\|\sum_{k=1}^n \mathbb{E}\left[X_k^T X_k\right]\right\|_{\text{op}}\right\}$. By some basic properties,

$$\left\|\sum_{k=1}^n \mathbb{E}\Big[X_k X_k^T\Big]\right\|_{\text{op}} \le \sum_{k=1}^n \mathbb{E}\left\|X_k X_k^T\right\|_{\text{op}} \le \sum_{k=1}^n \mathbb{E}\Big[\|X_k\|_{\text{op}}\cdot\|X_k^T\|_{\text{op}}\Big] \le nL^2.$$

Similarly, $\left\|\sum_{k=1}^{n} \mathbb{E}\left[X_k^T X_k\right]\right\|_{\text{op}} \le nL^2$. So the V defined above is at most nL^2. Consequently,

$$\mathbb{P}\left(\left\|\sum_{k=1}^{n} X_k\right\|_{\text{op}} \ge t\right) \le (d_1+d_2)\cdot\exp\left(\frac{-t^2/2}{V+Lt/3}\right) \le (d_1+d_2)\cdot\exp\left(\frac{-t^2/2}{nL^2+Lt/3}\right)$$

for any $t \ge 0$. With $t = 2L\sqrt{n\log\left(\frac{d_1+d_2}{\delta}\right)}$, note that $\frac{2}{3}Lt = \frac{4L^2}{3}\sqrt{n\log\left(\frac{d_1+d_2}{\delta}\right)} \le 2nL^2$, where the last step is from the assumption that $n \ge \frac{4}{9}\log\left(\frac{d_1+d_2}{\delta}\right)$. The claim then follows directly from (4.25). □

4.2.2 Technical Bounds for Estimators

Proposition 4.9. *Suppose Assumption 3.2 holds. For an arbitrary fixed pair $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ and any $\delta \in (0,1)$, assume $\widetilde{g}_i^k$ and $\widetilde{J}_i^k$ are constructed by (Est$_0$). (i) If $A \ge \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$, then the following holds with probability at least $1-\delta$,*

$$\left\|\widetilde{g}_i^k - g(x_i^k)\right\|_2 \le \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{m+1}{\delta}\right)};$$

(ii) If $B \ge \frac{4}{9}\log\left(\frac{m+n}{\delta}\right)$, then the following holds with probability at least $1-\delta$,

$$\left\|\widetilde{J}_i^k - g'(x_i^k)\right\|_{\text{op}} \le \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log\left(\frac{m+n}{\delta}\right)}.$$

Proof of Proposition 4.9. We first prove part (i). From the construction of $\widetilde{g}_i^k$ in (Est$_0$), $\widetilde{g}_i^k - g(x_i^k) = \frac{1}{A}\sum_{\xi\in\mathcal{A}_i^k}\left(g_\xi(x_i^k) - g(x_i^k)\right)$. Suppose $\mathcal{A}_i^k = \{\xi_1, ..., \xi_A\}$, then $\xi_1, ..., \xi_A$ are independently drawn from distribution D. For each $r = 1, ..., A$, denote $Y_r = g_{\xi_r}(x_i^k) - g(x_i^k)$. So $\widetilde{g}_i^k - g(x_i^k) = \frac{1}{A}\sum_{r=1}^{A} Y_r$.

Use $\mathbf{x_i^k}$ to denote the sequence of iterates $\{x_0^0, x_1^0, ..., x_i^k\}$ in the rest of this proof. Then if conditioning on $\mathbf{x_i^k}$, all the randomness at the (k,i)-th iteration comes from the sampling of $\mathcal{A}_i^k$. So $Y_1, ..., Y_A$ are independent conditioning on $\mathbf{x_i^k}$. Since $g = \mathbb{E}_{\xi\sim D}[g_\xi]$, we immediately have $\mathbb{E}[Y_r|\mathbf{x_i^k}] = 0$. By Assumption 3.2, σ_g is a constant upper bound of $\|Y_r\|_2$. Note Y_r is an $m\times 1$ matrix, $\|Y_r\|_{\text{op}} = \|Y_r\|_2$. Then we can apply Lemma 4.8 to $Y_1, ..., Y_A$ (conditioning on $\mathbf{x_i^k}$): if $A \ge \frac{4}{9}\log(\frac{m+1}{\delta})$ for some $\delta \in (0,1)$, then

$$\mathbb{P}\left(\|\widetilde{g}_i^k - g(x_i^k)\|_2 \ge t \mid \mathbf{x_i^k}\right) \le \delta$$

where $t = \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}$. This implies the unconditional probability is also upper bounded by δ,[5] i.e.,

$$\mathbb{P}\left(\|\widetilde{g}_i^k - g(x_i^k)\|_2 \ge t\right) \le \delta$$

which finishes the proof for part (i).

The proof for part (ii) is similar. Suppose $\mathcal{B}_0^k = \{\xi_1', ..., \xi_B'\}$, and denote $Z_r = g'_{\xi_r'}(x_i^k) - g'(x_i^k)$ for each $r = 1, ..., B$. Then $\widetilde{J}_i^k - g'(x_i^k) = \frac{1}{B}\sum_{r=1}^{B} Z_r$. Applying Lemma 4.8 to $Z_1, ..., Z_B$ (conditioning on $\mathbf{x_i^k}$) finishes the proof, since $\|Z_r\|_{\text{op}} \le \sigma_{g'}$. □

[5] To see this, we can rewrite all the probabilities as the expectations of corresponding indicator functions, then use the law of total expectation. Let $\mathbf{I} = 1$ if $\|\widetilde{g}_i^k - g(x_i^k)\|_2 \ge t$, and $\mathbf{I} = 0$ otherwise. It follows that

$$\mathbb{P}\left(\left\|\widetilde{g}_i^k - g(x_i^k)\right\|_2 \ge t\right) = \mathbb{E}\left[\mathbf{I}\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{I} \mid \mathbf{x_i^k}\right]\right] = \mathbb{E}\left[\mathbb{P}\left(\left\|\widetilde{g}_i^k - g(x_i^k)\right\|_2 \ge t \mid \mathbf{x_i^k}\right)\right] \le \mathbb{E}\left[\delta\right] = \delta.$$

Proposition 4.10. *For an arbitrary fixed pair $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ and any $\delta \in (0,1)$, if $\widetilde{g}_i^k$ is constructed by (Est₁) or (Est₃), and $a \geq \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$, then the following holds with probability at least $1-\delta$:*

$$\left\|\widetilde{g}_i^k - g(x_i^k)\right\|_2 \leq \left\|\widetilde{g}_0^k - g(x_0^k)\right\|_2 + \frac{4l_g}{\sqrt{a}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}\|x_i^k - x_0^k\|_2.$$

Proof of Proposition 4.10. Noting that $\widetilde{g}_i^k - g(x_i^k) = \left(\widetilde{g}_0^k - g(x_0^k)\right) + \frac{1}{a}\sum_{r=1}^a Y_r$ where $Y_r = g_{\xi_r}(x_i^k) - g_{\xi_r}(x_0^k) - g(x_i^k) + g(x_0^k)$ for each $r = 1, ..., a$, it suffices to bound $\|\sum_{r=1}^a Y_r\|_2$ and then apply the triangle inequality. The claimed bound then follows directly from Lemma 4.8 as $\|Y_r\|_{\text{op}} \leq 2l_g\|x_i^k - x_0^k\|_2$. □

Proposition 4.11. *For an arbitrary fixed pair $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ and any $\delta \in (0,1)$, if $\widetilde{J}_i^k$ is constructed by any method among (Est₁), (Est₂), (Est₃), (Est₄), and $b \geq \frac{4}{9}\log\left(\frac{m+n}{\delta}\right)$, then the following holds with probability at least $1-\delta$:*

$$\left\|\widetilde{J}_i^k - g'(x_i^k)\right\|_{\text{op}} \leq \left\|\widetilde{J}_0^k - g'(x_0^k)\right\|_{\text{op}} + \frac{4L_g}{\sqrt{b}}\sqrt{\log\left(\frac{m+n}{\delta}\right)}\|x_i^k - x_0^k\|_2.$$

Proof of Proposition 4.11. Noting that $\widetilde{J}_i^k - g'(x_i^k) = \left(\widetilde{J}_0^k - g'(x_0^k)\right) + \frac{1}{b}\sum_{r=1}^b Z_r$ where $Z_r = g'_{\xi_r}(x_i^k) - g'_{\xi_r}(x_0^k) - g'(x_i^k) + g'(x_0^k)$ for each $r = 1, ..., b$, it suffices to bound $\|\sum_{r=1}^b Z_r\|_{\text{op}}$ and then apply the triangle inequality. The claimed bound then follows directly from Lemma 4.8 as $\|Z_r\|_{\text{op}} \leq 2L_g\|x_i^k - x_0^k\|_2$. □

4.2.3 Proofs of Corollaries 3.6 and 3.7 Using the technical propositions above, we can prove the lemmas and corollaries for $\texttt{estimator}_1$ claimed in Section 3.1.

Proof of Lemma 3.5. For any $K \in \mathbb{N}_+$, $\boldsymbol{\tau} \in \mathbb{N}_+^K$ and $\Delta \in (0,1)$, let $\delta = \frac{\Delta}{2\Sigma_\tau}$. By Proposition 4.9, for an arbitrary $k \in \{0, ..., K-1\}$, any $A \geq \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$ and any $B \geq \frac{4}{9}\log\left(\frac{m+n}{\delta}\right)$, the following two inequalities hold with probability at least $1-2\delta$,

$$\left\|\widetilde{g}_0^k - g(x_0^k)\right\|_2 \leq \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}, \quad \left\|\widetilde{J}_0^k - g'(x_0^k)\right\|_{\text{op}} \leq \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log\left(\frac{m+n}{\delta}\right)}. \tag{4.26}$$

By Proposition 4.10, for an arbitrary $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ and any $a \geq \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$, the following holds with probability at least $1-\delta$:

$$\left\|\widetilde{g}_i^k - g(x_i^k)\right\|_2 \leq \left\|\widetilde{g}_0^k - g(x_0^k)\right\|_2 + \frac{4l_g}{\sqrt{a}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}\|x_i^k - x_0^k\|_2. \tag{4.27}$$

By Proposition 4.11, for an arbitrary $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ and any $b \geq \frac{4}{9}\log\left(\frac{m+n}{\delta}\right)$, the following holds with probability at least $1-\delta$:

$$\left\|\widetilde{J}_i^k - g'(x_i^k)\right\|_{\text{op}} \leq \left\|\widetilde{J}_0^k - g'(x_0^k)\right\|_{\text{op}} + \frac{4L_g}{\sqrt{b}}\sqrt{\log\left(\frac{m+n}{\delta}\right)}\|x_i^k - x_0^k\|_2. \tag{4.28}$$

Let $\mathcal{C}(K,\boldsymbol{\tau},\Delta) = \{(A,B,a,b) \in \mathbb{N}_+^4 : A, a \geq \frac{4}{9}\log(\frac{2(m+1)\Sigma_\tau}{\Delta}), \text{and } B, b \geq \frac{4}{9}\log(\frac{2(m+n)\Sigma_\tau}{\Delta})\}$. Then for any $(A,B,a,b) \in \mathcal{C}(K,\boldsymbol{\tau},\Delta)$, by using a union probability bound, (4.26), (4.27) and (4.28) hold for all $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ with probability at least $1 - 2\Sigma_\tau\delta$. Note that $1 - 2\Sigma_\tau\delta = 1 - \Delta$, so we can set $\gamma_0(K,\boldsymbol{\tau},\theta,\Delta) = \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log(\frac{2(m+1)\Sigma_\tau}{\Delta})}$, $\gamma_1(K,\boldsymbol{\tau},\theta,\Delta) = \frac{4l_g}{\sqrt{a}}\sqrt{\log(\frac{2(m+1)\Sigma_\tau}{\Delta})}$, $\gamma_2 = 0$, $\lambda_0(K,\boldsymbol{\tau},\theta,\Delta) = \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log(\frac{2(m+n)\Sigma_\tau}{\Delta})}$ and $\lambda_1(K,\boldsymbol{\tau},\theta,\Delta) = \frac{4L_g}{\sqrt{b}}\sqrt{\log(\frac{2(m+n)\Sigma_\tau}{\Delta})}$ to satisfy Assumption 2.3. □

Proof of Corollary 3.6. We can obtain the explicit form of $\mathcal{C}(K, \boldsymbol{\tau}, \Delta)$, $\{\gamma_l\}_{l=0}^2$ and $\{\lambda_l\}_{l=0}^1$ from Lemma 3.5, and plug them into Theorem 3.1. Then by Theorem 3.1, to get $\frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \le \epsilon$ with probability at least $1-\Delta$, we need $\overline{\delta} \le \Delta/(2K\tau)$, $\overline{\epsilon} \le \epsilon/(5\cdot 30M)$, and the following inequalities:

$$\begin{cases} A, a \ge \frac{4}{9}\log\left(\frac{4(m+1)K\tau}{\Delta}\right), \text{and } B, b \ge \frac{4}{9}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \\ K\tau \ge 5\cdot 30M(\Phi(x_0^0) - \Phi^*)/\epsilon \\ \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{4(m+1)K\tau}{\Delta}\right)} \le \epsilon/(5\cdot 125 l_f M) \\ \frac{4\sigma_{g'}^2}{B}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \le L_g\epsilon/(5\cdot 95 l_f M) \\ (1+\tau)^2\frac{16 l_g^2}{a}\log\left(\frac{4(m+1)K\tau}{\Delta}\right) \le L_g\epsilon/(5\cdot 525 l_f M) \\ (1+\tau)^2\frac{16 L_g^2}{b}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \le 3L_g^2/38 \end{cases}$$

which reduce to

$$\begin{cases} K\tau \ge C_\Sigma \cdot \epsilon^{-1} \\ A \ge C_A \cdot \epsilon^{-2} \cdot \log\left(\frac{4(m+1)K\tau}{\Delta}\right) \\ B \ge C_B \cdot \epsilon^{-1} \cdot \log\left(\frac{4(m+n)K\tau}{\Delta}\right) \\ a \ge C_a \cdot (1+\tau)^2 \cdot \epsilon^{-1} \cdot \log\left(\frac{4(m+1)K\tau}{\Delta}\right) \\ b \ge C_b \cdot (1+\tau)^2 \cdot \log\left(\frac{4(m+n)K\tau}{\Delta}\right) \end{cases}$$

providing that $1/\epsilon$ is sufficiently large. Here $C_\Sigma, C_A, C_B, C_a, C_b$ are some constants.

For any positive integer τ, let $K = \lceil\frac{C_\Sigma\cdot\epsilon^{-1}}{\tau}\rceil$, $A = \lceil C_A\cdot\epsilon^{-2}\cdot\log(\frac{4(m+1)K\tau}{\Delta})\rceil$, $B = \lceil C_B\cdot\epsilon^{-1}\cdot\log(\frac{4(m+n)K\tau}{\Delta})\rceil$, $a = \lceil C_a\cdot(1+\tau)^2\cdot\epsilon^{-1}\cdot\log(\frac{4(m+1)K\tau}{\Delta})\rceil$, $b = \lceil C_b\cdot(1+\tau)^2\cdot\log(\frac{4(m+n)K\tau}{\Delta})\rceil$, then the conditions above hold for sufficiently small ϵ. So Theorem 3.1 guarantees that $\frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \le \epsilon$ with probability at least $1-\Delta$.

In (Est_1), at the $(k,0)$-th iteration, we evaluate $g_\xi(\cdot)$ for A times and $g'_\xi(\cdot)$ for B times. At the (k,i)-th iteration (with $i > 0$), we evaluate $g_\xi(\cdot)$ for $2a$ times and $g'_\xi(\cdot)$ for $2b$ times. Supposing $\tau = O(\epsilon^{-1})$, the oracle complexity for evaluations of $g_\xi(\cdot)$ is

$$KA + 2K(\tau-1)a \le KA + 2K\tau a = \widetilde{\Theta}\left((\epsilon^{-3}\tau^{-1} + \epsilon^{-2}\tau^2)\log(1/\Delta)\right),$$

and the oracle complexity for evaluations of Jacobians $g'_\xi(\cdot)$ is

$$KB + 2K(\tau-1)b \le KB + 2K\tau b = \widetilde{\Theta}\left((\epsilon^{-2}\tau^{-1} + \epsilon^{-1}\tau^2)\log(1/\Delta)\right).$$

□

Proof of Corollary 3.7. Suppose $\tau = \Theta(\epsilon^{-\beta})$ for some $\beta \ge 0$. Then (3.10) can be simplified as

$$\widetilde{\Theta}((\epsilon^{-3}\tau^{-1} + \epsilon^{-2}\tau^2)\log(1/\Delta)) = \widetilde{\Theta}(\epsilon^{-\max\{3-\beta, 2+2\beta\}}\log(1/\Delta)),$$

and (3.11) can be simplified as

$$\widetilde{\Theta}((\epsilon^{-2}\tau^{-1} + \epsilon^{-1}\tau^2)\log(1/\Delta)) = \widetilde{\Theta}(\epsilon^{-\max\{2-\beta, 1+2\beta\}}\log(1/\Delta)).$$

The asymptotic rates for two bounds are both minimized by $\beta = \frac{1}{3}$. At $\tau = \Theta(\epsilon^{-1/3})$, the two bounds become $\widetilde{\Theta}(\epsilon^{-8/3}\log(1/\Delta))$ and $\widetilde{\Theta}(\epsilon^{-5/3}\log(1/\Delta))$ respectively. □

Acknowledgements. This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-23-1-0531. Benjamin Grimmer was additionally supported as a fellow of the Alfred P. Sloan Foundation.

References

[1] Junyu Zhang and Lin Xiao. Stochastic variance-reduced prox-linear algorithms for nonconvex composite optimization. *Math. Program.*, 195(1–2):649–691, October 2021.

[2] Quoc Tran-Dinh, Nhan Pham, and Lam Nguyen. Stochastic Gauss-Newton algorithms for nonconvex compositional optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9572–9582. PMLR, 13–18 Jul 2020.

[3] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[4] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced cubic regularized Newton methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5990–5999. PMLR, 10–15 Jul 2018.

[5] Damek Davis. Variance reduction for root-finding problems. *Math. Program.*, 197(1):375–410, January 2022.

[6] Yin Liu and Sam Davanloo Tajbakhsh. Stochastic composition optimization of functions without lipschitz continuous gradient. *Journal of Optimization Theory and Applications*, 198:239–289, 2022.

[7] Robert M. Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

[8] J. V. Burke and M. C. Ferris. A gauss-newton method for convex composite optimization. *Math. Program.*, 71(2):179–194, December 1995.

[9] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM J. Optim.*, 21(4):1721–1739, December 2011.

[10] Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.*, 43:919–948, 2016.

[11] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Math. Program.*, 178(1–2):503–558, November 2019.

[12] Adrian S. Lewis and Stephen J. Wright. A proximal method for composite minimization. *Math. Program.*, 158(1-2):501–546, 2016.

[13] YU. Nesterov. Modified gauss–newton scheme with worst case guarantees for global performance. *Optimization Methods and Software*, 22(3):469–483, 2007.

[14] Dmitriy Drusvyatskiy. The proximal point method revisited. *arXiv:1712.06038*, 2017.

[15] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[16] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: a novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2613–2621, 2017.

[17] Jakub Konečný and Peter Richtárik. Semi-stochastic gradient descent methods. *Front. Appl. Math. Stat.*, 3, May 2017.

[18] Liu Liu, Ji Liu, and Dacheng Tao. Variance reduced methods for non-convex composition optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5813–5825, 2022.

[19] Mengdi Wang, Ji Liu, and Ethan X. Fang. Accelerating stochastic composition optimization. *J. Mach. Learn. Res.*, 18(1):3721–3743, January 2017.

[20] Benjamin Grimmer and Danlin Li. Some primal-dual theory for subgradient methods for strongly convex optimization. *arXiv:2305.17323*, 2024.

[21] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29:207–239, 2018.

[22] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1–2):1–230, 2015.

[23] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

A Derivations of Remaining Corollaries

A.1 More Concentration Inequality and Technical Bounds

Lemma A.1 (Hoeffding's inequality)**.** *Let $Y_1, ..., Y_n$ be independent random variables bounded by $a_i \leq Y_i \leq b_i$. Then for any $t \geq 0$, $S_n = \sum_{i=1}^n Y_n$ has*

$$\mathbb{P}\left(S_n \leq \mathbb{E}[S_n] - t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Proof of Lemma A.1. This is a restatement of Hoeffding's inequality [23]. □

Proposition A.2. *For an arbitrary fixed pair $(k, i) \in \mathcal{I}(K, \boldsymbol{\tau})$ and any $\delta \in (0, 1)$, if $\widetilde{g}_i^k$ is constructed by (Est$_2$) or (Est$_4$), and $a \geq \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$, then the following holds with probability at least $1 - \delta$:*

$$\left\|\widetilde{g}_i^k - g(x_i^k)\right\|_2 \leq \left\|\widetilde{g}_0^k - g(x_0^k)\right\|_2 + \left\|\widetilde{J}_0^k - g'(x_0^k)\right\|_{\mathrm{op}} \|x_i^k - x_0^k\|_2 + \frac{2L_g}{\sqrt{a}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}\|x_i^k - x_0^k\|_2^2.$$

Proof of Proposition A.2. The proof is similar to the proofs in Section 4.2.2. Suppose $\mathcal{A}_i^k = \{\xi_1, ..., \xi_a\}$, and denote $Y_r = g_{\xi_r}(x_i^k) - g_{\xi_r}(x_0^k) - g'_{\xi_r}(x_0^k)(x_i^k - x_0^k) - g(x_i^k) + g(x_0^k) + g'(x_0^k)(x_i^k - x_0^k)$ for each $r = 1, ..., a$. Then we have

$$\widetilde{g}_i^k - g(x_i^k) = \left(\widetilde{g}_0^k - g(x_0^k)\right) + \left(\widetilde{J}_0^k - g'(x_0^k)\right)(x_i^k - x_0^k) + \frac{1}{a}\sum_{r=1}^a Y_r. \tag{A.1}$$

Use $\mathbf{x_i^k}$ to denote the sequence of iterates $\{x_0^0, x_1^0, ..., x_i^k\}$ in the rest of this proof. Then $Y_1, ..., Y_a$ are independent conditioning on $\mathbf{x_i^k}$, and $\mathbb{E}[Y_r|\mathbf{x_i^k}] = 0$. By the setting in Section 2 and Proposition 2.1, $g'_{\xi_r}(\cdot)$ and $g'(\cdot)$ are both L_g-Lipschitz, which implies $\|g_{\xi_r}(x_i^k) - g_{\xi_r}(x_0^k) - g'_{\xi_r}(x_0^k)(x_i^k - x_0^k)\|_2 \leq \frac{L_g}{2}\|x_i^k - x_0^k\|_2^2$ and $\|g(x_i^k) - g(x_0^k) - g'(x_0^k)(x_i^k - x_0^k)\|_2 \leq \frac{L_g}{2}\|x_i^k - x_0^k\|_2^2$. So $L_g\|x_i^k - x_0^k\|_2^2$ is an upper bound of $\|Y_r\|_2$. Applying Lemma 4.8 to $Y_1, ..., Y_a$ (conditioning on $\mathbf{x_i^k}$) and following a similar proof as Proposition 4.9, we have

$$\left\|\frac{1}{a}\sum_{r=1}^a Y_r\right\|_2 \leq \frac{2L_g}{\sqrt{a}}\|x_i^k - x_0^k\|_2^2\sqrt{\log\left(\frac{m+1}{\delta}\right)}$$

with probability at least $1 - \delta$. Combining it with (A.1) completes the proof. □

A.2 Proofs of Remaining Corollaries

Using the technical Propositions from Section 4.2.2 and Appendix A.1, we can prove the Lemmas and Corollaries for $\texttt{estimator}_0$ and $\texttt{estimator}_2$–$\texttt{estimator}_4$.

Proofs for $\texttt{estimator}_0$.

Proof of Lemma 3.3. For any $K \in \mathbb{N}_+$, $\boldsymbol{\tau} \in \mathbb{N}_+^K$ and $\Delta \in (0,1)$, let $\delta = \frac{\Delta}{2\Sigma_\tau}$. By Proposition 4.9, for an arbitrary $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$, any $A \geq \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$ and any $B \geq \frac{4}{9}\log\left(\frac{m+n}{\delta}\right)$, the following two inequalities hold with probability at least $1-2\delta$,

$$\left\|\widetilde{g}_i^k - g(x_i^k)\right\|_2 \leq \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}, \quad \left\|\widetilde{J}_i^k - g'(x_i^k)\right\|_{\text{op}} \leq \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log\left(\frac{m+n}{\delta}\right)}. \tag{A.2}$$

Let $\mathcal{C}(K,\boldsymbol{\tau},\Delta) = \{(A,B) \in \mathbb{N}_+^4 : A \geq \frac{4}{9}\log(\frac{2(m+1)\Sigma_\tau}{\Delta}), \text{and } B \geq \frac{4}{9}\log(\frac{2(m+n)\Sigma_\tau}{\Delta})\}$. Then for any $(A,B) \in \mathcal{C}(K,\boldsymbol{\tau},\Delta)$, by using a union probability bound, (A.2) holds for all $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ with probability at least $1-2\Sigma_\tau\delta$. Note that $1-2\Sigma_\tau\delta = 1-\Delta$, so we can set $\gamma_0(K,\boldsymbol{\tau},\theta,\Delta) = \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log(\frac{2(m+1)\Sigma_\tau}{\Delta})}$, $\lambda_0(K,\boldsymbol{\tau},\theta,\Delta) = \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log(\frac{2(m+n)\Sigma_\tau}{\Delta})}$ and $\gamma_1 = \gamma_2 = \lambda_1 = 0$ to satisfy Assumption 2.3. □

Proof of Corollary 3.4. We can obtain the explicit form of $\mathcal{C}(K,\boldsymbol{\tau},\Delta)$, $\{\gamma_l\}_{l=0}^2$ and $\{\lambda_l\}_{l=0}^1$ from Lemma 3.3, and plug them into Theorem 3.1. Then by Theorem 3.1, to get $\frac{1}{\Sigma_\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau_k-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ with probability at least $1-\Delta$, we need $\overline{\delta} \leq \Delta/(2\Sigma_\tau)$, $\overline{\epsilon} \leq \epsilon/(5\cdot 30M)$, and the following inequalities:

$$\begin{cases} A \geq \frac{4}{9}\log\left(\frac{4(m+1)\Sigma_\tau}{\Delta}\right), \text{and } B \geq \frac{4}{9}\log\left(\frac{4(m+n)\Sigma_\tau}{\Delta}\right) \\ \Sigma_\tau \geq 5\cdot 30M(\Phi(x_0^0)-\Phi^*)/\epsilon \\ \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{4(m+1)\Sigma_\tau}{\Delta}\right)} \leq \epsilon/(5\cdot 125 l_f M) \\ \frac{4\sigma_{g'}^2}{B}\log\left(\frac{4(m+n)\Sigma_\tau}{\Delta}\right) \leq L_g\epsilon/(5\cdot 95 l_f M) \end{cases}$$

which reduces to

$$\begin{cases} \Sigma_\tau \geq C_\Sigma \cdot \epsilon^{-1} \\ A \geq C_A \cdot \epsilon^{-2} \cdot \log\left(\frac{4(m+1)\Sigma_\tau}{\Delta}\right) \\ B \geq C_B \cdot \epsilon^{-1} \cdot \log\left(\frac{4(m+n)\Sigma_\tau}{\Delta}\right) \end{cases}$$

providing that $1/\epsilon$ is sufficiently large. Here C_Σ, C_A, C_B are some constants.

Let $\Sigma_\tau = \lceil C_\Sigma \cdot \epsilon^{-1}\rceil$, $A = \lceil C_A \cdot \epsilon^{-2} \cdot \log(\frac{4(m+1)\Sigma_\tau}{\Delta})\rceil$, $B = \lceil C_B \cdot \epsilon^{-1} \cdot \log(\frac{4(m+n)\Sigma_\tau}{\Delta})\rceil$, then the conditions above hold for sufficiently small ϵ. So Theorem 3.1 guarantees that $\frac{1}{\Sigma_\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau_k-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ with probability at least $1-\Delta$.

In (Est$_0$), at the (k,i)-th iteration, we evaluate $g_\xi(\cdot)$ for A times and $g'_\xi(\cdot)$ for B times. Then the oracle complexity for evaluations of $g_\xi(\cdot)$ is

$$\Sigma_\tau A = \Theta(\epsilon^{-1}) \cdot \widetilde{\Theta}\left(\epsilon^{-2}\log(1/\Delta)\right) = \widetilde{\Theta}\left(\epsilon^{-3}\log(1/\Delta)\right),$$

and the oracle complexity for evaluations of Jacobians $g'_\xi(\cdot)$ is

$$\Sigma_\tau B = \Theta(\epsilon^{-1}) \cdot \widetilde{\Theta}\left(\epsilon^{-1}\log(1/\Delta)\right) = \widetilde{\Theta}\left(\epsilon^{-2}\log(1/\Delta)\right).$$

□

Proofs for `estimator`$_2$**.**

Proof of Lemma 3.8. For any $K \in \mathbb{N}_+$, $\boldsymbol{\tau} \in \mathbb{N}_+^K$ and $\Delta \in (0,1)$, let $\delta = \frac{\Delta}{2\Sigma_\tau}$. By Proposition 4.9, for an arbitrary $k \in \{0, ..., K-1\}$, any $A \geq \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$ and any $B \geq \frac{4}{9}\log\left(\frac{m+n}{\delta}\right)$, the following two inequalities hold with probability at least $1-2\delta$,

$$\left\|\widetilde{g}_0^k - g(x_0^k)\right\|_2 \leq \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}, \quad \left\|\widetilde{J}_0^k - g'(x_0^k)\right\|_{\mathrm{op}} \leq \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log\left(\frac{m+n}{\delta}\right)}. \tag{A.3}$$

By Proposition A.2, for an arbitrary $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ and any $a \geq \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$, the following holds with probability at least $1-\delta$:

$$\left\|\widetilde{g}_i^k - g(x_i^k)\right\|_2 \leq \left\|\widetilde{g}_0^k - g(x_0^k)\right\|_2 + \left\|\widetilde{J}_0^k - g'(x_0^k)\right\|_{\mathrm{op}}\|x_i^k - x_0^k\|_2 + \frac{2L_g}{\sqrt{a}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}\|x_i^k - x_0^k\|_2^2. \tag{A.4}$$

By Proposition 4.11, for an arbitrary $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ and any $b \geq \frac{4}{9}\log\left(\frac{m+n}{\delta}\right)$, the following holds with probability at least $1-\delta$:

$$\left\|\widetilde{J}_i^k - g'(x_i^k)\right\|_{\mathrm{op}} \leq \left\|\widetilde{J}_0^k - g'(x_0^k)\right\|_{\mathrm{op}} + \frac{4L_g}{\sqrt{b}}\sqrt{\log\left(\frac{m+n}{\delta}\right)}\|x_i^k - x_0^k\|_2. \tag{A.5}$$

Let $\mathcal{C}(K,\boldsymbol{\tau},\Delta) = \{(A,B,a,b) \in \mathbb{N}_+^4 : A, a \geq \frac{4}{9}\log(\frac{2(m+1)\Sigma_\tau}{\Delta}), \text{and } B, b \geq \frac{4}{9}\log(\frac{2(m+n)\Sigma_\tau}{\Delta})\}$. Then for any $(A,B,a,b) \in \mathcal{C}(K,\boldsymbol{\tau},\Delta)$, by using a union probability bound, (A.3), (A.4) and (A.5) hold for all $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ with probability at least $1 - 2\Sigma_\tau\delta$. Note that $1 - 2\Sigma_\tau\delta = 1-\Delta$, so we can set $\gamma_0(K,\boldsymbol{\tau},\theta,\Delta) = \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log(\frac{2(m+1)\Sigma_\tau}{\Delta})}$, $\gamma_1(K,\boldsymbol{\tau},\theta,\Delta) = \lambda_0(K,\boldsymbol{\tau},\theta,\Delta) = \frac{2\sigma_{g'}}{\sqrt{B}}\sqrt{\log(\frac{2(m+n)\Sigma_\tau}{\Delta})}$, $\gamma_2(K,\boldsymbol{\tau},\theta,\Delta) = \frac{2L_g}{\sqrt{a}}\sqrt{\log(\frac{2(m+1)\Sigma_\tau}{\Delta})}$ and $\lambda_1(K,\boldsymbol{\tau},\theta,\Delta) = \frac{4L_g}{\sqrt{b}}\sqrt{\log(\frac{2(m+n)\Sigma_\tau}{\Delta})}$ to satisfy Assumption 2.3. □

Proof of Corollary 3.9. We can obtain the explicit form of $\mathcal{C}(K,\boldsymbol{\tau},\Delta)$, $\{\gamma_l\}_{l=0}^2$ and $\{\lambda_l\}_{l=0}^1$ from Lemma 3.8, and plug them into Theorem 3.1. Then by Theorem 3.1, to get $\frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ with probability at least $1-\Delta$, we need $\overline{\delta} \leq \Delta/(2K\tau)$, $\overline{\epsilon} \leq \epsilon/(5\cdot 30M)$, and the following inequalities:

$$\begin{cases} A, a \geq \frac{4}{9}\log\left(\frac{4(m+1)K\tau}{\Delta}\right), \text{and } B, b \geq \frac{4}{9}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \\ K\tau \geq 5\cdot 30M(\Phi(x_0^0) - \Phi^*)/\epsilon \\ \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{4(m+1)K\tau}{\Delta}\right)} \leq \epsilon/(5\cdot 125 l_f M) \\ \frac{4\sigma_{g'}^2}{B}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \leq L_g\epsilon/(5\cdot 95 l_f M) \\ (1+\tau)^2\frac{4\sigma_{g'}^2}{B}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \leq L_g\epsilon/(5\cdot 525 l_f M) \\ (1+\tau)^2\frac{2L_g}{\sqrt{a}}\sqrt{\log\left(\frac{4(m+1)K\tau}{\Delta}\right)} \leq 3L_g/50 \\ (1+\tau)^2\frac{16L_g^2}{b}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \leq 3L_g^2/38 \end{cases}$$

So it reduces to

$$\begin{cases} K\tau \geq C_\Sigma \cdot \epsilon^{-1} \\ A \geq C_A \cdot \epsilon^{-2} \cdot \log\left(\frac{4(m+1)K\tau}{\Delta}\right) \\ B \geq C_B \cdot (1+\tau)^2 \cdot \epsilon^{-1} \cdot \log\left(\frac{4(m+n)K\tau}{\Delta}\right) \\ a \geq C_a \cdot (1+\tau)^4 \cdot \log\left(\frac{4(m+1)K\tau}{\Delta}\right) \\ b \geq C_b \cdot (1+\tau)^2 \cdot \log\left(\frac{4(m+n)K\tau}{\Delta}\right) \end{cases}$$

providing that $1/\epsilon$ is sufficiently large. Here $C_\Sigma, C_A, C_B, C_a, C_b$ are some constants.

For any positive integer τ, let $K = \lceil \frac{C_\Sigma \cdot \epsilon^{-1}}{\tau} \rceil$, $A = \lceil C_A \cdot \epsilon^{-2} \cdot \log(\frac{4(m+1)K\tau}{\Delta}) \rceil$, $B = \lceil C_B \cdot (1+\tau)^2 \cdot \epsilon^{-1} \cdot \log(\frac{4(m+n)K\tau}{\Delta}) \rceil$, $a = \lceil C_a \cdot (1+\tau)^4 \cdot \log(\frac{4(m+1)K\tau}{\Delta}) \rceil$, $b = \lceil C_b \cdot (1+\tau)^2 \cdot \log(\frac{4(m+n)K\tau}{\Delta}) \rceil$, then the conditions above hold for sufficiently small ϵ. So Theorem 3.1 guarantees that $\frac{1}{K\tau} \sum_{k=0}^{K-1} \sum_{i=0}^{\tau-1} \|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ with probability at least $1 - \Delta$.

In (Est$_2$), at the $(k,0)$-th iteration, we evaluate $g_\xi(\cdot)$ for A times and $g'_\xi(\cdot)$ for B times. At the (k,i)-th iteration (with $i > 0$), we evaluate $g_\xi(\cdot)$ for $2a$ times and $g'_\xi(\cdot)$ for $a + 2b$ times. Suppose $\tau = O(\epsilon^{-1})$, then the oracle complexity for evaluations of $g_\xi(\cdot)$ is

$$KA + 2K(\tau-1)a \leq KA + 2K\tau a = \widetilde{\Theta}\left((\epsilon^{-3}\tau^{-1} + \epsilon^{-1}\tau^4)\log(1/\Delta)\right),$$

and the oracle complexity for evaluations of Jacobians $g'_\xi(\cdot)$ is

$$KB + K(\tau-1)(a+2b) \leq KB + K\tau a + 2K\tau b = \widetilde{\Theta}\left((\epsilon^{-2}\tau + \epsilon^{-1}\tau^4)\log(1/\Delta)\right).$$

□

Proof of Corollary 3.10. Suppose $\tau = \Theta(\epsilon^{-\beta})$ for some $\beta \geq 0$. Then (3.12) can be simplified as

$$\widetilde{\Theta}\left((\epsilon^{-3}\tau^{-1} + \epsilon^{-1}\tau^4)\log(1/\Delta)\right) = \widetilde{\Theta}(\epsilon^{-\max\{3-\beta,1+4\beta\}}\log(1/\Delta)), \tag{A.6}$$

and (3.13) can be simplified as

$$\widetilde{\Theta}\left((\epsilon^{-2}\tau + \epsilon^{-1}\tau^4)\log(1/\Delta)\right) = \widetilde{\Theta}(\epsilon^{-\max\{2+\beta,1+4\beta\}}\log(1/\Delta)). \tag{A.7}$$

(A.6) is minimized by $\beta = \frac{2}{5}$. When $\beta = \frac{2}{5}$,(A.6) and (A.7) are $\widetilde{\Theta}(\epsilon^{-13/5}\log(1/\Delta))$. (A.7) is minimized by $\beta = 0$. When $\beta = 0$, (A.7) becomes $\widetilde{\Theta}(\epsilon^{-2}\log(1/\Delta))$ and (A.6) becomes $\widetilde{\Theta}(\epsilon^{-3}\log(1/\Delta))$. □

Proofs for `estimator`$_3$.

Proof of Lemma 3.12. We have $\tilde{g}_0^k = g(x_0^k)$ and $\tilde{J}_0^k = g'(x_0^k)$ from (Est$_3$). For any $K \in \mathbb{N}_+$, $\boldsymbol{\tau} \in \mathbb{N}_+^K$ and $\Delta \in (0,1)$, let $\delta = \frac{\Delta}{2\Sigma_\tau}$. By Proposition 4.10, for an arbitrary $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ and any $a \geq \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$, the following holds with probability at least $1 - \delta$:

$$\left\|\tilde{g}_i^k - g(x_i^k)\right\|_2 \leq \frac{4l_g}{\sqrt{a}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}\|x_i^k - x_0^k\|_2. \tag{A.8}$$

By Proposition 4.11, for an arbitrary $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ and any $b \geq \frac{4}{9}\log\left(\frac{m+n}{\delta}\right)$, the following holds with probability at least $1 - \delta$:

$$\left\|\tilde{J}_i^k - g'(x_i^k)\right\|_{\text{op}} \leq \frac{4L_g}{\sqrt{b}}\sqrt{\log\left(\frac{m+n}{\delta}\right)}\|x_i^k - x_0^k\|_2. \tag{A.9}$$

Let $\mathcal{C}(K,\boldsymbol{\tau},\Delta) = \{(a,b) \in \mathbb{N}_+^2 : a \geq \frac{4}{9}\log(\frac{2(m+1)\Sigma_\tau}{\Delta}), b \geq \frac{4}{9}\log(\frac{2(m+n)\Sigma_\tau}{\Delta})\}$. Then for any $(a,b) \in \mathcal{C}(K,\boldsymbol{\tau},\Delta)$, by using a union probability bound, (A.8) and (A.9) hold for all $(k,i) \in \mathcal{I}(K,\boldsymbol{\tau})$ with probability at least $1 - 2\Sigma_\tau\delta$. Note that $1 - 2\Sigma_\tau\delta = 1 - \Delta$, so we can set $\gamma_0 = \gamma_2 = \lambda_0 = 0$, $\gamma_1(K,\boldsymbol{\tau},\theta,\Delta) = \frac{4l_g}{\sqrt{a}}\sqrt{\log(\frac{2(m+1)\Sigma_\tau}{\Delta})}$ and $\lambda_1(K,\boldsymbol{\tau},\theta,\Delta) = \frac{4L_g}{\sqrt{b}}\sqrt{\log(\frac{2(m+n)\Sigma_\tau}{\Delta})}$ to satisfy Assumption 2.3. □

Proof of Corollary 3.13. We can obtain the explicit form of $\mathcal{C}(K,\boldsymbol{\tau},\Delta)$, $\{\gamma_l\}_{l=0}^2$ and $\{\lambda_l\}_{l=0}^1$ from Lemma 3.12, and plug them into Theorem 3.1. Then by Theorem 3.1, to get $\frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ with probability at least $1-\Delta$, we need $\overline{\delta} \leq \Delta/(2K\tau)$, $\overline{\epsilon} \leq \epsilon/(5{\cdot}30M)$, and the following inequalities:

$$\begin{cases} a \geq \frac{4}{9}\log\left(\frac{4(m+1)K\tau}{\Delta}\right), \text{and } b \geq \frac{4}{9}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \\ K\tau \geq 5 \cdot 30M(\Phi(x_0^0) - \Phi^*)/\epsilon \\ (1+\tau)^2\frac{16l_g^2}{a}\log\left(\frac{4(m+1)K\tau}{\Delta}\right) \leq L_g\epsilon/(5 \cdot 525 l_f M) \\ (1+\tau)^2\frac{16L_g^2}{b}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \leq 3L_g^2/38 \end{cases}$$

which reduces to

$$\begin{cases} K\tau \geq C_\Sigma \cdot \epsilon^{-1} \\ a \geq C_a \cdot (1+\tau)^2 \cdot \epsilon^{-1} \cdot \log\left(\frac{4(m+1)K\tau}{\Delta}\right) \\ b \geq C_b \cdot (1+\tau)^2 \cdot \log\left(\frac{4(m+n)K\tau}{\Delta}\right) \end{cases}$$

providing that $1/\epsilon$ is sufficiently large. Here C_Σ, C_a, C_b are some constants.

For any positive integer τ, let $K = \lceil\frac{C_\Sigma \cdot \epsilon^{-1}}{\tau}\rceil$, $a = \lceil C_a \cdot (1+\tau)^2 \cdot \epsilon^{-1} \cdot \log(\frac{4(m+1)K\tau}{\Delta})\rceil$, $b = \lceil C_b \cdot (1+\tau)^2 \cdot \log(\frac{4(m+n)K\tau}{\Delta})\rceil$. Then the conditions above are satisfied, so Theorem 3.1 guarantees that $\frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ with probability at least $1 - \Delta$.

In (Est$_3$), at the $(k,0)$-th iteration, we evaluate $g_j(\cdot)$ for N times and $g_j'(\cdot)$ for N times. At the (k,i)-th iteration (with $i > 0$), we evaluate $g_j(\cdot)$ for a times and $g_j'(\cdot)$ for b times. So the oracle complexity for evaluations of $g_\xi(\cdot)$ is

$$\begin{aligned} KN + K(\tau-1)a \leq K(N + \tau \cdot a) &\leq (1 + \frac{C_\Sigma}{\tau\epsilon}) \cdot (N + \widetilde{\Theta}(\epsilon^{-1}\tau^3)) \\ &= \widetilde{\Theta}\left(N + \epsilon^{-1}\tau^3 + N\epsilon^{-1}\tau^{-1} + \epsilon^{-2}\tau^2\right), \end{aligned}$$

and the oracle complexity for evaluations of Jacobians $g_\xi'(\cdot)$ is

$$\begin{aligned} KN + K(\tau-1)b \leq K(N + \tau \cdot b) &\leq (1 + \frac{C_\Sigma}{\tau\epsilon}) \cdot (N + \widetilde{\Theta}(\tau^3)) \\ &= \widetilde{\Theta}\left(N + \tau^3 + N\epsilon^{-1}\tau^{-1} + \epsilon^{-1}\tau^2\right). \end{aligned}$$

□

Proof of Corollary 3.14. Note that $\tau \geq 1$, so

$$N + \epsilon^{-1}\tau^3 + N\epsilon^{-1}\tau^{-1} + \epsilon^{-2}\tau^2 \geq N + \epsilon^{-2}\tau^2 \geq N + \epsilon^{-2} \geq 0.$$

We also have

$$N + \epsilon^{-1}\tau^3 + N\epsilon^{-1}\tau^{-1} + \epsilon^{-2}\tau^2 \geq \max\{N\epsilon^{-1}\tau^{-1}, \epsilon^{-2}\tau^2\} \geq N^{2/3}\epsilon^{-4/3} \geq 0$$

Combine the two inequalities above, the rate in (3.14) is at least $\widetilde{\Theta}\left(\max\{N + \epsilon^{-2}, N^{2/3}\epsilon^{-4/3}\}\right) = \widetilde{\Theta}\left(N + \epsilon^{-2} + N^{2/3}\epsilon^{-4/3}\right)$. We claim that it can be attained by $\tau = \Theta\left(\max\{1, N^{1/3}\epsilon^{1/3}\}\right)$. We consider the following two subcases under this choice of τ:

- If $N = O(\epsilon^{-1})$, then $\tau = \Theta(1)$, (3.14) becomes $\widetilde{\Theta}\left(N + \epsilon^{-1} + N\epsilon^{-1} + \epsilon^{-2}\right) = \widetilde{\Theta}\left(\epsilon^{-2}\right) = \widetilde{\Theta}\left(N + N^{2/3}\epsilon^{-4/3} + \epsilon^{-2}\right)$, and (3.15) becomes $\widetilde{\Theta}\left(N + 1 + N\epsilon^{-1} + \epsilon^{-1}\right) = \widetilde{\Theta}\left(N\epsilon^{-1}\right)$ $= \widetilde{\Theta}\left(\min\{N\epsilon^{-1}, N + N^{2/3}\epsilon^{-4/3}\}\right)$.
- If $N = \Omega(\epsilon^{-1})$, then $\tau = \Theta(N^{1/3}\epsilon^{1/3})$, (3.14) becomes $\widetilde{\Theta}\left(N + N^{2/3}\epsilon^{-4/3}\right) = \widetilde{\Theta}\left(N + N^{2/3}\epsilon^{-4/3} + \epsilon^{-2}\right)$, and (3.15) becomes $\widetilde{\Theta}\left(N + N\epsilon + N^{2/3}\epsilon^{-4/3} + N^{2/3}\epsilon^{-1/3}\right) = \widetilde{\Theta}\left(N + N^{2/3}\epsilon^{-4/3}\right)$ $= \widetilde{\Theta}\left(\min\{N\epsilon^{-1}, N + N^{2/3}\epsilon^{-4/3}\}\right)$.

By the two subcases above, when $\tau = \Theta\left(\max\{1, N^{1/3}\epsilon^{1/3}\}\right)$, (3.14) attains the aforementioned minimal asymptotic rate, and (3.15) becomes $\widetilde{\Theta}\left(\min\{N\epsilon^{-1}, N + N^{2/3}\epsilon^{-4/3}\}\right)$, which finishes the proof of part (i).

Next, we consider minimizing the asymptotic rate of (3.15). Note that $\tau \geq 0$ and $N\epsilon^{-1}\tau^{-1} + \epsilon^{-1}\tau^{2} \geq \max\{N\tau^{-1}, \tau^{2}\}\epsilon^{-1} \geq N^{2/3}\epsilon^{-1}$, so we have $N + \tau^3 + N\epsilon^{-1}\tau^{-1} + \epsilon^{-1}\tau^{2} \geq N + N^{2/3}\epsilon^{-1} \geq 0$. So the rate in (3.15) is at least $\widetilde{\Theta}\left(N + N^{2/3}\epsilon^{-1}\right)$. When $\tau = \Theta\left(N^{1/3}\right)$, (3.15) attains $\widetilde{\Theta}\left(N + N^{2/3}\epsilon^{-1}\right)$, and (3.14) becomes $\widetilde{\Theta}\left(N + N\epsilon^{-1} + N^{2/3}\epsilon^{-1} + N^{2/3}\epsilon^{-2}\right) = \widetilde{\Theta}\left(N\epsilon^{-1} + N^{2/3}\epsilon^{-2}\right)$. □

Proofs for $\texttt{estimator}_4$.

Proof of Lemma 3.15. We have $\widetilde{g}_0^k = g(x_0^k)$ and $\widetilde{J}_0^k = g'(x_0^k)$ from (Est$_4$). For any $K \in \mathbb{N}_+$, $\boldsymbol{\tau} \in \mathbb{N}_+^K$ and $\Delta \in (0, 1)$, let $\delta = \frac{\Delta}{2\Sigma_\tau}$. By Proposition A.2, for an arbitrary $(k, i) \in \mathcal{I}(K, \boldsymbol{\tau})$ and any $a \geq \frac{4}{9}\log\left(\frac{m+1}{\delta}\right)$, the following holds with probability at least $1 - \delta$:

$$\left\|\widetilde{g}_i^k - g(x_i^k)\right\|_2 \leq \frac{2L_g}{\sqrt{a}}\sqrt{\log\left(\frac{m+1}{\delta}\right)}\|x_i^k - x_0^k\|_2^2. \tag{A.10}$$

By Proposition 4.11, for an arbitrary $(k, i) \in \mathcal{I}(K, \boldsymbol{\tau})$ and any $b \geq \frac{4}{9}\log\left(\frac{m+n}{\delta}\right)$, the following holds with probability at least $1 - \delta$:

$$\left\|\widetilde{J}_i^k - g'(x_i^k)\right\|_{\text{op}} \leq \frac{4L_g}{\sqrt{b}}\sqrt{\log\left(\frac{m+n}{\delta}\right)}\|x_i^k - x_0^k\|_2. \tag{A.11}$$

Let $\mathcal{C}(K, \boldsymbol{\tau}, \Delta) = \{(a, b) \in \mathbb{N}_+^2 : a \geq \frac{4}{9}\log(\frac{2(m+1)\Sigma_\tau}{\Delta}), b \geq \frac{4}{9}\log(\frac{2(m+n)\Sigma_\tau}{\Delta})\}$. Then for any $(a, b) \in \mathcal{C}(K, \boldsymbol{\tau}, \Delta)$, by using a union probability bound, (A.10) and (A.11) hold for all $(k, i) \in \mathcal{I}(K, \boldsymbol{\tau})$ with probability at least $1 - 2\Sigma_\tau\delta$. Note that $1 - 2\Sigma_\tau\delta = 1 - \Delta$, so we can set $\gamma_0 = \gamma_1 = \lambda_0 = 0$, $\gamma_2(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{2L_g}{\sqrt{a}}\sqrt{\log(\frac{2(m+1)\Sigma_\tau}{\Delta})}$ and $\lambda_1(K, \boldsymbol{\tau}, \theta, \Delta) = \frac{4L_g}{\sqrt{b}}\sqrt{\log(\frac{2(m+n)\Sigma_\tau}{\Delta})}$ to satisfy Assumption 2.3. □

Proof of Corollary 3.16. We can obtain the explicit form of $\mathcal{C}(K, \boldsymbol{\tau}, \Delta)$, $\{\gamma_l\}_{l=0}^2$ and $\{\lambda_l\}_{l=0}^1$ from Lemma 3.15, and plug them into Theorem 3.1. Then by Theorem 3.1, to get $\frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ with probability at least $1 - \Delta$, we need $\overline{\delta} \leq \Delta/(2K\tau)$, $\overline{\epsilon} \leq \epsilon/(5{\cdot}30M)$, and the following inequalities:

$$\left\{\begin{array}{l} a \geq \frac{4}{9}\log\left(\frac{4(m+1)K\tau}{\Delta}\right), \text{and } b \geq \frac{4}{9}\log\left(\frac{4(m+n)K\tau}{\Delta}\right) \\ K\tau \geq 5 \cdot 30M(\Phi(x_0^0) - \Phi^*)/\epsilon \\ (1+\tau)^2\frac{2L_g}{\sqrt{a}}\sqrt{\log\left(\frac{4(m+1)K\tau}{\Delta}\right)} \leq 3L_g/50 \\ (1+\tau)^2\frac{16L_g^2}{b}\log\left(\frac{4(m+n)\Sigma_\tau}{\Delta}\right) \leq 3L_g^2/38 \end{array}\right.$$

So it reduces to

$$\begin{cases} K\tau \geq C_\Sigma \cdot \epsilon^{-1} \\ a \geq C_a \cdot (1+\tau)^4 \cdot \log\left(\frac{4(m+1)K\tau}{\Delta}\right) \\ b \geq C_b \cdot (1+\tau)^2 \cdot \log\left(\frac{4(m+n)K\tau}{\Delta}\right) \end{cases}$$

where C_Σ, C_a, C_b are some constants.

For any positive integer τ, let $K = \lceil \frac{C_\Sigma \cdot \epsilon^{-1}}{\tau} \rceil$, $a = \lceil C_a \cdot (1+\tau)^4 \cdot \log(\frac{4(m+1)K\tau}{\Delta}) \rceil$, $b = \lceil C_b \cdot (1+\tau)^2 \cdot \log(\frac{4(m+n)K\tau}{\Delta}) \rceil$. Then the conditions above are satisfied, so Theorem 3.1 guarantees that $\frac{1}{K\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ with probability at least $1-\Delta$.

In (Est$_4$), at the $(k,0)$-th iteration, we evaluate $g_j(\cdot)$ for N times and $g_j'(\cdot)$ for N times. At the (k,i)-th iteration (with $i>0$), we evaluate $g_j(\cdot)$ for a times and $g_j'(\cdot)$ for b times. So the oracle complexity for evaluations of $g_\xi(\cdot)$ is

$$\begin{aligned} KN + K(\tau-1)a \leq K(N+\tau\cdot a) &\leq (1+\frac{C_\Sigma}{\tau\epsilon})\cdot(N+\widetilde{\Theta}(\tau^5)) \\ &= \widetilde{\Theta}\left(N+\tau^5+N\epsilon^{-1}\tau^{-1}+\epsilon^{-1}\tau^4\right), \end{aligned}$$

and the oracle complexity for evaluations of Jacobians $g_\xi'(\cdot)$ is

$$\begin{aligned} KN + K(\tau-1)b \leq K(N+\tau\cdot b) &\leq (1+\frac{C_\Sigma}{\tau\epsilon})\cdot(N+\widetilde{\Theta}(\tau^3)) \\ &= \widetilde{\Theta}\left(N+\tau^3+N\epsilon^{-1}\tau^{-1}+\epsilon^{-1}\tau^2\right). \end{aligned}$$

□

Proof of Corollary 3.17. Note that $N\epsilon^{-1}\tau^{-1}+\epsilon^{-1}\tau^4 \geq N^{4/5}\epsilon^{-1}$, so we have $N+\tau^5+N\epsilon^{-1}\tau^{-1}+\epsilon^{-1}\tau^4 \geq N+N^{4/5}\epsilon^{-1} \geq 0$. So the rate in (3.16) is at least $\widetilde{\Theta}\left(N+N^{4/5}\epsilon^{-1}\right)$. When $\tau = \Theta\left(N^{1/5}\right)$, (3.16) attains $\widetilde{\Theta}\left(N+N^{4/5}\epsilon^{-1}\right)$ and (3.17) is also $\widetilde{\Theta}\left(N+N^{4/5}\epsilon^{-1}\right)$.

Similarly, by the fact $N\epsilon^{-1}\tau^{-1}+\epsilon^{-1}\tau^2 \geq N^{2/3}\epsilon^{-1}$, we have $N+\tau^3+N\epsilon^{-1}\tau^{-1}+\epsilon^{-1}\tau^2 \geq N+N^{2/3}\epsilon^{-1} \geq 0$. So the rate in (3.17) is at least $\widetilde{\Theta}\left(N+N^{2/3}\epsilon^{-1}\right)$. When $\tau = \Theta\left(N^{1/3}\right)$, (3.17) attains $\widetilde{\Theta}\left(N+N^{2/3}\epsilon^{-1}\right)$ and (3.16) becomes $\widetilde{\Theta}\left(N^{5/3}+N^{4/3}\epsilon^{-1}\right)$. □

A.3 Proof for Randomized Epoch Durations

Proof of Corollary 3.20. We first analyze part (i) in a similar way as the proof of Corollary 3.6. Use the explicit form of $\mathcal{C}(K,\boldsymbol{\tau},\Delta)$, $\{\gamma_l\}_{l=0}^2$ and $\{\lambda_l\}_{l=0}^1$ from Lemma 3.5, and plug them into Theorem 3.1. Then by Theorem 3.1, to get $\frac{1}{\Sigma_\tau}\sum_{k=0}^{K-1}\sum_{i=0}^{\tau_k-1}\|\mathcal{G}_M(x_i^k)\|_2^2 \leq \epsilon$ with probability at least $1-\Delta$, we need

$$\overline{\delta} \leq \Delta/(2\Sigma_\tau) \quad \text{and} \quad \overline{\epsilon} \leq \epsilon/(5\cdot 30M) \tag{A.12}$$

and the following inequalities:

$$\begin{cases} A, a \geq \frac{4}{9}\log\left(\frac{4(m+1)\Sigma_\tau}{\Delta}\right), \text{and } B, b \geq \frac{4}{9}\log\left(\frac{4(m+n)\Sigma_\tau}{\Delta}\right) \\ \Sigma_\tau \geq 5 \cdot 30M(\Phi(x_0^0) - \Phi^*)/\epsilon \\ \frac{2\sigma_g}{\sqrt{A}}\sqrt{\log\left(\frac{4(m+1)\Sigma_\tau}{\Delta}\right)} \leq \epsilon/(5 \cdot 125 l_f M) \\ \frac{4\sigma_{g'}^2}{B}\log\left(\frac{4(m+n)\Sigma_\tau}{\Delta}\right) \leq L_g\epsilon/(5 \cdot 95 l_f M) \\ (1+\tau_{\max})^2\frac{16 l_g^2}{a}\log\left(\frac{4(m+1)\Sigma_\tau}{\Delta}\right) \leq L_g\epsilon/(5 \cdot 525 l_f M) \\ (1+\tau_{\max})^2\frac{16 L_g^2}{b}\log\left(\frac{4(m+n)\Sigma_\tau}{\Delta}\right) \leq 3L_g^2/38 \end{cases} \tag{A.13}$$

By (3.18) and Assumption 3.19, the constructed $\boldsymbol{\tau}$ satisfies $\tau_{\max} \leq \tau_+$ and $S_\tau \leq \Sigma_\tau \leq S_\tau + \tau_+$. So the choices $\overline{\delta} = \frac{\Delta}{2(S_\tau+\tau_+)}$ and $\overline{\epsilon} = \epsilon/(5 \cdot 30M)$ imply (A.12), and the following inequalities suffice to imply (A.13):

$$\begin{cases} S_\tau \geq C_\Sigma \cdot \epsilon^{-1} \\ A \geq C_A \cdot \epsilon^{-2} \cdot \log\left(\frac{4(m+1)(S_\tau+\tau_+)}{\Delta}\right) \\ B \geq C_B \cdot \epsilon^{-1} \cdot \log\left(\frac{4(m+n)(S_\tau+\tau_+)}{\Delta}\right) \\ a \geq C_a \cdot (1+\tau_+)^2 \cdot \epsilon^{-1} \cdot \log\left(\frac{4(m+1)(S_\tau+\tau_+)}{\Delta}\right) \\ b \geq C_b \cdot (1+\tau_+)^2 \cdot \log\left(\frac{4(m+n)(S_\tau+\tau_+)}{\Delta}\right) \end{cases} \tag{A.14}$$

providing that $1/\epsilon$ is sufficiently large. Here $C_\Sigma, C_A, C_B, C_a, C_b$ are some constants.

Let $\tau_+ = \lceil \epsilon^{-1/3} \rceil$, $S_\tau = \lceil C_\Sigma \cdot \epsilon^{-1} \rceil$, $A = \lceil C_A \cdot \epsilon^{-2} \cdot \log(\frac{5(m+1)S_\tau}{\Delta}) \rceil$, $B = \lceil C_B \cdot \epsilon^{-1} \cdot \log(\frac{5(m+n)S_\tau}{\Delta}) \rceil$, $a = \lceil C_a \cdot (1+\tau_+)^2 \cdot \epsilon^{-1} \cdot \log(\frac{5(m+1)S_\tau}{\Delta}) \rceil$, $b = \lceil C_b \cdot (1+\tau_+)^2 \cdot \log(\frac{5(m+n)S_\tau}{\Delta}) \rceil$, then (A.14) holds for sufficiently small ϵ. So part (i) of Corollary 3.20 holds with probability at least $1 - \Delta$ by choosing these parameters.

In the rest of the proof, we analyze part (ii) of Corollary 3.20. We need to provide a high probability upper bound on K. For any positive integer M, (3.18) implies that $K > M$ if and only if $\sum_{k=0}^{M-1} \tau_k < S_\tau$, so $\mathbb{P}(K > M) = \mathbb{P}\left(\sum_{k=0}^{M-1} \tau_k < S_\tau\right)$. Note that the random variables $\{\tau_k\}$ are independent and bounded between $[0, \tau_+]$, so we can use Hoeffding's Inequality. Denote $\mu_\tau := \mathbb{E}_{\tau \sim D_\tau(\cdot;\tau_+,\theta_\tau)}[\tau]$. By Lemma A.1, for any $t \geq 0$,

$$\mathbb{P}\left(\sum_{k=0}^{M-1} \tau_k \leq M\mu_\tau - t\right) \leq \exp\left(-\frac{2t^2}{M\tau_+^2}\right). \tag{A.15}$$

Let $M = \lceil \frac{2C_\tau S_\tau}{\tau_+} \rceil$. By Assumption 3.19, $\mu_\tau > 0$ and $C_\tau \mu_\tau \geq \tau_+ > 0$, so $M \geq \frac{2C_\tau S_\tau}{\tau_+} \geq \frac{2S_\tau}{\mu_\tau} > \frac{S_\tau}{\mu_\tau}$. Let $t = M\mu_\tau - S_\tau \geq 0$ in (A.15), then we get

$$\mathbb{P}(K > M) = \mathbb{P}\left(\sum_{k=0}^{M-1} \tau_k < S_\tau\right) \leq \mathbb{P}\left(\sum_{k=0}^{M-1} \tau_k \leq S_\tau\right) \leq \exp\left(-\frac{2(M\mu_\tau - S_\tau)^2}{M\tau_+^2}\right). \tag{A.16}$$

Note that the mapping $\phi(x) = \frac{1}{x}(x\mu_\tau - S_\tau)^2$ is increasing on the interval $[\frac{S_\tau}{\mu_\tau}, +\infty)$, so $\phi(M) \geq \phi(\frac{2C_\tau S_\tau}{\tau_+})$, which further implies

$$\begin{aligned} \exp\left(-\frac{2(M\mu_\tau - S_\tau)^2}{M\tau_+^2}\right) &\leq \exp\left(-\frac{2}{(\frac{2C_\tau S_\tau}{\tau_+})\tau_+^2}\left(\frac{2C_\tau S_\tau}{\tau_+}\mu_\tau - S_\tau\right)^2\right) \\ &= \exp\left(-\frac{S_\tau}{C_\tau \tau_+}\left(\frac{2C_\tau \mu_\tau}{\tau_+} - 1\right)^2\right) \leq \exp\left(-\frac{S_\tau}{C_\tau \tau_+}\right) \end{aligned} \tag{A.17}$$

where the last step is because $\frac{2C_\tau\mu_\tau}{\tau_+}-1\geq 1$. By (A.16) and (A.17), $\mathbb{P}(K>\lceil\frac{2C_\tau S_\tau}{\tau_+}\rceil)\leq \exp\left(-\frac{S_\tau}{C_\tau\tau_+}\right)$. So $K\leq\lceil\frac{2C_\tau S_\tau}{\tau_+}\rceil=\Theta(\epsilon^{-2/3})$ with probability at least $1-\exp(-\frac{S_\tau}{C_\tau\tau_+})\geq 1-\exp(-C_p\epsilon^{-2/3})$ for some constant C_p.

In (Est$_1$), at the $(k,0)$-th iteration, we evaluate $g_\xi(\cdot)$ for A times and $g'_\xi(\cdot)$ for B times. At the (k,i)-th iteration (with $i>0$), we evaluate $g_\xi(\cdot)$ for $2a$ times and $g'_\xi(\cdot)$ for $2b$ times. On the high probability set where $K=O(\epsilon^{-2/3})$, the oracle complexity for evaluations of $g_\xi(\cdot)$ is

$$KA+2(\Sigma_\tau-K)a\leq KA+2\Sigma_\tau\cdot a\leq KA+2(S_\tau+\tau_+)a=\widetilde{O}(\epsilon^{-8/3}\log(1/\Delta)),$$

and the oracle complexity for evaluations of Jacobians $g'_\xi(\cdot)$ is

$$KB+2(\Sigma_\tau-K)b\leq KB+2\Sigma_\tau\cdot b\leq KB+2(S_\tau+\tau_+)b=\widetilde{O}(\epsilon^{-5/3}\log(1/\Delta)).$$

Using a union probability bound for part (i) and part (ii), they hold simultaneously with probability at least $1-\Delta-\exp(-C_p\epsilon^{-2/3})$, which finishes the proof. □