HeteRSGD: Tackling Heterogeneous Sampling Costs via Optimal Reweighted Stochastic Gradient Descent

Ziang Chen Duke University
Jianfeng Lu Duke University
Huajie Qian Alibaba Group
Xinshang Wang Alibaba Group
Wotao Yin Alibaba Group

Abstract

One implicit assumption in current stochastic gradient descent (SGD) algorithms is the identical cost for sampling each component function of the finite-sum objective. However, there are applications where the costs differ substantially, for which SGD schemes with uniform sampling invoke a high sampling load. We investigate the use of importance sampling (IS) as a cost saver in this setting, in contrast to its traditional use for variance reduction. The key ingredient is a novel efficiency metric for IS that advocates low sampling costs while penalizing high gradient variances. We then propose HeteRSGD, an SGD scheme that performs gradient sampling according to optimal probability weights stipulated by the metric, and establish theories on its optimal asymptotic and finite-time convergence rates among all possible IS-based SGD schemes. We show that the relative efficiency gain of HeteRSGD can be arbitrarily large regardless of the problem dimension and number of components. Our theoretical results are validated numerically for both convex and nonconvex problems.

1 INTRODUCTION

We consider the finite-sum optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1.1}$$

where each f_i itself can be in the form of a finite sum. Such problems are ubiquitous in machine learning (Bishop and Nasrabadi, 2006), operations research (Birge and Louveaux, 2011), and statistics (Box and Tiao, 2011). Unlike most works that implicitly assume the identical cost for sampling each component f_i, we consider the case of heterogeneous sampling costs, i.e., some components can be considerably more costly to sample than others. Such sampling cost heterogeneity arises in various applications. In automated algorithm configuration (e.g., Hoos, 2011), parameters of an algorithm are optimized on a set of problem instances that are of different difficulty levels and hence incur highly varying time or resource costs when fed to the algorithm. In stochastic programming (Birge and Louveaux, 2011), the evaluation of the components requires solving subproblems among which some can be harder to solve than others. Lastly, in federated learning (Diao et al., 2020; Luo et al., 2022), the components can represent clients with varying model architectures, data sizes, and computation and communication capabilities.

In the case of homogeneous sampling costs, when the full gradient is expensive to evaluate, the preferred optimization method is stochastic gradient descent (SGD) (Robbins and Monro, 1951) that samples one or several functions uniformly at random to approximate the full gradient at each iteration. In the heterogeneous setting, however, SGD can be inefficient because it can sample costly components with a substantial chance at each iteration and thus incur high sampling costs and consequently slow convergence, or even become impractical when the sampling costs differ drastically. This paper thus aims to address the problem:

How do we design gradient-based schemes with a much lighter sampling burden than SGD under heterogeneous sampling costs? (P)

We attempt to tackle (P) using importance sampling (IS), a technique from Monte Carlo computation (Rubinstein and Kroese, 2016, Chapter 5) that samples from a different distribution than the original one and then corrects biases by reweighting. In contrast to the typical use of IS as a variance reducer, we use IS as a cost saver by sampling costly components less frequently. To explain, in our finite-sum setting, IS samples the components according to possibly non-uniform probability weights $\{p_i\}_{i=1}^n$ and then reweights the samples with factors $\{\frac{1}{np_i}\}_{i=1}^n$. In order to achieve the least average cost per gradient evaluation, a naive scheme is to use $p_i \approx 1$ for the cheapest component

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

and $p_i \approx 0$ otherwise; however, this will blow up the gradient variance and hinder convergence. Relatedly, IS has traditionally been used in SGD to reduce gradient estimation variance in order to accelerate convergence (Needell et al., 2016; Papa et al., 2015; El Hanchi et al., 2022, etc.). Despite their better control of the variance and consequently faster convergence than standard SGD, these IS schemes are designed based on only the magnitude of the gradients or smoothness constants of the components, but not their sampling costs, and thus can still be inefficient under heterogeneous sampling costs.

Our main contribution thus lies in a novel IS scheme for (P) that directly reduces sampling costs using as small sampling weights on costly components as possible while controlling the variance. The key ingredient is a judiciously designed efficiency metric that, for a given sampling distribution, balances the impacts of the average cost in sampling a component and the gradient estimation variance on the overall sampling requirement. Specifically, it takes the the form of the product of the cost and the variance, and hence penalizes both high costs and variances. This particular form is motivated from an estimation of the sampling effort needed to achieve a certain amount of error reduction over a single SGD iteration. Our importance sampler is then obtained by optimizing the metric, for which we provide efficient routines.

Based on the proposed efficiency metric, we design HeteRSGD, a new SGD algorithm that adaptively estimates the optimal sampling weights in each iteration and performs gradient sampling according to the estimated weights. To properly characterize its convergence in our heterogeneous setting, we establish novel central limit theorems (CLTs) that scale the solution error with the (random) cumulative sampling cost instead of the number of iterations in previous CLTs. It turns out that the asymptotic errors of the Polyak-Ruppert (Polyak and Juditsky, 1992; Ruppert, 1988) and the α-suffix (Rakhlin et al., 2012) averaged solutions exactly match our efficiency metric, implying the optimality of HeteRSGD among all IS-based SGD schemes in the sense that it achieves the least asymptotic solution error under a given sampling budget. Moreover, the efficiency gain relative to the standard and other IS-based SGD can be arbitrarily large, regardless of the dimension d and the number of components n.

Lastly, we further generalize our efficiency metric to a parametric family with varying preferences between cost reduction and variance reduction, and each of them matches the asymptotic error of an individual SGD iterate under a corresponding decay rate in the step size. This gives rise to a family of HeteRSGD variants with each being optimal for individual SGD iterates instead of averaged ones.

We summarize our main contributions in this paper:

1. We propose a novel family of efficiency metrics for IS that balance sampling cost reduction and variance reduction under sampling cost heterogeneity.
2. We design a family of IS-based SGD algorithms, called HeteRSGD, and develop novel asymptotic CLTs and finite-time convergence bounds in the strongly convex and smooth case that reveal the optimality of HeteRSGD in attaining the least sampling complexity among all possible IS-based schemes.
3. We conduct experiments on both convex and non-convex examples that demonstrate a 40-70% reduction in sampling cost compared to existing SGD methods in order to achieve similar solution accuracy.

Related Work There have been extensive studies on integrating IS into SGD for variance reduction. Needell et al. (2016); Zhao and Zhang (2015); Gower et al. (2019); Csiba and Richtárik (2018); Katharopoulos and Fleuret (2018) design importance samplers based on global smoothness information such as Lipschitz constants and bounds of gradient norms to obtain improved convergence rates. El Hanchi et al. (2022); Papa et al. (2015); Yuan et al. (2016); He et al. (2021); Liu et al. (2021); Gopal (2016); Alain et al. (2015); Stich et al. (2017); Johnson and Guestrin (2018) develop sampling methods that adaptively approximate the ideal sampler (2.3) as the iteration progresses to further reduce the gradient estimation variance and sampling complexities. Another orthogonal line of works (Borsos et al., 2018; El Hanchi and Stephens, 2020; Namkoong et al., 2017; Salehi et al., 2017) pose the search of optimal sampling weights as an online learning problem and provide regret bounds with respect to the best weights in hindsight. However, these works assume homogeneous sampling costs and thus can be inefficient in our heterogeneous setting. Recently, An and Ying (2021) also utilizes IS to balance the gradient variance across solutions in order to escape from flat local optima in non-convex settings.

Apart from using IS, a large family of variance-reduced SGD algorithms (e.g, Defazio et al., 2014; Schmidt et al., 2017; Johnson and Zhang, 2013; Allen-Zhu, 2017) build on the idea of control variates. Besides the assumed homogeneity in sampling costs, these methods also incur a significant overhead for either storing gradients of all components or periodic evaluation of the full gradient, whereas our approach only needs to maintain norms of gradients and possibly an estimated full gradient. Other works (e.g., Horváth and Richtárik, 2019; Qian et al., 2021; Shen et al., 2016) combine IS with variance-reduced SGD to further speed up convergence.

Lastly, similar sampling heterogeneity also arises in federated learning (e.g, Diao et al., 2020; Shen et al., 2022; Cho et al., 2022). In particular, Luo et al. (2022) considers system and statistical heterogeneity among clients and proposes adaptive sampling to minimize global convergence time. Their sampling designs are specialized to federated

learning rather than general finite-sum optimization.

Organization The rest of this paper will be organized as follows. Section 2 discusses our novel IS efficiency metric under sampling heterogeneity and the corresponding importance sampler. Section 3 presents our SGD algorithm, with theories on its asymptotic optimality and finite-time convergence rates in Section 4, and Section 5 further discusses extensions to a family of importance samplers and their optimality theories. Section 6 contains experimental results, and Section 7 concludes the paper.

2 IS UNDER HETEROGENEOUS COSTS

We first introduce the design rationale of importance samplers in the homogeneous setting and then present our novel sampling efficiency metric and the associated optimal weights for heterogeneous sampling costs.

Notations Throughout this paper, we use $\|\cdot\|$ as the ℓ_2-norm on $\mathbb{R}^d$ and $\langle\cdot,\cdot\rangle$ as the associated inner product. Denote by $\Delta_n = \{(p_1, p_2, \ldots, p_n) \in [0,1]^n : \sum_{i=1}^n p_i = 1\}$ the probability simplex in $\mathbb{R}^n$. We always set $0/0 = 0$.

Given the k-th SGD iterate x_k, we choose a probability distribution $p^k = (p_1^k, p_2^k, \ldots, p_n^k) \in \Delta_n$, and sample a multi-set $\mathcal{I}_k$ in which each index is i.i.d. drawn with replacement from $\{1, 2, \ldots, n\}$ with probability $\mathbb{P}(\cdot = i) = p_i^k, \forall\, i$. Then the gradient $\nabla f(x_k)$ can be estimated using

$$g_k = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \frac{1}{n p_i^k} \nabla f_i(x_k). \tag{2.1}$$

g_k is an unbaised estimator of $\nabla f(x_k)$, i.e., $\mathbb{E}[g_k | \mathcal{F}_{k-1}] = \nabla f(x_k)$, where $\mathcal{F}_{k-1}$ is the σ-algebra generated by $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_{k-1}$. We summarize the general framework for SGD with adaptive sampling in Algorithm 1. A natural idea for finding the optimal weights p^k is to minimize the variance of the gradient estimator (2.1), which can be computed as

Algorithm 1 SGD with adaptive sampling

Require: initial point x_1 and stepsize $\{\alpha_k\}_{k=1}^\infty$.
1: **for** $k = 1, 2, \ldots$ **do**
2: Choose $p^k \in \Delta_n$ and sample the index set $\mathcal{I}_k$.
3: Compute the gradient estimate g_k via (2.1).
4: Update the iterate $x_{k+1} = x_k - \alpha_k g_k$.
5: **end for**

$$\begin{aligned} &\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2 \middle| \mathcal{F}_{k-1}\right] \\ =&\frac{1}{|\mathcal{I}_k| n^2} \sum_{i=1}^n \frac{1}{p_i^k} \|\nabla f_i(x_k)\|^2 - \frac{1}{|\mathcal{I}_k|} \|\nabla f(x_k)\|^2. \end{aligned} \tag{2.2}$$

The minimizing weights can then be shown to be (Zhao and Zhang, 2015):

$$p_i^k \propto \|\nabla f_i(x_k)\|, \tag{2.3}$$

which have been extensively studied to improve convergence rates (e.g., El Hanchi et al., 2022; Papa et al., 2015). Note that computing such ideal weights requires knowledge of the gradient of every single component and hence SGD algorithms that use IS for variance reduction rely on approximations of (2.3).

The derivation of (2.3) implicitly assumes the identical cost in sampling each component gradient. In our setting with heterogeneous sampling costs, however, the cost of evaluating $\nabla f_i(x)$ varies in i. Thus, minimizing the variance solely does not necessarily lead to less sampling effort, e.g., when components with large gradient norms happen to be costly to sample, and one has to jointly consider the gradient variance and the incurred sampling costs. To proceed, we consider the following cost model:

Assumption 2.1. *The cost for evaluating each ∇f_i, $i = 1, \ldots, n$, is a random variable $\hat{c}_i > 0$ with $c_i := \mathbb{E}[\hat{c}_i] < \infty$, and the total sampling cost is cumulative.*

To find a metric that can meaningfully measure the efficiency of a given sampling distribution p_k under Assumption 2.1, we examine the solution error reduction in a single SGD step, as done in earlier works (e.g., Papa et al., 2015; Johnson and Guestrin, 2018)

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x^*\|^2 | \mathcal{F}_{k-1}] = &\|x_k - x^* - \alpha_k \nabla f(x_k)\|^2 \\ &+ \alpha_k^2 \mathbb{E}[\|g_k - \nabla f(x_k)\|^2 | \mathcal{F}_{k-1}], \end{aligned}$$

where x^* is an optimum, and only the last term, with the gradient variance given by (2.2), depends on p^k. The average cost of sampling a single gradient according to p^k is $\sum_{i=1}^n p_i^k c_i$. We aim to minimize the average cost while ensuring a certain amount of reduction, or formally

$$\begin{aligned} &\min_{p^k, |\mathcal{I}_k|} |\mathcal{I}_k| \sum_{i=1}^n p_i^k c_i, \\ &\text{s.t. } \frac{\alpha_k^2}{|\mathcal{I}_k|} \left(\sum_{i=1}^n \frac{1}{n^2 p_i^k} \|\nabla f_i(x_k)\|^2 - \|\nabla f(x_k)\|^2 \right) \le \epsilon, \end{aligned}$$

for a fixed $\epsilon > 0$. By relaxing the integrality constraint on $|\mathcal{I}_k|$ and optimizing out $|\mathcal{I}_k|$, we see immediately that the above is equivalent to minimizing

$$\left(\sum_{i=1}^n p_i^k c_i \right) \left(\sum_{i=1}^n \frac{\|\nabla f_i(x_k)\|^2}{n^2 p_i^k} - \|\nabla f(x_k)\|^2 \right), \tag{2.4}$$

after dropping the constants ϵ and α_k. Compared to the homogeneous case, our new efficiency metric (2.4) penalizes both high variance and high sampling cost, and thus balances their impacts on the overall sampling efficiency. (2.4) is computationally more challenging though due to its non-convexity. Fortunately it turns out readily solvable:

Proposition 2.2. *Let $c_i > 0, b_i \ge 0$ for all $i = 1, \ldots, n$, $0 \le b_0 \le (\sum_{i=1}^n \sqrt{b_i}/n)^2$, and consider*

$$\min_{p \in \Delta_n} \left(\sum_{i=1}^n p_i c_i \right) \left(\sum_{i=1}^n \frac{b_i}{n^2 p_i} - b_0 \right). \tag{2.5}$$

If at least one $b_i > 0$, then there exists a unique minimizer

p^* for (2.5) *and is given by*

$$p_i^* = \sqrt{\frac{b_i/n^2}{\kappa^* c_i + b_0}}, \quad i = 1, 2, \ldots, n, \tag{2.6}$$

where $\kappa^* \geq 0$ uniquely solves $\sum_{i=1}^n \sqrt{\frac{b_i/n^2}{\kappa^* c_i + b_0}} = 1$. Otherwise if all $b_i = 0$, then (2.5) *is constantly* 0.

Since the $\sum_{i=1}^n \sqrt{\frac{b_i/n^2}{\kappa^* c_i + b_0}}$ is strictly monotonic in κ^*, the desired κ^* can be computed by bisection, and then the optimal weights under the metric (2.4) can be computed from (2.6) with $b_i = \|\nabla f_i(x_k)\|^2$ and $b_0 = \|\nabla f(x_k)\|^2$. The proof of Proposition 2.2 is deferred to Appendix F.

3 THE HeteRSGD ALGORITHM

The oracle efficiency metric (2.4) that guides the choice of p^k is constructed with perfect knowledge about c_i, $\|\nabla f_i(x_k)\|$, $1 \leq i \leq n$, and $\|\nabla f(x_k)\|$, which may not be available in practice. Therefore, we propose a practical approximation of the oracle metric from which the sampling distribution p^k is then derived.

Estimation of c_i: The estimated cost vector $\tilde{c}^k = (\tilde{c}_1^k, \ldots, \tilde{c}_n^k)$, where each $\tilde{c}_i^k$ is the average cost incurred by all the sampled gradients from f_i. Specifically, let $s_i^k, i = 1, \ldots, n$ be the number of times that each ∇f_i has been sampled so far at the beginning of the k-th iteration. Let $\hat{c}_{i,j}$ be the random cost of the j-th sample taken for ∇f_i throughout the algorithm. The cost vector is updated via $s_i^{k+1} = s_i^k + \sum_{j \in \mathcal{I}_k} \mathbf{1}(j = i)$ and

$$\tilde{c}_i^{k+1} = \frac{1}{s_i^{k+1}} \left(s_i^k \tilde{c}_i^k + \sum_{j=s_i^k+1}^{s_i^{k+1}} \hat{c}_{i,j} \right). \tag{3.1}$$

Estimation of $\|\nabla f_i(x_k)\|$: Each $\|\nabla f_i(x_k)\|$ is estimated with the most recently sampled gradient from f_i. We use a vector $\tilde{g}^k = (\tilde{g}_1^k, \ldots, \tilde{g}_n^k)$ to store the estimates which is updated via

$$\tilde{g}_i^{k+1} = \begin{cases} \|\nabla f_i(x_k)\|, & \text{if } i \in \mathcal{I}_k, \\ \tilde{g}_i^k, & \text{otherwise.} \end{cases} \tag{3.2}$$

Estimation of $\|\nabla f(x_k)\|$: We estimate $\nabla f(x_k)$ with the averaged gradient $\tilde{G}_k = \frac{1}{k-1}(g_1 + g_2 + \cdots + g_{k-1}) \in \mathbb{R}^d$. To avoid the storage of past gradients, we update $\tilde{G}_k$ via

$$\tilde{G}_{k+1} = \frac{k-1}{k} \tilde{G}_k + \frac{1}{k} g_k. \tag{3.3}$$

Now one can construct an empirical version of (2.4):

$$\left(\sum_{i=1}^n p_i^k \tilde{c}_i^k \right) \left(\sum_{i=1}^n \frac{(\tilde{g}_i^k)^2}{n^2 p_i^k} - \min \left(\|\tilde{G}_k\|, \sum_{i=1}^n \frac{\tilde{g}_i^k}{n} \right)^2 \right), \tag{3.4}$$

where the minimum of $\|\tilde{G}_k\|$ and $\sum_{i=1}^n \tilde{g}_i^k/n$ instead of simply $\|\tilde{G}_k\|$ ensures non-negativeness of the variance term. By minimizing (3.4) as stated in Proposition 2.2, we obtain the estimated optimal sampling distribution. In addition, to ensure a sufficient chance for each component to be sampled and a controlled gradient variance, we slightly mix the estimated weights with the uniform weights to keep them away from zero as in some earlier works (e.g., El Hanchi et al., 2022; Papa et al., 2015; Delyon and Portier, 2021). We summarize our SGD algorithm in Algorithm 2, which is an implementation of the template Algorithm 1.

Algorithm 2 HeteRSGD: SGD under heterogeneous costs

Require: initial point x_1, initial estimates $\tilde{c}^1, \tilde{g}^1, \tilde{G}_1 = 0$, stepsizes $\{\alpha_k\}_{k=1}^\infty$, and mixing weight $\{w_k\}_{k=1}^\infty$.

1: **for** $k = 1, 2, \ldots$ **do**
2: Compute the $\tilde{p}^k \in \Delta_n$ minimizing (3.4).
3: Set $p^k = (1 - w_k)\tilde{p}^k + w_k(1/n, \ldots, 1/n)$.
4: Sample the index set $\mathcal{I}_k$ and incur sampling costs.
5: Compute the gradient estimate g_k via (2.1).
6: Update the iterate $x_{k+1} = x_k - \alpha_k g_k$.
7: Update $\tilde{c}^{k+1}$, $\tilde{g}^{k+1}$, and $\tilde{G}_{k+1}$ according to (3.1), (3.2), and (3.3).
8: **end for**

4 CONVERGENCE ANALYSIS

In this section, we provide convergence analysis for Algorithms 1 and 2, and demonstrate the optimal convergence rate and sampling complexity of HeteRSGD among all IS-based schemes. We first present our asymptotic analysis (Subsections 4.1-4.2) on exact convergence rates of the template Algorithm 1, and specialize the result to different IS-based algorithms to illustrate the optimality of HeteRSGD (Subsection 4.3). We then investigate the non-asymptotic behavior of HeteRSGD and compare it with that of the standard SGD (Section 4.4). We assume that the objective function is μ-strongly convex and L-smooth:

Assumption 4.1. *We assume*

(i) f is μ-strongly convex, i.e, it holds that $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$ for any $x, y \in \mathbb{R}^d$.

(ii) f_i is L-smooth for any $i \in \{1, 2, \ldots, n\}$, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$ holds for any $x, y \in \mathbb{R}^d$. As a consequence, f is also L-smooth.

4.1 Global Convergence

We state the result of global convergence in this subsection. Denote by

$$\xi_k = g_k - \nabla f(x_k),$$

the noise or error in the gradient estimator. If ξ_k satisfies some summable property, then one can show that Algorithm 1 converges to the global minimum x^*, both in L^2 and almost surely, with proper diminishing stepsizes:

Theorem 4.2. *Suppose that Assumption 4.1,* $\sum_{k=1}^{\infty}\alpha_k = \infty$, *and* $\sum_{k=1}^{\infty}\alpha_k^2\mathbb{E}[\|\xi_k\|^2] < \infty$ *hold. Then the solution* x_k *from Algorithm 1 satisfies* $\lim_{k\to\infty}\mathbb{E}\left[\|x_k - x^*\|^2\right] = 0$ *and* $x_k \to x^*$ *a.s..*

A key condition in Theorem 4.2 is $\sum_{k=1}^{\infty}\alpha_k^2\mathbb{E}[\|\xi_k\|^2] < \infty$. Although this is not easy to verify in general, we provide transparent sufficient conditions in the lemma below.

Lemma 4.3. *Suppose that Assumption 4.1 holds. If there exists a sequence* $\{w_k\}_{k=1}^{\infty} \subset (0,1]$ *satisfying* $p_i^k \geq w_k/n,\ \forall\ i \in \{1,2,\ldots,n\},\ k \geq 1$, *and* $\sum_{k=1}^{\infty}\alpha_k^2/w_k < \infty$, *then it holds that* $\sum_{k=1}^{\infty}\alpha_k^2\mathbb{E}[\|\xi_k\|^2] < \infty$ *for Algorithm 1.*

Thanks to the mixing weight w_k with the uniform sampling weights in Algorithm 2, we immediately see that our HeteRSGD converges globally as long as w_k is chosen to decay sufficiently slowly as described in Lemma 4.3. Proofs of Theorem 4.2 and Lemma 4.3 are left to Appendix A.

4.2 Local Convergence

In this subsection, we investigate exact local/asymptotic convergence rates of Algorithm 1 for averaged SGD solutions, which lay the foundation for comparing different IS-based SGD algorithms in next subsection. Before proceeding, let us introduce some notations:

- $\mathrm{cost}_k := \sum_{i=1}^{n}\sum_{j=1}^{s_i^k}\hat{c}_{i,j}$ is the cumulative sampling cost of the first $k-1$ iterations, i.e., the cost to generate $\{x_1, x_2, \ldots, x_k\}$.
- $c(p) := \sum_{i=1}^{n} p_i c_i$ is the expected sampling cost of a single gradient evaluation for a distribution $p \in \Delta_n$.
- The covariance matrix of an importance sampled gradient at x^* according to a distribution $p \in \Delta_n$ is:
$$G(p) := \sum_{i=1}^{n}\frac{1}{n^2 p_i}\nabla f_i(x^*)\nabla f_i(x^*)^T.$$
- The averaged iterate of $x_{[\gamma k]+1}, x_{[\gamma k]+1}, \ldots, x_k$ is:
$$\bar{x}_{k,\gamma} = \frac{1}{(1-\gamma)k}\sum_{j=[\gamma k]+1}^{k} x_j,$$
where $\gamma \in [0,1)$ and $[\gamma k]$ is the largest integer that is smaller than or equal to γk. $\gamma = 0$ corresponds to the Polyak-Ruppert averaging, and $\gamma \in (0,1)$ corresponds to the α-suffix averaging.
- $H = \nabla^2 f(x^*)$ is the Hessian matrix of f at x^*.

We make two more assumptions. The first is on the non-degeneracy of the gradient noise at the optimum x^*:

Assumption 4.4. *There exists at least one* $i \in \{1,2,\ldots,n\}$ *such that* $\nabla f_i(x^*) \neq 0$.

The second assumption is the convergence of the sampling distribution p^k:

Assumption 4.5. *The sequence* $\{p^k\}_{k=1}^{\infty} \subset \Delta_n$ *converges almost surely to some fixed* $p^* \in \Delta_n$, *and* $p_i^* > 0$ *for every* i *such that* $\nabla f_i(x^*) \neq 0$.

The positiveness condition on the limit weights ensures that the limit is an eligible importance sampler at x^*. This assumption trivially holds for the standard SGD that performs uniform sampling throughout, as well as for many SGD variants that adaptively approximate the optimal importance weights (2.3) at x^*, e.g., those proposed in Papa et al. (2015); El Hanchi et al. (2022). It also holds for our HeteRSGD:

Proposition 4.6. *Assume the same conditions in Lemma 4.3,* $\lim_{k\to\infty} w_k = 0$, $\sum_{k=1}^{\infty} w_k = \infty$, $\inf_{k\geq 1} k\alpha_k > 0$, *and Assumption 4.4 holds. Then* $p^k \to p_{Hete}^*$ *a.s. for Algorithm 2, where*
$$p_{Hete}^* := \left(\frac{\|\nabla f_i(x^*)\|/\sqrt{c_i}}{\sum_{j=1}^{n}\|\nabla f_j(x^*)\|/\sqrt{c_j}}\right)_{i=1}^{n}$$
minimizes the sampling efficiency metric at x^*
$$\rho(p) := \left(\sum_{i=1}^{n} p_i c_i\right)\left(\sum_{i=1}^{n}\frac{1}{n^2 p_i}\|\nabla f_i(x^*)\|^2\right). \quad (4.1)$$

Proposition 4.6 can be shown by proving $\lim_{k\to\infty}\tilde{c}^k \to (c_1,\ldots,c_n)$, $\lim_{k\to\infty}\tilde{g}_i^k = \|\nabla f_i(x^*)\|$, $\forall\ i \in \{1,2,\ldots,n\}$, and $\lim_{k\to\infty}\tilde{G}_k = 0$, almost surely, with the details deferred to Appendix C.

An immediate consequence of Assumption 4.5 is the convergence of the average cost per gradient sample (with proof in Appendix C):

Proposition 4.7. *If Assumptions 2.1 and 4.5 hold, then* $\mathrm{cost}_k/(\sum_{j=1}^{k-1}|\mathcal{I}_j|) \to c(p^*)$ *a.s. for Algorithm 1.*

We then have the following asymptotic convergence rate for Algorithm 1:

Theorem 4.8. *Suppose Assumptions 2.1, 4.1, 4.4 and 4.5 hold. Suppose in addition that* $\alpha_k = \alpha_1/k^{\beta}$, *where* $\beta \in (1/2, 1)$, $|\mathcal{I}_k| = |\mathcal{I}|$ *is fixed for any* $k \geq 1$, *and* f *is twice continuously differentiable in a neighbourhood of* x^*. *If there exists a non-increasing sequence* $\{w_k\}_{k=1}^{\infty} \subset (0,1]$ *satisfying* $p_i^k \geq w_k/n,\ \forall\ i \in \{1,2,\ldots,n\}$ *and* $k \geq 1$, $\lim_{k\to\infty}\alpha_k/w_k^2 = 0$ *and* $\sum_{k=1}^{\infty}\alpha_k/(w_k\sqrt{k}) < \infty$, *then for Algorithm 1 the following holds*
$$\sqrt{\mathrm{cost}_k}\cdot(\bar{x}_{k,\gamma} - x^*) \Rightarrow \mathcal{N}\left(0, \frac{c(p^*)}{1-\gamma}H^{-1}G(p^*)H^{-1}\right),$$
$$\sqrt{\mathrm{cost}_k}\cdot\nabla f(\bar{x}_{k,\gamma}) \Rightarrow \mathcal{N}\left(0, \frac{c(p^*)}{1-\gamma}G(p^*)\right),$$
$$\mathrm{cost}_k\cdot(f(\bar{x}_{k,\gamma}) - f(x^*)) \Rightarrow \left\|\mathcal{N}\left(0, \frac{c(p^*)}{2(1-\gamma)}H^{-\frac{1}{2}}G(p^*)H^{-\frac{1}{2}}\right)\right\|^2,$$
where $\mathcal{N}(0,\cdot)$ *denotes the multivariate Gaussian distribution with mean zero and covariance matrix* $\cdot$, *and* $\Rightarrow$ *denotes convergence in distribution.*

Theorem 4.8 states that the solution error (in terms of difference with the optimum, gradient norm, or optimality gap) scaled by the cumulative sampling cost converges in distribution to a multivariate Gaussian whose covariance depends on the limit sampling distribution p^*. The sampling efficiencies of different IS schemes thus are determined by their respective limit sampling distributions. Results of such type allow asymptotically exact quantification of the solution error under a pre-specified sampling budget, and subsequently transparent comparisons of sampling efficiencies of different SGD algorithms. Notably, the mixing weight w_k is allowed to approach zero at a sufficiently slow rate via a delicate control of the gradient variance near the optimum, and hence the limit p^* does not need to be restricted to the interior of Δ_n (Papa et al., 2015).

Theorem 4.8 is established by proving a CLT for the solution error $\bar{x}_{k,\gamma} - x^*$ followed by an application of the Slutsky's theorem. Compared to the analysis of classical SGD CLTs, one additional challenge here is that not only the solution but also the sampling distribution keeps changing over iterations. To simultaneously handle the non-stationarity in the solution and sampling distribution we use a probabilistic coupling argument, instead of uniform-integrability-type assumptions or the i.i.d assumption (Polyak and Juditsky, 1992, Assumption 3.3 and Assumption 4.2), that explicitly links the gradient noise $\{\xi_k\}_{k=1}^{\infty}$ at each iterate to an oracle noise incurred when sampling at the optimum according to the limit distribution p^*. The full proof is deferred to Appendix B.

4.3 Comparison with Existing SGD Variants

This subsection utilizes Theorem 4.8 to compare our HeteRSGD with existing SGD algorithms including the standard SGD and the stochastic reweighted gradient descent (SRG) (El Hanchi et al., 2022), and thereby establishes the asymptotic optimality of HeteRSGD among all IS-based SGD schemes encompassed by Algorithm 1.

Since the asymptotic efficiency is determined by the limit sampling weights p^* as Theorem 4.8 suggests, we now quantify the efficiency realized by an arbitrary limit, and specialize to different algorithms with distinct limits later. With the limit p^*, we consider running Algorithm 1 for $C/(|\mathcal{I}|c(p^*))$ iterations, with a fixed minibatch size $|\mathcal{I}_k| = |\mathcal{I}|$, to approximately reach a fixed sampling budget C (see Proposition 4.7). Let $\bar{x}_C$ be the Polyak-Ruppert or α-suffix average, then Theorem 4.8 entails that $\sqrt{C}\nabla f(\bar{x}_C)$ is approximately $\mathcal{N}(0, c(p^*)G(p^*)/(1-\gamma))$, therefore

$$\begin{aligned}\mathbb{E}[\|\nabla f(\bar{x}_C)\|^2] &\approx \frac{1}{C(1-\gamma)} c(p^*)\mathrm{Tr}(G(p^*)) \\ &= \frac{1}{C(1-\gamma)}\rho(p^*),\end{aligned} \tag{4.2}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace, and ρ is the efficiency metric from (4.1), therefore the asymptotic solution error of an SGD algorithm boils down to the metric value $\rho(p^*)$ realized by its limit sampling distribution.

We then compare the asymptotic error (4.2) realized by different SGD algorithms. As special cases of Algorithm 1, the standard SGD with uniform sampling and the SRG have limit sampling distributions

$$p^*_{SGD} := (1/n, \ldots, 1/n)\,, \text{ and}$$
$$p^*_{SRG} := \left(\frac{\|\nabla f_i(x^*)\|}{\sum_{j=1}^n \|\nabla f_j(x^*)\|}\right)_{i=1}^n$$

respectively. The limit sampling distribution of HeteRSGD is p^*_{Hete} as given in Proposition 4.6. Since p^*_{Hete} optimizes the efficiency metric $\rho(\cdot)$, we immediately see that HeteRSGD achieves the minimum asymptotic error (4.2). Therefore HeteRSGD is optimal in the sense that, with a fixed sampling budget, it achieves the least asymptotic error in the gradient of the averaged solution among all possible IS-based SGD schemes. The standard SGD and SRG are in general suboptimal. The following result (with proof in Appendix C) shows that the efficiency gain of HeteRSGD can be arbitrarily large relative to both SGD and SRG.

Proposition 4.9. *For any dimension $d \geq 1$, number of components $n \geq 3$, and $\epsilon > 0$, there exist examples with $\rho(p^*_{Hete})/\rho(p^*_{SGD}) < \epsilon$ and $\rho(p^*_{Hete})/\rho(p^*_{SRG}) < \epsilon$.*

Lastly, we briefly compare the asymptotic errors in terms of $\bar{x}_C - x^*$ and $f(\bar{x}_C) - f(x^*)$. From Theorem 4.8 we can obtain the following characterizations similar to (4.2):

$$\mathbb{E}[\|\bar{x}_C - x^*\|^2] \approx \frac{c(p^*)}{C(1-\gamma)}\mathrm{Tr}(H^{-1}G(p^*)H^{-1}),$$
$$\mathbb{E}[f(\bar{x}_C)] - f(x^*) \approx \frac{c(p^*)}{2C(1-\gamma)}\mathrm{Tr}(H^{-\frac{1}{2}}G(p^*)H^{-\frac{1}{2}}).$$

The errors now depend on the Hessian H, in addition to p^*, and hence HeteRSGD may not be optimal. However, HeteRSGD is still optimal in following minmax sense. Consider all Hessian such that $H^{-1} \preceq \mu^{-1} I_d$, and calculate the worst errors

$$\begin{aligned}&\sup_{H^{-1}\preceq\mu^{-1}I_d} c(p^*)\mathrm{Tr}(H^{-1}G(p^*)H^{-1}) \\ =\ &\sup_{H^{-1}\preceq\mu^{-1}I_d} c(p^*)\mathrm{Tr}(G(p^*)^{\frac{1}{2}}H^{-2}G(p^*)^{\frac{1}{2}}) \\ =\ &\frac{c(p^*)}{\mu^2}\mathrm{Tr}(G(p^*)^{\frac{1}{2}}I_d G(p^*)^{\frac{1}{2}}) = \frac{1}{\mu^2}\rho(p^*),\end{aligned}$$

and similarly $\sup_{H^{-1}\preceq\mu^{-1}I_d} c(p^*)\mathrm{Tr}(H^{-\frac{1}{2}}G(p^*)H^{-\frac{1}{2}}) = \rho(p^*)/\mu$. HeteRSGD therefore achieves the least worst-case asymptotic errors in $\bar{x}_C - x^*$ and $f(\bar{x}_C) - f(x^*)$.

4.4 Finite-Time Bounds and Comparison

This subsection complements the asymptotic theories on HeteRSGD with non-asymptotic convergence bounds and a finite-time comparison with the standard SGD.

We need two more assumptions stated as follows.

Assumption 4.10. *Each $f_i, i = 1, \dots, n$ is twice differentiable with Lipschitz continuous second-order derivatives, i.e., $\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \le L_2\|x-y\|$ for any $x, y \in \mathbb{R}^d$ for some $L_2 < \infty$.*

Assumption 4.11. *Each sampling cost $\hat{c}_i$ has a finite second moment $\mathrm{Var}(\hat{c}_i) < \infty$, and there exists a constant $\underline{c} > 0$ such that each $\hat{c}_i \ge \underline{c}$ almost surely.*

Our main finite-time result is the following theorem.

Theorem 4.12 (Finite-time bounds)**.** *Suppose Assumptions 2.1, 4.1, 4.4, 4.10 and 4.11 hold. Suppose that in Algorithm 2 $\alpha_k = \alpha_1/k^\beta$, where $\beta \in (1/2, 1)$ and $0 < \alpha_1 < \min\left(1/\mu, \mu/L^2\right)$, $w_k = w_1/k^\eta$ with $w_1 \in (0, 1]$ and $0 < \eta < \min(\beta/7, 1-\beta, \beta - 1/2)$, and that $|\mathcal{I}_k| = |\mathcal{I}|$ is fixed for all k. Then for every $\gamma \in [0, 1)$, we have for HeteRSGD that*

$$\sqrt{\mathbb{E}[\mathrm{cost}_k]} \cdot \mathbb{E}\left[\|H(\bar{x}_{k,\gamma} - x^*)\|\right] \le \sqrt{\frac{\rho(p^*_{Hete})}{1-\gamma}} + C_1\left(\frac{\sqrt{|\mathcal{I}|}}{k^{c_{\beta,\eta}}} + \left(\frac{n}{|\mathcal{I}|k^{1-6\eta}}\right)^{\frac{1}{4}} + \left(\frac{n}{|\mathcal{I}|k^{1-2\eta}}\right)^{\frac{1}{8}}\right), \tag{4.3}$$

whenever $k \ge \left(\frac{n}{|\mathcal{I}|}\right)^{1/(1-6\eta)}$, where

$$c_{\beta,\eta} = \min\left(\frac{\beta - 7\eta}{4}, \frac{1-\beta-\eta}{2}, \beta - \frac{1}{2} - \eta, \frac{\beta - 3\eta}{8}, \frac{\eta}{2}\right),$$

and $C_1 := C_1(\gamma, \alpha_1, w_1, \beta, \eta, x_1, f_i, \hat{c}_i, \tilde{g}_i^1, \tilde{c}_i^1, i = 1, \dots, n)$ does not explicitly depend on n.

Under the same conditions, for the standard SGD we have for all $k \ge 1$ that

$$\sqrt{\mathbb{E}[\mathrm{cost}_k]} \cdot \mathbb{E}\left[\|H(\bar{x}_{k,\gamma} - x^*)\|\right] \le \sqrt{\frac{\rho(p^*_{SGD})}{1-\gamma}} + C_2\frac{\sqrt{|\mathcal{I}|}}{k^{c_\beta}},$$

where the constant $c_\beta := \min(1/2 - \beta/2, \beta - 1/2)$ and $C_2 := C_2(\gamma, \alpha_1, \beta, x_1, f_i, \hat{c}_i, i = 1, \dots, n)$ does not explicitly depend on n.

The key step of the proof of Theorem 4.12 is to control the error of the estimated sampling weights p^k in approximating the limit p^*_{Hete}, which is carried out by first bounding the estimation errors of the quantities $\tilde{c}^k, \tilde{g}_k, \tilde{G}_k$ needed in the efficiency metric (2.4) and then propagating the errors to the sampling weights via a novel sensitivity analysis of the mapping from these quantities to the resulting sampling weights. The details can be found in Appendix D.

The finite-time bound (4.3) for the cost-scaled error consists of a constant term $\sqrt{\rho(p^*_{Hete})/(1-\gamma)}$ that matches the asymptotic size of $\sqrt{\mathrm{cost}_k}\nabla f(\bar{x}_{k,\gamma})$ given in Theorem 4.8 and several polynomially decaying high-order terms. To compare the finite-time behavior of HeteRSGD with the standard SGD, suppose the constant term dominates the bound (4.3), i.e., $\mathbb{E}\left[\|H(\bar{x}_{k,\gamma} - x^*)\|\right] \le \sqrt{\rho(p^*_{Hete})/((1-\gamma)\mathbb{E}[\mathrm{cost}_k])}$ approximately holds, and for the standard SGD we approximately have $\mathbb{E}\left[\|H(\bar{x}_{k,\gamma} - x^*)\|\right] \le \sqrt{\rho(p^*_{SGD})/((1-\gamma)\mathbb{E}[\mathrm{cost}_k])}$. By the optimality of p^*_{Hete}, HeteRSGD achieves lower solution errors than the standard SGD under the same sampling budget in this regime. It can be verified that the depicted condition $k \ge (n/|\mathcal{I}|)^{1/(1-6\eta)}$ is sufficient for making the high-order terms negligible in (4.3). Note that this is roughly $k \ge n/|\mathcal{I}|$ when η is chosen small, therefore HeteRSGD outperforms the standard SGD after each f_i has been sampled at least once on average and has a reasonable estimate for its sampling cost.

5 EXTENSION

The previous section shows the optimal sampling complexity of HeteRSGD for averaged solutions, and this section extends HeteRSGD to a family of algorithms that are optimal for individual SGD iterates.

We begin with designing a new family of efficiency metrics. One crucial fact that makes the sampling weights determined by (2.4) optimal for averaged solutions is that the cost term $\sum_{i=1}^n p_i^k c_i$ in (2.4) comes with an exponent of 1 that matches the k^{-1} convergence rate of the error of an averaged solution. In the strongly convex and smooth case, the error of the final iterate is of order $k^{-\beta}$ (e.g., Papa et al., 2015) if the step size $\alpha_k = \alpha_1/k^\beta, \beta \in (1/2, 1)$ is used, which motivates the following counterpart of (2.4)

$$\left(\sum_{i=1}^n p_i^k c_i\right)^\beta \left(\sum_{i=1}^n \frac{\|\nabla f_i(x_k)\|^2}{n^2 p_i^k} - \|\nabla f(x_k)\|^2\right). \tag{5.1}$$

Compared to (2.4), (5.1) is slightly less sensitive to surges in sampling costs due to the smaller exponent β. The family of efficiency metrics (5.1) parameterized by $\beta \in (1/2, 1)$ therefore induce a family of importance samplers with varying levels of awareness of cost heterogeneity.

Our HeteRSGD variant using the new sampling metric (5.1), called HeteRSGD$_\beta$, is the same as Algorithm 2 except that $\tilde{p}^k$ is now calculated by minimizing

$$\left(\sum_{i=1}^n p_i^k \tilde{c}_i^k\right)^\beta \left(\sum_{i=1}^n \frac{(\tilde{g}_i^k)^2}{n^2 p_i^k} - \min\left(\|\tilde{G}_k\|, \sum_{i=1}^n \frac{\tilde{g}_i^k}{n}\right)^2\right),$$

in place of (3.4). The optimal weights here can be computed efficiently via a nested bisection, the details of which are left to Appendix F. We have the following counterpart of Theorem 4.8 for individual SGD iterates:

Theorem 5.1. *Assume all the conditions in Theorem 4.8. Assume further that f is thrice continuously differentiable in a neighborhood of x^*, and that the sequence $\{w_k\}_{k=1}^\infty$ satisfies $\sup_k \alpha_k/w_k^{3+\delta} < \infty$ for some $\delta > 0$, and $\lim_{k\to\infty} \alpha_k \sum_{j=1}^k \alpha_j/w_j^2 = 0$. Then it holds for Algorithm*

1 that

$$\mathrm{cost}_k^{\frac{\beta}{2}} \cdot (x_k - x^*) \Rightarrow \mathcal{N}\left(0, \frac{\alpha_1}{|\mathcal{I}|^{1-\beta}} c(p^*)^\beta \Sigma(p^*)\right),$$

where $\Sigma(p^*)$ *satisfies* $\Sigma(p^*)H + H\Sigma(p^*) = G(p^*)$*, and*

$$\mathrm{cost}_k^{\frac{\beta}{2}} \cdot \nabla f(x_k) \Rightarrow \mathcal{N}\left(0, \frac{\alpha_1}{|\mathcal{I}|^{1-\beta}} c(p^*)^\beta H\Sigma(p^*)H\right),$$

$$\mathrm{cost}_k^{\beta} \cdot (f(x_k) - f(x^*)) \Rightarrow \left\|\mathcal{N}\left(0, \frac{\alpha_1}{2|\mathcal{I}|^{1-\beta}} c(p^*)^\beta H^{\frac{1}{2}}\Sigma(p^*)H^{\frac{1}{2}}\right)\right\|^2.$$

The proof of Theorem 5.1 builds on a general CLT from Fort (2015) for controlled Markov chains and can be found in Appendix E.

To demonstrate the applicability of Theorem 5.1 to HeteRSGD$_\beta$, we need to verify Assumption 4.5. Denote by

$$\rho_\beta(p) := \left(\sum_{i=1}^n p_i c_i\right)^\beta \left(\sum_{i=1}^n \frac{1}{n^2 p_i}\|\nabla f_i(x^*)\|^2\right) \quad (5.2)$$

the counterpart of (4.1), i.e., the metric (5.1) at the optimum. Then the sampling distribution in HeteRSGD$_\beta$ converges to the optimal one stipulated by (5.2) (see proof in Appendix E):

Proposition 5.2. *Let* $p^*_{Hete_\beta} := \mathrm{argmin}_{p\in\Delta_n}\rho_\beta(p)$*. Under the same conditions of Proposition 4.6, we have* $p^k \to p^*_{Hete_\beta}$ *almost surely for HeteRSGD*$_\beta$*.*

We demonstrate the optimality of HeteRSGD$_\beta$ for individual iterates based on Theorem 5.1. For an SGD scheme with limit sampling weights p^*, a similar analysis as in Section 4.3 leads to the following asymptotic error of its last iterate x_C when a fixed sampling budget C is consumed

$$\begin{aligned}\mathbb{E}[f(x_C)] - f(x^*) &\approx \frac{\alpha_1}{2C^\beta|\mathcal{I}|^{1-\beta}} c(p^*)^\beta \mathrm{Tr}(H^{\frac{1}{2}}\Sigma(p^*)H^{\frac{1}{2}}) \\ &= \frac{\alpha_1}{2C^\beta|\mathcal{I}|^{1-\beta}} c(p^*)^\beta \mathrm{Tr}(H\Sigma(p^*)) \\ &= \frac{\alpha_1}{4C^\beta|\mathcal{I}|^{1-\beta}} c(p^*)^\beta \mathrm{Tr}(G(p^*)) \\ &= \frac{\alpha_1}{4C^\beta|\mathcal{I}|^{1-\beta}} \rho_\beta(p^*),\end{aligned}$$

where the second equality follows from $\Sigma(p^*)H + H\Sigma(p^*) = G(p^*)$. Since $p^*_{\mathrm{Hete}_\beta}$ minimizes the efficiency metric ρ_β, by Proposition 5.2 HeteRSGD$_\beta$ with β matching the decay rate of the step size α_1/k^β achieves the least asymptotic error in the objective of the last (and hence each individual) SGD iterate among all possible IS-based schemes. Similarly, one can argue similar minmax optimality as in Section 4.3 for the errors $x_C - x^*$ and $\nabla f(x_C)$.

Besides the type of solution (averaged versus individual ones) that is concerned for optimality, another notable distinction between HeteRSGD$_\beta$ and HeteRSGD is that HeteRSGD$_\beta$ becomes optimal only if the step size decays as $1/k^\beta$ so that the order of the solution error matches the exponent β in the efficiency metric. Averaged solutions always have errors of order $1/k$, and thus optimality of HeteRSGD holds regardless of the step size choice.

6 NUMERICAL EXPERIMENTS

We present the numerical results in this section. In each experiment, we compare HeteRSGD (Algorithm 2) and HeteRSGD$_\beta$ with several baselines:

- SGD: The standard stochastic gradient descent with uniform sampling.
- SRG: The stochastic reweighted gradient descent (El Hanchi et al., 2022, Algorithm 1) that switches between the uniform distribution and an IS distribution induced from estimates of gradient norms, and updates gradient norm estimates only if the uniform distribution is used.
- SRG-m: A modified version of SRG that draws samples according to a weighted average of the two probability distributions used in SRG, and updates gradient norm estimates in each iteration. We consider this version because the original SRG seldom updates the gradient norms.

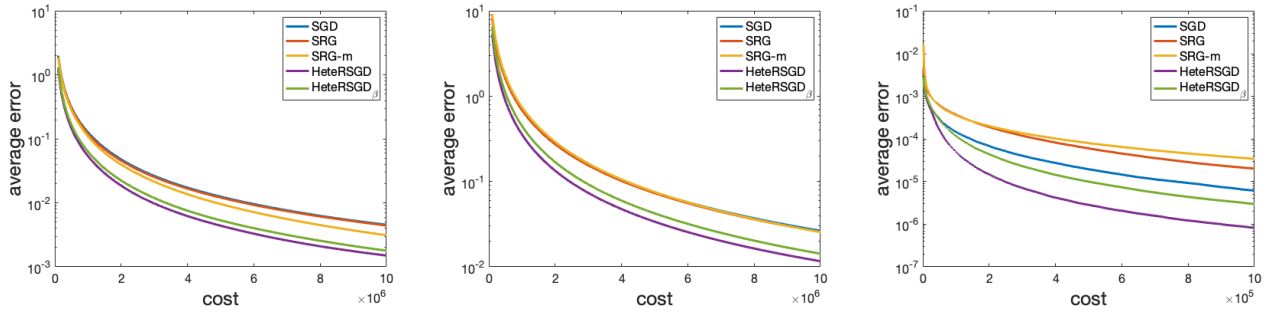
The problems we test on are all in the form of (1.1). For each problem, we run the algorithms until some pre-fixed sampling cost budget is reached. To mitigate the effect of algorithmic randomness, we run each algorithm for 10 times and report the average error of the Polyak-Ruppert averaged solution. More precisely, we report the average of the error $\|\frac{1}{k}\sum_{j=1}^k x_j - x^*\|^2$ for convex problems or $\|\nabla f(\frac{1}{k}\sum_{j=1}^k x_j)\|^2$ for non-convex problems. Other implementation details can be found in Appendix G. We use the following test problems.

A synthetic example: We consider a finite-sum of $n = 100$ components, and each

$$f_i(x_1, x_2) = \frac{1}{2}(x_1 + a_i)^2 + \frac{1}{2}(x_2 + b_i)^2.$$

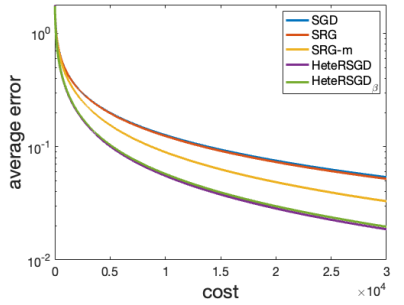
Let $\theta_j, 1 \le j \le n/2$ be independently drawn from $\mathrm{Uniform}(0, 2\pi)$. For each $1 \le j \le n/4$, let $-a_{2j} = a_{2j-1} = \sin(\theta_j)$, $-b_{2j} = b_{2j-1} = \cos(\theta_j)$, and $c_{2j} = c_{2j-1} = \epsilon^2$ for $\epsilon \in \{0.01, 0.3\}$. For each $n/4 + 1 \le j \le n/2$, let $-a_{2j} = a_{2j-1} = 0.01\sin(\theta_j), -b_{2j} = b_{2j-1} = 0.01\cos(\theta_j)$, and $c_{2j} = c_{2j-1} = 1$. The optimal solution is $x_1^* = x_2^* = 0$. ϵ controls the degree of heterogeneity in sampling costs for this example, and a smaller ϵ induces higher heterogeneity. When $\epsilon = 0.01$, the estimated speedup from using HeteRSGD compared to SGD and SRG can be calculated to be $\rho(p^*_{Hete})/\rho(p^*_{SGD}) \approx 4 \times 10^{-4}$ and $\rho(p^*_{Hete})/\rho(p^*_{SRG}) \approx 0.039$. Results are shown in Figure 2.

ℓ_2-regularized logistic regression: The ℓ_2-regularized logistic regression is a strongly convex binary classification

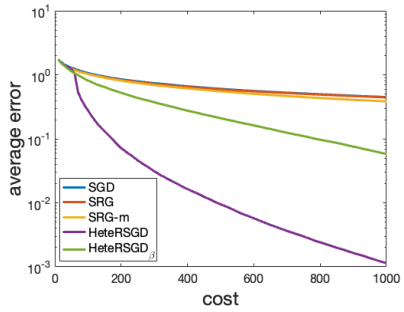


(a) Logistic regression on CINA. (b) Logistic regression on gisette. (c) Nonconvex classification on CINA.

Figure 1: Results on real data sets.



(a) $\epsilon = 0.3$.



(b) $\epsilon = 0.01$.

Figure 2: Results on the synthetic example.

problem with

$$f_i(x) = \log\left(1 + \exp(-b_i a_i^T x)\right) + \frac{\mu}{2}\|x\|^2,$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, 1\}$. Each a_i is normalized such that $\|a_i\| = 1$ and $\mu = 0.001$. Each cost c_i is drawn i.i.d. from Lognormal$(0, 1)$. We use the dataset CINA[1] ($n = 16033$, $d = 132$) and gisette[2] ($n = 6000$, $d = 5000$). The relative efficiency is calculated to be $\rho(p^*_{Hete})/\rho(p^*_{SGD}) = 0.28$, $\rho(p^*_{Hete})/\rho(p^*_{SRG}) = 0.78$ for CINA, and $\rho(p^*_{Hete})/\rho(p^*_{SGD}) = 0.53$, $\rho(p^*_{Hete})/\rho(p^*_{SRG}) = 0.78$ for gisette. Results are shown in Figure 1a and Figure 1b, respectively.

A non-convex example: Although our theory focuses on the convex setting, we also test a nonconvex binary classification problem (Mason et al., 1999; Wang et al., 2017) with f_i being

$$f_i(x) = 1 - \tanh(b_i a_i^T x) + \frac{\mu}{2}\|x\|^2,$$

and all other settings being the same as the logistic regression problem. Results on CINA are in Figure 1c.

Figures 1 and 2 show that the two proposed methods, especially HeteRSGD, outperform the three baselines by achieving the same level of solution error as the best baseline with roughly 40% less sampling cost in all convex cases and 70% less in the nonconvex example. Notably, Figure 2b shows almost an order-of-magnitude improvement from SGD/SRG to HeteRSGD$_\beta$ and an even more significant speedup when HeteRSGD is used. Specifically, HeteRSGD and HeteRSGD$_\beta$ achieve the same accuracy as SGD/SRG/SRG-m with roughly 95% and 70% less sampling costs respectively, consolidating the advantage of our methods in the presence of high sampling heterogeneity. As a side note, the relative ranking among HeteRSGD, SGD and SRG-m at the largest sampling cost roughly match their theoretical efficiencies in each case. For example, in Figure 1a the ranking of the solution error HeteRSGD $<$ SRG-m $<$ SGD matches that of their efficiency metrics $\rho(p^*_{Hete}) < \rho(p^*_{SRG}) < \rho(p^*_{SGD})$.

Comparing HeteRSGD and HeteRSGD$_\beta$, we see that in all the cases the sampling cost reduced by HeteRSGD$_\beta$ is not as significant as HeteRSGD. This is consistent with the theories that HeteRSGD is optimal for averaged iterates whereas HeteRSGD$_\beta$ is optimal for individual iterates.

7 CONCLUSION

In this work, we investigate the use of importance sampling (IS) as a cost saver to accelerate stochastic gradient descent (SGD) under heterogeneous sampling costs. We propose a novel family of sampling efficiency metrics for IS design that balance cost reduction and variance reduction. Our proposed algorithm HeteRSGD draws samples according to probability weights derived from an empirical version of the efficiency metric in each iteration, and is provably more efficient than any other IS-based SGD scheme. Encouraging numerical results are discussed.

[1] http://www.causality.inf.ethz.ch/data/CINA.html

[2] https://archive.ics.uci.edu/ml/datasets/Gisette

Acknowledgements

A major part of the work of Z. Chen was completed during his internship at Alibaba US DAMO Academy.

References

Alain, G., Lamb, A., Sankar, C., Courville, A., and Bengio, Y. (2015). Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*.

Allen-Zhu, Z. (2017). Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244.

An, J. and Ying, L. (2021). Combining resampling and reweighting for faithful stochastic optimization. *arXiv preprint arXiv:2105.14694*.

Birge, J. R. and Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Borsos, Z., Krause, A., and Levy, K. Y. (2018). Online variance reduction for stochastic optimization. In *Conference On Learning Theory*, pages 324–357. PMLR.

Box, G. E. and Tiao, G. C. (2011). *Bayesian inference in statistical analysis*. John Wiley & Sons.

Cho, Y. J., Wang, J., and Joshi, G. (2022). Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10351–10375. PMLR.

Csiba, D. and Richtárik, P. (2018). Importance sampling for minibatches. *The Journal of Machine Learning Research*, 19(1):962–982.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27.

Delyon, B. and Portier, F. (2021). Safe adaptive importance sampling: A mixture approach. *The Annals of Statistics*, 49(2):885–917.

Diao, E., Ding, J., and Tarokh, V. (2020). Heterofl: Computation and communication efficient federated learning for heterogeneous clients. *arXiv preprint arXiv:2010.01264*.

El Hanchi, A. and Stephens, D. (2020). Adaptive importance sampling for finite-sum optimization and sampling with decreasing step-sizes. *Advances in Neural Information Processing Systems*, 33:15702–15713.

El Hanchi, A., Stephens, D., and Maddison, C. (2022). Stochastic reweighted gradient descent. In *International Conference on Machine Learning*, pages 8359–8374. PMLR.

Fort, G. (2015). Central limit theorems for stochastic approximation with controlled markov chain dynamics. *ESAIM: Probability and Statistics*, 19:60–80.

Gopal, S. (2016). Adaptive sampling for sgd by exploiting side information. In *International Conference on Machine Learning*, pages 364–372. PMLR.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR.

He, S., Jiang, G., Lam, H., and Fu, M. C. (2021). Adaptive importance sampling for efficient stochastic root finding and quantile estimation. *arXiv preprint arXiv:2102.10631*.

Hoos, H. H. (2011). Automated algorithm configuration and parameter tuning. In *Autonomous search*, pages 37–71. Springer.

Horváth, S. and Richtárik, P. (2019). Nonconvex variance reduced optimization with arbitrary sampling. In *International Conference on Machine Learning*, pages 2781–2789. PMLR.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26.

Johnson, T. B. and Guestrin, C. (2018). Training deep models faster with robust, approximate importance sampling. *Advances in Neural Information Processing Systems*, 31.

Katharopoulos, A. and Fleuret, F. (2018). Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pages 2525–2534. PMLR.

Liu, H., Wang, X., Li, J., and So, A. M.-C. (2021). Low-cost lipschitz-independent adaptive importance sampling of stochastic gradients. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2150–2157. IEEE.

Luo, B., Xiao, W., Wang, S., Huang, J., and Tassiulas, L. (2022). Tackling system and statistical heterogeneity for federated learning with adaptive client sampling. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1739–1748. IEEE.

Mason, L., Baxter, J., Bartlett, P., and Frean, M. (1999). Boosting algorithms as gradient descent. *Advances in neural information processing systems*, 12.

Namkoong, H., Sinha, A., Yadlowsky, S., and Duchi, J. C. (2017). Adaptive sampling probabilities for non-smooth optimization. In *International Conference on Machine Learning*, pages 2574–2583. PMLR.

Needell, D., Srebro, N., and Ward, R. (2016). Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming*, 155(1):549–573.

Papa, G., Bianchi, P., and Clémençon, S. (2015). Adaptive sampling for incremental optimization using stochastic gradient descent. In *International Conference on Algorithmic Learning Theory*, pages 317–331. Springer.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855.

Qian, X., Qu, Z., and Richtárik, P. (2021). L-svrg and l-katyusha with arbitrary sampling. *Journal of Machine Learning Research*, 22:1–49.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1571–1578.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rubinstein, R. Y. and Kroese, D. P. (2016). *Simulation and the Monte Carlo method*. John Wiley & Sons.

Ruppert, D. (1988). Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Salehi, F., Celis, L. E., and Thiran, P. (2017). Stochastic optimization with bandit sampling. *arXiv preprint arXiv:1708.02544*.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112.

Shen, G., Gao, D., Yang, L., Zhou, F., Song, D., Lou, W., and Pan, S. (2022). Variance-reduced heterogeneous federated learning via stratified client selection. *arXiv preprint arXiv:2201.05762*.

Shen, Z., Qian, H., Zhou, T., and Mu, T. (2016). Adaptive variance reducing for stochastic gradient descent. In *IJCAI*, pages 1990–1996.

Stich, S. U., Raj, A., and Jaggi, M. (2017). Safe adaptive importance sampling. *Advances in Neural Information Processing Systems*, 30.

Wang, X., Ma, S., Goldfarb, D., and Liu, W. (2017). Stochastic quasi-newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956.

Yuan, K., Ying, B., Vlaski, S., and Sayed, A. H. (2016). Stochastic gradient descent with finite samples sizes. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Zhao, P. and Zhang, T. (2015). Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pages 1–9. PMLR.

HeteRSGD: Tackling Heterogeneous Sampling Costs via Optimal Reweighted Stochastic Gradient Descent: Supplementary Materials

A PROOFS FOR RESULTS IN SECTION 4.1

Proof of Theorem 4.2. It follows from $\sum_{k=1}^{\infty}\alpha_k^2<\infty$ that $\lim_{k\to\infty}\alpha_k=0$. Without loss of generality, we assume that $\alpha_k<\min\{1/\mu,\mu/L^2\},\ \forall\, k\geq 1$. It can be computed that

$$\begin{aligned}\|x_{k+1}-x^*\|^2&=\|x_k-x^*-\alpha_k\nabla f(x_k)-\alpha_k\xi_k\|^2\\&=\|x_k-x^*-\alpha_k\nabla f(x_k)\|^2-2\alpha_k\langle x_k-x^*-\alpha_k\nabla f(x_k),\xi_k\rangle+\alpha_k^2\|\xi_k\|^2,\end{aligned}$$

which combined with $\mathbb{E}[\xi_k|\mathcal{F}_{k-1}]=0$ yields that

$$\mathbb{E}\left[\|x_{k+1}-x^*\|^2|\mathcal{F}_{k-1}\right]=\|x_k-x^*-\alpha_k\nabla f(x_k)\|^2+\alpha_k^2\mathbb{E}\left[\|\xi_k\|^2|\mathcal{F}_{k-1}\right].$$

By strong convexity, one has that

$$\begin{aligned}\|x_k-x^*-\alpha_k\nabla f(x_k)\|^2&=\|x_k-x^*\|^2-2\alpha_k\langle\nabla f(x_k),x_k-x^*\rangle+\alpha_k^2\|\nabla f(x_k)\|^2\\&\leq\|x_k-x^*\|^2+2\alpha_k\left(f(x^*)-f(x_k)-\frac{\mu}{2}\|x_k-x^*\|^2\right)+\alpha_k^2\|\nabla f(x_k)\|^2\\&\leq(1-\alpha_k\mu)\|x_k-x^*\|^2-2\alpha_k(f(x_k)-f(x^*))+\alpha_k^2\|\nabla f(x_k)\|^2,\end{aligned}$$

and that

$$f(x_k)-f(x^*)\geq\frac{\mu}{2}\|x_k-x^*\|^2\geq\frac{\mu}{2L^2}\|\nabla f(x_k)\|^2\geq\frac{\alpha_k}{2}\|\nabla f(x_k)\|^2,$$

where the last inequality is guaranteed by $\alpha_k\leq\mu/L^2$. Therefore, combining the above calculations, one obtains that

$$\mathbb{E}\left[\|x_{k+1}-x^*\|^2|\mathcal{F}_{k-1}\right]\leq(1-\alpha_k\mu)\|x_k-x^*\|^2+\alpha_k^2\mathbb{E}\left[\|\xi_k\|^2|\mathcal{F}_{k-1}\right],\tag{A.1}$$

which then implies that

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1}-x^*\|^2\right]&\leq(1-\alpha_k\mu)\mathbb{E}\left[\|x_k-x^*\|^2\right]+\alpha_k^2\mathbb{E}\left[\|\xi_k\|^2\right]\\&\leq(1-\alpha_k\mu)(1-\alpha_{k-1}\mu)\mathbb{E}\left[\|x_{k-1}-x^*\|^2\right]+(1-\alpha_k\mu)\alpha_{k-1}^2\mathbb{E}\left[\|\xi_{k-1}\|^2\right]+\alpha_k^2\mathbb{E}\left[\|\xi_k\|^2\right]\\&\leq\cdots\\&\leq\left(\prod_{j=1}^{k}(1-\alpha_j\mu)\right)\mathbb{E}\left[\|x_1-x^*\|^2\right]+\sum_{j=1}^{k}\left(\prod_{l=j+1}^{k}(1-\alpha_l\mu)\right)\alpha_j^2\mathbb{E}\left[\|\xi_j\|^2\right].\end{aligned}$$

It follows from $\sum_{k=1}^{\infty}\alpha_k=\infty$ that $\prod_{k=1}^{\infty}(1-\alpha_k\mu)=0$. For any $\epsilon>0$, since $\sum_{k=1}^{\infty}\alpha_k^2\mathbb{E}[\|\xi_k\|^2]<\infty$, there exists $k_0\in\mathbb{N}_+$ such that $\sum_{k=k_0}^{\infty}\alpha_k^2\mathbb{E}[\|\xi_k\|^2]<\epsilon$, and $k_1\in\mathbb{N}_+$ with $k_1>k_0$ such that $\prod_{k=k_0}^{k_1}(1-\alpha_k\mu)<\epsilon$. Then for any $k\geq k_1$, it holds that

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1}-x^*\|^2\right]&\leq\left(\prod_{j=k_0}^{k_1}(1-\alpha_j\mu)\right)\mathbb{E}\left[\|x_1-x^*\|^2\right]+\sum_{j=1}^{k_0-1}\left(\prod_{l=k_0}^{k_1}(1-\alpha_l\mu)\right)\alpha_j^2\mathbb{E}\left[\|\xi_j\|^2\right]+\sum_{j=k_0}^{k}\alpha_j^2\mathbb{E}\left[\|\xi_j\|^2\right]\\&\leq\epsilon\cdot\mathbb{E}\left[\|x_1-x^*\|^2\right]+\epsilon\cdot\sum_{j=1}^{k_0-1}\alpha_j^2\mathbb{E}\left[\|\xi_j\|^2\right]+\epsilon\end{aligned}$$

$$\leq \epsilon \cdot \left(\mathbb{E}\left[\|x_1 - x^*\|^2\right] + \sum_{j=1}^{\infty} \alpha_j^2 \mathbb{E}\left[\|\xi_j\|^2\right] + 1 \right),$$

which proves that $\lim_{k\to\infty} \mathbb{E}\left[\|x_k - x^*\|^2\right] = 0$, i.e., $x_k \to x^*$ in L^2.

Then we consider the almost sure convergence. According to (A.1), it holds that

$$\sum_{k=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 | \mathcal{F}_{k-1}\right]_+\right] \leq \sum_{k=1}^{\infty} \alpha_k^2 \mathbb{E}\left[\|\xi_k\|^2\right] < \infty.$$

Then using the martingale convergence theorem (Blum, 1954, Corollary in Section 3), one can conclude that $\|x_k - x^*\|^2$ converges almost surely to a random variable Z. It follows from $\lim_{k\to\infty} \mathbb{E}\left[\|x_k - x^*\|^2\right] = 0$ that $\{\|x_k - x^*\|^2\}_{k=1}^{\infty}$ has a subsequence that converges to 0 almost surely. Therefore, $\|x_k - x^*\|^2 \to 0$, a.s., which leads to $x_k \to x^*$, a.s.. □

Proof of Lemma 4.3. Similarly, we can assume that $\alpha_k < \min\{1/\mu, \mu/L^2\}$, $\forall\, k \geq 1$. It follows from

$$\xi_k = \frac{1}{|\mathcal{I}_k|} \sum_{i\in\mathcal{I}_k} \frac{1}{np_i^k} \nabla f_i(x_k) - \nabla f(x_k) = \frac{1}{|\mathcal{I}_k|} \sum_{i\in\mathcal{I}_k} \left(\frac{1}{np_i^k} \nabla f_i(x_k) - \nabla f(x_k) \right),$$

that

$$\begin{aligned}
\mathbb{E}\left[\|\xi_k\|^2 | \mathcal{F}_{k-1}\right] &= \frac{1}{|\mathcal{I}_k|} \left(\frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{p_i^k} \|\nabla f_i(x_k)\|^2 - \|\nabla f(x_k)\|^2 \right) \\
&\leq \frac{1}{|\mathcal{I}_k| \cdot n w_k} \sum_{i=1}^{n} \|\nabla f_i(x_k)\|^2 \\
&\leq \frac{1}{|\mathcal{I}_k| \cdot n w_k} \sum_{i=1}^{n} (\|\nabla f_i(x^*)\| + L\|x_k - x^*\|)^2 \\
&\leq \frac{2}{|\mathcal{I}_k| \cdot n w_k} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2 + \frac{2L^2}{|\mathcal{I}_k| \cdot w_k} \|x_k - x^*\|^2,
\end{aligned}$$

where we used the L-smoothness of f_i and $p_i^k \geq w_k/n$. Therefore, one has that

$$\mathbb{E}\left[\|\xi_k\|^2\right] \leq \frac{C_f}{|\mathcal{I}_k| w_k} \left(1 + \mathbb{E}\left[\|x_k - x^*\|^2\right]\right), \tag{A.2}$$

with $C_f = \max\left\{\frac{2}{n} \sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2, 2L^2\right\}$, which combined with (A.1) yields that

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] \leq \left(1 - \alpha_k \mu + \frac{\alpha_k^2 C_f}{|\mathcal{I}_k| w_k}\right) \mathbb{E}\left[\|x_k - x^*\|^2\right] + \frac{\alpha_k^2 C_f}{|\mathcal{I}_k| w_k} \leq \left(1 + \frac{\alpha_k^2 C_f}{|\mathcal{I}_k| w_k}\right) \mathbb{E}\left[\|x_k - x^*\|^2\right] + \frac{\alpha_k^2 C_f}{|\mathcal{I}_k| w_k},$$

i.e.,

$$\mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] + 1 \leq \left(1 + \frac{\alpha_k^2 C_f}{|\mathcal{I}_k| w_k}\right) \left(\mathbb{E}\left[\|x_k - x^*\|^2\right] + 1\right).$$

It follows from $\sum_{k=1}^{\infty} \frac{\alpha_k^2}{w_k} < \infty$ that $\prod_{k=1}^{\infty} \left(1 + \frac{\alpha_k^2 C_f}{|\mathcal{I}_k| w_k}\right) < \infty$. Then it holds that

$$\sup_{k\geq 1} \mathbb{E}\left[\|x_k - x^*\|^2\right] < \infty. \tag{A.3}$$

Combining (A.2) and (A.3), one can conclude that

$$\sum_{k=1}^{\infty} \alpha_k^2 \mathbb{E}\left[\|\xi_k\|^2\right] \leq \sum_{k=1}^{\infty} \frac{\alpha_k^2 C_f}{|\mathcal{I}_k| w_k} \cdot \left(1 + \sup_{k\geq 1} \mathbb{E}\left[\|x_k - x^*\|^2\right]\right) < \infty,$$

where we used $\sum_{k=1}^{\infty} \frac{\alpha_k^2}{w_k} < \infty$. □

B PROOF FOR THEOREM 4.8

This section is for the proof of Theorem 4.8. We use similar framework and techniques as in Polyak and Juditsky (1992): we first establish in Theorem B.1 a central limit theorem for the linearized system $\{y_k\}_{k=1}^{\infty}$ defined via $y_1 = x_1$, and

$$y_{k+1} - x^* = (y_k - x^*) - \alpha_k H(y_k - x^*) - \alpha_k \xi_k, \quad k \geq 1,$$

where $H = \nabla^2 f(x^*)$, and then prove in Theorem B.2 that the Polyak-Ruppert averaging or α-suffix averaging of

$$\delta_k = x_k - y_k, \quad k \geq 1,$$

converges to 0 in probability.

Theorem B.1. *Suppose Assumptions 4.1, 4.4 and 4.5 hold. Suppose in addition that $\alpha_k = \alpha_1/k^\beta$, where $\beta \in (1/2, 1)$ and that $|\mathcal{I}_k| = |\mathcal{I}|$ is fixed for any $k \geq 1$. If there exists a non-increasing sequence $\{w_k\}_{k=1}^{\infty} \subset (0, 1]$ satisfying $p_i^k \geq w_k/n$, $\forall\, i \in \{1, 2, \ldots, n\}$ and $k \geq 1$, $\lim_{k\to\infty} \alpha_k/w_k^2 = 0$, and $\sum_{k=1}^{\infty} \alpha_k^2/w_k < \infty$, then the sequence $\{y_k\}_{k=1}^{\infty}$ generated by the linearized system satisfies*

$$\sqrt{(1-\gamma)k} \cdot (\bar{y}_{k,\gamma} - x^*) \Rightarrow \mathcal{N}\left(0, \frac{1}{|\mathcal{I}|} H^{-1} G(p^*) H^{-1}\right), \tag{B.1}$$

where $\bar{y}_{k,\gamma} = \frac{1}{(1-\gamma)k} \sum_{j=[\gamma k]+1}^{k} y_j$ and $\gamma \in [0, 1)$.

Theorem B.2. *Suppose Assumptions 4.1, 4.4 and 4.5 hold. Suppose in addition that $\alpha_k = \alpha_1/k^\beta$, where $\beta \in (1/2, 1)$, and f is twice continuously differentiable in a neighbourhood of x^*. If there exists a non-increasing sequence $\{w_k\}_{k=1}^{\infty} \subset (0, 1]$ satisfying $p_i^k \geq w_k/n$, $\forall\, i \in \{1, 2, \ldots, n\}$ and $k \geq 1$, $\lim_{k\to\infty} \alpha_k/w_k^2 = 0$ and $\sum_{k=1}^{\infty} \alpha_k/(w_k\sqrt{k}) < \infty$. Then $\sqrt{(1-\gamma)k} \cdot \bar{\delta}_{k,\gamma} \to 0$ in probability, where $\bar{\delta}_{k,\gamma} = \frac{1}{(1-\gamma)k} \sum_{j=[\gamma k]+1}^{k} \delta_j$ and $\gamma \in [0, 1)$.*

The proof of Theorem 4.8 can now be presented based on Theorem B.1 and Theorem B.2 whose proofs can be found in Subsection B.1 and Subsection B.2, respectively.

Proof of Theorem 4.8. Combining Theorem B.1 and Theorem B.2, we have that

$$\sqrt{k} \cdot (\bar{x}_{k,\gamma} - x^*) \Rightarrow \mathcal{N}\left(0, \frac{1}{(1-\gamma)|\mathcal{I}|} H^{-1} G(p^*) H^{-1}\right).$$

Note that $|\mathcal{I}_k| = |\mathcal{I}|$ is a constant for any $k \in \mathbb{N}_+$. Proposition 4.7 yields that

$$\frac{\text{cost}_k}{k} \to |\mathcal{I}| \cdot c(p^*), \quad \text{almost surely.}$$

According to Slutsky's theorem, the above two convergence results immediately imply that

$$\sqrt{\text{cost}_k} \cdot (\bar{x}_{k,\gamma} - x^*) \Rightarrow \mathcal{N}\left(0, \frac{c(p^*)}{1-\gamma} H^{-1} G(p^*) H^{-1}\right).$$

Therefore, it holds that

$$\sqrt{\text{cost}_k} \cdot H(\bar{x}_{k,\gamma} - x^*) \Rightarrow \mathcal{N}\left(0, \frac{c(p^*)}{1-\gamma} G(p^*)\right),$$

and that

$$\text{cost}_k \cdot \frac{1}{2}(\bar{x}_{k,\gamma} - x^*)^T H(\bar{x}_{k,\gamma} - x^*) \Rightarrow \left\| \mathcal{N}\left(0, \frac{c(p^*)}{2(1-\gamma)} H^{-\frac{1}{2}} G(p^*) H^{-\frac{1}{2}}\right) \right\|^2.$$

Note also that $x_k \to x^*$ a.s. by Theorem 4.2 and that $\nabla f(x) = H(x - x^*) + o(\|x - x^*\|)$ and $f(x) - f(x^*) = \frac{1}{2}(x - x^*)^T H(x - x^*) + o(\|x - x^*\|^2)$ as $x \to x^*$. We can thus conclude that

$$\sqrt{\text{cost}_k} \cdot \nabla f(\bar{x}_{k,\gamma}) \Rightarrow \mathcal{N}\left(0, \frac{c(p^*)}{1-\gamma} G(p^*)\right),$$

and that

$$\text{cost}_k \cdot (f(\bar{x}_{k,\gamma}) - f(x^*)) \Rightarrow \left\| \mathcal{N}\left(0, \frac{c(p^*)}{2(1-\gamma)} H^{-\frac{1}{2}} G(p^*) H^{-\frac{1}{2}}\right) \right\|^2.$$

□

B.1 Proof of Theorem B.1

To present the proof of Theorem B.1, we need the following lemma and the notations therein.

Lemma B.3 (Lemma 1 from Polyak and Juditsky (1992)). *Suppose that $H \in \mathbb{R}^{d\times d}$ is a symmetric matrix with $H \succeq \mu I$, $\mu > 0$. Define $\{A_j^k\}_{k\geq j\geq 1} \subset \mathbb{R}^{d\times d}$ via:*

$$A_j^j = I, \quad \textit{and} \quad A_j^{k+1} = (I - \alpha_k H)A_j^k,\ k \geq j+1.$$

Set

$$S_j^k = \sum_{l=j}^{k} A_j^l,\ k \geq j \geq 1.$$

If the stepsize satisfies $\alpha_k \to 0$, $k\alpha_k \uparrow \infty$, and $\frac{\alpha_k - \alpha_{k+1}}{\alpha_k \alpha_{k+1}} \downarrow 0$, as $k \to \infty$, then the followings hold:

(i) There exists some constant $C_S > 0$, such that $\left\|H^{-1} - \alpha_j S_{j+1}^k\right\| \leq C_S,\ \forall\, k \geq j+1, j \in \mathbb{N}_+$.

(ii) $\lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k-1} \left\|H^{-1} - \alpha_j S_{j+1}^k\right\| = 0$.

Proof of Theorem B.1. The proof uses techniques from Polyak and Juditsky (1992), with some new technical lemmas. For any $k \geq k_0 \geq 1$, it can be computed that

$$\begin{aligned}
y_k - x^* &= (I - \alpha_{k-1}H)(y_{k-1} - x^*) - \alpha_{k-1}\xi_{k-1} \\
&= (I - \alpha_{k-1}H)(I - \alpha_{k-2}H)(y_{k-2} - x^*) - (I - \alpha_{k-1}H)\alpha_{k-2}\xi_{k-2} - \alpha_{k-1}\xi_{k-1} \\
&= \cdots \\
&= \left(\prod_{j=1}^{k-k_0}(I - \alpha_{k-j}H)\right)(y_{k_0} - x^*) - \sum_{j=k_0}^{k-1}\left(\prod_{l=1}^{k-j-1}(I - \alpha_{k-l}H)\right)\alpha_j\xi_j \\
&= A_{k_0}^k(y_{k_0} - x^*) - \sum_{j=k_0}^{k-1} A_{j+1}^k \alpha_j \xi_j.
\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}
\bar{y}_{k,\gamma} - x^* &= \frac{1}{(1-\gamma)k}\sum_{j=[\gamma k]+1}^{k} y_j \\
&= \frac{1}{(1-\gamma)k}\sum_{j=[\gamma k]+1}^{k}\left(A_{[\gamma k]+1}^j(y_{[\gamma k]+1} - x^*) - \sum_{l=[\gamma k]+1}^{j-1} A_{l+1}^j \alpha_l \xi_l\right) \\
&= \frac{1}{(1-\gamma)k}\left(\sum_{j=[\gamma k]+1}^{k} A_{[\gamma k]+1}^j\right)(y_{[\gamma k]+1} - x^*) - \frac{1}{(1-\gamma)k}\sum_{l=[\gamma k]+1}^{k-1}\alpha_l\left(\sum_{j=l+1}^{k} A_{l+1}^j\right)\xi_l \\
&= \frac{1}{(1-\gamma)k} S_{[\gamma k]+1}^k (y_{[\gamma k]+1} - x^*) - \frac{1}{(1-\gamma)k}\sum_{j=[\gamma k]+1}^{k-1}\alpha_j S_{j+1}^k \xi_j \\
&= \frac{1}{(1-\gamma)k} S_{[\gamma k]+1}^k (y_{[\gamma k]+1} - x^*) - \frac{1}{(1-\gamma)k}\sum_{j=[\gamma k]+1}^{k-1} H^{-1}\xi_j + \frac{1}{(1-\gamma)k}\sum_{j=[\gamma k]+1}^{k-1}\left(H^{-1} - \alpha_j S_{j+1}^k\right)\xi_j,
\end{aligned}$$

i.e.,

$$\begin{aligned}
&\sqrt{(1-\gamma)k}\cdot(\bar{y}_{k,\gamma} - x^*) = \\
&\quad \frac{1}{\sqrt{(1-\gamma)k}} S_{[\gamma k]+1}^k (y_{[\gamma k]+1} - x^*) - \frac{1}{\sqrt{(1-\gamma)k}}\sum_{j=[\gamma k]+1}^{k-1} H^{-1}\xi_j + \frac{1}{\sqrt{(1-\gamma)k}}\sum_{j=[\gamma k]+1}^{k-1}\left(H^{-1} - \alpha_j S_{j+1}^k\right)\xi_j.
\end{aligned}$$

The limiting behaviour of the three terms in $\sqrt{(1-\gamma)k}\cdot(\bar{y}_{k,\gamma}-x^*)$ are established in the Lemma B.4, Lemma B.6, and Lemma B.7, respectively. Then one can conclude (B.1). □

Lemma B.4. *Suppose that Assumption 4.1, Assumption 4.4, and Assumption 4.5 holds and that $\alpha_k = \alpha_1/k^\beta$ for $\beta \in (1/2,1)$. Suppose further that there exists a sequence $\{w_k\}_{k=1}^\infty \subset (0,1]$ satisfying $p_i^k \geq w_k/n$, $\forall\, i \in \{1,2,\ldots,n\}$, $k \geq 1$, $\sum_{k=1}^\infty \alpha_k^2/w_k < \infty$, and $\lim_{k\to\infty} \alpha_k/w_k^2 = 0$. Then*

$$\frac{1}{\sqrt{k}} S^k_{[\gamma k]+1}(y_{[\gamma k]+1} - x^*) \to 0,$$

in probability.

We need another lemma for proving Lemma B.4.

Lemma B.5. *Suppose that Assumption 4.1 holds. If there exists a sequence $\{w_k\}_{k=1}^\infty \subset (0,1]$ satisfying $p_i^k \geq w_k/n$, $\forall\, i \in \{1,2,\ldots,n\}$, $k \geq 1$, $\sum_{k=1}^\infty \alpha_k^2/w_k < \infty$, and $\lim_{k\to\infty} \frac{\alpha_k - \alpha_{k+1}}{\alpha_k^2} = 0$, then there exists $C_x > 0$, such that $\mathbb{E}\left[\|x_k - x^*\|^2\right] \leq C_x \alpha_k / w_k$.*

Proof. The proof of Lemma 4.3 implies that $\mathbb{E}\left[\|\xi_k\|^2\right] \leq C_\xi/(|\mathcal{I}_k| w_k)$, $\forall\, k \in \mathbb{N}_+$ holds for some constant $C_\xi > 0$. There exists $k_0 \in \mathbb{N}_+$, such that $\alpha_k - \alpha_{k+1} \leq \frac{\mu}{2}\alpha_k^2$, $\forall\, k \geq k_0$. Choose $C_x > 0$ such that $C_\xi/C_x < \mu/2$ and $\mathbb{E}\left[\|x_k - x^*\|^2\right] \leq C_x \alpha_k/w_k$, $\forall\, k \leq k_0$. We then prove by induction that $\mathbb{E}\left[\|x_k - x^*\|^2\right] \leq C_x \alpha_k/w_k$ holds for all $k \in \mathbb{N}_+$. Assume that $\mathbb{E}\left[\|x_k - x^*\|^2\right] \leq C_x \alpha_k/w_k$ for some k, then by (A.1), one has that

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^*\|^2\right] &\leq (1-\alpha_k\mu)\mathbb{E}\left[\|x_k - x^*\|^2\right] + \alpha_k^2 \frac{C_\xi}{|\mathcal{I}_k| w_k} \leq \frac{1}{w_k}\left(C_x\alpha_k - C_x\mu\alpha_k^2 + C_x\frac{\mu}{2}\alpha_k^2\right) \\
&= \frac{C_x}{w_k}\left(\alpha_k - \frac{\mu}{2}\alpha_k^2\right) \leq \frac{C_x \alpha_{k+1}}{w_{k+1}},
\end{aligned}$$

which completes the proof. □

Proof of Lemma B.4. Since $\alpha_{[\gamma k]} S^k_{[\gamma k]+1}$ is bounded by Lemma B.3, it suffices to show that $\frac{y_{[\gamma k]+1} - x^*}{\sqrt{k}\alpha_{[\gamma k]+1}} \to 0$ in probability, which is equivalent to $\frac{y_k - x^*}{\sqrt{k}\alpha_k} \to 0$ in probability. Note that

$$\frac{1}{\sqrt{k}\alpha_k}(y_k - x^*) = \frac{1}{\sqrt{k}\alpha_k} A_1^k (y_1 - x^*) - \frac{1}{\sqrt{k}\alpha_k}\sum_{j=1}^{k-1} A_{j+1}^k \alpha_j \xi_j.$$

Therefore, it suffices to prove that

$$\frac{1}{\sqrt{k}\alpha_k} A_1^k (y_1 - x^*) \to 0, \tag{B.2}$$

and that

$$\frac{1}{\sqrt{k}\alpha_k}\sum_{j=1}^{k-1} A_{j+1}^k \alpha_j \xi_j \to 0, \quad \text{in probability.} \tag{B.3}$$

We first prove (B.2). Let $H = U\Sigma U^\top$ be the singular value decomposition of $H \succeq \mu I$, where $U \in O(d)$ and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$ satisfies $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq \mu$. Without loss of generality, we can assume that $\alpha_1 \leq 1/\sigma_1$. Then it holds that

$$0 \preceq \frac{1}{\sqrt{k}\alpha_k} A_1^k = U\mathrm{diag}\left(\frac{1}{\sqrt{k}\alpha_k}\prod_{j=1}^{k-1}(1-\alpha_j\sigma_1), \ldots, \frac{1}{\sqrt{k}\alpha_k}\prod_{j=1}^{k-1}(1-\alpha_j\sigma_d)\right) U^T \preceq \frac{1}{\sqrt{k}\alpha_k}\prod_{j=1}^{k-1}(1-\alpha_j\mu)\cdot I.$$

Since

$$\frac{1}{\sqrt{k}\alpha_k}\prod_{j=1}^{k-1}(1-\alpha_j\mu) \leq \frac{k^{\beta-1/2}}{\alpha_1}\exp\left(-\mu\sum_{j=1}^{k-1}\alpha_j\right) \leq \frac{k^{\beta-1/2}}{\alpha_1}\exp\left(-\mu\alpha_1\int_1^k t^{-\beta}dt\right)$$

$$= \frac{k^{\beta-1/2}}{\alpha_1} \exp\left(-\frac{\mu\alpha_1}{1-\beta}(k^{1-\beta}-1)\right) \to 0, \quad \text{as } k \to \infty,$$

we get $\frac{1}{\sqrt{k}\alpha_k} A_1^k \to 0$ which implies (B.2).

We then consider (B.3). Let $\mathcal{I}_+^* = \{i : p_i^* > 0\}$ and $\delta^* = \min\{p_i^*/2 : i \in \mathcal{I}_+^*\}$. By Assumption 4.4, we know that $\mathcal{I}_+^* \neq \emptyset$ and $\delta^* > 0$. Define

$$\Omega_k = \{p_i^k \geq \delta^*, \ \forall\, i \in \mathcal{I}_+^*\}, \tag{B.4}$$

which is $\mathcal{F}_{k-1}$-measurable, and

$$\Omega_{\geq k} = \{p_i^j \geq \delta^*, \ \forall\, i \in \mathcal{I}_+^*, \ j \geq k\} = \bigcap_{j \geq k} \Omega_j, \quad \text{and} \quad \Omega_{T:k} = \bigcap_{T \leq j \leq k} \Omega_j.$$

Assumption 4.5 and the continuity of probability guarantee that

$$\lim_{k\to\infty} \mathbb{P}(\Omega_{\geq k}) = \mathbb{P}\left(\bigcup_{k\geq 0} \Omega_{\geq k}\right) = 1. \tag{B.5}$$

Even if we do not assume the L^2-boundedness of $\{\xi_k\}_{k=1}^\infty$, $\{\xi_k \mathbb{I}_{\Omega_k}\}_{k=1}^\infty$ can be proved as bounded in the L^2 sense:

$$\begin{aligned}
\mathbb{E}\left[\|\xi_k\|^2 \mathbb{I}_{\Omega_k}\right] &= \mathbb{E}\left[\mathbb{E}\left[\|\xi_k\|^2 \mathbb{I}_{\Omega_k} | \mathcal{F}_{k-1}\right]\right] \\
&= \mathbb{E}\left[\left(\frac{1}{|\mathcal{I}_k| n^2} \sum_{i=1}^n \frac{1}{p_i^k} \|\nabla f_i(x_k)\|^2 - \frac{1}{|\mathcal{I}_k|} \|\nabla f(x_k)\|^2\right) \mathbb{I}_{\Omega_k}\right] \\
&\leq \frac{1}{n^2} \mathbb{E}\left[\sum_{i \in \mathcal{I}_+^*} \frac{\|\nabla f_i(x_k)\|^2}{\delta^*} + \sum_{i \notin \mathcal{I}_+^*} \frac{\|\nabla f_i(x_k)\|^2}{p_i^k}\right] \\
&\leq \frac{1}{n^2} \mathbb{E}\left[\sum_{i \in \mathcal{I}_+^*} \frac{(\|\nabla f_i(x^*)\| + L\|x_k - x^*\|)^2}{\delta^*} + \sum_{i \notin \mathcal{I}_+^*} \frac{\|x_k - x^*\|^2}{w_k}\right] \\
&\leq C'_{\xi,\Omega}\left(1 + \mathbb{E}\left[\|x_k - x^*\|^2\right] + \frac{1}{w_k} \mathbb{E}\left[\|x_k - x^*\|^2\right]\right) \\
&\leq C'_{\xi,\Omega}\left(1 + \frac{C_x \alpha_k}{w_k} + \frac{C_x \alpha_k}{w_k^2}\right) \\
&\leq C_{\xi,\Omega},
\end{aligned}$$

i.e.,

$$\mathbb{E}\left[\|\xi_k\|^2 \mathbb{I}_{\Omega_k}\right] \leq C_{\xi,\Omega}, \tag{B.6}$$

for some constant $C'_{\xi,\Omega}, C_{\xi,\Omega} > 0$, where we used Lemma B.5 and $\lim_{k\to\infty} \alpha_k / w_k^2 = 0$.

Consider any $k \geq T$. It holds that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=1}^{k-1} A_{j+1}^k \alpha_j \xi_j \mathbb{I}_{\Omega_{\geq T}}\right\|^2\right] &\leq \mathbb{E}\left[\left\|\sum_{j=1}^{k-1} A_{j+1}^k \alpha_j \xi_j \mathbb{I}_{\Omega_{T:k-1}}\right\|^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{j=1}^{k-2} A_{j+1}^k \alpha_j \xi_j \mathbb{I}_{\Omega_{T:k-1}} + A_k^k \alpha_{k-1} \xi_{k-1} \mathbb{I}_{\Omega_{T:k-1}}\right\|^2 \middle| \mathcal{F}_{k-2}\right]\right] \\
&= \mathbb{E}\left[\left\|\sum_{j=1}^{k-2} A_{j+1}^k \alpha_j \xi_j \mathbb{I}_{\Omega_{T:k-1}}\right\|^2\right] + \mathbb{E}\left[\|A_k^k \alpha_{k-1} \xi_{k-1} \mathbb{I}_{\Omega_{T:k-1}}\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left\|\sum_{j=1}^{k-2} A_{j+1}^k \alpha_j \xi_j \mathbb{I}_{\Omega_{T:k-2}}\right\|^2\right] + \alpha_{k-1}^2 \|A_k^k\|^2 \mathbb{E}\left[\|\xi_{k-1}\|^2 \mathbb{I}_{\Omega_{k-1}}\right] \\
&\leq \cdots \\
&\leq \mathbb{E}\left[\left\|\sum_{j=1}^{T-1} A_{j+1}^k \alpha_j \xi_j\right\|^2\right] + \sum_{j=T}^{k-1} \alpha_j^2 \|A_{j+1}^k\|^2 \mathbb{E}\left[\|\xi_j\|^2 \mathbb{I}_{\Omega_j}\right] \\
&\leq \sum_{j=1}^{T-1} \alpha_j^2 \|A_{j+1}^k\|^2 \mathbb{E}\left[\|\xi_j\|^2\right] + C_{\xi,\Omega} \sum_{j=T}^{k-1} \alpha_j^2 \|A_{j+1}^k\|^2,
\end{aligned}
$$

i.e.,

$$
\mathbb{E}\left[\left\|\frac{1}{\sqrt{k}\alpha_k} \sum_{j=1}^{k-1} A_{j+1}^k \alpha_j \xi_j \mathbb{I}_{\Omega_{\geq T}}\right\|^2\right] \leq \frac{1}{k\alpha_k^2} \sum_{j=1}^{T-1} \alpha_j^2 \|A_{j+1}^k\|^2 \mathbb{E}\left[\|\xi_j\|^2\right] + \frac{C_{\xi,\Omega}}{k\alpha_k^2} \sum_{j=T}^{k-1} \alpha_j^2 \|A_{j+1}^k\|^2. \tag{B.7}
$$

Recall that from the singular value decomposition of H, one can compute that

$$
A_T^k = U \mathrm{diag}\left(\prod_{l=T}^{k-1}(1-\alpha_l\sigma_1), \ldots, \prod_{l=T}^{k-1}(1-\alpha_l\sigma_d)\right) U^T.
$$

Thus, for any fixed T with $\alpha_T < 1/\sigma_1$, we have for

$$
\|A_T^k\| \leq \prod_{l=T}^{k-1}(1-\alpha_l\mu) \leq \exp\left(-\mu \sum_{l=T}^{k-1} \alpha_l\right) \leq \exp\left(-\mu\alpha_1 \int_T^k t^{-\beta} dt\right) \leq \exp\left(-\frac{\mu\alpha_1}{1-\beta}\left(k^{1-\beta} - T^{1-\beta}\right)\right), \tag{B.8}
$$

which implies that

$$
\begin{aligned}
\frac{1}{k\alpha_k^2} \sum_{j=1}^{T-1} \alpha_j^2 \|A_{j+1}^k\|^2 \mathbb{E}\left[\|\xi_j\|^2\right] &\leq \sum_{j=1}^{T-1} \alpha_j^2 \|A_{j+1}^T\|^2 \mathbb{E}\left[\|\xi_j\|^2\right] \cdot \frac{1}{k\alpha_k^2} \|A_T^k\|^2 \\
&\leq \sum_{j=1}^{T-1} \alpha_j^2 \|A_{j+1}^T\|^2 \mathbb{E}\left[\|\xi_j\|^2\right] \cdot \frac{k^{2\beta-1}}{\alpha_1^2} \exp\left(-\frac{2\mu\alpha_1}{1-\beta}(k^{1-\beta} - T^{1-\beta})\right) \to 0,
\end{aligned} \tag{B.9}
$$

as $k \to \infty$. Consider $h(x) = x^{1-\beta}$, which is increasing and concave. So $h(k - k^{\frac{\beta+1}{2}}) \leq h(k) - k^{\frac{\beta+1}{2}} h'(k) = h(k) - (1-\beta)k^{\frac{1-\beta}{2}}$, which implies

$$
k^{1-\beta} - (k - k^{\frac{\beta+1}{2}})^{1-\beta} \geq (1-\beta)k^{\frac{1-\beta}{2}}.
$$

For any fixed T with $\alpha_T < 1/\sigma_1$ and any $j \geq T$, it can be estimated in a way similar to (B.8) that

$$
\|A_{j+1}^k\| \leq \exp\left(-\frac{\mu\alpha_1}{1-\beta}\left(k^{1-\beta} - (j+1)^{1-\beta}\right)\right) \leq \exp\left(\frac{\mu\alpha_1}{1-\beta}(T+1)^{1-\beta}\right).
$$

Thus, one has that

$$
\begin{aligned}
\frac{1}{k\alpha_k^2} \sum_{j=T}^{k-1} \alpha_j^2 \|A_{j+1}^k\|^2 \leq &\sum_{T \leq j < k-k^{\frac{\beta+1}{2}}} \frac{\alpha_j^2}{k\alpha_k^2} \exp\left(-\frac{\mu\alpha_1}{1-\beta}\left(k^{1-\beta} - (j+1)^{1-\beta}\right)\right) \\
&+ \exp\left(\frac{\mu\alpha_1}{1-\beta}(T+1)^{1-\beta}\right) \cdot \frac{1}{k} \sum_{k-k^{\frac{\beta+1}{2}} \leq j < k} \frac{\alpha_j^2}{\alpha_k^2} \\
\leq &\sum_{T \leq j < k-k^{\frac{\beta+1}{2}}} \frac{\alpha_j^2}{k\alpha_k^2} \exp\left(-\frac{\mu\alpha_1}{1-\beta}\left(k^{1-\beta} - (k - k^{\frac{\beta+1}{2}})^{1-\beta}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\exp\left(\frac{\mu\alpha_1}{1-\beta}(T+1)^{1-\beta}\right)\cdot\frac{1}{k}\sum_{k-k^{\frac{\beta+1}{2}}\le j<k}\left(\frac{k}{k-k^{\frac{\beta+1}{2}}}\right)^{2\beta}\\
\le& k^{2\beta}\exp\left(-\frac{\mu\alpha_1}{1-\beta}\cdot(1-\beta)k^{\frac{1-\beta}{2}}\right)\\
&+\exp\left(\frac{\mu\alpha_1}{1-\beta}(T+1)^{1-\beta}\right)\cdot\frac{k^{\frac{\beta+1}{2}}}{k}\left(\frac{k}{k-k^{\frac{\beta+1}{2}}}\right)^{2\beta}\to 0,
\end{aligned}
$$

which combined with (B.7) and (B.9) yields that

$$
\lim_{k\to\infty}\mathbb{E}\left[\left\|\frac{1}{\sqrt{k}\alpha_k}\sum_{j=1}^{k-1}A_{j+1}^k\alpha_j\xi_j\mathbb{I}_{\Omega_{\ge T}}\right\|^2\right]\to 0.
$$

Then we are ready to prove (B.3). For any $\epsilon,\delta>0$, according to (B.5), there exists T such that $\mathbb{P}(\Omega_{\ge T})\ge 1-\delta$. We can further require that $\alpha_T<1/\sigma_1$. Thus,

$$
\begin{aligned}
\mathbb{P}\left[\left\|\frac{1}{\sqrt{k}\alpha_k}\sum_{j=1}^{k-1}A_{j+1}^k\alpha_j\xi_j\right\|>\epsilon\right]&\le\mathbb{P}\left[\left\|\frac{1}{\sqrt{k}\alpha_k}\sum_{j=1}^{k-1}A_{j+1}^k\alpha_j\xi_j\mathbb{I}_{\Omega_{\ge T}}\right\|>\epsilon\right]+\mathbb{P}(\Omega_{\ge T}^C)\\
&\le\frac{1}{\epsilon^2}\mathbb{E}\left[\left\|\frac{1}{\sqrt{k}\alpha_k}\sum_{j=1}^{k-1}A_{j+1}^k\alpha_j\xi_j\mathbb{I}_{\Omega_{\ge T}}\right\|^2\right]+\delta<2\delta,
\end{aligned}
$$

for sufficiently large k, which implies that

$$
\lim_{k\to\infty}\mathbb{P}\left[\left\|\frac{1}{\sqrt{k}\alpha_k}\sum_{j=1}^{k-1}A_{j+1}^k\alpha_j\xi_j\right\|>\epsilon\right]=0,
$$

and hence the convergence in probability (B.3). The proof is completed. □

Lemma B.6. *Suppose Assumptions 4.1, 4.4 and 4.5 hold, and the minibatch size $|\mathcal{I}_k|=|\mathcal{I}|$ is fixed for any $k\ge 1$. If there exists a non-increasing sequence $\{w_k\}_{k=1}^\infty\subset(0,1]$ satisfying $p_i^k\ge w_k/n,\ \forall\ i\in\{1,2,\dots,n\}$ and $k\ge 1$, $\sum_{k=1}^\infty\alpha_k^2/w_k<\infty$, $\lim_{k\to\infty}\frac{\alpha_k-\alpha_{k+1}}{\alpha_k^2}=0$, and $\lim_{k\to\infty}\alpha_k/w_k^2=0$, then*

$$
\frac{1}{\sqrt{(1-\gamma)k}}\sum_{j=[\gamma k]+1}^{k-1}\xi_j\Rightarrow\mathcal{N}\left(0,\frac{1}{|\mathcal{I}|}G(p^*)\right).
$$

Proof. At the j-th iteration, we consider sampling the gradients in the following way: We generate $|\mathcal{I}|$ independent $\mathrm{Uniform}(0,1)$ variables, $U_{j,s},\ s=1,\dots,|\mathcal{I}|$. Let $P_i^j=\sum_{l=1}^i p_l^j$ be the sum of the sampling weights of the first i components, and form

$$
\xi_j'=\frac{1}{|\mathcal{I}|}\sum_{s=1}^{|\mathcal{I}|}\sum_{i=1}^n\mathbb{I}_{P_{i-1}^j\le U_{j,s}<P_i^j}\left(\frac{1}{np_i^j}\nabla f_i(x_j)-\nabla f(x_j)\right),\tag{B.10}
$$

then $\frac{1}{\sqrt{(1-\gamma)k}}\sum_{j=[\gamma k]+1}^{k-1}\xi_j$ and $\frac{1}{\sqrt{(1-\gamma)k}}\sum_{j=[\gamma k]+1}^{k-1}\xi_j'$ are equal in distribution for any k. We couple (B.10) with the following counterpart at the optimum x^*

$$
\xi_j^*=\frac{1}{|\mathcal{I}|}\sum_{s=1}^{|\mathcal{I}|}\sum_{i=1}^n\mathbb{I}_{P_{i-1}^*\le U_{j,s}<P_i^*}\frac{1}{np_i^*}\nabla f_i(x^*),\tag{B.11}
$$

where each $P_i^*=\sum_{l=1}^i p_l^*$ is the sum of the limit sampling weights p^* of the first i components.

Since $\xi_j^*, j \geq 1$ are i.i.d., by the standard multivariate central limit theorem we have

$$\frac{1}{\sqrt{(1-\gamma)k}} \sum_{j=[\gamma k]+1}^{k-1} \xi_j^* \Rightarrow \mathcal{N}\left(0, \frac{1}{|\mathcal{I}|} G(p^*)\right),$$

therefore it suffices to show that

$$\frac{1}{\sqrt{(1-\gamma)k}} \sum_{j=[\gamma k]+1}^{k-1} (\xi_j' - \xi_j^*) = o_p(1).$$

Recall the event $\Omega_k, \Omega_{\geq k}$ from the proof of Lemma B.4, and using the same argument therein we see that the above is equivalent to

$$\frac{1}{\sqrt{(1-\gamma)k}} \sum_{j=[\gamma k]+1}^{k-1} (\xi_j' - \xi_j^*) \mathbb{I}_{\Omega_{\geq T}} = o_p(1) \text{ for each fixed } T.$$

By the expressions (B.10) and (B.11) for ξ_j' and ξ_j^*, both in the form of finite sum, it suffices to show

$$\frac{1}{\sqrt{(1-\gamma)k}} \sum_{j=[\gamma k]+1}^{k-1} \left(\mathbb{I}_{P_{i-1}^j \leq U_{j,s} < P_i^j} \left(\frac{1}{np_i^j} \nabla f_i(x_j) - \nabla f(x_j) \right) - \mathbb{I}_{P_{i-1}^* \leq U_{j,s} < P_i^*} \frac{1}{np_i^*} \nabla f_i(x^*) \right) \mathbb{I}_{\Omega_{\geq T}} = o_p(1),$$

for each fixed T, i and s. To show this, we write

$$\begin{aligned}
& \mathbb{E}\left[\left\| \frac{1}{\sqrt{(1-\gamma)k}} \sum_{j=[\gamma k]+1}^{k-1} \left(\mathbb{I}_{P_{i-1}^j \leq U_{j,s} < P_i^j} \left(\frac{1}{np_i^j} \nabla f_i(x_j) - \nabla f(x_j) \right) - \mathbb{I}_{P_{i-1}^* \leq U_{j,s} < P_i^*} \frac{1}{np_i^*} \nabla f_i(x^*) \right) \mathbb{I}_{\Omega_{\geq T}} \right\|^2 \right] \\
= & \frac{1}{(1-\gamma)k} \mathbb{E}\left[\left\| \left(\sum_{j=[\gamma k]+1}^{T-1} \left(\mathbb{I}_{P_{i-1}^j \leq U_{j,s} < P_i^j} \left(\frac{1}{np_i^j} \nabla f_i(x_j) - \nabla f(x_j) \right) - \mathbb{I}_{P_{i-1}^* \leq U_{j,s} < P_i^*} \frac{1}{np_i^*} \nabla f_i(x^*) \right) \right. \right. \right. \\
& \left. \left. \left. + \sum_{j=T}^{k-1} \left(\mathbb{I}_{P_{i-1}^j \leq U_{j,s} < P_i^j} \left(\frac{1}{np_i^j} \nabla f_i(x_j) - \nabla f(x_j) \right) - \mathbb{I}_{P_{i-1}^* \leq U_{j,s} < P_i^*} \frac{1}{np_i^*} \nabla f_i(x^*) \right) \mathbb{I}_{\Omega_j} \right) \mathbb{I}_{\Omega_{\geq T}} \right\|^2 \right] \\
\leq & \frac{1}{(1-\gamma)k} \mathbb{E}\left[\left\| \sum_{j=[\gamma k]+1}^{T-1} \left(\mathbb{I}_{P_{i-1}^j \leq U_{j,s} < P_i^j} \left(\frac{1}{np_i^j} \nabla f_i(x_j) - \nabla f(x_j) \right) - \mathbb{I}_{P_{i-1}^* \leq U_{j,s} < P_i^*} \frac{1}{np_i^*} \nabla f_i(x^*) \right) \right. \right. \\
& \left. \left. + \sum_{j=T}^{k-1} \left(\mathbb{I}_{P_{i-1}^j \leq U_{j,s} < P_i^j} \left(\frac{1}{np_i^j} \nabla f_i(x_j) - \nabla f(x_j) \right) - \mathbb{I}_{P_{i-1}^* \leq U_{j,s} < P_i^*} \frac{1}{np_i^*} \nabla f_i(x^*) \right) \mathbb{I}_{\Omega_j} \right\|^2 \right] \\
= & \frac{1}{(1-\gamma)k} \sum_{j=[\gamma k]+1}^{T-1} \mathbb{E}\left[\left\| \mathbb{I}_{P_{i-1}^j \leq U_{j,s} < P_i^j} \left(\frac{1}{np_i^j} \nabla f_i(x_j) - \nabla f(x_j) \right) - \mathbb{I}_{P_{i-1}^* \leq U_{j,s} < P_i^*} \frac{1}{np_i^*} \nabla f_i(x^*) \right\|^2 \right] \\
& + \frac{1}{(1-\gamma)k} \sum_{j=T}^{k-1} \mathbb{E}\left[\left\| \left(\mathbb{I}_{P_{i-1}^j \leq U_{j,s} < P_i^j} \left(\frac{1}{np_i^j} \nabla f_i(x_j) - \nabla f(x_j) \right) - \mathbb{I}_{P_{i-1}^* \leq U_{j,s} < P_i^*} \frac{1}{np_i^*} \nabla f_i(x^*) \right) \mathbb{I}_{\Omega_j} \right\|^2 \right],
\end{aligned}$$

where the last equality follows from the martingale increment property of the gradient errors. Since the first sum above eventually becomes 0 as $k \to \infty$ for a fixed T, we focus on the second sum. To show that the second sum approaches 0 as $k \to \infty$, it suffices to show that

$$a_k := \mathbb{E}\left[\left\| \left(\mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k} \left(\frac{1}{np_i^k} \nabla f_i(x_k) - \nabla f(x_k) \right) - \mathbb{I}_{P_{i-1}^* \leq U_{k,s} < P_i^*} \frac{1}{np_i^*} \nabla f_i(x^*) \right) \mathbb{I}_{\Omega_k} \right\|^2 \right] \to 0, \quad \text{as } k \to \infty. \tag{B.12}$$

We consider two cases:

(i) $p_i^* > 0$. In this case, by Minkowski inequality we can bound a_k as

$$\begin{aligned}
\sqrt{a_k} &\leq \sqrt{\mathbb{E}\left[\left\|\mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k}\left(\frac{1}{np_i^k}\nabla f_i(x_k) - \nabla f(x_k) - \frac{1}{np_i^*}\nabla f_i(x^*)\right)\right\|^2 \mathbb{I}_{\Omega_k}\right]} \\
&\quad + \sqrt{\mathbb{E}\left[\left\|\left(\mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k} - \mathbb{I}_{P_{i-1}^* \leq U_{k,s} < P_i^*}\right)\frac{1}{np_i^*}\nabla f_i(x^*)\right\|^2 \mathbb{I}_{\Omega_k}\right]} \\
&= \sqrt{\mathbb{E}\left[\left\|\frac{1}{np_i^k}\left(\nabla f_i(x_k) - \nabla f_i(x^*)\right) - \nabla f(x_k) + \left(\frac{1}{np_i^k} - \frac{1}{np_i^*}\right)\nabla f_i(x^*)\right\|^2 \mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k}\mathbb{I}_{\Omega_k}\right]} \\
&\quad + \sqrt{\mathbb{E}\left[\left\|\left(\mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k} - \mathbb{I}_{P_{i-1}^* \leq U_{k,s} < P_i^*}\right)\frac{1}{np_i^*}\nabla f_i(x^*)\right\|^2 \mathbb{I}_{\Omega_k}\right]} \\
&\leq \sqrt{\mathbb{E}\left[\left(\frac{L\|x_k - x^*\|}{n\delta^*} + L\|x_k - x^*\| + \frac{|p_i^k - p_i^*|\|\nabla f_i(x^*)\|}{n{\delta^*}^2}\right)^2 \mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k}\mathbb{I}_{\Omega_k}\right]} \\
&\quad + \sqrt{\mathbb{E}\left[\left\|\left(\mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k} - \mathbb{I}_{P_{i-1}^* \leq U_{k,s} < P_i^*}\right)\frac{1}{np_i^*}\nabla f_i(x^*)\right\|^2\right]} \\
&\leq \sqrt{\mathbb{E}\left[\left(\frac{L}{n\delta^*} + L\right)^2 \|x_k - x^*\|^2\right]} + \sqrt{\mathbb{E}\left[\frac{(p_i^k - p_i^*)^2\|\nabla f_i(x^*)\|^2}{n^2{\delta^*}^4}\right]} \qquad \text{(B.13)} \\
&\quad + \sqrt{\mathbb{E}\left[\left\|\left(\mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k} - \mathbb{I}_{P_{i-1}^* \leq U_{k,s} < P_i^*}\right)\frac{1}{np_i^*}\nabla f_i(x^*)\right\|^2\right]}, \qquad \text{(B.14)}
\end{aligned}$$

where the second inequality follows the definition of Ω_k and Assumption 4.1. The first term in (B.13) converges to 0 since $\mathbb{E}[\|x_k - x^*\|^2] \leq C_x\alpha_k/w_k \to 0$ by Lemma B.5. The second term in (B.13) and the term in (B.14) both converge to 0 by that $p_i^k \to p_i^*$ and $P_i^k \to P_i^*$ a.s. for each i and the bounded convergence theorem. Therefore, we have $a_k \to 0$.

(ii) $p_i^* = 0$. In this case, $\nabla f_i(x^*) = 0$ by Assumption 4.5, and we can bound a_k as follows

$$\begin{aligned}
a_k &= \mathbb{E}\left[\left\|\mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k}\left(\frac{1}{np_i^k}\nabla f_i(x_k) - \nabla f(x_k)\right)\mathbb{I}_{\Omega_k}\right\|^2\right] \\
&\leq \mathbb{E}\left[\left(\frac{L\|x_k - x^*\|}{np_i^k} + L\|x_k - x^*\|\right)^2 \mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{L\|x_k - x^*\|}{np_i^k} + L\|x_k - x^*\|\right)^2 \mathbb{I}_{P_{i-1}^k \leq U_{k,s} < P_i^k}\middle|\mathcal{F}_{k-1}\right]\right] \\
&= \mathbb{E}\left[\left(\frac{L\|x_k - x^*\|}{np_i^k} + L\|x_k - x^*\|\right)^2 p_i^k\right] \\
&\leq 2\mathbb{E}\left[\frac{L^2\|x_k - x^*\|^2}{n^2 p_i^k}\right] + 2\mathbb{E}\left[\|x_k - x^*\|^2\right] \quad \text{by Young's inequality} \\
&\leq C_x\left(\frac{2L^2\alpha_k}{nw_k^2} + \frac{2\alpha_k}{w_k}\right) \quad \text{by Lemma B.5.} \qquad \text{(B.15)}
\end{aligned}$$

The assumed condition $\alpha_k/w_k^2 \to 0$ then immediately implies that (B.15), hence a_k, approaches 0 as $k \to \infty$.

This concludes (B.12), and hence completes the proof. □

Lemma B.7. *Under the same assumptions as in Lemma B.4, it holds that*

$$\frac{1}{\sqrt{k}} \sum_{j=[\gamma k]+1}^{k-1} \left(H^{-1} - \alpha_j S_{j+1}^k\right) \xi_j \to 0,$$

in probability.

Proof. Similarly to (B.7), one has for $k, T \in \mathbb{N}_+$ with $[\gamma k] + 1 \geq T$ that

$$\mathbb{E}\left[\left\| \frac{1}{\sqrt{k}} \sum_{j=[\gamma k]+1}^{k-1} \left(H^{-1} - \alpha_j S_{j+1}^k\right) \xi_j \mathbb{I}_{\Omega_{\geq T}} \right\|^2\right] \leq \frac{C_{\xi,\Omega}}{k} \sum_{j=[\gamma k]+1}^{k-1} \|H^{-1} - \alpha_j S_{j+1}^k\|^2.$$

Then using Lemma B.3, one has that

$$\mathbb{E}\left[\left\| \frac{1}{\sqrt{k}} \sum_{j=[\gamma k]+1}^{k-1} \left(H^{-1} - \alpha_j S_{j+1}^k\right) \xi_j \mathbb{I}_{\Omega_{\geq T}} \right\|^2\right] \leq \frac{C_{\xi,\Omega} C_S}{k} \sum_{j=1}^{k-1} \|H^{-1} - \alpha_j S_{j+1}^k\| \to 0,$$

as $k \to \infty$. The rest of the proof is similar to the last part in the proof of Lemma B.4. □

B.2 Proof of Theorem B.2

We present the proof of Theorem B.2 in this subsection.

Proof of Theorem B.2. This proof also uses some techniques from Polyak and Juditsky (1992) with some new technical lemmas. It can be computed that

$$\begin{aligned}
\delta_{k+1} &= x_{k+1} - y_{k+1} \\
&= (x_k - \alpha_k \nabla f(x_k) - \alpha_k \xi_k) - (y_k - \alpha_k H(y_k - x^*) - \alpha_k \xi_k) \\
&= (x_k - y_k) - \alpha_k H(x_k - y_k) - \alpha_k (\nabla f(x_k) - H(x_k - x^*)) \\
&= (I - \alpha_k H)(x_k - y_k) - \alpha_k (\nabla f(x_k) - H(x_k - x^*)) \\
&= (I - \alpha_k H)\delta_k - \alpha_k (\nabla f(x_k) - H(x_k - x^*)).
\end{aligned}$$

Using the same techniques in the calculation of $\bar{y}_{k,\gamma} - x^*$, one obtains that

$$\sqrt{(1-\gamma)k} \cdot \bar{\delta}_{k,\gamma} = \frac{1}{\sqrt{(1-\gamma)k}} S_{[\gamma k]+1}^k \delta_{[\gamma k]+1} - \frac{1}{\sqrt{(1-\gamma)k}} \sum_{j=[\gamma k]+1}^{k-1} \alpha_j S_{j+1}^k (\nabla f(x_j) - H(x_j - x^*)). \tag{B.16}$$

The first part in (B.16) converges to 0 in probability by Lemma B.4 and Lemma B.8. Then we estimate the second part. By Fatou's lemma and Lemma B.5, it holds that

$$\mathbb{E}\left[\sum_{k=1}^{\infty} \frac{\|x_k - x^*\|^2}{\sqrt{k}}\right] \leq \sum_{k=1}^{\infty} \frac{\mathbb{E}\left[\|x_k - x^*\|^2\right]}{\sqrt{k}} \leq C_x \sum_{k=1}^{\infty} \frac{\alpha_k}{w_k \sqrt{k}} < \infty,$$

which implies that

$$\sum_{k=1}^{\infty} \frac{\|x_k - x^*\|^2}{\sqrt{k}} < \infty, \quad \text{a.s..}$$

almost surely. Since f is twice continuously differentiable in a neighbourhood of x^*, it holds that

$$\|\nabla f(x) - H(x - x^*)\| \leq C_H \|x - x^*\|^2,$$

for some constant $C_H > 0$ and all x in some neighbourhood of x^*. According to Theorem 4.2, almost surely, the sequence x_j converges to x^* and always stays in the neighbourhood after some finite time. Then for sufficiently large k, one can conclude that

$$\left\| \frac{1}{\sqrt{k}} \sum_{j=[\gamma k]+1}^{k-1} \alpha_j S_{j+1}^k (\nabla f(x_j) - H(x_j - x^*)) \right\| \leq C_H (C_S + \|H^{-1}\|) \cdot \frac{1}{\sqrt{k}} \sum_{j=[\gamma k]+1}^{k-1} \|x_j - x^*\|^2 \to 0,$$

where we used the Kronecker's lemma and Lemma B.3. This proves that the second part in (B.16) almost surely converges to 0. □

Lemma B.8. *Suppose that Assumption 4.1, Assumption 4.4, and Assumption 4.5 hold and that $\alpha_k = \alpha_1/k^\beta$ for $\beta \in (1/2, 1)$. Suppose further that there exists a sequence $\{w_k\}_{k=1}^\infty \subset (0, 1]$ satisfying $p_i^k \geq w_k/n$, $\forall\, i \in \{1, 2, \ldots, n\}$, $k \geq 1$, $\sum_{k=1}^\infty \alpha_k^2/w_k < \infty$, and $\lim_{k\to\infty} \alpha_k/w_k^2 = 0$. Then*

$$\frac{1}{\sqrt{k}} S_{[\gamma k]+1}^k (x_{[\gamma k]+1} - x^*) \to 0,$$

in probability.

Proof. Similar to Lemma B.4, it suffices to show that $\frac{x_k - x^*}{\sqrt{k}\alpha_k} \to 0$ in probability. Consider any fixed $T \in \mathbb{N}_+$ with $\alpha_T < \min\{1/\mu, \mu/L^2\}$. Similar to (A.1), it holds for $k > T$ that

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mathbb{I}_{\Omega_{T:k}} | \mathcal{F}_{k-1}\right] &= \mathbb{E}\left[\|x_k - x^* - \alpha_k \nabla f(x_k) - \alpha_k \xi_k\|^2 \mathbb{I}_{\Omega_{T:k}} | \mathcal{F}_{k-1}\right] \\
&= \|x_k - x^* - \alpha_k \nabla f(x_k)\|^2 \mathbb{I}_{\Omega_{T:k}} + \alpha_k^2 \mathbb{E}\left[\|\xi_k\|^2 \mathbb{I}_{\Omega_{T:k}} | \mathcal{F}_{k-1}\right] \\
&\leq (1 - \alpha_k \mu) \|x_k - x^*\|^2 \mathbb{I}_{\Omega_{T:k}} + \alpha_k^2 \mathbb{E}\left[\|\xi_k\|^2 \mathbb{I}_{\Omega_{T:k}} | \mathcal{F}_{k-1}\right],
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x^*\|^2 \mathbb{I}_{\Omega_{T:k}}\right] &\leq (1 - \alpha_k \mu) \mathbb{E}\left[\|x_k - x^*\|^2 \mathbb{I}_{\Omega_{T:k}}\right] + \alpha_k^2 \mathbb{E}\left[\|\xi_k\|^2 \mathbb{I}_{\Omega_{T:k}}\right] \\
&\leq (1 - \alpha_k \mu) \mathbb{E}\left[\|x_k - x^*\|^2 \mathbb{I}_{\Omega_{T:k-1}}\right] + C_{\xi,\Omega} \alpha_k^2 \\
&\leq (1 - \alpha_k \mu)(1 - \alpha_{k-1} \mu) \mathbb{E}\left[\|x_{k-1} - x^*\|^2 \mathbb{I}_{\Omega_{T:k-2}}\right] + C_{\xi,\Omega} \alpha_{k-1}^2 (1 - \alpha_k \mu) + C \alpha_k^2 \\
&\leq \prod_{j=T}^k (1 - \alpha_j \mu) \mathbb{E}[\|x_T - x^*\|^2] + C_{\xi,\Omega} \sum_{j=T}^k \alpha_j^2 \prod_{l=j+1}^k (1 - \alpha_j \mu),
\end{aligned}$$

where we used (B.6). Therefore, using similar arguments as in the proof of Lemma B.4, one can establish that

$$\lim_{k\to\infty} \mathbb{E}\left[\frac{1}{\alpha_k \sqrt{k}} \|x_{k+1} - x^*\|^2 \mathbb{I}_{\Omega_{T:k}}\right] = 0,$$

and hence that the convergence in probability $\frac{x_k - x^*}{\sqrt{k}\alpha_k} \to 0$. □

C PROOFS FOR PROPOSITIONS 4.6, 4.7, AND 4.9

Proof of Proposition 4.6. According to Proposition 2.2, the optimal solution to the subproblem (2.5) is continuous in its coefficients. Note also that $\lim_{k\to\infty} w_k = 0$. Therefore, it suffices to show that the followings hold almost surely:

- $\lim_{k\to\infty} \tilde{c}^k \to (c_1, \ldots, c_n)$.
- $\lim_{k\to\infty} \tilde{g}_i^k = \|\nabla f_i(x^*)\|$, $\forall\, i \in \{1, 2, \ldots, n\}$.
- $\lim_{k\to\infty} \tilde{G}_k = 0$.

(i) Since $p_i^k \geq w_k$ and $\sum_{k=1}^\infty w_k = \infty$, every index i will be sampled for infinitely many times almost surely, which leads to $\lim_{k\to\infty} \tilde{c}^k \to (c_1, \ldots, c_n)$ by (C.2).

(ii) It follows from Theorem 4.2 that $x_k \to x^*$ almost surely. Then for any $i \in \{1, 2, \ldots, n\}$, one has $\lim_{k\to\infty} \tilde{g}_i^k = \|\nabla f_i(x^*)\|$ a.s. since i will be sampled for infinitely many times.

(iii) One has that

$$\tilde{G}_k = \frac{1}{k-1}\sum_{j=1}^{k-1} g_j = \frac{1}{k-1}\sum_{j=1}^{k-1}\xi_j + \frac{1}{k-1}\sum_{j=1}^{k-1}\nabla f(x_j). \tag{C.1}$$

Set $Y_k = \|\frac{1}{k}\sum_{j=1}^{k}\xi_j\|^2$. Then one can compute that

$$\mathbb{E}[Y_{k+1}|\mathcal{F}_k] = \mathbb{E}\left[\left\|\frac{1}{k+1}\xi_{k+1} + \frac{k}{k+1}\cdot\frac{1}{k}\sum_{j=1}^{k}\xi_j\right\|^2\middle|\mathcal{F}_k\right] = \frac{1}{(k+1)^2}\mathbb{E}[\|\xi_{k+1}\|^2|\mathcal{F}_k] + \frac{k^2}{(k+1)}Y_k,$$

which implies that

$$\mathbb{E}[Y_{k+1} - Y_k|\mathcal{F}_k]_+ \le \frac{1}{(k+1)^2}\mathbb{E}[\|\xi_{k+1}\|^2|\mathcal{F}_k],$$

and hence that

$$\sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{E}[Y_{k+1} - Y_k|\mathcal{F}_k]_+\right] \le \sum_{k=1}^{\infty}\frac{1}{(k+1)^2}\mathbb{E}[\|\xi_{k+1}\|^2] < \infty,$$

where we use $\sum_{k=1}^{\infty}\alpha_k^2\mathbb{E}[\|\xi_k\|^2]$ from Lemma 4.3 and $\inf_{k\ge1} k\alpha_k > 0$. Therefore, by martingale convergence theorem (Blum, 1954, Corollary in Section 3), Y_k converges almost surely to some random variable Y. On the other hand, it follows from $\sum_{k=1}^{\infty}\frac{1}{k^2}\mathbb{E}[\|\xi_k\|^2] < \infty$ and the Kronecker's lemma that

$$\mathbb{E}Y_k = \frac{1}{k^2}\sum_{j=1}^{k}\mathbb{E}[\|\xi_j\|^2] \to 0,$$

which guarantees that $\{Y_k\}_{k=1}^{\infty}$ has a subsequence that converges to 0 almost surely. Therefore, $Y_k = \|\frac{1}{k}\sum_{j=1}^{k}\xi_j\|^2 \to 0$ a.s., which implies that the first part in (C.1) converges to 0 almost surely. In addition, since $x_k \to x^*$ a.s., we immediately have that $\frac{1}{k-1}\sum_{j=1}^{k-1}\nabla f(x_j) \to 0$ a.s.. Thus, it holds that $\lim_{k\to\infty}\tilde{G}_k = 0$ almost surely. □

Proof for Proposition 4.7. Define $\eta_{i,s} = 1$ if the s-th sampled gradient throughout the algorithm is from f_i, and 0 otherwise. Note that the index s does not necessarily correspond to the iteration index as multiple samples can be drawn in each iteration (i.e., $|\mathcal{I}_k| > 1$ in Algorithm 1). Let $s_k = \sum_{j=1}^{k-1}|\mathcal{I}_j|$ be the total number of gradient samples, and $s_i^k = \sum_{s=1}^{s_k}\eta_{i,s}$ be the total number of samples from f_i at the beginning of the k-th iteration. Let $\hat{c}_{i,j}$ be the random cost of the j-th sample collected from f_i throughout the algorithm. Then we can express

$$\frac{\text{cost}_k}{\sum_{j=1}^{k-1}|\mathcal{I}_j|} = \sum_{i=1}^{n}\frac{1}{s_k}\sum_{j=1}^{s_i^k}\hat{c}_{i,j}.$$

To proceed, note that by Assumption 2.1 and the strong law of large numbers we have for each i

$$\frac{1}{s_i^k}\sum_{j=1}^{s_i^k}\hat{c}_{i,j} \to c_i, \quad \text{if } s_i^k \to \infty \text{ as } k \to \infty. \tag{C.2}$$

Now suppose we can show that

$$\frac{s_i^k}{s_k} \to p_i^*, \quad \text{a.s. for each } i = 1, \ldots, n. \tag{C.3}$$

Then if $p_i^* > 0$, we have $s_i^k \to \infty$, and hence $\frac{1}{s_k}\sum_{j=1}^{s_i^k}\hat{c}_{i,j} \to p_i^* c_i$ by (C.2). Otherwise, if $p_i^* = 0$, then no matter whether $s_i^k \to \infty$ or not, $\frac{1}{s_i^k}\sum_{j=1}^{s_i^k}\hat{c}_{i,j}$ converges to some finite number a.s., and hence $\frac{1}{s_k}\sum_{j=1}^{s_i^k}\hat{c}_{i,j} \to 0 = p_i^* c_i$. Therefore $\frac{1}{s_k}\sum_{j=1}^{s_i^k}\hat{c}_{i,j} \to p_i^* c_i$ in either case, and the desired conclusion immediately follows.

It remains to prove (C.3). We prove its stronger version

$$\frac{\sum_{s=1}^{S}\eta_{i,s}}{S}\to p_i^*,\quad \text{a.s. as } S\to\infty \text{ for each } i=1,\dots,n, \tag{C.4}$$

where S denotes the total number of sampled gradients so far. (C.3) is a subsequence of (C.4) with the index S restricted to $\sum_{j=1}^{k-1}|\mathcal{I}_j|, k\geq 1$. We slightly abuse the notation to denote by p^s the sampling weights according to which the s-th sample is drawn, and by $\mathcal{F}_s, s\geq 0$ the filtration generated by $\eta_{i,j}, i=1,\dots,n, j=1,\dots,s$, then $\mathbb{P}(\eta_{i,s}=1|\mathcal{F}_{s-1})=p_i^s$ and $p_i^s\to p_i^*$ a.s. as $s\to\infty$ for all $i=1,\dots,n$. We now show (C.4) by martingale convergence. Set $Y_S:=\big(\sum_{s=1}^{S}(\eta_{i,s}-p_i^s)\big)^2/S^2, S\geq 1$ with $Y_0=0$, then

$$\begin{aligned}\mathbb{E}[Y_S] &= \frac{1}{S^2}\sum_{s=1}^{S}\mathbb{E}[(\eta_{i,s}-p_i^s)^2] \quad \text{by martingale increment property}\\ &\leq \frac{1}{S}\to 0,\end{aligned}$$

therefore $Y_S=o_p(1)$. On the other hand, we have

$$\begin{aligned}\mathbb{E}[Y_{S+1}|\mathcal{F}_S] &= \frac{\big(\sum_{s=1}^{S}(\eta_{i,s}-p_i^s)\big)^2}{(S+1)^2}+\frac{\mathbb{E}\left[(\eta_{i,S+1}-p_i^{S+1})^2|\mathcal{F}_S\right]}{(S+1)^2}\\ &= \frac{S^2}{(S+1)^2}Y_S+\frac{p_i^{S+1}(1-p_i^{S+1})}{(S+1)^2},\end{aligned}$$

and hence

$$\sum_{s=0}^{\infty}\mathbb{E}[\mathbb{E}[Y_{s+1}-Y_s|\mathcal{F}_s]_+]\leq\sum_{s=1}^{\infty}\frac{p_i^s(1-p_i^s)}{s^2}\leq\sum_{s=1}^{\infty}\frac{1}{s^2}<\infty.$$

Therefore, by martingale convergence theorem (Blum, 1954, Corollary in Section 3), there exists some finite random variable Y_∞ such that $Y_S\to Y_\infty$ a.s.. Since $Y_S=o_p(1)$, there must exist a subsequence converging to 0 a.s., which entails that $Y_\infty=0$, i.e., $Y_S\to 0$ a.s.. Now we have $\sum_{s=1}^{S}\eta_{i,s}/S=\sqrt{Y_S}+\sum_{s=1}^{S}p_i^s/S$, and $\sum_{s=1}^{S}p_i^s/S\to p_i^*$ a.s., and hence conclude (C.4). This completes the proof. □

Proof of Proposition 4.9. When $n\geq 3$, for any $\epsilon>0$ we argue that there exist $f_i, i=1,\dots,n$ such that

$$\|\nabla f_1(x^*)\|=\|\nabla f_2(x^*)\|=1 \text{ and } \|\nabla f_i(x^*)\|=\epsilon,\quad \text{for all } i=3,\dots,n, \text{ and } \nabla f(x^*)=0.$$

Specifically, if n is even, we let $\nabla f_{2k}(x^*)=-\nabla f_{2k-1}(x^*)$ for $k=1,\dots,n/2$. If n is odd, we let $\nabla f_{2k}(x^*)=-\nabla f_{2k-1}(x^*)$ for $k=2,\dots,(n-1)/2$, and let $\nabla f_1(x^*)$ lie on the unit sphere near $-\nabla f_2(x^*)$ so that $\|\nabla f_n(x^*)\|=\|\nabla f_1(x^*)+\nabla f_2(x^*)\|=\epsilon$. Correspondingly, we consider the cost

$$c_i=\epsilon^2,\quad \text{for } i=1,\dots,n-1, \text{ and } c_n=1.$$

Then we can calculate for small ϵ and fixed n that

$$\begin{aligned}\rho(p_{Hete}^*)&=\left(\sum_{i=1}^{n}\frac{\|\nabla f_i(x^*)\|\sqrt{c_i}}{n}\right)^2=\frac{(3\epsilon+(n-3)\epsilon^2)^2}{n^2}\sim\frac{9\epsilon^2}{n^2},\\ \rho(p_{SGD}^*)&=\left(\sum_{i=1}^{n}\frac{c_i}{n}\right)\left(\sum_{i=1}^{n}\frac{\|\nabla f_i(x^*)\|^2}{n}\right)=\frac{((n-1)\epsilon^2+1)((n-2)\epsilon^2+2)}{n^2}\sim\frac{2}{n^2},\\ \rho(p_{SRG}^*)&=\left(\sum_{i=1}^{n}\frac{\|\nabla f_i(x^*)\|c_i}{n}\right)\left(\sum_{i=1}^{n}\frac{\|\nabla f_i(x^*)\|}{n}\right)=\frac{(\epsilon+2\epsilon^2+(n-3)\epsilon^3)((n-2)\epsilon+2)}{n^2}\sim\frac{2\epsilon}{n^2}.\end{aligned}$$

Therefore

$$\frac{\rho(p_{Hete}^*)}{\rho(p_{SGD}^*)}\to 0,\ \ \frac{\rho(p_{Hete}^*)}{\rho(p_{SRG}^*)}\to 0,\quad \text{as } \epsilon\to 0.$$

This completes the proof. □

D PROOF FOR THEOREM 4.12

This section presents the proof for Theorem 4.12. We first introduce a series of lemmas (Lemma D.1-D.4) that characterize the approximation errors of the quantities in our sampling efficiency metric. We then propagate these errors to control the approximation errors of the sampling weights (Proposition D.6) via a sensitivity result (Lemma D.5), and obtain the key intermediate result Theorem D.7 on the finite-time bounds of the solution error. The main proof is presented based on Theorem D.7 at the end of this section.

Lemma D.1 (An explicit version of Lemma B.5)**.** *Suppose that Assumption 4.1 holds and that $\alpha_k = \alpha_1/k^\beta$ with $\alpha_1 < \min\{1/\mu, \mu/L^2\}$ and $\beta \in (1/2, 1)$. Suppose in addition that $p_i^k \geq w_k/n$ with $\alpha_k/w_k \downarrow 0$, $\alpha_k^2/w_k \downarrow 0$, and $\sum_{k=1}^\infty \alpha_k^2/w_k < \infty$. Then $\mathbb{E}[\|x_k - x^*\|^2] \leq C_x\alpha_k/w_k$ holds for all $k \in \mathbb{N}_+$ with*

$$C_x = C_\xi \max\left\{\frac{2}{|\mathcal{I}|\mu}, \frac{k_0^\beta w_{k_0}}{C_f\alpha_1}\right\}, \tag{D.1}$$

where $C_f = \max\left\{\frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2, 2L^2\right\}$, $C_\xi = C_f\prod_{k=1}^\infty\left(1+\frac{\alpha_k^2 C_f}{|\mathcal{I}|w_k}\right)\cdot\left(1+\mathbb{E}\left[\|x_1-x^*\|^2\right]\right)$, and $k_0 = \left\lceil\left(\frac{2\beta}{\mu\alpha_1}\right)^{\frac{1}{1-\beta}}\right\rceil$. In particular, if $w_k = w_1/k^\eta$, we can let

$$C_\xi = C_f e^{\frac{C_f\alpha_1^2}{|\mathcal{I}|w_1}\left(1+\frac{1}{2\beta-\eta-1}\right)}\left(1+\mathbb{E}\left[\|x_1-x^*\|^2\right]\right).$$

Proof. According to the proof of Lemma 4.3, we have $\mathbb{E}\left[\|x_k - x^*\|^2\right] \leq C_\xi/C_f$ and $\mathbb{E}[\|\xi_k\|^2] \leq C_\xi/(|\mathcal{I}|w_k)$. For any $k \geq k_0$, it holds that

$$\alpha_k - \alpha_{k+1} = \alpha_1\frac{(k+1)^\beta - k^\beta}{k^\beta(k+1)^\beta} \leq \frac{\alpha_1\beta k^{\beta-1}}{k^{2\beta}} \leq \frac{\mu\alpha_1^2}{2k^{2\beta}} = \frac{\mu\alpha_k^2}{2}.$$

Then by setting (D.1), we could have $C_\xi/(C_x|\mathcal{I}|) \leq \mu/2$ and $\mathbb{E}\left[\|x_k - x^*\|^2\right] \leq C_x\alpha_k/w_k$, $\forall\, k \leq k_0$. Then by the proof of Lemma B.5, $\mathbb{E}[\|x_k - x^*\|^2] \leq C_x\alpha_k/w_k$ holds for all $k \in \mathbb{N}_+$. □

Lemma D.2. *Under the same assumptions as in Lemma D.1, suppose further that $|\mathcal{I}_k| = |\mathcal{I}| \leq n$ and $w_k = w_1/k^\eta$ for $\forall\, k \geq 1$. It holds for all k that*

$$\mathbb{E}\left[\sum_{i=1}^n |\tilde{g}_i^k - \|\nabla f_i(x^*)\|\|\right] \leq C_1 nL\sqrt{C_x\frac{\alpha_1}{w_1^3}}\cdot\frac{1}{k^{\frac{\beta}{2}-\frac{3\eta}{2}}} + \mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^1 - \nabla f_i(x^*)\|\right]\cdot e^{-\frac{|\mathcal{I}|w_1C_2}{n}k^{1-\eta}}, \tag{D.2}$$

where C_1, C_2 are universal constants, and C_x is the constant from Lemma D.1.

Proof. In this proof, we slightly abuse the notation $\tilde{g}_i^k$ for convenience to represent the gradient used to update the gradient norm estimate rather than the norm estimate itself. We then have

$$\begin{aligned}
\mathbb{E}[\|\tilde{g}_i^{k+1} - \nabla f_i(x^*)\||\mathcal{F}_{k-1}] &= \mathbb{P}(i\in\mathcal{I}_k)\|\nabla f_i(x_k) - \nabla f_i(x^*)\| + \mathbb{P}(i\notin\mathcal{I}_k)\|\tilde{g}_i^k - \nabla f_i(x^*)\| \\
&\leq |\mathcal{I}|p_i^k\|\nabla f_i(x_k) - \nabla f_i(x^*)\| + (1-p_i^k)^{|\mathcal{I}|}\|\tilde{g}_i^k - \nabla f_i(x^*)\| \\
&\leq |\mathcal{I}|p_i^kL\|x_k - x^*\| + (1-p_i^k)^{|\mathcal{I}|}\|\tilde{g}_i^k - \nabla f_i(x^*)\| \quad \text{by Assumption 4.1} \\
&\leq |\mathcal{I}|p_i^kL\|x_k - x^*\| + \left(1-\frac{w_k}{n}\right)^{|\mathcal{I}|}\|\tilde{g}_i^k - \nabla f_i(x^*)\| \quad \text{since } p_i^k \geq w_k/n, \\
&\leq |\mathcal{I}|p_i^kL\|x_k - x^*\| + \left(1-\frac{|\mathcal{I}|w_k}{2n}\right)\|\tilde{g}_i^k - \nabla f_i(x^*)\|,
\end{aligned}$$

where the last inequality follows from the fact that $(1-x)^\ell \leq 1 - \frac{\ell x}{2}$ holds for any $\ell \in \mathbb{N}_+$ and any $x \in [0, 1/\ell]$. Therefore

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{k+1} - \nabla f_i(x^*)\|\middle|\mathcal{F}_{k-1}\right] &\leq |\mathcal{I}|\sum_{i=1}^n p_i^kL\|x_k - x^*\| + \left(1-\frac{|\mathcal{I}|w_k}{2n}\right)\sum_{i=1}^n\|\tilde{g}_i^k - \nabla f_i(x^*)\| \\
&= |\mathcal{I}|L\|x_k - x^*\| + \left(1-\frac{|\mathcal{I}|w_k}{2n}\right)\sum_{i=1}^n\|\tilde{g}_i^k - \nabla f_i(x^*)\|.
\end{aligned}$$

Taking full expectation on both sides of the above equation gives

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{k+1} - \nabla f_i(x^*)\|\right] \\
\leq\ & \left(1 - \frac{|\mathcal{I}|w_k}{2n}\right)\mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{k} - \nabla f_i(x^*)\|\right] + |\mathcal{I}|L\sqrt{C_x\frac{\alpha_k}{w_k}} \\
\leq\ & \left(1 - \frac{|\mathcal{I}|w_k}{2n}\right)\left(1 - \frac{|\mathcal{I}|w_{k-1}}{2n}\right)\mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{k-1} - \nabla f_i(x^*)\|\right] + \left(1 - \frac{|\mathcal{I}|w_k}{2n}\right)|\mathcal{I}|L\sqrt{C_x\frac{\alpha_{k-1}}{w_{k-1}}} + |\mathcal{I}|L\sqrt{C_x\frac{\alpha_k}{w_k}} \\
&\vdots \\
\leq\ & \mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{1} - \nabla f_i(x^*)\|\right]\prod_{j=1}^k\left(1 - \frac{|\mathcal{I}|w_j}{2n}\right) + \sum_{j=1}^k |\mathcal{I}|L\sqrt{C_x\frac{\alpha_j}{w_j}}\prod_{s=j+1}^k\left(1 - \frac{|\mathcal{I}|w_s}{2n}\right) \\
=\ & \mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{1} - \nabla f_i(x^*)\|\right]\prod_{j=1}^k\left(1 - \frac{|\mathcal{I}|w_1}{2nj^\eta}\right) + \sum_{j=1}^k |\mathcal{I}|L\sqrt{C_x\frac{\alpha_1}{w_1}}\frac{1}{j^{\frac{\beta}{2}-\frac{\eta}{2}}}\prod_{s=j+1}^k\left(1 - \frac{|\mathcal{I}|w_1}{2ns^\eta}\right) \\
\leq\ & \mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{1} - \nabla f_i(x^*)\|\right]e^{-\frac{|\mathcal{I}|w_1}{2n}\sum_{j=1}^k j^{-\eta}} + |\mathcal{I}|L\sqrt{C_x\frac{\alpha_1}{w_1}}\sum_{j=1}^k\frac{1}{j^{\frac{\beta}{2}-\frac{\eta}{2}}}e^{-\frac{|\mathcal{I}|w_1}{2n}\sum_{s=j+1}^k s^{-\eta}} \\
\leq\ & \mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{1} - \nabla f_i(x^*)\|\right]e^{-\frac{|\mathcal{I}|w_1}{4n(1-\eta)}k^{1-\eta}} + |\mathcal{I}|L\sqrt{C_x\frac{\alpha_1}{w_1}}\sum_{j=1}^k\frac{1}{j^{\frac{\beta}{2}-\frac{\eta}{2}}}e^{-\frac{|\mathcal{I}|w_1}{2n}\sum_{s=j+1}^k s^{-\eta}}.
\end{aligned}
\tag{D.3}
$$

To bound the sum $\sum_{j=1}^k \frac{1}{j^{\frac{\beta}{2}-\frac{\eta}{2}}}e^{-\frac{|\mathcal{I}|w_1}{2n}\sum_{s=j+1}^k s^{-\eta}}$ we write

$$
\begin{aligned}
\sum_{j=1}^k \frac{1}{j^{\frac{\beta}{2}-\frac{\eta}{2}}}e^{-\frac{|\mathcal{I}|w_1}{2n}\sum_{s=j+1}^k s^{-\eta}} &\leq \sum_{1\leq j\leq \frac{k}{2}}\frac{1}{j^{\frac{\beta}{2}-\frac{\eta}{2}}}e^{-\frac{|\mathcal{I}|w_1}{2n}\sum_{s=j+1}^k s^{-\eta}} + \sum_{\frac{k}{2}<j\leq k}\frac{1}{j^{\frac{\beta}{2}-\frac{\eta}{2}}}e^{-\frac{|\mathcal{I}|w_1}{2n}\sum_{s=j+1}^k s^{-\eta}} \\
&\leq e^{-\frac{|\mathcal{I}|w_1}{2n}\sum_{\frac{k}{2}<s\leq k} s^{-\eta}}\sum_{1\leq j\leq \frac{k}{2}}\frac{1}{j^{\frac{\beta}{2}-\frac{\eta}{2}}} + \frac{2}{k^{\frac{\beta}{2}-\frac{\eta}{2}}}\sum_{\frac{k}{2}<j\leq k} e^{-\frac{|\mathcal{I}|w_1}{2n}k^{-\eta}(k-j)} \\
&\leq e^{-\frac{|\mathcal{I}|w_1}{4n(1-\eta)}(1-2^{\eta-1})k^{1-\eta}}\frac{1}{1-\frac{\beta}{2}+\frac{\eta}{2}}k^{1-\frac{\beta}{2}+\frac{\eta}{2}} + \frac{2}{k^{\frac{\beta}{2}-\frac{\eta}{2}}}\sum_{j=-\infty}^k e^{-\frac{|\mathcal{I}|w_1}{2n}k^{-\eta}(k-j)} \\
&\leq e^{-\frac{|\mathcal{I}|w_1 C}{n}k^{1-\eta}}\frac{1}{1-\frac{\beta}{2}+\frac{\eta}{2}}k^{1-\frac{\beta}{2}+\frac{\eta}{2}} + \frac{2}{k^{\frac{\beta}{2}-\frac{\eta}{2}}}\frac{1}{1-e^{-\frac{|\mathcal{I}|w_1}{2n}k^{-\eta}}} \\
&\leq e^{-\frac{|\mathcal{I}|w_1 C}{n}k^{1-\eta}}\frac{1}{1-\frac{\beta}{2}+\frac{\eta}{2}}k^{1-\frac{\beta}{2}+\frac{\eta}{2}} + \frac{2}{k^{\frac{\beta}{2}-\frac{\eta}{2}}}\cdot\frac{2n}{(1-e^{-1})|\mathcal{I}|w_1}k^{\eta} \\
&\qquad \text{since } 1-e^{-x}\geq(1-e^{-1})x \text{ for } x\in[0,1] \\
&= \frac{1}{1-\frac{\beta}{2}+\frac{\eta}{2}}e^{-\frac{|\mathcal{I}|w_1 C}{n}k^{1-\eta}}k^{1-\frac{\beta}{2}+\frac{\eta}{2}} + \frac{4n}{(1-e^{-1})|\mathcal{I}|w_1}\cdot\frac{1}{k^{\frac{\beta}{2}-\frac{3\eta}{2}}}
\end{aligned}
\tag{D.4}
$$

where $C := \frac{1}{4}(1-\frac{1}{\sqrt{2}})$ is a universal constant. Substituting (D.4) back into (D.3) gives

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{k} - \nabla f_i(x^*)\|\right] \leq\ & C_1 nL\sqrt{C_x\frac{\alpha_1}{w_1^3}}\cdot\frac{1}{k^{\frac{\beta}{2}-\frac{3\eta}{2}}} \\
&+ \left(|\mathcal{I}|L\sqrt{C_x\frac{\alpha_1}{w_1}}\frac{1}{1-\frac{\beta}{2}+\frac{\eta}{2}}k^{1-\frac{\beta}{2}+\frac{\eta}{2}} + \mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^{1} - \nabla f_i(x^*)\|\right]\right)e^{-\frac{|\mathcal{I}|w_1 C_2}{n}k^{1-\eta}},
\end{aligned}
$$

where C_1, C_2 are universal constants. Note that

$$
k^{1-\frac{\beta}{2}+\frac{\eta}{2}}\cdot e^{-\frac{|\mathcal{I}|w_1 C_2}{n}k^{1-\eta}} = \frac{1}{k^{\frac{\beta}{2}-\frac{3\eta}{2}}}\cdot k^{1-\eta}e^{-\frac{|\mathcal{I}|w_1 C_2}{n}k^{1-\eta}} \leq \frac{1}{k^{\frac{\beta}{2}-\frac{3\eta}{2}}}\cdot\sup_{x>0} xe^{-\frac{|\mathcal{I}|w_1 C_2}{n}x} = \frac{1}{k^{\frac{\beta}{2}-\frac{3\eta}{2}}}\cdot\frac{ne^{-1}}{|\mathcal{I}|w_1 C_2},
$$

and that $1-\frac{\beta}{2}+\frac{\eta}{2}\geq\frac{1}{2}$, hence we can further bound $\mathbb{E}\big[\sum_{i=1}^n\|\tilde{g}_i^k-\nabla f_i(x^*)\|\big]$ as

$$\mathbb{E}\left[\sum_{i=1}^n\|\tilde{g}_i^k-\nabla f_i(x^*)\|\right]\leq C_1nL\sqrt{C_x\frac{\alpha_1}{w_1^3}}\cdot\frac{1}{k^{\frac{\beta}{2}-\frac{3\eta}{2}}}+\mathbb{E}\left[\sum_{i=1}^n\|\tilde{g}_i^1-\nabla f_i(x^*)\|\right]\cdot e^{-\frac{|\mathcal{I}|w_1C_2}{n}k^{1-\eta}}$$

where C_1 is properly enlarged if needed but still universal. The desired conclusion then follows immediately from an application of the triangular inequality to $|\|\tilde{g}_i^k\|-\|\nabla f_i(x^*)\||$ for each i. □

Lemma D.3. *Under the same assumptions as in Lemma D.1, suppose further that $|\mathcal{I}_k|=|\mathcal{I}|\leq n$ and $w_k=w_1/k^\eta$ for $\forall\, k\geq 1$. It holds that*

$$\mathbb{E}[\|\tilde{G}_k\|^2]\leq C\left(\frac{C_\xi}{|\mathcal{I}|}+\frac{L^2C_x\alpha_1}{w_1}\right)\frac{1}{k^{\beta-\eta}},$$

where C is a universal constant, and $C_\xi=\prod_{k=1}^\infty\left(1+\frac{\alpha_k^2C_f}{|\mathcal{I}_k|w_k}\right)\left(\mathbb{E}\left[\|x_1-x^*\|^2\right]+1\right)$ with $C_f=\max\left\{\frac{2}{n}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2,2L^2\right\}$.

Proof. On one hand we have

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\left\|\sum_{j=1}^{k-1}\nabla f(x_j)\right\|^2\right]} &\leq \sqrt{\mathbb{E}\left[\left(\sum_{j=1}^{k-1}\|\nabla f(x_j)\|\right)^2\right]} \quad\text{by triangular inequality}\\
&\leq \sum_{j=1}^{k-1}\sqrt{\mathbb{E}[\|\nabla f(x_j)\|^2]} \quad\text{by Minkowski inequality}\\
&\leq L\sum_{j=1}^{k-1}\sqrt{\mathbb{E}[\|x_j-x^*\|^2]} \quad\text{by Assumption 4.1}\\
&\leq L\sum_{j=1}^{k-1}\sqrt{C_x\frac{\alpha_k}{w_k}} \quad\text{by Lemma D.1}\\
&\leq L\sum_{j=1}^{k-1}\sqrt{C_x\frac{\alpha_1}{w_1}\cdot\frac{1}{k^{\beta-\eta}}}\\
&\leq L\sqrt{C_x\frac{\alpha_1}{w_1}}\cdot\frac{1}{1-\beta/2+\eta/2}k^{1-\beta/2+\eta/2}.
\end{aligned}$$

On the other hand, recall that $\xi_k=g_k-\nabla f(x_k)$, and we can write

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{j=1}^{k-1}\xi_j\right\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\sum_{j=1}^{k-1}\xi_j\right\|^2\middle|\mathcal{F}_{k-2}\right]\right]\\
&= \mathbb{E}\left[\left\|\sum_{j=1}^{k-2}\xi_j\right\|^2\right]+\mathbb{E}\left[\mathbb{E}\left[\|\xi_{k-1}\|^2\middle|\mathcal{F}_{k-2}\right]\right] \quad\text{by conditional unbiasedness}\\
&\leq \mathbb{E}\left[\left\|\sum_{j=1}^{k-2}\xi_j\right\|^2\right]+\frac{C_\xi}{|\mathcal{I}|w_{k-1}} \quad\text{by the proof of Lemma 4.3}\\
&\ \ \vdots\\
&\leq \frac{C_\xi w_1}{|\mathcal{I}|}\sum_{j=1}^{k-1}j^{-\eta}\leq\frac{C_\xi w_1}{|\mathcal{I}|(1-\eta)}k^{1-\eta}.
\end{aligned}$$

From the above two bounds it follows that

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\|\tilde{G}_k\|^2\right]} &\leq \frac{1}{k-1}\sqrt{\mathbb{E}\left[\left(\left\|\sum_{j=1}^{k-1}\xi_j\right\| + \left\|\sum_{j=1}^{k-1}\nabla f(x_j)\right\|\right)^2\right]} \\
&\leq \frac{1}{k-1}\left(\sqrt{\mathbb{E}\left[\left\|\sum_{j=1}^{k-1}\xi_j\right\|^2\right]} + \sqrt{\mathbb{E}\left[\left\|\sum_{j=1}^{k-1}\nabla f(x_j)]\right\|^2\right]}\right) \\
&\leq \frac{1}{k-1}\left(\sqrt{\frac{C_\xi w_1}{|\mathcal{I}|(1-\eta)}}k^{\frac{1-\eta}{2}} + L\sqrt{C_x\frac{\alpha_1}{w_1}}\cdot\frac{1}{1-\beta/2+\eta/2}k^{1-\beta/2+\eta/2}\right) \\
&\leq \frac{1}{k-1}\left(\sqrt{\frac{C_\xi w_1}{|\mathcal{I}|(1-\eta)}} + L\sqrt{C_x\frac{\alpha_1}{w_1}}\cdot\frac{1}{1-\beta/2+\eta/2}\right)k^{1-\beta/2+\eta/2} \\
&\qquad \text{since } 1-\beta/2+\eta/2 > \frac{1-\eta}{2} > 0 \\
&= 2\left(\sqrt{\frac{C_\xi w_1}{|\mathcal{I}|(1-\eta)}} + L\sqrt{C_x\frac{\alpha_1}{w_1}}\cdot\frac{1}{1-\beta/2+\eta/2}\right)k^{-\beta/2+\eta/2}.
\end{aligned}$$

Squaring both sides of the above bound and applying Young's inequality give

$$\mathbb{E}[\|\tilde{G}_k\|^2] \leq 4\left(\frac{2C_\xi w_1}{|\mathcal{I}|(1-\eta)} + \frac{2L^2C_x\alpha_1}{w_1(1-(\beta-\eta)/2)^2}\right)k^{-\beta+\eta}.$$

Noticing that $1/(1-\eta)\leq 2$, $1/(1-\beta/2+\eta/2)^2\leq 4$, and $w_1\leq 1$ completes the proof. □

Lemma D.4. *Suppose that assumptions made in Lemma D.1 and Assumption 2.1 and 4.11 hold and suppose further that $|\mathcal{I}_k| = |\mathcal{I}| \leq n$ and $w_k = w_1/k^\eta$ for $\forall\, k\geq 1$. It holds that*

$$\mathbb{E}\left[\sum_{i=1}^n |\tilde{c}_i^k - c_i|\right] \leq C_1\left(\frac{n^{3/2}\sqrt{\max_i \mathrm{Var}(\hat{c}_i)}}{\sqrt{|\mathcal{I}|}w_1}\cdot\frac{1}{k^{\frac{1}{2}-\eta}} + \sqrt{n\sum_{i=1}^n(\mathbb{E}[|\tilde{c}_i^1 - c_i|])^2 e^{-\frac{C_2|\mathcal{I}|w_1}{n}k^{1-\eta}}}\right),$$

where C_1, C_2 are universal constants.

Proof. We define $I_{i,j}=1$ if ∇f_i is the j-th sampled gradient and otherwise $I_{i,j}=0$. For each i, let $\hat{c}_{i,j}, j\geq 1$ be a sequence of i.i.d. random cost for the i-th function. Denote by

$$Y_k := \sum_{j=1}^{|\mathcal{I}|(k-1)} I_{i,j} \tag{D.5}$$

the cumulative number of samples from the i-th function at the beginning of the k-th iteration, and

$$X_k := \sum_{j=1}^{|\mathcal{I}|(k-1)} (\hat{c}_{i,j} - c_i)\mathbb{I}(I_{i,j}=1). \tag{D.6}$$

Then we can represent

$$\tilde{c}_i^k - c_i = \frac{X_k}{Y_k}\cdot\mathbb{I}(Y_k>0) + (\tilde{c}_i^1 - c_i)\cdot\mathbb{I}(Y_k=0),$$

hence by triangular inequality and the independence between $\tilde{c}_i^1$ and Y_k we have

$$\mathbb{E}\big[|\tilde{c}_i^k - c_i|\big] \leq \mathbb{E}\left[\left|\frac{X_k}{Y_k}\cdot\mathbb{I}(Y_k>0)\right|\right] + \mathbb{E}[|\tilde{c}_i^1 - c_i|]\cdot\mathbb{P}(Y_k=0). \tag{D.7}$$

The second term above can be handled by

$$\begin{aligned}
\mathbb{P}(Y_k = 0) &= \mathbb{E}\left[\prod_{j=1}^{|\mathcal{I}|(k-1)} \mathbb{I}(I_{i,j} = 0)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^{|\mathcal{I}|(k-1)} \mathbb{I}(I_{i,j} = 0)\middle| \mathcal{F}_{k-2}\right]\right] \\
&= \mathbb{E}\left[\left(1 - p_i^{k-1}\right)^{|\mathcal{I}|} \prod_{j=1}^{|\mathcal{I}|(k-2)} \mathbb{I}(I_{i,j} = 0)\right] \\
&\leq \left(1 - \frac{w_{k-1}}{n}\right)^{|\mathcal{I}|} \mathbb{E}\left[\prod_{j=1}^{|\mathcal{I}|(k-2)} \mathbb{I}(I_{i,j} = 0)\right] \quad \text{since } p_i^{k-1} \geq \frac{w_{k-1}}{n} \\
&\vdots \\
&\leq \prod_{j=1}^{k-1} \left(1 - \frac{w_j}{n}\right)^{|\mathcal{I}|} \leq e^{-|\mathcal{I}|\sum_{j=1}^{k-1} \frac{w_j}{n}} \leq e^{-\frac{C|\mathcal{I}|w_1}{n} k^{1-\eta}}
\end{aligned} \tag{D.8}$$

where C is a universal constant. To handle the first term in (D.7), we use Cauchy-Schwarz inequality to write

$$\mathbb{E}\left[\left|\frac{X_k}{Y_k} \cdot \mathbb{I}(Y_k > 0)\right|\right] \leq \sqrt{\mathbb{E}[X_k^2]\mathbb{E}\left[\frac{1}{Y_k^2} \cdot \mathbb{I}(Y_k > 0)\right]},$$

and analyze the two expectations on the right-hand side in the next two paragraphs.

Bound for $\mathbb{E}[X_k^2]$: By the independence between $\hat{c}_{i,j}$ and $I_{i,j}$ we have that

$$\begin{aligned}
\mathbb{E}[X_k^2] &= \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{j=1}^{|\mathcal{I}|(k-1)} (\hat{c}_{i,j} - c_i)\mathbb{I}(I_{i,j} = 1)\right)^2 \middle| \mathcal{F}_{k-2}\right]\right] \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{|\mathcal{I}|(k-2)} (\hat{c}_{i,j} - c_i)\mathbb{I}(I_{i,j} = 1)\right)^2\right] + \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{j=|\mathcal{I}|(k-2)+1}^{|\mathcal{I}|(k-1)} (\hat{c}_{i,j} - c_i)\mathbb{I}(I_{i,j} = 1)\right)^2 \middle| \mathcal{F}_{k-2}\right]\right] \\
&\quad \text{by martingale property} \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{|\mathcal{I}|(k-2)} (\hat{c}_{i,j} - c_i)\mathbb{I}(I_{i,j} = 1)\right)^2\right] + \mathrm{Var}(\hat{c}_i)|\mathcal{I}|\mathbb{E}[p_i^{k-1}] \\
&\vdots \\
&\leq \mathrm{Var}(\hat{c}_i)|\mathcal{I}|\sum_{j=1}^{k-1} \mathbb{E}[p_i^j].
\end{aligned} \tag{D.9}$$

Bound for $\mathbb{E}[\frac{1}{Y_k^2}\mathbb{I}(Y_k \neq 0)]$: Let $I'_{i,j}$, $1 \leq i \leq k-1, 1 \leq j \leq |\mathcal{I}|$ be independent random variables with distributions $\mathbb{P}(I'_{i,j} = 1) = w_i/n$ and $\mathbb{P}(I'_{i,j} = 0) = 1 - w_i/n$. Define $Z_k = \sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|} I'_{i,j}$. It follows from $p_i^k \geq w_k/n$ that $\mathbb{P}(Y_k \geq y) \geq \mathbb{P}(Z_k \geq y)$, $\forall\, y \in \mathbb{N}$. It holds that

$$\mathbb{E}[Y_k] \geq \mathbb{E}[Z_k] = \sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|} \mathbb{E}[I'_{i,j}] = \frac{|\mathcal{I}|}{n}\sum_{i=1}^{k-1} w_i,$$

and that

$$\begin{aligned}
\mathbb{E}\left[|Z_k - \mathbb{E}[Z_k]|^4\right] &= \mathbb{E}\left[\left|\sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|}(I'_{i,j} - \mathbb{E}[I'_{i,j}])\right|^4\right] \\
&= \sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|}\mathbb{E}\left[|I'_{i,j} - \mathbb{E}[I'_{i,j}]|^4\right] + 3\sum_{(i,j)\neq(i',j')}\mathbb{E}\left[|I'_{i,j} - \mathbb{E}[I'_{i,j}]|^2\right]\cdot\mathbb{E}\left[|I'_{i',j'} - \mathbb{E}[I'_{i',j'}]|^2\right] \\
&\leq \sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|}\mathbb{E}\left[|I'_{i,j} - \mathbb{E}[I'_{i,j}]|^4\right] + 3\left(\sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|}\mathbb{E}\left[|I'_{i,j} - \mathbb{E}[I'_{i,j}]|^2\right]\right)^2 \\
&= \sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|}\left(\frac{w_i}{n}\cdot\left(1-\frac{w_i}{n}\right)^4 + \left(1-\frac{w_i}{n}\right)\cdot\left(\frac{w_i}{n}\right)^4\right) \\
&\qquad + 3\left(\sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|}\left(\frac{w_i}{n}\cdot\left(1-\frac{w_i}{n}\right)^2 + \left(1-\frac{w_i}{n}\right)\cdot\left(\frac{w_i}{n}\right)^2\right)\right)^2 \\
&\leq \sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|}\frac{w_i}{n}\cdot\left(1-\frac{w_i}{n}\right) + 3\left(\sum_{i=1}^{k-1}\sum_{j=1}^{|\mathcal{I}|}\frac{w_i}{n}\cdot\left(1-\frac{w_i}{n}\right)\right)^2 \\
&\leq \frac{|\mathcal{I}|}{n}\sum_{i=1}^{k-1}w_i + 3\left(\frac{|\mathcal{I}|}{n}\sum_{i=1}^{k-1}w_i\right)^2.
\end{aligned}$$

Thus, using Markov's inequality we can estimate that

$$\begin{aligned}
\mathbb{P}\left(Y_k \leq \frac{|\mathcal{I}|}{2n}\sum_{i=1}^{k-1}w_i\right) &\leq \mathbb{P}\left(Z_k \leq \frac{|\mathcal{I}|}{2n}\sum_{i=1}^{k-1}w_i\right) \leq \mathbb{P}\left(|Z_k - \mathbb{E}[Z_k]| \geq \frac{|\mathcal{I}|}{2n}\sum_{i=1}^{k-1}w_i\right) \\
&\leq \frac{\mathbb{E}\left[|Z_k - \mathbb{E}[Z_k]|^4\right]}{\left(\frac{|\mathcal{I}|}{2n}\sum_{i=1}^{k-1}w_i\right)^4} \leq \frac{16}{\left(\frac{|\mathcal{I}|}{n}\sum_{i=1}^{k-1}w_i\right)^3} + \frac{48}{\left(\frac{|\mathcal{I}|}{n}\sum_{i=1}^{k-1}w_i\right)^2},
\end{aligned}$$

and that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{Y_k^2}\mathbb{I}(Y_k \neq 0)\right] &\leq 1\cdot\mathbb{P}\left(Y_k \leq \frac{|\mathcal{I}|}{2n}\sum_{i=1}^{k-1}w_i\right) + \frac{1}{\left(\frac{|\mathcal{I}|}{2n}\sum_{i=1}^{k-1}w_i\right)^2}\cdot\mathbb{P}\left(Y_k > \frac{|\mathcal{I}|}{2n}\sum_{i=1}^{k-1}w_i\right) \\
&\leq \frac{16}{\left(\frac{|\mathcal{I}|}{n}\sum_{i=1}^{k-1}w_i\right)^3} + \frac{52}{\left(\frac{|\mathcal{I}|}{n}\sum_{i=1}^{k-1}w_i\right)^2}.
\end{aligned}$$

Since it must hold that $\mathbb{E}\left[\frac{1}{Y_k^2}\mathbb{I}(Y_k \neq 0)\right] \leq 1$, we can assume $\frac{|\mathcal{I}|}{n}\sum_{i=1}^{k-1}w_i \geq 1$ in the above upper bound without loss of generality, giving rise to

$$\mathbb{E}\left[\frac{1}{Y_k^2}\mathbb{I}(Y_k \neq 0)\right] \leq \frac{68}{\left(\frac{|\mathcal{I}|}{n}\sum_{i=1}^{k-1}w_i\right)^2} \leq \frac{Cn^2}{|\mathcal{I}|^2 w_1^2}\cdot\frac{1}{k^{2-2\eta}} \tag{D.10}$$

where C is a universal constant. Combining (D.9) and (D.10) gives

$$\mathbb{E}\left[\left|\frac{X_k}{Y_k}\cdot\mathbb{I}(Y_k > 0)\right|\right] \leq \sqrt{\mathrm{Var}(\hat{c}_i)|\mathcal{I}|\sum_{j=1}^{k-1}\mathbb{E}[p_i^j]\cdot\frac{Cn^2}{|\mathcal{I}|^2 w_1^2}\cdot\frac{1}{k^{2-2\eta}}}. \tag{D.11}$$

We now derive bounds for the aggregated error $\mathbb{E}[\sum_{i=1}^n |\tilde{c}_i^k - c_i|]$. We first note that by Jensen's inequality

$$\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[|\tilde{c}_i^k - c_i|\right]\right)^2 \leq \frac{1}{n}\sum_{i=1}^n\left(\mathbb{E}\left[|\tilde{c}_i^k - c_i|\right]\right)^2,$$

therefore we can write

$$\begin{aligned}\left(\sum_{i=1}^n \mathbb{E}\left[|\tilde{c}_i^k - c_i|\right]\right)^2 &\leq n\sum_{i=1}^n \left(\mathbb{E}\left[|\tilde{c}_i^k - c_i|\right]\right)^2 \\ &\leq 2n\sum_{i=1}^n \left(\mathrm{Var}(\hat{c}_i)|\mathcal{I}|\sum_{j=1}^{k-1}\mathbb{E}[p_i^j]\cdot\frac{C_1 n^2}{|\mathcal{I}|^2 w_1^2}\cdot\frac{1}{k^{2-2\eta}} + (\mathbb{E}[|\tilde{c}_i^1 - c_i|])^2 e^{-\frac{2C_2|\mathcal{I}|w_1}{n}k^{1-\eta}}\right) \\ &\qquad \text{by (D.7) and the bounds (D.8) and (D.11), where } C_1, C_2 \text{ are universal constants} \\ &\leq 2n\left(\max_i \mathrm{Var}(\hat{c}_i)k\cdot\frac{C_1 n^2}{|\mathcal{I}|w_1^2}\cdot\frac{1}{k^{2-2\eta}} + \sum_{i=1}^n (\mathbb{E}[|\tilde{c}_i^1 - c_i|])^2 e^{-\frac{2C_2|\mathcal{I}|w_1}{n}k^{1-\eta}}\right) \\ &\leq 2n\left(\frac{C_1 n^2 \max_i \mathrm{Var}(\hat{c}_i)}{|\mathcal{I}|w_1^2}\cdot\frac{1}{k^{1-2\eta}} + \sum_{i=1}^n (\mathbb{E}[|\tilde{c}_i^1 - c_i|])^2 e^{-\frac{2C_2|\mathcal{I}|w_1}{n}k^{1-\eta}}\right).\end{aligned}$$

Taking the square root of the above bound then gives

$$\sum_{i=1}^n \mathbb{E}\left[|\tilde{c}_i^k - c_i|\right] \leq C_1\left(\frac{n^{3/2}\sqrt{\max_i \mathrm{Var}(\hat{c}_i)}}{\sqrt{|\mathcal{I}|}w_1}\cdot\frac{1}{k^{\frac{1}{2}-\eta}} + \sqrt{n\sum_{i=1}^n (\mathbb{E}[|\tilde{c}_i^1 - c_i|])^2 e^{-\frac{C_2|\mathcal{I}|w_1}{n}k^{1-\eta}}}\right).$$

This completes the proof. □

In the next lemma we consider the error bound of p_i^* in terms of the errors in c_i, b_i, and b_0, i.e., perturbation analysis of Proposition 2.2.

Lemma D.5. *Let p_i^* and κ^* be defined in Proposition 2.2 with c_i, b_i, and b_0. Let $p_i^* + \Delta p_i^*$ be the new probability weights when c_i, b_i, and b_0 are perturbed to $c_i + \Delta c_i > 0$, $b_i + \Delta b_i \geq 0$, and $b_0 + \Delta b_0 \geq 0$. Suppose that $\min(c_i + \Delta c_i, c_i) \geq \underline{c}$ for each i and some constant $\underline{c} > 0$, and $b_0 + \Delta b_0 \geq \frac{b_0}{2}$, then*

$$\sum_{i=1}^n |\Delta p_i^*| \leq C\max\left(\sqrt{c_{\max}}, 1\right)\cdot\frac{c_{\max}}{\sqrt{\underline{c}}c_{\min}\sum_{i=1}^n\sqrt{b_i}}\cdot\left(\sum_{i=1}^n\sqrt{|\Delta b_i|} + \sum_{i=1}^n\sqrt{b_i|\Delta c_i|} + n\sqrt{|\Delta b_0|}\right),$$

where $c_{\max} = \max_{1\leq j\leq n} c_j$, $c_{\min} = \min_{1\leq j\leq n} c_j$, and C is a universal constant.

Proof. Let us denote $q_i = \sqrt{\frac{(b_i+\Delta b_i)/n^2}{\kappa^*(c_i+\Delta c_i)+(b_0+\Delta b_0)}}$. It can be computed that

$$\begin{aligned}|q_i - p_i^*| &= \left|\sqrt{\frac{(b_i+\Delta b_i)/n^2}{\kappa^*(c_i+\Delta c_i)+(b_0+\Delta b_0)}} - \sqrt{\frac{b_i/n^2}{\kappa^* c_i + b_0}}\right| \\ &\leq \sqrt{\left|\frac{(b_i+\Delta b_i)/n^2}{\kappa^*(c_i+\Delta c_i)+(b_0+\Delta b_0)} - \frac{b_i/n^2}{\kappa^* c_i + b_0}\right|} \\ &\leq \sqrt{\frac{\frac{|\Delta b_i|}{n^2}(\kappa^* c_i + b_0) + \frac{b_i}{n^2}(\kappa^*|\Delta c_i| + |\Delta b_0|)}{(\kappa^*(c_i+\Delta c_i)+(b_0+\Delta b_0))(\kappa^* c_i + b_0)}} \\ &\leq \sqrt{\frac{\frac{|\Delta b_i|}{n^2}(\kappa^* c_i + b_0) + \frac{b_i}{n^2}(\kappa^*|\Delta c_i| + |\Delta b_0|)}{\frac{\underline{c}}{2c_i}(\kappa^* c_i + b_0)^2}} \\ &\leq \frac{\sqrt{2c_{\max}/\underline{c}}}{n}\sqrt{\frac{|\Delta b_i|}{\kappa^* c_i + b_0} + \frac{b_i\kappa^*|\Delta c_i|}{(\kappa^* c_i + b_0)^2} + \frac{b_i|\Delta b_0|}{(\kappa^* c_i + b_0)^2}} \\ &\leq \frac{\sqrt{2c_{\max}/\underline{c}}}{n}\left(\sqrt{\frac{|\Delta b_i|}{\kappa^* c_i + b_0}} + \sqrt{\frac{b_i\kappa^*|\Delta c_i|}{(\kappa^* c_i + b_0)^2}} + \sqrt{\frac{b_i|\Delta b_0|}{(\kappa^* c_i + b_0)^2}}\right)\end{aligned}$$

$$\leq \frac{\sqrt{2c_{\max}/\underline{c}}}{n}\left(\sqrt{\frac{|\Delta b_i|}{\kappa^* c_{\min}+b_0}}+\sqrt{\frac{b_i|\Delta c_i|}{c_{\min}(\kappa^* c_{\min}+b_0)}}+\sqrt{\frac{b_i|\Delta b_0|}{(\kappa^* c_{\min}+b_0)^2}}\right).$$

It follows from

$$1=\sum_{i=1}^n\sqrt{\frac{b_i/n^2}{\kappa^* c_i+b_0}}\geq\sum_{i=1}^n\sqrt{\frac{b_i/n^2}{\kappa^*\|c\|_\infty+b_0}}\geq\frac{1}{\sqrt{\kappa^* c_{\min}+b_0}}\cdot\frac{\sqrt{c_{\min}}\sum_{i=1}^n\sqrt{b_i}}{n\sqrt{\|c\|_\infty}},$$

that

$$\frac{1}{\sqrt{\kappa^* c_{\min}+b_0}}\leq\frac{n\sqrt{\|c\|_\infty}}{\sqrt{c_{\min}}\sum_{i=1}^n\sqrt{b_i}}.$$

Therefore, it holds that

$$\begin{aligned}|q_i-p_i^*|\leq&\frac{\sqrt{2c_{\max}/\underline{c}}}{n}\left(\sqrt{\frac{|\Delta b_i|}{\kappa^* c_{\min}+b_0}}+\sqrt{\frac{b_i|\Delta c_i|}{c_{\min}(\kappa^* c_{\min}+b_0)}}+\sqrt{\frac{b_i|\Delta b_0|}{(\kappa^* c_{\min}+b_0)^2}}\right)\\ \leq&\sqrt{\frac{2c_{\max}}{\underline{c}}}\left(\frac{\sqrt{\|c\|_\infty}}{\sqrt{c_{\min}}\sum_{i=1}^n\sqrt{b_i}}\sqrt{|\Delta b_i|}+\frac{\sqrt{b_i\|c\|_\infty}}{c_{\min}\sum_{i=1}^n\sqrt{b_i}}\sqrt{|\Delta c_i|}+\frac{n\|c\|_\infty\sqrt{b_i}}{c_{\min}(\sum_{i=1}^n\sqrt{b_i})^2}\sqrt{|\Delta b_0|}\right),\end{aligned}$$

which implies that

$$\begin{aligned}\left|1-\sum_{i=1}^n q_i\right|\leq&\sum_{i=1}^n|q_i-p_i^*|\\ \leq&\sqrt{\frac{2c_{\max}}{\underline{c}}}\left(\frac{\sqrt{\|c\|_\infty}}{\sqrt{c_{\min}}\sum_{i=1}^n\sqrt{b_i}}\sum_{i=1}^n\sqrt{|\Delta b_i|}+\frac{\sqrt{\|c\|_\infty}}{c_{\min}\sum_{i=1}^n\sqrt{b_i}}\sum_{i=1}^n\sqrt{b_i|\Delta c_i|}+\frac{n\|c\|_\infty}{c_{\min}\sum_{i=1}^n\sqrt{b_i}}\sqrt{|\Delta b_0|}\right).\end{aligned}$$

Note that $p_i^*+\Delta p_i^*=\sqrt{\frac{(b_i+\Delta b_i)/n^2}{(\kappa^*+\Delta\kappa^*)(c_i+\Delta c_i)+(b_0+\Delta b_0)}}$, where $\kappa=\kappa^*+\Delta\kappa^*$ is the unique solution to the equation

$$\sum_{i=1}^n\sqrt{\frac{(b_i+\Delta b_i)/n^2}{\kappa(c_i+\Delta c_i)+(b_0+\Delta b_0)}}=1.$$

We can thus know that $p_i^*+\Delta p_i^*\geq q_i$ for all $i\in\{1,2,\ldots,n\}$ if $\Delta\kappa^*\leq 0$ and that $p_i^*+\Delta p_i^*\leq q_i$ for all $i\in\{1,2,\ldots,n\}$ if $\Delta\kappa^*\geq 0$. In both cases, we have that

$$\left|1-\sum_{i=1}^n q_i\right|=\left|\sum_{i=1}^n(p_i^*+\Delta p_i^*-q_i)\right|=\sum_{i=1}^n|p_i^*+\Delta p_i^*-q_i|.$$

Combining all estimations above, we can conclude that

$$\begin{aligned}\sum_{i=1}^n|\Delta p_i^*|\leq&\sum_{i=1}^n(|p_i^*+\Delta p_i^*-q_i|+|q_i-p_i^*|)\leq\left|1-\sum_{i=1}^n q_i\right|+\sum_{i=1}^n|q_i-p_i^*|\\ \leq&2\sqrt{\frac{c_{\max}}{\underline{c}}}\left(\frac{\sqrt{\|c\|_\infty}}{\sqrt{c_{\min}}\sum_{i=1}^n\sqrt{b_i}}\sum_{i=1}^n\sqrt{|\Delta b_i|}+\frac{\sqrt{\|c\|_\infty}}{c_{\min}\sum_{i=1}^n\sqrt{b_i}}\sum_{i=1}^n\sqrt{b_i|\Delta c_i|}+\frac{n\|c\|_\infty}{c_{\min}\sum_{i=1}^n\sqrt{b_i}}\sqrt{|\Delta b_0|}\right)\\ \leq&C\max\left(\sqrt{2c_{\max}},1\right)\cdot\frac{c_{\max}}{\sqrt{\underline{c}}c_{\min}\sum_{i=1}^n\sqrt{b_i}}\cdot\left(\sum_{i=1}^n\sqrt{|\Delta b_i|}+\sum_{i=1}^n\sqrt{b_i|\Delta c_i|}+n\sqrt{|\Delta b_0|}\right),\end{aligned}$$

which is the desired result. □

Proposition D.6. *Suppose that the assumptions made in Lemma D.1 and Assumption 2.1 and 4.11 hold and suppose further that* $|\mathcal{I}_k|=|\mathcal{I}|\leq n$ *and* $w_k=w_1/k^\eta$ *with* $w_1\leq 1$ *for* $\forall\ k\geq 1$. *Denote by* $c_{\max}=\max_i c_i$, $c_{\min}=\min_i c_i$

respectively the maximum and minimum sampling costs per gradient evaluation, and by $\overline{G} = (1/n) \cdot \sum_{i=1}^{n} \|\nabla f_i(x^*)\|$ *the averaged gradient norm at the optimum. Then we have the following error bound for the estimated sampling weights*

$$\mathbb{E}[\|\tilde{p}^k - p^*_{Hete}\|_1] \le C_1 \max\left(\sqrt{c_{\max}}, 1\right) \cdot \frac{c_{\max}}{\sqrt{\underline{c}} c_{\min}} \Big[\Big(\frac{1}{\overline{G}} \sqrt{C_x \frac{\mu}{w_1^3}} + 1 \Big) \cdot \frac{1}{k^{\frac{\beta}{4} - \frac{3\eta}{4}}} + \frac{1}{\overline{G}} \Big(\frac{C_f^2 \max_i \mathrm{Var}(\hat{c}_i)}{w_1^2} \Big)^{\frac{1}{4}} \cdot \frac{(n/|\mathcal{I}|)^{\frac{1}{4}}}{k^{\frac{1}{4} - \frac{\eta}{2}}} \Big] + C_3(f_i, \tilde{g}_i^1, \tilde{c}_i^1, c_i, i = 1, \ldots, n) \cdot e^{-\frac{C_2 |\mathcal{I}| w_1}{n} k^{1-\eta}},$$

where C_1, C_2 *are universal constants, and* C_3 *depends on* $f_i, \tilde{g}_i^1, \tilde{c}_i^1, c_i, i = 1, \ldots, n$ *only but does not explicitly depend on* n*.*

Proof. Substituting the true values $b_i = \|\nabla f_i(x^*)\|^2$, $b_0 = 0$, and the estimates $b_i + \Delta b_i = (\tilde{g}_i^k)^2$, $b_0 + \Delta b_0 = \min\left(\|\tilde{G}_k\|, \sum_{i=1}^n \frac{\tilde{g}_i^k}{n}\right)^2$, $c_i + \Delta c_i = \tilde{c}_i^k$ into Lemma D.5 gives

$$\begin{aligned} \|\tilde{p}^k - p^*_{Hete}\|_1 \le & C \max\left(\sqrt{c_{\max}}, 1\right) \cdot \frac{c_{\max}}{\sqrt{\underline{c}} c_{\min} \sum_{i=1}^n \|\nabla f_i(x^*)\|} \\ & \cdot \left(\sum_{i=1}^n \sqrt{|(\tilde{g}_i^k)^2 - \|\nabla f_i(x^*)\|^2|} + \sum_{i=1}^n \|\nabla f_i(x^*)\| \sqrt{|\tilde{c}_i^k - c_i|} + n \|\tilde{G}_k\| \right). \end{aligned} \tag{D.12}$$

We bound the expectation of each term on the right-hand side.

Bound for $\mathbb{E}\left[\sum_{i=1}^n \sqrt{|(\tilde{g}_i^k)^2 - \|\nabla f_i(x^*)\|^2|}\right]$**:** Since $|x^2 - y^2|^{1/2} \le ((x-y)^2 + 2|x-y||y|)^{1/2} \le |x-y| + \sqrt{2|x-y||y|}$, we can write

$$\begin{aligned} \sum_{i=1}^n \sqrt{|(\tilde{g}_i^k)^2 - \|\nabla f_i(x^*)\|^2|} &\le \sum_{i=1}^n |\tilde{g}_i^k - \|\nabla f_i(x^*)\|| + \sqrt{2} \sum_{i=1}^n \sqrt{|\tilde{g}_i^k - \|\nabla f_i(x^*)\|| \cdot \|\nabla f_i(x^*)\|} \\ &\le \sum_{i=1}^n |\tilde{g}_i^k - \|\nabla f_i(x^*)\|| + \sqrt{2} \sqrt{\sum_{i=1}^n |\tilde{g}_i^k - \|\nabla f_i(x^*)\|| \cdot \sum_{i=1}^n \|\nabla f_i(x^*)\|}, \end{aligned}$$

where in the second inequality we use Cauchy-Schwarz inequality. Using Jensen's inequality to swap the expectation and square root operations we have

$$\begin{aligned} & \mathbb{E}\left[\sum_{i=1}^n \sqrt{|(\tilde{g}_i^k)^2 - \|\nabla f_i(x^*)\|^2|}\right] \\ \le\ & \mathbb{E}\left[\sum_{i=1}^n |\tilde{g}_i^k - \|\nabla f_i(x^*)\||\right] + \sqrt{2 \sum_{i=1}^n \|\nabla f_i(x^*)\|} \cdot \sqrt{\mathbb{E}\left[\sum_{i=1}^n |\tilde{g}_i^k - \|\nabla f_i(x^*)\||\right]} \\ \le\ & C_1 \left(nL \sqrt{C_x \frac{\alpha_1}{w_1^3}} + \sqrt{nL \sum_{i=1}^n \|\nabla f_i(x^*)\|} \cdot \left(C_x \frac{\alpha_1}{w_1^3} \right)^{\frac{1}{4}} \right) \cdot \frac{1}{k^{\frac{\beta}{4} - \frac{3\eta}{4}}} \\ & + \left(\mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^1 - \nabla f_i(x^*)\|\right] + \sqrt{2 \sum_{i=1}^n \|\nabla f_i(x^*)\| \cdot \mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^1 - \nabla f_i(x^*)\|\right]} \right) \cdot e^{-\frac{|\mathcal{I}| w_1 C_2}{n} k^{1-\eta}} \\ \le\ & C_1 \left(nL \sqrt{C_x \frac{\alpha_1}{w_1^3}} + \sum_{i=1}^n \|\nabla f_i(x^*)\| \right) \cdot \frac{1}{k^{\frac{\beta}{4} - \frac{3\eta}{4}}} \\ & + 2 \left(\mathbb{E}\left[\sum_{i=1}^n \|\tilde{g}_i^1 - \nabla f_i(x^*)\|\right] + \sum_{i=1}^n \|\nabla f_i(x^*)\| \right) \cdot e^{-\frac{C_2 |\mathcal{I}| w_1}{n} k^{1-\eta}}, \end{aligned}$$

where the second inequality follows from Lemma D.2 and the universal constants C_1, C_2 are properly adjusted if needed.

Bound for $\mathbb{E}\left[\sum_{i=1}^n \|\nabla f_i(x^*)\|\sqrt{|\tilde{c}_i^k - c_i|}\right]$: Similar to the above bound, we use Cauchy-Schwarz inequality and Jensen's inequality to write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^n \|\nabla f_i(x^*)\|\sqrt{|\tilde{c}_i^k - c_i|}\right] \\
\leq\ &\mathbb{E}\left[\sqrt{\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \sum_{i=1}^n |\tilde{c}_i^k - c_i|}\right] \leq \sqrt{\sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \mathbb{E}\left[\sum_{i=1}^n |\tilde{c}_i^k - c_i|\right]} \\
\leq\ &C_3\sqrt{\sum_{i=1}^n \|\nabla f_i(x^*)\|^2}\left(\frac{n^{3/4}\max_i(\mathrm{Var}(\hat{c}_i))^{1/4}}{|\mathcal{I}|^{1/4}\sqrt{w_1}}\cdot\frac{1}{k^{\frac{1}{4}-\frac{\eta}{2}}} + \left(n\sum_{i=1}^n(\mathbb{E}[|\tilde{c}_i^1 - c_i|])^2\right)^{\frac{1}{4}} e^{-\frac{C_4|\mathcal{I}|w_1}{n}k^{1-\eta}}\right),
\end{aligned}$$

where the last inequality follows from Lemma D.4 and C_3, C_4 are universal constants.

Bound for $\mathbb{E}\left[\|\tilde{G}_k\|\right]$: It follows directly from Lemma D.3 and Jensen's inequality that

$$\mathbb{E}\left[\|\tilde{G}_k\|\right] \leq C_5\left(\sqrt{\frac{C_\xi}{|\mathcal{I}|}} + L\sqrt{\frac{C_x\alpha_1}{w_1}}\right)\frac{1}{k^{\frac{\beta}{2}-\frac{\eta}{2}}}$$

where C_5 is a universal constant.

Substituting all these bounds into (D.12) gives

$$\begin{aligned}
&\frac{1}{\sum_{i=1}^n\|\nabla f_i(x^*)\|}\cdot\mathbb{E}\left[\sum_{i=1}^n\sqrt{|(\tilde{g}_i^k)^2 - \|\nabla f_i(x^*)\|^2|} + \sum_{i=1}^n\|\nabla f_i(x^*)\|\sqrt{|\tilde{c}_i^k - c_i|} + n\|\tilde{G}_k\|\right] \\
\leq\ &\frac{1}{\sum_{i=1}^n\|\nabla f_i(x^*)\|}\cdot\left[C_1\left(nL\sqrt{C_x\frac{\alpha_1}{w_1^3}} + \sum_{i=1}^n\|\nabla f_i(x^*)\|\right)\cdot\frac{1}{k^{\frac{\beta}{4}-\frac{3\eta}{4}}}\right. \\
&+C_2\left(\sqrt{\sum_{i=1}^n\|\nabla f_i(x^*)\|^2}\frac{n^{3/4}\max_i(\mathrm{Var}(\hat{c}_i))^{1/4}}{|\mathcal{I}|^{1/4}\sqrt{w_1}} + n\sqrt{\frac{C_\xi}{|\mathcal{I}|}} + nL\sqrt{\frac{C_x\alpha_1}{w_1}}\right)\cdot\frac{1}{k^{\frac{1}{4}-\frac{\eta}{2}}} \\
&\left.+nC_4(f_i,\tilde{g}_i^1,\tilde{c}_i^1,c_i,i=1,\dots,n)\cdot e^{-\frac{C_3|\mathcal{I}|w_1}{n}k^{1-\eta}}\right] \\
&\quad\text{where the } 1/k^{\beta/2-\eta/2} \text{ term is absorbed since } \frac{\beta}{2}-\frac{\eta}{2} > \frac{1}{4}-\frac{\eta}{2} \\
\leq\ &C_1\left(\frac{L}{\overline{G}}\sqrt{C_x\frac{\alpha_1}{w_1^3}}+1\right)\cdot\frac{1}{k^{\frac{\beta}{4}-\frac{3\eta}{4}}} + C_2\left(\frac{1}{\overline{G}}\left(\frac{C_f^2 n\max_i\mathrm{Var}(\hat{c}_i)}{|\mathcal{I}|w_1^2}\right)^{\frac{1}{4}} + \frac{1}{\overline{G}}\sqrt{\frac{C_\xi}{|\mathcal{I}|}}\right)\cdot\frac{1}{k^{\frac{1}{4}-\frac{\eta}{2}}} \\
&+\frac{1}{\overline{G}}C_4(f_i,\tilde{g}_i^1,\tilde{c}_i^1,c_i,i=1,\dots,n)\cdot e^{-\frac{C_3|\mathcal{I}|w_1}{n}k^{1-\eta}},
\end{aligned}$$

where C_1, C_2, C_3 are universal constants, and C_4 depends $f_i, \tilde{g}_i^1, \tilde{c}_i^1, c_i, i = 1, \dots, n$ only but does not explicitly depend on n. Since $C_\xi \leq C_x|\mathcal{I}|\mu/2$, $\alpha_1 \leq \mu/L^2$ and $w_1 \leq 1$, the coefficient $\frac{1}{\overline{G}}\sqrt{\frac{C_\xi}{|\mathcal{I}|}}$ is no larger than $\frac{L}{\overline{G}}\sqrt{C_x\frac{\alpha_1}{w_1^3}}$ hence can be absorbed into the first term. $\frac{1}{\overline{G}}$ can be absorbed into C_4. This completes the proof. □

Theorem D.7 (Finite-time bounds for Polyak-Ruppert and α-suffix averaging)**.** *Suppose Assumptions 2.1, 4.1, 4.4, 4.10 and 4.11 hold. Suppose that in Algorithm 2 $\alpha_k = \alpha_1/k^\beta$, where $\beta \in (1/2, 1)$ and $0 < \alpha_1 < \min\left(1/\mu, \mu/L^2\right)$, $w_k = w_1/k^\eta$ with $w_1 \in (0, 1]$ and $0 < \eta < \min\left(\beta/7, 1-\beta, \beta-1/2\right)$, and that $|\mathcal{I}_k| = |\mathcal{I}|$ is fixed for all k. Then for every $\gamma \in [0, 1)$ and every non-singular matrix $A \in \mathbb{R}^{d\times d}$, we have the decomposition*

$$A(\bar{x}_{k,\gamma} - x^*) = \mathcal{L}_{k,\gamma} + \mathcal{E}_{k,\gamma},$$

where $\mathcal{L}_k$ is the leading error term that satisfies

$$\mathbb{E}\left[\mathcal{L}_{k,\gamma}\right] = 0,$$

$$\mathbb{E}[\|\mathcal{L}_{k,\gamma}\|^2] = \frac{1}{(1-\gamma)k|\mathcal{I}|}\mathrm{Tr}(AH^{-1}G(p^*_{Hete})H^{-1}A^T),$$

and the high-order error $\mathcal{E}_{k,\gamma}$ satisfies

$$\begin{aligned}\mathbb{E}\left[\|\mathcal{E}_{k,\gamma}\|\right] \quad &\leq \quad C_1(A,\gamma,\alpha_1,w_1,\beta,\eta,x_1,f_i,\hat{c}_i,\tilde{g}_i^1,\tilde{c}_i^1,i=1,\ldots,n)\cdot\frac{1}{\sqrt{k}}\\ &\quad\cdot\left(\frac{1}{k^{\min(\frac{\beta}{4}-\frac{7\eta}{4},\eta,\frac{1}{2}-\frac{\beta}{2}-\frac{\eta}{2},\beta-\frac{1}{2}-\eta)}}+\frac{1}{\sqrt{|\mathcal{I}|}}\cdot\left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}}\frac{1}{k^{\frac{1}{4}-\frac{3\eta}{2}}}+\frac{1}{\sqrt{|\mathcal{I}|}}\cdot\left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1+\eta}{1-\eta}}\frac{1}{k}\right).\end{aligned}$$

The cumulative sampling cost satisfies

$$\begin{aligned}|\mathbb{E}\left[\mathrm{cost}_k\right]-c(p^*_{Hete})|\mathcal{I}|(k-1)| \quad &\leq \quad C_2(\alpha_1,w_1,\beta,\eta,x_1,f_i,\hat{c}_i,\tilde{g}_i^1,\tilde{c}_i^1,i=1,\ldots,n)\cdot|\mathcal{I}|k\\ &\quad\cdot\left(\frac{1}{k^{\min(\frac{\beta}{4}-\frac{3\eta}{4},\eta)}}+\left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}}\frac{1}{k^{\frac{1}{4}-\frac{\eta}{2}}}+\left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{1-\eta}}\frac{1}{k}\right).\end{aligned}$$

Here the constants C_1, C_2 depend on the quantities specified respectively and do not explicitly depend on n.

Proof. We have $g_k = \frac{1}{\alpha_k}(x_k - x_{k+1})$ and

$$\begin{aligned}g_k \quad &= \quad \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_k}\frac{1}{np_i^k}\nabla f_i(x_k) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_k}\frac{1}{np_i^k}\nabla f_i(x^*)+\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_k}\frac{1}{np_i^k}\left(\int_0^1\nabla^2 f_i((1-\theta)x^*+\theta x_k)d\theta\right)(x_k-x^*)\\ &= \quad \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_k}\frac{1}{np_i^k}\nabla f_i(x^*)+\nabla^2 f(x^*)(x_k-x^*)+\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_k}\frac{1}{np_i^k}\nabla^2 f_i(x^*)-\nabla^2 f(x^*)\right)(x_k-x^*)\\ &\quad +\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_k}\frac{1}{np_i^k}\left(\int_0^1\left(\nabla^2 f_i((1-\theta)x^*+\theta x_k)-\nabla^2 f_i(x^*)\right)d\theta\right)(x_k-x^*).\end{aligned}$$

Therefore summing up the above equality over k and rearranging terms give

$$\begin{aligned}&(1-\gamma)k\nabla^2 f(x^*)(\bar{x}_{k,\gamma}-x^*) \qquad\qquad (D.13)\\ = \quad &-\sum_{j=[\gamma k]+1}^{k}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}\nabla f_i(x^*)\\ &+\sum_{j=[\gamma k]+1}^{k}\frac{1}{\alpha_j}(x_j-x_{j+1})-\sum_{j=[\gamma k]+1}^{k}\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}\nabla^2 f_i(x^*)-\nabla^2 f(x^*)\right)(x_j-x^*)\\ &-\sum_{j=[\gamma k]+1}^{k}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}\left(\int_0^1\left(\nabla^2 f_i((1-\theta)x^*+\theta x_j)-\nabla^2 f_i(x^*)\right)d\theta\right)(x_j-x^*).\end{aligned}$$

Next we bound each term on the right-hand side of the above representation.

Bound for $\sum_{j=[\gamma k]+1}^{k}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}\nabla f_i(x^*)$: For an arbitrary given $d\times d$ matrix M, we can use the martingale property to write

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{j=[\gamma k]+1}^{k}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}M\nabla f_i(x^*)\right\|^2\right] \quad &= \quad \sum_{j=[\gamma k]+1}^{k}\mathbb{E}\left[\left\|\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}M\nabla f_i(x^*)\right\|^2\right]\\ &= \quad \frac{1}{|\mathcal{I}|}\sum_{j=[\gamma k]+1}^{k}\mathbb{E}\left[\sum_{i=1}^{n}\frac{1}{n^2p_i^j}\|M\nabla f_i(x^*)\|^2\right]. \qquad (D.14)\end{aligned}$$

To further bound (D.14), we consider the perturbed optimal sampling weights $p^{*,k} := (1-w_k)p^*_{Hete} + w_k(1/n,\ldots,1/n)$ and note that $p^k = (1-w_k)\tilde{p}^k + w_k\cdot(1/n,\ldots,1/n)$. Denote by $\mathcal{I}^*_+ = \{i : p^*_{Hete,i} > 0\}$ and $\delta^* = \min\{p^*_{Hete,i}/2 : i \in$

$\mathcal{I}_+^*\}$. We then have

$$\begin{aligned}
&\left|\mathbb{E}\left[\sum_{i=1}^n \frac{1}{n^2 p_i^k}\|M\nabla f_i(x^*)\|^2\right] - \sum_{i=1}^n \frac{1}{n^2 p_i^{*,k}}\|M\nabla f_i(x^*)\|^2\right| \\
= &\left|\mathbb{E}\left[\sum_{i\in\mathcal{I}_+^*} \frac{1}{n^2 p_i^k}\|M\nabla f_i(x^*)\|^2\right] - \sum_{i\in\mathcal{I}_+^*} \frac{1}{n^2 p_i^{*,k}}\|M\nabla f_i(x^*)\|^2\right| \\
\le &\frac{\max_i \|M\nabla f_i(x^*)\|^2}{n^2}\mathbb{E}\left[\sum_{i\in\mathcal{I}_+^*} \frac{|p_i^k - p_i^{*,k}|}{p_i^k p_i^{*,k}}\right] \\
\le &\frac{\max_i \|M\nabla f_i(x^*)\|^2}{n\delta^* w_k}\mathbb{E}\left[\sum_{i\in\mathcal{I}_+^*} |p_i^k - p_i^{*,k}|\right] \quad \text{since } p_i^k \ge \frac{w_k}{n} \text{ and } p_i^{*,k} \ge \delta^* \\
\le &\frac{\max_i \|M\nabla f_i(x^*)\|^2}{n\delta^* w_k}\cdot(1-w_k)\mathbb{E}\left[\sum_{i\in\mathcal{I}_+^*} |\tilde{p}_i^k - p_{Hete,i}^*|\right] \\
\le &\frac{\max_i \|M\nabla f_i(x^*)\|^2}{n\delta^* w_k}\mathbb{E}\left[\|\tilde{p}^k - p_{Hete}^*\|_1\right].
\end{aligned}$$

On the other hand, since both $p_i^{*,k} \ge \delta^*$ and $p_{Hete,i}^* \ge \delta^*$ for all $i \in \mathcal{I}_+^*$, by a similar argument as above we have

$$\begin{aligned}
\left|\sum_{i=1}^n \frac{1}{n^2 p_i^{*,k}}\|M\nabla f_i(x^*)\|^2 - \sum_{i=1}^n \frac{1}{n^2 p_{Hete,i}^*}\|M\nabla f_i(x^*)\|^2\right| &\le \frac{\max_i \|M\nabla f_i(x^*)\|^2}{(n\delta^*)^2}\|p^{*,k} - p_{Hete}^*\|_1 \\
&\le \frac{\max_i \|M\nabla f_i(x^*)\|^2}{(n\delta^*)^2}\cdot 2w_k.
\end{aligned}$$

Combining the two error bounds we obtain the overall error

$$\begin{aligned}
&\left|\mathbb{E}\left[\sum_{i=1}^n \frac{1}{n^2 p_i^k}\|M\nabla f_i(x^*)\|^2\right] - \sum_{i=1}^n \frac{1}{n^2 p_{Hete,i}^*}\|M\nabla f_i(x^*)\|^2\right| \\
&\qquad\le \frac{\max_i \|M\nabla f_i(x^*)\|^2}{n\delta^*}\left(\frac{1}{w_k}\mathbb{E}\left[\|\tilde{p}^k - p^*\|_1\right] + \frac{2w_k}{n\delta^*}\right).
\end{aligned}$$

Using this bound we can bound (D.14) as

$$\begin{aligned}
&\left|\mathcal{|I|}\cdot\mathbb{E}\left[\left\|\sum_{j=[\gamma k]+1}^k \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}M\nabla f_i(x^*)\right\|^2\right] - (1-\gamma)k\sum_{i=1}^n \frac{1}{n^2 p_{Hete,i}^*}\|M\nabla f_i(x^*)\|^2\right| \\
\le &\frac{\max_i \|M\nabla f_i(x^*)\|^2}{n\delta^*}\left(\sum_{j=[\gamma k]+1}^k \frac{1}{w_j}\mathbb{E}\left[\|\tilde{p}^j - p_{Hete}^*\|_1\right] + \frac{2\sum_{j=[\gamma k]+1}^k w_j}{n\delta^*}\right). \qquad \text{(D.15)}
\end{aligned}$$

Note that (D.14) also have the lower bound

$$\sum_{i=1}^n \frac{1}{n^2 p_i^k}\|M\nabla f_i(x^*)\|^2 \ge \min_{p\in\Delta}\sum_{i=1}^n \frac{1}{n^2 p_i}\|M\nabla f_i(x^*)\|^2 \ge \left(\frac{1}{n}\sum_{i=1}^n \|M\nabla f_i(x^*)\|\right)^2.$$

Therefore we also have the following lower bound

$$\mathbb{E}\left[\left\|\sum_{j=[\gamma k]+1}^k \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}M\nabla f_i(x^*)\right\|^2\right] \ge \frac{(1-\gamma)k}{|\mathcal{I}|}\left(\frac{1}{n}\sum_{i=1}^n \|M\nabla f_i(x^*)\|\right)^2. \qquad \text{(D.16)}$$

Bound for $\sum_{j=[\gamma k]+1}^{k} \frac{1}{\alpha_j}(x_j - x_{j+1})$: We can rearrange terms to write

$$\begin{aligned}\sum_{j=[\gamma k]+1}^{k} \frac{1}{\alpha_j}(x_j - x_{j+1}) &= \sum_{j=[\gamma k]+1}^{k} \frac{1}{\alpha_j}(x_j - x^*) - \sum_{j=[\gamma k]+2}^{k+1} \frac{1}{\alpha_{j-1}}(x_j - x^*) \\ &= \sum_{j=[\gamma k]+2}^{k} (\frac{1}{\alpha_j} - \frac{1}{\alpha_{j-1}})(x_j - x^*) + \frac{1}{\alpha_{[\gamma k]+1}}(x_{[\gamma k]+1} - x^*) - \frac{1}{\alpha_k}(x_{k+1} - x^*),\end{aligned}$$

therefore we can use Lemma D.1 to bound the first moment as

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{j=[\gamma k]+1}^{k} \frac{1}{\alpha_j}(x_j - x_{j+1})\right\|\right] \\ \leq\ & \sum_{j=[\gamma k]+2}^{k} (\frac{1}{\alpha_j} - \frac{1}{\alpha_{j-1}})\mathbb{E}[\|x_j - x^*\|] + \frac{1}{\alpha_{[\gamma k]+1}}\mathbb{E}[\|x_{[\gamma k]+1} - x^*\|] + \frac{1}{\alpha_k}\mathbb{E}[\|x_{k+1} - x^*\|] \\ \leq\ & \frac{1}{\alpha_1}\left(\sum_{j=[\gamma k]+2}^{k} \beta(j-1)^{\beta-1}\sqrt{C_x\frac{\alpha_j}{w_j}} + ([\gamma k]+1)^{\beta}\sqrt{C_x\frac{\alpha_{[\gamma k]+1}}{w_{[\gamma k]+1}}} + k^{\beta}\sqrt{C_x\frac{\alpha_{k+1}}{w_{k+1}}}\right) \\ \leq\ & \sqrt{\frac{CC_x}{\alpha_1 w_1}}k^{\frac{\beta}{2}+\frac{\eta}{2}},\end{aligned}$$

where C is a universal constant.

Bound for $\sum_{j=[\gamma k]+1}^{k}\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}\nabla^2 f_i(x^*) - \nabla^2 f(x^*)\right)(x_j - x^*)$: Similar to the analysis of $\mathbb{E}[\|\xi_j\|^2]$, we can write

$$\begin{aligned}&\mathbb{E}\left[\left\|\sum_{j=[\gamma k]+1}^{k}\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}\nabla^2 f_i(x^*) - \nabla^2 f(x^*)\right)(x_j - x^*)\right\|^2\right] \\ =\ & \sum_{j=[\gamma k]+1}^{k}\mathbb{E}\left[\left\|\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}\nabla^2 f_i(x^*) - \nabla^2 f(x^*)\right)(x_j - x^*)\right\|^2\right] \\ \leq\ & \frac{1}{|\mathcal{I}|}\sum_{j=[\gamma k]+1}^{k}\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{p_i^j}\|\nabla^2 f_i(x^*)(x_j - x^*)\|^2 - \|\nabla^2 f(x^*)(x_j - x^*)\|^2\right] \\ \leq\ & \frac{1}{|\mathcal{I}|}\sum_{j=[\gamma k]+1}^{k}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_j}L^2\|(x_j - x^*)\|^2\right] \quad \text{since } \|\nabla^2 f_i(x^*)\| \leq L \text{ and } p_i^j \geq \frac{w_j}{n} \\ =\ & \frac{L^2}{|\mathcal{I}|}\sum_{j=[\gamma k]+1}^{k}\frac{1}{w_j}\mathbb{E}\left[\|(x_j - x^*)\|^2\right] \\ \leq\ & \frac{L^2 C_x\alpha_1}{|\mathcal{I}|w_1^2}\sum_{j=[\gamma k]+1}^{k} j^{-\beta+2\eta} \quad \text{by Lemma D.1} \\ \leq\ & \frac{L^2 C_x\alpha_1}{|\mathcal{I}|w_1^2}\cdot\frac{1}{1-\beta+2\eta}k^{1-\beta+2\eta}.\end{aligned}$$

Bound for $\sum_{j=[\gamma k]+1}^{k}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}\left(\int_0^1\left(\nabla^2 f_i((1-\theta)x^* + \theta x_j) - \nabla^2 f_i(x^*)\right)d\theta\right)(x_j - x^*)$: By smoothness of the second-order derivatives we have $\|\nabla^2 f_i((1-\theta)x^* + \theta x_j) - \nabla^2 f_i(x^*)\| \leq L_2\theta\|x_j - x^*\|$, hence we can write

$$\mathbb{E}\left[\left\|\sum_{j=[\gamma k]+1}^{k}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}\left(\int_0^1\left(\nabla^2 f_i((1-\theta)x^* + \theta x_j) - \nabla^2 f_i(x^*)\right)d\theta\right)(x_j - x^*)\right\|\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sum_{j=[\gamma k]+1}^{k}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}L_2\|x_j-x^*\|^2\right]\\
&= L_2\mathbb{E}\left[\sum_{j=[\gamma k]+1}^{k}\|x_j-x^*\|^2\right]\\
&\leq \frac{L_2C_x\alpha_1}{w_1}\sum_{j=[\gamma k]+1}^{k}j^{-\beta+\eta}\quad\text{by Lemma D.1}\\
&\leq \frac{L_2C_x\alpha_1}{w_1}\cdot\frac{1}{1-\beta+\eta}k^{1-\beta+\eta}.
\end{aligned}
$$

Then we combine all the above bounds to characterize (D.13). We consider multiplying both sides of (D.13) by AH^{-1} (recall that $H=\nabla^2 f(x^*)$) for some matrix $A\in\mathbb{R}^{d\times d}$, and get

$$
A(\bar{x}_{k,\gamma}-x^*)=-\frac{1}{(1-\gamma)k}\sum_{j=[\gamma k]+1}^{k}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}AH^{-1}\nabla f_i(x^*)+\frac{1}{(1-\gamma)k}AH^{-1}\cdot\text{Remainder},\tag{D.17}
$$

where the remainder aggregates the remaining terms from (D.13), and has the following first order moment by aggregating the bounds derived above

$$
\begin{aligned}
\mathbb{E}[\|\text{Remainder}\|] &\leq C\sqrt{\frac{C_x}{\alpha_1 w_1}}k^{\frac{\beta}{2}+\frac{\eta}{2}}+L\sqrt{\frac{C_x\alpha_1}{|\mathcal{I}|w_1^2(1-\beta+2\eta)}}k^{\frac{1}{2}-\frac{\beta}{2}+\eta}+\frac{L_2C_x\alpha_1}{w_1}\cdot\frac{1}{1-\beta+\eta}k^{1-\beta+\eta}\\
&\leq C\sqrt{\frac{C_x}{\alpha_1 w_1}}k^{\frac{\beta}{2}+\frac{\eta}{2}}+\frac{(L+L_2)(C_x\alpha_1+\sqrt{C_x\alpha_1})}{w_1(1-\beta+\eta)}\cdot k^{1-\beta+\eta}.
\end{aligned}\tag{D.18}
$$

We now deal with the leading term in (D.17) which we denote by

$$
\widetilde{\mathcal{L}}_{k,\gamma}:=-\frac{1}{(1-\gamma)k}\sum_{j=[\gamma k]+1}^{k}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}_j}\frac{1}{np_i^j}AH^{-1}\nabla f_i(x^*)
$$

for convenience. We also denote by

$$
M_{k,\gamma}^{Hete}:=\frac{1}{(1-\gamma)|\mathcal{I}|k}\sum_{i=1}^{n}\frac{1}{n^2p_{Hete,i}^*}\|AH^{-1}\nabla f_i(x^*)\|^2
$$

the target variance of the leading error. Note that $M_{k,\gamma}^{Hete}=\frac{1}{(1-\gamma)k|\mathcal{I}|}\text{Tr}(AH^{-1}G(p_{Hete}^*)H^{-1}A^T)$. Then by (D.15) $\widetilde{\mathcal{L}}_{k,\gamma}$ satisfies

$$
\begin{aligned}
&\left|\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right]-M_{k,\gamma}^{Hete}\right| \qquad\qquad (D.19)\\
&\leq \frac{\max_i\|AH^{-1}\nabla f_i(x^*)\|^2}{(1-\gamma)^2k^2|\mathcal{I}|n\delta^*}\left(\sum_{j=[\gamma k]+1}^{k}\frac{1}{w_j}\mathbb{E}\left[\|\tilde{p}^j-p^*\|_1\right]+\frac{2\sum_{j=[\gamma k]+1}^{k}w_j}{n\delta^*}\right)\\
&\leq C_1\max\left(\sqrt{c_{\max}},1\right)\cdot\frac{c_{\max}}{\sqrt{\underline{c}}c_{\min}}\cdot\frac{\max_i\|AH^{-1}\nabla f_i(x^*)\|^2}{(1-\gamma)^2|\mathcal{I}|n\delta^*w_1}\Big[\Big(\frac{1}{\overline{G}}\sqrt{C_x\frac{\mu}{w_1^3}}+1\Big)\cdot\frac{1}{k^{1+\frac{\beta}{4}-\frac{7\eta}{4}}}\\
&\quad+\frac{1}{\overline{G}}\Big(\frac{C_f^2\max_i\text{Var}(\hat{c}_i)}{w_1^2}\Big)^{\frac{1}{4}}\cdot\frac{(n/|\mathcal{I}|)^{\frac{1}{4}}}{k^{\frac{5}{4}-\frac{3\eta}{2}}}\Big]\\
&\quad+C_3(f_i,\tilde{g}_i^1,\tilde{c}_i^1,c_i,i=1,\ldots,n)\cdot\frac{\max_i\|AH^{-1}\nabla f_i(x^*)\|^2}{(1-\gamma)^2k^2|\mathcal{I}|n\delta^*w_1}\sum_{j=[\gamma k]+1}^{k}j^{\eta}e^{-\frac{C_2|\mathcal{I}|w_1}{n}j^{1-\eta}}\\
&\quad+\frac{C_4w_1}{n\delta^*}\cdot\frac{\max_i\|AH^{-1}\nabla f_i(x^*)\|^2}{(1-\gamma)^2|\mathcal{I}|n\delta^*}\frac{1}{k^{1+\eta}},
\end{aligned}
$$

where C_4 is a universal constant and we used results from Proposition D.6. To handle the summation of the exponential term, we can calculate that

$$\begin{aligned}\sum_{j=[\gamma k]+1}^{k} j^{\eta} e^{-\frac{C_2|\mathcal{I}|w_1}{n} j^{1-\eta}} &\leq C_5 \int_0^{\infty} x^{\eta} e^{-\frac{C_2|\mathcal{I}|w_1}{n} x^{1-\eta}} dx \quad \text{where } C_5 \text{ is a universal constant} \\ &= C_5 \cdot \left(\frac{n}{C_2|\mathcal{I}|w_1}\right)^{\frac{1+\eta}{1-\eta}} \int_0^{\infty} y^{\eta} e^{-y^{1-\eta}} dy \quad \text{with } x = \left(\frac{n}{C_2|\mathcal{I}|w_1}\right)^{\frac{1}{1-\eta}} y \\ &\leq C_5 \cdot \left(\frac{n}{C_2|\mathcal{I}|w_1}\right)^{\frac{1+\eta}{1-\eta}} \leq C_5 \cdot \left(\frac{n}{|\mathcal{I}|w_1}\right)^{\frac{1+\eta}{1-\eta}}, \end{aligned} \tag{D.20}$$

where in the last inequality the integral is finite and continuous in η and hence is uniformly bounded for $\eta \in [0, 1/2]$, therefore the integral can be absorbed into the universal constant C_5. The universal constant C_2 is also absorbed into C_5. Substituting this bound back into (D.19) and rearranging the terms finally give

$$\begin{aligned}&\left|\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right] - M_{k,\gamma}^{Hete}\right| \\ \leq\ & C_1 \frac{\max_i \|AH^{-1}\nabla f_i(x^*)\|^2}{(1-\gamma)^2|\mathcal{I}|} \\ &\cdot \left[\frac{\max\left(\sqrt{c_{\max}}, 1\right) c_{\max}}{\sqrt{\underline{c}} c_{\min} n\delta^* w_1} \left(\left(\frac{1}{\overline{G}}\sqrt{C_x \frac{\mu}{w_1^3}} + 1\right) \cdot \frac{1}{k^{1+\frac{\beta}{4}-\frac{7\eta}{4}}} + \frac{1}{\overline{G}}\left(\frac{C_f^2 \max_i \mathrm{Var}(\hat{c}_i)}{w_1^2}\right)^{\frac{1}{4}} \cdot \frac{(n/|\mathcal{I}|)^{\frac{1}{4}}}{k^{\frac{5}{4}-\frac{3\eta}{2}}}\right)\right. \\ &\left. + \frac{w_1}{(n\delta^*)^2} \cdot \frac{1}{k^{1+\eta}} + C_3(f_i, \tilde{g}_i^1, \tilde{c}_i^1, c_i, i = 1, \ldots, n) \cdot \frac{1}{n\delta^* w_1^{2/(1-\eta)}} \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1+\eta}{1-\eta}} \cdot \frac{1}{k^2}\right]. \end{aligned} \tag{D.21}$$

On the other hand, from (D.16) we have the lower bound

$$\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right] \geq M_{k,\gamma}^{*} := \frac{1}{(1-\gamma)|\mathcal{I}|k}\left(\frac{1}{n}\sum_{i=1}^{n}\|AH^{-1}\nabla f_i(x^*)\|\right)^2 > 0, \tag{D.22}$$

where the positiveness is due to Assumption 4.4. Therefore we can define

$$\mathcal{L}_{k,\gamma} := \sqrt{\frac{M_{k,\gamma}^{Hete}}{\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right]}} \cdot \widetilde{\mathcal{L}}_{k,\gamma},$$

then it is clear that $\mathbb{E}\left[\left\|\mathcal{L}_{k,\gamma}\right\|^2\right] = M_{k,\gamma}^{Hete}$ and $\mathbb{E}\left[\mathcal{L}_{k,\gamma}\right] = \sqrt{M_{k,\gamma}^{Hete}/\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right]}\,\mathbb{E}\left[\widetilde{\mathcal{L}}_{k,\gamma}\right] = 0$. We also need to control the difference

$$\begin{aligned}\mathbb{E}\left[\left\|\mathcal{L}_{k,\gamma} - \widetilde{\mathcal{L}}_{k,\gamma}\right\|\right] &= \left|\frac{\sqrt{M_{k,\gamma}^{Hete}} - \sqrt{\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right]}}{\sqrt{\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right]}}\right| \mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|\right] \\ &\leq \left|\sqrt{M_{k,\gamma}^{Hete}} - \sqrt{\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right]}\right| \quad \text{by Jensen's inequality} \\ &= \frac{\left|\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right] - M_{k,\gamma}^{Hete}\right|}{\sqrt{\mathbb{E}\left[\left\|\widetilde{\mathcal{L}}_{k,\gamma}\right\|^2\right]} + \sqrt{M_{k,\gamma}^{Hete}}} \end{aligned}$$

$$\leq \quad \frac{1}{2\sqrt{M^*_{k,\gamma}}} \cdot \left| \mathbb{E}\left[\left\| \widetilde{\mathcal{L}}_{k,\gamma} \right\|^2 \right] - M^{Hete}_{k,\gamma} \right| \quad \text{by (D.22)} \tag{D.23}$$

Finally we combine the error bounds in (D.18), (D.21), and (D.23) to conclude the moment bound for the high-order error. Specifically, applying these error bounds and rearranging terms we obtain

$$\begin{aligned}
& \mathbb{E}\left[\|A(\bar{x}_{k,\gamma} - x^*) - \mathcal{L}_{k,\gamma}\|\right] \\
\leq\ & \mathbb{E}\left[\left\| \mathcal{L}_{k,\gamma} - \widetilde{\mathcal{L}}_{k,\gamma} \right\| \right] + \frac{1}{(1-\gamma)k} \|AH^{-1}\| \cdot \mathbb{E}\left[\|\text{Remainder}\|\right] \\
\leq\ & \frac{1}{2\sqrt{M^*_{k,\gamma}}} \cdot \left| \mathbb{E}\left[\left\| \widetilde{\mathcal{L}}_{k,\gamma} \right\|^2 \right] - M^{Hete}_{k,\gamma} \right| + \frac{1}{(1-\gamma)k} \|AH^{-1}\| \cdot \mathbb{E}\left[\|\text{Remainder}\|\right] \\
\leq\ & C_1 \frac{\max_i \|AH^{-1}\nabla f_i(x^*)\|^2}{(1-\gamma)^{\frac{3}{2}} \sqrt{|\mathcal{I}|} \frac{1}{n} \sum_{i=1}^n \|AH^{-1}\nabla f_i(x^*)\|} \\
& \quad \cdot \left[\frac{\max\left(\sqrt{c_{\max}}, 1\right) c_{\max}}{\sqrt{\underline{c}} c_{\min} n \delta^* w_1} \left(\left(\frac{1}{\overline{G}} \sqrt{C_x \frac{\mu}{w_1^3}} + 1 \right) \cdot \frac{1}{k^{\frac{1}{2}+\frac{\beta}{4}-\frac{7\eta}{4}}} + \frac{1}{\overline{G}} \left(\frac{C_f^2 \max_i \mathrm{Var}(\hat{c}_i)}{w_1^2} \right)^{\frac{1}{4}} \cdot \frac{(n/|\mathcal{I}|)^{\frac{1}{4}}}{k^{\frac{3}{4}-\frac{3\eta}{2}}} \right) \right. \\
& \quad \left. + \frac{w_1}{(n\delta^*)^2} \cdot \frac{1}{k^{\frac{1}{2}+\eta}} + C_3(f_i, \tilde{g}_i^1, \tilde{c}_i^1, c_i, i = 1, \ldots, n) \cdot \frac{1}{n\delta^* w_1^{2/(1-\eta)}} \left(\frac{n}{|\mathcal{I}|} \right)^{\frac{1+\eta}{1-\eta}} \cdot \frac{1}{k^{\frac{3}{2}}} \right] \\
& \quad + \frac{\|AH^{-1}\|}{(1-\gamma)} \left(C \sqrt{\frac{C_x}{\alpha_1 w_1}} \cdot \frac{1}{k^{1-\frac{\beta}{2}-\frac{\eta}{2}}} + \frac{(L+L_2)(C_x\alpha_1 + \sqrt{C_x \alpha_1})}{w_1(1-\beta+\eta)} \cdot \frac{1}{k^{\beta-\eta}} \right) \\
\leq\ & C_1(A, \gamma, \delta^*, \underline{c}, C_x, \alpha_1, w_1, \beta, \eta, f_i, c_i, i = 1, \ldots, n) \cdot k^{-\min(\frac{1}{2}+\frac{\beta}{4}-\frac{7\eta}{4}, \frac{1}{2}+\eta, 1-\frac{\beta}{2}-\frac{\eta}{2}, \beta-\eta)} \\
& \quad + C_2(A, \gamma, \delta^*, \underline{c}, w_1, f_i, c_i, \mathrm{Var}(\hat{c}_i), i = 1, \ldots, n) \cdot \frac{1}{\sqrt{|\mathcal{I}|}} \cdot \left(\frac{n}{|\mathcal{I}|} \right)^{\frac{1}{4}} k^{-\frac{3}{4}+\frac{3\eta}{2}} \\
& \quad + C_3(A, \gamma, \delta^*, w_1, f_i, c_i, \tilde{g}_i^1, \tilde{c}_i^1, i = 1, \ldots, n) \cdot \frac{1}{\sqrt{|\mathcal{I}|}} \cdot \left(\frac{n}{|\mathcal{I}|} \right)^{\frac{1+\eta}{1-\eta}} k^{-\frac{3}{2}}.
\end{aligned}$$

It only remains to study the cumulative cost. By the conditional independence of the sampling cost $\hat{c}_{i,j}$ given the sampling weights, we have

$$\mathbb{E}\left[\text{cost}_k\right] = \mathbb{E}\left[\sum_{j=1}^{k-1} |\mathcal{I}| \sum_{i=1}^n c_i p_i^j \right],$$

therefore we can write

$$\begin{aligned}
& |\mathbb{E}\left[\text{cost}_k\right] - c(p^*_{Hete})|\mathcal{I}|(k-1)| \\
\leq\ & |\mathcal{I}| \sum_{j=1}^{k-1} \mathbb{E}\left[\left| \sum_{i=1}^n c_i p_i^j - \sum_{i=1}^n c_i p^*_{Hete,i} \right| \right] \\
\leq\ & |\mathcal{I}| \sum_{j=1}^{k-1} \mathbb{E}\left[\max_i c_i \sum_{i=1}^n |p_i^j - p^*_{Hete,i}| \right] \\
=\ & |\mathcal{I}| c_{\max} \sum_{j=1}^{k-1} \mathbb{E}\left[\|p^j - p^*_{Hete}\|_1 \right] \\
\leq\ & |\mathcal{I}| c_{\max} \left(\sum_{j=1}^{k-1} \mathbb{E}\left[\|\tilde{p}^j - p^*_{Hete}\|_1 \right] + \sum_{j=1}^{k-1} 2w_j \right) \quad \text{using } p^j = (1-w_j)\tilde{p}^j + w_j \left(\frac{1}{n}, \ldots, \frac{1}{n} \right) \\
\leq\ & |\mathcal{I}| c_{\max} \cdot C_1 \max\left(\sqrt{c_{\max}}, 1\right) \cdot \frac{c_{\max}}{\sqrt{\underline{c}} c_{\min}} \cdot \left[\left(\frac{1}{\overline{G}} \sqrt{C_x \frac{\mu}{w_1^3}} + 1 \right) \cdot k^{1-\frac{\beta}{4}+\frac{3\eta}{4}} \right.
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{G}\Big(\frac{C_f^2 \max_i \mathrm{Var}(\hat{c}_i)}{w_1^2}\Big)^{\frac{1}{4}} \cdot \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}} k^{\frac{3}{4}+\frac{\eta}{2}}\Big] \quad \text{using results from Proposition D.6} \\
&+C_2(f_i, \tilde{g}_i^1, \tilde{c}_i^1, c_i, i=1,\dots,n)\cdot |\mathcal{I}| c_{\max} \cdot \left(\frac{n}{|\mathcal{I}| w_1}\right)^{\frac{1}{1-\eta}} \quad \text{using a similar calculation as (D.20)} \\
&+C_3 |\mathcal{I}| c_{\max} \cdot w_1 k^{1-\eta} \\
\leq\ & C_4(\underline{c}, C_x, w_1, f_i, c_i, i=1,\dots,n)\cdot |\mathcal{I}| k^{\max(1-\frac{\beta}{4}+\frac{3\eta}{4}, 1-\eta)} \\
&+C_5(\underline{c}, w_1, f_i, c_i, \mathrm{Var}(\hat{c}_i), i=1,\dots,n)\cdot |\mathcal{I}| \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}} k^{\frac{3}{4}+\frac{\eta}{2}} \\
&+C_6(w_1, f_i, \tilde{g}_i^1, \tilde{c}_i^1, c_i, i=1,\dots,n)\cdot |\mathcal{I}| \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{1-\eta}}.
\end{aligned}
$$

This completes the proof. □

Now we can present the proof of Theorem 4.12.

Proof of Theorem 4.12. We first prove the bound for HeteRSGD. Theorem D.7 entails the following bounds for the solution error and cumulative sampling cost

$$
\begin{aligned}
\mathbb{E}[\|H(\bar{x}_{k,\gamma} - x^*)\|] &\leq \sqrt{\mathbb{E}[\|\mathcal{L}_{k,\gamma}\|^2]} + \mathbb{E}[\|\mathcal{E}_{k,\gamma}\|] \qquad \text{(D.24)} \\
&\leq \sqrt{\frac{1}{(1-\gamma)k|\mathcal{I}|}\mathrm{Tr}(G(p^*_{Hete}))} \\
&\quad + \frac{C_1}{\sqrt{k}} \cdot \left(\frac{1}{k^{\hat{c}_{\beta,\eta}}} + \frac{1}{\sqrt{|\mathcal{I}|}} \cdot \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}} \frac{1}{k^{\frac{1}{4}-\frac{3\eta}{2}}} + \frac{1}{\sqrt{|\mathcal{I}|}} \cdot \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1+\eta}{1-\eta}} \frac{1}{k} \right), \qquad \text{(D.25)}
\end{aligned}
$$

where $\hat{c}_{\beta,\eta} := \min(\frac{\beta}{4} - \frac{7\eta}{4}, \eta, \frac{1}{2} - \frac{\beta}{2} - \frac{\eta}{2}, \beta - \frac{1}{2} - \eta)$, C_1 depends on $\gamma, \alpha_1, w_1, \beta, \eta, x_1$ and $f_i, \hat{c}_i, \tilde{g}_i^1, \tilde{c}_i^1, i=1,\dots,n$, and

$$
\mathbb{E}\left[\mathrm{cost}_k\right] \leq c(p^*_{Hete})|\mathcal{I}|k + C_2|\mathcal{I}|k \left(\frac{1}{k^{\min(\frac{\beta}{4}-\frac{3\eta}{4}, \eta)}} + \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}} \frac{1}{k^{\frac{1}{4}-\frac{\eta}{2}}} + \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{1-\eta}} \frac{1}{k} \right)
$$

where the constant C_2 depends on $\alpha_1, w_1, \beta, \eta, x_1, f_i, \hat{c}_i, \tilde{g}_i^1, \tilde{c}_i^1, i=1,\dots,n$. Taking square root of both sides of the above inequality and using the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$ give

$$
\sqrt{\mathbb{E}\left[\mathrm{cost}_k\right]} \leq \sqrt{c(p^*_{Hete})|\mathcal{I}|k} + \sqrt{C_2|\mathcal{I}|k} \left(\frac{1}{k^{\min(\frac{\beta}{8}-\frac{3\eta}{8}, \frac{\eta}{2})}} + \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{8}} \frac{1}{k^{\frac{1}{8}-\frac{\eta}{4}}} + \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{2(1-\eta)}} \frac{1}{k^{\frac{1}{2}}} \right). \qquad \text{(D.26)}
$$

Under the stated conditions on η, we have $\hat{c}_{\beta,\eta} > 0$ and $\min(\frac{\beta}{8} - \frac{3\eta}{8}, \frac{\eta}{2}) > 0$, and hence $\frac{1}{k^{\hat{c}_{\beta,\eta}}} \leq 1$ in (D.25) and $\frac{1}{k^{\min(\frac{\beta}{8}-\frac{3\eta}{8}, \frac{\eta}{2})}} \leq 1$ in (D.26). When $k \geq \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{1-6\eta}}$, we have that

$$
\begin{aligned}
\left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}} \frac{1}{k^{\frac{1}{4}-\frac{3\eta}{2}}} \Big/ \left(\left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1+\eta}{1-\eta}} \frac{1}{k} \right) &= \left(\frac{n}{|\mathcal{I}|}\right)^{-\frac{3+5\eta}{4(1-\eta)}} \cdot k^{\frac{3+6\eta}{4}} \\
&\geq \left(\frac{n}{|\mathcal{I}|}\right)^{-\frac{3+6\eta}{4(1-6\eta)}} \cdot k^{\frac{3+6\eta}{4}} \\
&= \left(k \Big/ \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{1-6\eta}} \right)^{\frac{3+6\eta}{4}} \geq 1,
\end{aligned}
$$

and that

$$
\left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{8}} \frac{1}{k^{\frac{1}{8}-\frac{\eta}{4}}} \Big/ \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{2(1-\eta)}} \frac{1}{k^{\frac{1}{2}}} = \left(\frac{n}{|\mathcal{I}|}\right)^{-\frac{3+\eta}{8(1-\eta)}} \cdot k^{\frac{3+2\eta}{8}}
$$

$$\geq \left(\frac{n}{|\mathcal{I}|}\right)^{-\frac{3+2\eta}{8(1-6\eta)}} \cdot k^{\frac{3+2\eta}{8}}$$
$$= \left(k \Big/ \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{1-6\eta}}\right)^{\frac{3+2\eta}{8}} \geq 1.$$

Therefore, the second terms in (D.25) and (D.26) dominate the third terms respectively. One can further verify that $\left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}} \frac{1}{k^{\frac{1}{4}-\frac{3\eta}{2}}} \leq 1$ and $\left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{8}} \frac{1}{k^{\frac{1}{8}-\frac{\eta}{4}}} \leq 1$ when $k \geq \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{1-6\eta}}$. Multiplying (D.24) with (D.26) and leaving out the high-order terms, we obtain

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[\mathrm{cost}_k\right]} \cdot \mathbb{E}[\|H(\bar{x}_{k,\gamma} - x^*)\|] \\
\leq\ & \sqrt{\frac{c(p^*_{Hete})}{1-\gamma}\mathrm{Tr}(G(p^*_{Hete}))} + C_3\left(\frac{\sqrt{|\mathcal{I}|}}{k^{\hat{c}_{\beta,\eta}}} + \frac{1}{k^{\min((\beta-3\eta)/8,\eta/2)}} + \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}} \frac{1}{k^{\frac{1}{4}-\frac{3\eta}{2}}} + \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{8}} \frac{1}{k^{\frac{1}{8}-\frac{\eta}{4}}}\right) \\
\leq\ & \sqrt{\frac{c(p^*_{Hete})}{1-\gamma}\mathrm{Tr}(G(p^*_{Hete}))} + C_3\left(\frac{\sqrt{|\mathcal{I}|}}{k^{c_{\beta,\eta}}} + \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{4}} \frac{1}{k^{\frac{1}{4}-\frac{3\eta}{2}}} + \left(\frac{n}{|\mathcal{I}|}\right)^{\frac{1}{8}} \frac{1}{k^{\frac{1}{8}-\frac{\eta}{4}}}\right),
\end{aligned}$$

where $c_{\beta,\eta} := \min(\frac{\beta}{4} - \frac{7\eta}{4}, \frac{1}{2} - \frac{\beta}{2} - \frac{\eta}{2}, \beta - \frac{1}{2} - \eta, \frac{\beta}{8} - \frac{3\eta}{8}, \frac{\eta}{2})$, and C_3 depends on $\gamma, \alpha_1, w_1, \beta, \eta, x_1, f_i, \hat{c}_i, \tilde{g}_i^1, \tilde{c}_i^1, i = 1, \ldots, n$. The bound for HeteRSGD then follows by noticing that $\rho(p) = c(p)\mathrm{Tr}(G(p))$.

We now prove the bound for the standard SGD. By following the proof of Theorem D.7 with straightforward modifications, e.g., with $p^k = p^*_{SGD}$ and $w_k = 1$, we can easily obtain the following counterpart of Theorem D.7 for the standard SGD

$$A(\bar{x}_{k,\gamma} - x^*) = \mathcal{L}_{k,\gamma} + \mathcal{E}_{k,\gamma}$$

where the leading term $\mathcal{L}_{k,\gamma}$ satisfies

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}_{k,\gamma}] = 0, \\
&\mathbb{E}[\|\mathcal{L}_{k,\gamma}\|^2] = \frac{1}{(1-\gamma)k|\mathcal{I}|}\mathrm{Tr}(AH^{-1}G(p^*_{SGD})H^{-1}A^T)
\end{aligned}$$

and the high-order error satisfies

$$\mathbb{E}[\|\mathcal{E}_{k,\gamma}\|] \leq C_4(A, \gamma, \alpha_1, \beta, x_1, f_i, i = 1, \ldots, n) \cdot \frac{1}{\sqrt{k}} \cdot \frac{1}{k^{c_\beta}}$$

with $c_\beta := \min(\frac{1}{2} - \frac{\beta}{2},\ \beta - \frac{1}{2})$. Therefore, letting $A = H$, we obtain

$$\mathbb{E}[\|H(\bar{x}_{k,\gamma} - x^*)\|] \leq \sqrt{\mathbb{E}[\|\mathcal{L}_{k,\gamma}\|^2]} + \mathbb{E}[\|\mathcal{E}_{k,\gamma}\|] \leq \sqrt{\frac{1}{(1-\gamma)k|\mathcal{I}|}\mathrm{Tr}(G(p^*_{SGD}))} + \frac{C_4}{\sqrt{k}} \cdot \frac{1}{k^{c_\beta}}.$$

On the other hand $\mathbb{E}[\mathrm{cost}_k] = c(p^*_{SGD})|\mathcal{I}|(k-1) \leq c(p^*_{SGD})|\mathcal{I}|k$, hence

$$\sqrt{\mathbb{E}[\mathrm{cost}_k]} \cdot \mathbb{E}[\|\mathcal{E}_{k,\gamma}\|] \leq \sqrt{\frac{c(p^*_{SGD})}{1-\gamma}\mathrm{Tr}(G(p^*_{SGD}))} + C_4 c(p^*_{SGD}) \cdot \frac{|\mathcal{I}|}{k^{c_\beta}}.$$

This completes the proof. □

E PROOFS FOR RESULTS IN SECTION 5

Proof of Theorem 5.1. We need the following result, which is a straightforward application of a central limit theorem for controlled Markov chains from Fort (2015) and hence the proof is omitted.

Lemma E.1 (An application of Theorem 2.1 from Fort (2015))**.** *Consider the $\mathbb{R}^d$-valued sequence x_k generated by Algorithm 1*

$$x_{k+1} = x_k + \alpha_k \nabla f(x_k) + \alpha_k \xi_k.$$

If Assumptions 4.4 and 4.5 holds, $x_k \to x^*$ a.s.., and the following five conditions are satisfied:

1. *$\nabla f(x)$ is measurable and $f(x)$ has continuous third-order derivatives in a neighborhood of x^*,*
2. *The Hessian $H := \nabla^2 f(x^*)$ is positive definite,*
3. *$\sum_{k=1}^{\infty} \alpha_k = \infty$, and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, and $\lim_{k\to\infty} \log(\alpha_{k-1}/\alpha_k)/\alpha_k = 0$,*
4. *There exists a sequence of events $\{\mathcal{A}_k \in \mathcal{F}_k, k \geq 0\}$ and $\delta > 0$ such that $\sup_{k\geq 1} \mathbb{E}[\|\xi_k\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}}] < \infty$ and $\lim_{k\to\infty} \mathbb{I}_{\mathcal{A}_k} = 1$ a.s.,*
5. *$\mathbb{E}[\xi_k \xi_k^T | \mathcal{F}_{k-1}] = M(p^*) + D_{k,1} + D_{k,2}$, where $M(p^*)$ is a positive semidefinite matrix depending on the limit sampling distribution p^*. The matrices $D_{k,1}, D_{k,2}$ satisfy $\lim_{k\to\infty} D_{k,1} = 0$ a.s. and $\lim_{k\to\infty} \alpha_k \mathbb{E}[\| \sum_{j=1}^{k} D_{j,2} \|] \to 0$,*

then

$$\frac{1}{\sqrt{\alpha_k}} (x_k - x^*) \Rightarrow \mathcal{N}(0, \Sigma(p^*)), \tag{E.1}$$

where the covariance matrix $\Sigma(p^*)$ satisfies $\Sigma(p^*)H + H\Sigma(p^*) = M(p^*)$.

To use Lemma E.1, we first argue that $x_k \to x^*$ a.s.. Since $\alpha_k = \alpha_1/k^\beta$ with $\beta \in (1/2, 1)$, we have $\sum_{k=1}^{\infty} \alpha_k^2/w_k \leq \sum_{k=1}^{\infty} \alpha_k/w_k \cdot \frac{\alpha_1}{\sqrt{k}} = \alpha_1 \sum_{k=1}^{\infty} \alpha_k/(w_k\sqrt{k}) < \infty$, therefore the conditions of Lemma 4.3 are satisfied and by Theorem 4.2 we have $x_k \to x^*$ a.s..

We then verify the five conditions in Lemma E.1. Condition 1 is directly implied by Assumption 4.1 and the continuous differentiability condition in Theorem 5.1. Condition 2 is a consequence of strong convexity from Assumption 4.1. Condition 3 can be verified to be true for step sizes in the form of $\alpha_k = \alpha_1/k^\beta$ for $\beta \in (1/2, 1)$. To verify condition 4, we consider

$$\mathcal{A}_k = \{\|x_{k+1} - x^*\| \leq 1\} \cap \Omega_{k+1}$$

where Ω_{k+1} is the event (B.4) defined in the Proof of Theorem 4.8, then by the almost sure convergence of x_k and p^k we have $\lim_{k\to\infty} \mathbb{I}_{\mathcal{A}_k} = 1$ a.s.. We've assumed that $\sup_k \alpha_k/w_k^{3+\delta} < \infty$. We write

$$\begin{aligned}
\mathbb{E}[\|\xi_k\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}}] &\leq \frac{1}{|\mathcal{I}|} \mathbb{E}\left[\sum_{i\in\mathcal{I}_k} \left\| \frac{\nabla f_i(x_k)}{n p_i^k} - \nabla f(x_k) \right\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}} \right] \quad \text{by Jensen's inequality} \\
&= \mathbb{E}\left[\left\| \frac{\nabla f_{\tilde{i}}(x_k)}{n p_{\tilde{i}}^k} - \nabla f(x_k) \right\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}} \right],
\end{aligned}$$

where $\tilde{i} | \mathcal{F}_{k-1} \sim p^k$. To further bound the above expectation, we consider two cases. If $p_{\tilde{i}}^* > 0$, then by Assumption 4.1

$$\left\| \frac{\nabla f_{\tilde{i}}(x_k)}{n p_{\tilde{i}}^k} - \nabla f(x_k) \right\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}} \leq \left(\frac{\|\nabla f_{\tilde{i}}(x^*)\| + L\|x_k - x^*\|}{n\delta^*} + L\|x_k - x^*\| \right)^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}}.$$

Otherwise if $p_{\tilde{i}}^* = 0$, we have $\nabla f_{\tilde{i}}(x^*) = 0$ by Assumption 4.5, and hence

$$\left\| \frac{\nabla f_{\tilde{i}}(x_k)}{n p_{\tilde{i}}^k} - \nabla f(x_k) \right\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}} \leq \left(\frac{L\|x_k - x^*\|}{w_k} + L\|x_k - x^*\| \right)^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}}.$$

So we can bound $\mathbb{E}[\|\xi_k\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}}]$ as

$$\begin{aligned}
\left(\mathbb{E}[\|\xi_k\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}}]\right)^{\frac{1}{2+\delta}} &\leq \left(\mathbb{E}\left[\mathbb{E}\left[\left\| \frac{\nabla f_{\tilde{i}}(x_k)}{n p_{\tilde{i}}^k} - \nabla f(x_k) \right\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}} \Big| \mathcal{F}_{k-1} \right] \right] \right)^{\frac{1}{2+\delta}} \\
&\leq \left(\mathbb{E}\left[\left(\frac{\max_i \|\nabla f_i(x^*)\|}{n\delta^*} + \frac{L\|x_k - x^*\|}{\min(n\delta^*, w_k)} + L\|x_k - x^*\| \right)^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}} \right] \right)^{\frac{1}{2+\delta}} \\
&\leq \frac{\max_i \|\nabla f_i(x^*)\|}{n\delta^*} + \left(\frac{L}{\min(n\delta^*, w_k)} + L \right) \left(\mathbb{E}[\|x_k - x^*\|^{2+\delta} \mathbb{I}_{\mathcal{A}_{k-1}}]\right)^{\frac{1}{2+\delta}}
\end{aligned}$$

$$
\begin{aligned}
& \text{by Minkowski inequality} \\
\leq\ & \frac{\max_i \|\nabla f_i(x^*)\|}{n\delta^*} + \left(\frac{L}{\min(n\delta^*, w_k)} + L\right) (\mathbb{E}[\|x_k - x^*\|^2])^{\frac{1}{2+\delta}} \\
& \text{since } \|x_k - x^*\| \leq 1 \text{ on } \mathcal{A}_{k-1} \\
\leq\ & \frac{\max_i \|\nabla f_i(x^*)\|}{n\delta^*} + C_x \left(\frac{L}{\min(n\delta^*, w_k)} + L\right) \left(\frac{\alpha_k}{w_k}\right)^{\frac{1}{2+\delta}} \quad \text{by Lemma B.5.} \quad \text{(E.2)}
\end{aligned}
$$

Since $\sup_k \alpha_k / w_k^{3+\delta} < \infty$, we see that (E.2) is bounded as $k \to \infty$. Therefore condition 4 is satisfied.

We now verify condition 5 for $M(p^*) = \frac{1}{|\mathcal{I}|} G(p^*)$. We decompose

$$
\begin{aligned}
\mathbb{E}[\xi_k \xi_k^T | \mathcal{F}_{k-1}] &= \frac{1}{|\mathcal{I}|} \left(\sum_{i=1}^n \frac{\nabla f_i(x_k) \nabla f_i^T(x_k)}{n^2 p_i^k} - \nabla f(x_k) \nabla f^T(x_k) \right) \\
&= \frac{G(p^*)}{|\mathcal{I}|} + \frac{1}{|\mathcal{I}|} \left[\sum_{i: p_i^* > 0} \left(\frac{\nabla f_i(x_k) \nabla f_i^T(x_k)}{n^2 p_i^k} - \frac{\nabla f_i(x^*) \nabla f_i^T(x^*)}{n^2 p_i^*} \right) - \nabla f(x_k) \nabla f^T(x_k) \right] \quad \text{(E.3)} \\
&\quad + \frac{1}{|\mathcal{I}|} \sum_{i: p_i^* = 0} \frac{\nabla f_i(x_k) \nabla f_i^T(x_k)}{n^2 p_i^k}. \quad \text{(E.4)}
\end{aligned}
$$

Note that the remainder in (E.3) converges to 0 a.s. since each $p_i^k \to p_i^* > 0$ and $x_k \to x^*$ a.s. and can be regarded as $D_{k,1}$ in condition 5. We then let (E.4) be $D_{k,2}$ and verify the condition for $D_{k,2}$. We write

$$
\begin{aligned}
\alpha_k \mathbb{E}\left[\left\| \sum_{j=1}^k D_{j,2} \right\| \right] &\leq \alpha_k \sum_{j=1}^k \mathbb{E}[\|D_{j,2}\|] \\
&\leq \alpha_k \frac{1}{|\mathcal{I}|} \sum_{i: p_i^* = 0} \sum_{j=1}^k \mathbb{E}\left[\left\| \frac{\nabla f_i(x_j) \nabla f_i^T(x_j)}{n^2 p_i^j} \right\| \right] \quad \text{by triangular inequality} \\
&\leq \alpha_k \frac{1}{|\mathcal{I}|} \sum_{i: p_i^* = 0} \sum_{j=1}^k \mathbb{E}\left[\frac{L^2 \|x_j - x^*\|^2}{n w_j} \right] \quad \text{by Assumptions 4.1 and 4.5} \\
&\leq \frac{C_x L^2}{|\mathcal{I}| n} \sum_{i: p_i^* = 0} \alpha_k \sum_{j=1}^k \frac{\alpha_j}{w_j^2} \quad \text{by Lemma B.5} \\
&\to 0 \quad \text{by the assumed condition } \alpha_k \sum_{j=1}^k \frac{\alpha_j}{w_j^2} \to 0.
\end{aligned}
$$

Therefore condition 5 is satisfied.

The above verification proves (E.1). Since $\mathrm{cost}_k / (|\mathcal{I}| k) \to c(p^*)$ a.s. by Proposition 4.7, then by Slutsky's theorem we can conclude

$$
\mathrm{cost}_k^{\frac{\beta}{2}} (x_k - x^*) = \sqrt{|\mathcal{I}|^\beta \alpha_1} \cdot \left(\frac{\mathrm{cost}_k}{|\mathcal{I}| k} \right)^{\frac{\beta}{2}} \cdot \frac{1}{\sqrt{\alpha_k}} (x_k - x^*) \Rightarrow \mathcal{N}\left(0, \alpha_1 |\mathcal{I}|^\beta c(p^*)^\beta \Sigma(p^*)\right). \quad \text{(E.5)}
$$

Since $\nabla f(x^*) = 0$ and the Hessian H at x^* is positive definite, the CLTs for $\nabla f(x_k)$ and $f(x_k) - f(x^*)$ then follow from the delta method. Specifically, $\nabla f(x_k) = H(x_k - x^*) + o_p(\|x_k - x^*\|)$ and $f(x_k) - f(x^*) = \frac{1}{2}(x_k - x^*)^T H (x_k - x^*) + o_p(\|x_k - x^*\|^2)$, hence the delta method implies that

$$
\begin{aligned}
& \frac{1}{\sqrt{\alpha_k}} \nabla f(x_k) \Rightarrow \mathcal{N}\left(0, H \Sigma(p^*) H\right), \\
& \frac{1}{\alpha_k} (f(x_k) - f(x^*)) \Rightarrow \left\| \mathcal{N}\left(0, \frac{1}{2} H^{\frac{1}{2}} \Sigma(p^*) H^{\frac{1}{2}}\right) \right\|^2.
\end{aligned}
$$

A similar application of Slutsky's theorem as in (E.5) then concludes

$$\text{cost}_k^{\frac{\beta}{2}} \nabla f(x_k) = \sqrt{|\mathcal{I}|^\beta \alpha_1} \cdot \left(\frac{\text{cost}_k}{|\mathcal{I}|k}\right)^{\frac{\beta}{2}} \cdot \frac{1}{\sqrt{\alpha_k}} \nabla f(x_k) \Rightarrow \mathcal{N}\left(0, \alpha_1 |\mathcal{I}|^\beta c(p^*)^\beta H \Sigma(p^*) H\right),$$

$$\text{cost}_k^{\beta}(f(x_k) - f(x^*)) = |\mathcal{I}|^\beta \alpha_1 \cdot \left(\frac{\text{cost}_k}{|\mathcal{I}|k}\right)^{\beta} \cdot \frac{1}{\alpha_k}(f(x_k) - f(x^*)) \Rightarrow \left\| \mathcal{N}\left(0, \frac{1}{2}\alpha_1 |\mathcal{I}|^\beta c(p^*)^\beta H^{\frac{1}{2}} \Sigma(p^*) H^{\frac{1}{2}}\right) \right\|^2.$$

This completes the proof. □

Proof of Proposition 5.2. According to Proposition F.1 in Section F, the optimal sampling distribution that minimizes (F.5) is continuous in the coefficients b_i, c_i, b_0. Therefore by the same argument in the proof of Proposition 4.6, p^k converges to the optimal weights $p^*_{Hete_\beta}$ a.s.. □

F EFFICIENT ROUTINES FOR OPTIMIZING (2.4) AND (5.1)

We first provide the proof for Proposition 2.2, and then present an efficient nested bisection approach (Proposition F.1 below) for optimizing efficiency metrics in the form of (5.1).

Proof of Proposition 2.2. The case that all $b_i = 0$ is trivial, so we assume at least one $b_i > 0$. Let p^* be an optimal solution to (2.5). If $p_i^* = 0$ then the corresponding b_i must be 0, since otherwise (2.5) becomes ∞. Therefore, if we only consider the nonzero p_i^*'s, then they sum up to 1 and minimize

$$\left(\sum_{i:p_i^*>0} p_i c_i\right)\left(\sum_{i:p_i^*>0} \frac{b_i}{n^2 p_i} - b_0\right),$$

which is in the same form of (2.5). For this reason, we assume all $p_i^* > 0$ without loss of generality.

Consider the Lagrangian

$$\mathcal{L}(p_1, \ldots, p_n, \lambda) = \left(\sum_{i=1}^n c_i p_i\right)\left(\sum_{i=1}^n \frac{b_i}{n^2 p_i} - b_0\right) + \lambda\left(\sum_{i=1}^n p_i - 1\right).$$

Since all $p_i^* > 0$ the following KKT condition is necessary

$$\frac{\partial \mathcal{L}}{\partial p_i}\Big|_{p=p^*, \lambda=\lambda^*} = -\left(\sum_{j=1}^n c_j p_j^*\right)\frac{b_i}{n^2 {p_i^*}^2} + c_i\left(\sum_{j=1}^n \frac{b_j}{n^2 p_j^*} - b_0\right) + \lambda^* = 0 \text{ for } i = 1, \ldots, n, \tag{F.1}$$

$$\sum_{i=1}^n p_i^* = 1. \tag{F.2}$$

where λ^* is the corresponding optimal dual variable associated with the constraint $\sum_{i=1}^n p_i = 1$. Therefore we have

$$\begin{aligned} 0 = \sum_{i=1}^n p_i^* \frac{\partial \mathcal{L}}{\partial p_i}\Big|_{p=p^*, \lambda=\lambda^*} &= -\sum_{i=1}^n c_i p_i^* \sum_{i=1}^n \frac{b_i}{n^2 p_i^*} + \sum_{i=1}^n c_i p_i^* \left(\sum_{i=1}^n \frac{b_i}{n^2 p_i^*} - b_0\right) + \lambda^* \\ &= -b_0 \sum_{i=1}^n c_i p_i^* + \lambda^*. \end{aligned}$$

Denote by $c := \sum_{i=1}^n c_i p_i^* > 0$, then $\lambda^* = b_0 c$. Denoting by $v := \sum_{i=1}^n \frac{b_i}{n^2 p_i^*} - b_0 \geq \min_{p \in \Delta_n} \sum_{i=1}^n \frac{b_i}{n^2 p_i} - b_0 = (\sum_{i=1}^n \frac{\sqrt{b_i}}{n})^2 - b_0 \geq 0$ and plugging λ^*, c, v into (F.1) we get

$$c\left(b_0 - \frac{b_i}{n^2 {p_i^*}^2}\right) + v c_i = 0, \quad \text{for } i = 1, \ldots, n.$$

We let $\kappa := \frac{v}{c} \geq 0$, and obtain

$$p_i^* = \sqrt{\frac{b_i/n^2}{\kappa c_i + b_0}}, \quad \text{for } i = 1, \ldots, n, \tag{F.3}$$

and by the feasibility condition in (F.2), κ must satisfy

$$\sum_{i=1}^n \sqrt{\frac{b_i/n^2}{\kappa c_i + b_0}} = 1. \tag{F.4}$$

From the expression (F.3) we see that in fact $p_i^* = 0$ if and only if $b_i = 0$, therefore (F.3) holds true even without assuming all $p_i^* > 0$. Finally, the uniqueness of p^* follows because the equation (F.4) has a unique root κ^* due to the strict monotonicity of its left-hand side in κ. □

We can efficiently optimize (5.1) using a nested bisection described in the next result:

Proposition F.1. *Let $c_i > 0, b_i \geq 0$ for all $i = 1, \ldots, n$, $0 \leq b_0 \leq (\sum_{i=1}^n \sqrt{b_i}/n)^2$, and consider*

$$\min_{p \in \Delta_n} \left(\sum_{i=1}^n p_i c_i \right)^\beta \left(\sum_{i=1}^n \frac{b_i}{n^2 p_i} - b_0 \right), \tag{F.5}$$

where $\beta \in (0, 1)$. If at least one $b_i > 0$, then there exists a unique minimizer p^* for (F.5) *and is given by*

$$p_i^* = \sqrt{\frac{b_i/n^2}{(1-\beta)\kappa^* + \beta c_i \kappa^*/c^* + b_0}}, \quad i = 1, 2, \ldots, n, \tag{F.6}$$

where $c^* > 0$ uniquely solves

$$\sum_{i=1}^n \sqrt{\frac{b_i/n^2 \cdot c_i^2}{(1-\beta)\kappa^*(c^*) + \beta c_i \kappa^*(c^*)/c^* + b_0}} = c^*,$$

with $\kappa^*(c) \geq 0$ for each fixed $c > 0$ uniquely solving

$$\sum_{i=1}^n \sqrt{\frac{b_i/n^2}{(1-\beta)\kappa^*(c) + \beta c_i \kappa^*(c)/c + b_0}} = 1,$$

and $\kappa^* = \kappa^*(c^*)$. Otherwise if all $b_i = 0$, then (F.5) *is constantly* 0.

We observe that the optimal weights (F.6) can be viewed as an interpolation between the variance-minimizing weights $p_i^* \propto \sqrt{b_i}$ ($\beta = 0$) and the optimal weights (2.6) for $\beta = 1$. Here is the proof for Proposition F.1:

Proof. Similar to the proof of Proposition 2.2, we assume all $p_i^* > 0$ without loss of generality. We consider the Lagrangian

$$\mathcal{L}(p_1, \ldots, p_n, \lambda) = \left(\sum_{i=1}^n c_i p_i \right)^\beta \left(\sum_{i=1}^n \frac{b_i}{n^2 p_i} - b_0 \right) + \lambda \left(\sum_{i=1}^n p_i - 1 \right),$$

and following KKT condition is necessary

$$\left. \frac{\partial \mathcal{L}}{\partial p_i} \right|_{p=p^*, \lambda=\lambda^*} = -\left(\sum_{j=1}^n c_j p_j^* \right)^\beta \frac{b_i}{n^2 {p_i^*}^2} + \beta c_i \left(\sum_{j=1}^n c_j p_j^* \right)^{\beta-1} \left(\sum_{j=1}^n \frac{b_j}{n^2 p_j^*} - b_0 \right) + \lambda^* = 0, \quad \text{for } i = 1, \ldots, n, \tag{F.7}$$

$$\sum_{i=1}^n p_i^* = 1. \tag{F.8}$$

Multiplying each side of (F.7) by p_i^* and summing up gives rise to

$$0 = \sum_{i=1}^n p_i^* \frac{\partial \mathcal{L}}{\partial p_i}\Big|_{p=p^*,\lambda=\lambda^*} = -\left(\sum_{i=1}^n c_i p_i^*\right)^\beta \left(\sum_{i=1}^n \frac{b_i}{n^2 p_i^*}\right) + \beta \left(\sum_{i=1}^n c_i p_i^*\right)^\beta \left(\sum_{i=1}^n \frac{b_i}{n^2 p_i^*} - b_0\right) + \lambda^*.$$

Again we denote by $c := \sum_{i=1}^n c_i p_i^* > 0$ and $v := \sum_{i=1}^n \frac{b_i}{n^2 p_i^*} - b_0 \geq 0$, we have $\lambda^* = ((1-\beta)v + b_0)c^\beta$. Plugging λ^*, c, v into (F.7) and dividing each side by $c^{\beta-1}$ gives

$$-\frac{c b_i}{n^2 p_i^{*2}} + \beta c_i v + c((1-\beta)v + b_0) = 0, \quad \text{for } i = 1, \ldots, n,$$

that is

$$p_i^* = \sqrt{\frac{b_i/n^2}{(1-\beta)v + \beta c_i v/c + b_0}}, \quad \text{for } i = 1, \ldots, n, \tag{F.9}$$

where the unknowns c, v must satisfy

$$\sum_{i=1}^n \sqrt{\frac{b_i/n^2}{(1-\beta)v + \beta c_i v/c + b_0}} = 1, \quad \text{by (F.8)}, \tag{F.10}$$

$$\sum_{i=1}^n \sqrt{\frac{b_i/n^2 \cdot c_i^2}{(1-\beta)v + \beta c_i v/c + b_0}} = c, \quad \text{by the definition of } c. \tag{F.11}$$

Note that (F.9) entails that $p_i^* = 0$ if and only if $b_i = 0$, therefore (F.9) continues to hold even if some $p_i^* = 0$ as in the proof of Proposition 2.2. It remains to show uniqueness. For a given $c > 0$, there exists a unique $v^*(c)$ that solves (F.10) due to strict monotonicity in v, therefore it suffices to show that (F.11) is solved by a unique c^* and the corresponding $v^*(c^*)$. Suppose that $c_1^*, v^*(c_1^*)$ and $c_2^*, v^*(c_2^*)$ both solve (F.11). Suppose $c_2^* = \eta c_1^* > c_1^* > 0$ with $\eta > 1$ without loss of generality, then we must have $v^*(c_2^*) \geq v^*(c_1^*) \geq 0$ since the left-hand side of (F.10) is increasing in c and strictly decreasing in v, therefore

$$\begin{aligned}
c_2^* &= \sum_{i=1}^n \sqrt{\frac{b_i/n^2 \cdot c_i^2}{(1-\beta)v^*(c_2^*) + \beta c_i v^*(c_2^*)/c_2^* + b_0}} \\
&\leq \sum_{i=1}^n \sqrt{\frac{b_i/n^2 \cdot c_i^2}{(1-\beta)v^*(c_1^*) + \beta c_i v^*(c_1^*)/(\eta c_1^*) + b_0}} \\
&\leq \sum_{i=1}^n \sqrt{\frac{b_i/n^2 \cdot c_i^2}{(1-\beta)v^*(c_1^*)/\eta + \beta c_i v^*(c_1^*)/(\eta c_1^*) + b_0/\eta}} \\
&= \sqrt{\eta} \sum_{i=1}^n \sqrt{\frac{b_i/n^2 \cdot c_i^2}{(1-\beta)v^*(c_1^*) + \beta c_i v^*(c_1^*)/c_1^* + b_0}} = \sqrt{\eta} c_1^* < \eta c_1^*,
\end{aligned}$$

contradicting with the starting assumption that $c_2^* = \eta c_1^*$. Hence the optimal p^* must be unique. □

G ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

G.1 Implementation Details

This subsection contains some implementation details for numerical experiments in Section 6. In all experiments, the cost for evaluating ∇f_i is fixed as c_i, not a random variable. In HeteRSGD and HeteRSGD$_\beta$, we always set $\tilde{G}_k = 0$ which makes the subproblem (3.4) (or the counterpart for HeteRSGD$_\beta$) easier to solve. We use the following parameters when implementing the algorithms:

- The synthetic example: We use $x_1 = (1, 1)$, $\alpha_k = \frac{1}{10 \cdot k^{0.8}}$, and $|\mathcal{I}_k| = 10$ for all algorithms implemented. For SRG, SRG-m, HeteRSGD and HeteRSGD$_\beta$, we set $w_k = \frac{1}{100 \cdot k^{0.4}}$.

- ℓ_2-regularized logistic regression: We use $x_1 = (0, 0, \ldots, 0)$, $\alpha_k = \frac{100}{k^{0.8}}$, and $|\mathcal{I}_k| = 10$ for all algorithms implemented. For SRG, SRG-m, HeteRSGD and HeteRSGD$_\beta$, we set $w_k = \frac{1}{100 \cdot k^{0.2}}$.
- The nonconvex example: We use $x_1 = (0, 0, \ldots, 0)$, $\alpha_k = \frac{100}{k^{0.8}}$, and $|\mathcal{I}_k| = 100$ for all algorithms implemented. For SRG, SRG-m, HeteRSGD and HeteRSGD$_\beta$, we set $w_k = \frac{1}{100 \cdot k^{0.2}}$.

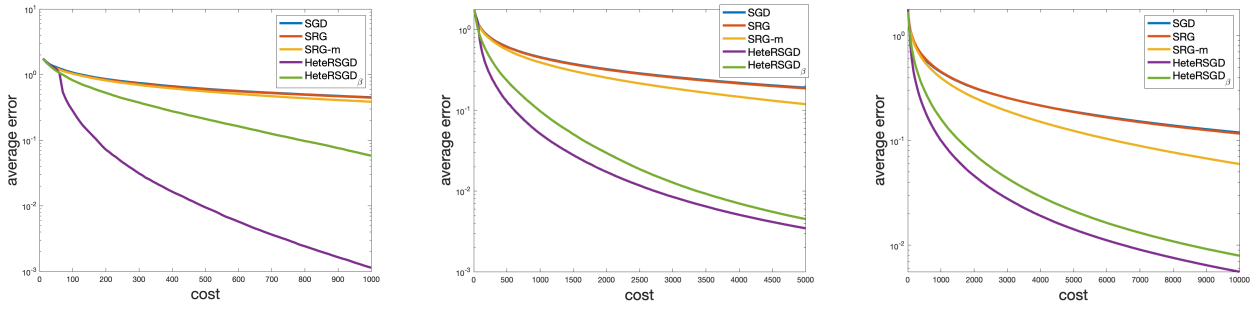
For HeteRSGD$_\beta$, we use $\beta = 0.8$ in the sampling efficiency metric (5.1) to match the decay rate of the step size used.

G.2 Additional Numerical Results

We present additional results in this subsection.

Efficiency gain under varying degrees of heterogeneity: We consider the same synthetic example from Section 6 but study a wider range of degree of heterogeneity using $\epsilon \in \{0.01, 0.05, 0.1, 0.3, 0.5, 1\}$.

As in Section 6, Figure 3 shows the average error of the Polyak-Ruppert averaged solution over 10 independent runs versus the sampling cost incurred, as the parameter ϵ increases from 0.01 to 1 (the homogenenous setting). We see a strong



(a) $\epsilon = 0.01$. (b) $\epsilon = 0.05$. (c) $\epsilon = 0.1$.

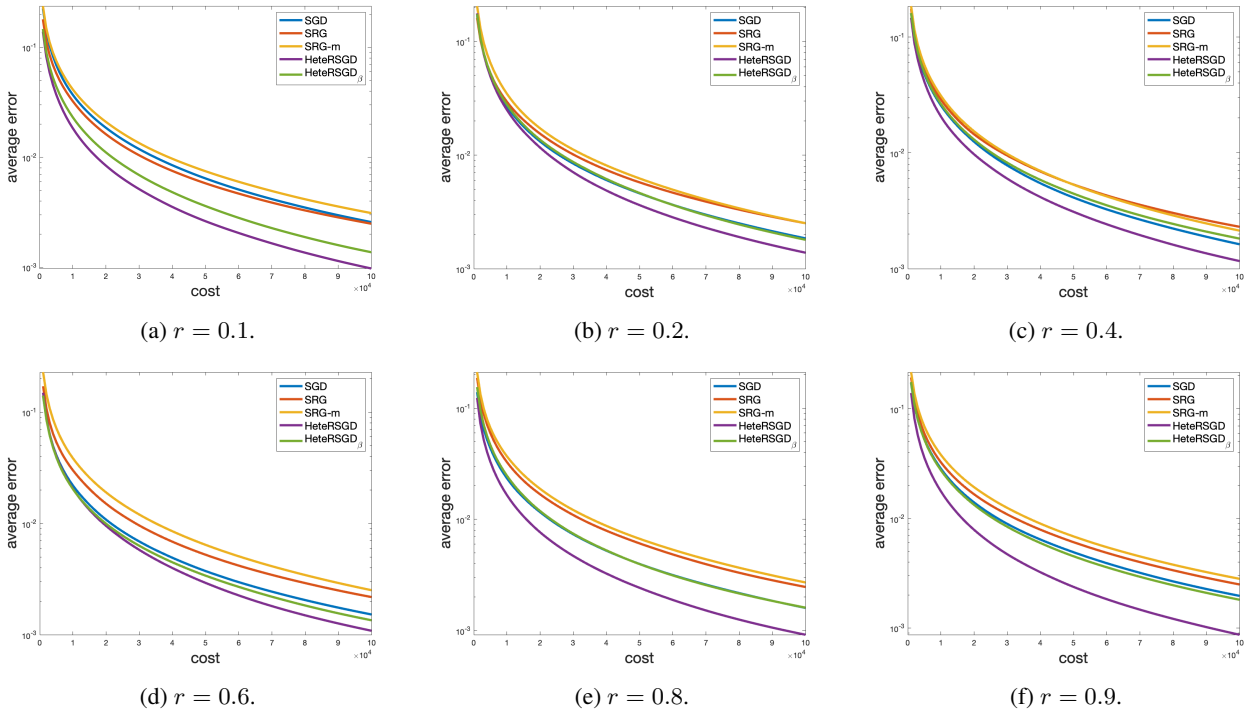
(d) $\epsilon = 0.3$. (e) $\epsilon = 0.5$. (f) $\epsilon = 1$.

Figure 3: The synthetic example from Section 6 with varying degrees of heterogeneity in sampling costs.

correlation between the degree of heterogeneity and the efficiency gain from our HeteRSGD schemes. When $\epsilon = 1$, which is essentially the homogeneous setting, all methods are indistinguishable. The efficiency gain of our HeteRSGD variants over existing schemes immediately shows up as ϵ decreases. In particular, in the case of highest heterogeneity $\epsilon = 0.01$, we observe a speedup by an order of magnitude from our HeteRSGD algorithms, especially HeteRSGD, compared to existing SGD schemes. Specifically, HeteRSGD and HeteRSGD$_\beta$ achieve the same accuracy as SGD/SRG/SRG-m with roughly 95% and 70% less sampling costs respectively. This further speedup from HeteRSGD compared to the synthetic case in Section 6 is consistent with the changes in the relative efficiency from 0.96 to 4×10^{-4} with respect to SGD and from 0.84 to 0.039 with respect to SRG. All these show that our HeteRSGD can reduce the required sampling cost by a significant amount or even an order of magnitude depending on the degree of heterogeneity in the costs.

Robustness against random sampling costs: In order to test the methods in the case that the cost of sampling each component is random rather than deterministic, we consider the same synthetic example from Section 6 but now let each sample from f_i incur a random cost $\hat{c}_i \sim \text{Uniform}((1-r)c_i, (1+r)c_i)$ for a parameter $r \in [0, 1]$. Figure 4 shows the results for $r \in \{0.1, 0.2, 0.4, 0.6, 0.8, 0.9\}$. HeteRSGD consistently outperforms existing methods for all the considered r values.

HeteRSGD$_\beta$ and SGD perform similarly, with HeteRSGD$_\beta$ outperforming SGD for $r = 0.1$. The similarity between HeteRSGD$_\beta$ and SGD is suggested by their similar asymptotic efficiencies for the averaged solution: $\rho(p^*_{Hete})/\rho(p^*_{SGD}) = 0.96$ and $\rho(p^*_{Hete})/\rho(p^*_{Hete_\beta}) = 0.98$. Nevertheless, HeteRSGD robustly outperforms existing methods under different levels of randomness in sampling costs.



(a) $r = 0.1$. (b) $r = 0.2$. (c) $r = 0.4$.

(d) $r = 0.6$. (e) $r = 0.8$. (f) $r = 0.9$.

Figure 4: The synthetic example from Section 6 under increasing randomness in sampling costs.

Comparing HeteRSGD and HeteRSGD$_\beta$, we have the same observation as in Section 6, i.e., HeteRSGD outperforms HeteRSGD$_\beta$ in almost all the cases in terms of the achieved accuracy of the averaged iterate. Having said that, we find that even if we use errors of individual iterates in place of averaged ones for comparison, the results remain similar, i.e., HeteRSGD continues to outperform HeteRSGD$_\beta$ and both HeteRSGD variants outperform existing SGD schemes. This suggests that it may take a large number of iterations in practice for the asymptotic errors of individual iterates to take effect and hence the optimality of HeteRSGD$_\beta$ appears more of theoretical interest. Based on these observations, we recommend HeteRSGD over HeteRSGD$_\beta$ for better finite-sample performance.

References

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

Fort, G. (2015). Central limit theorems for stochastic approximation with controlled markov chain dynamics. *ESAIM: Probability and Statistics*, 19:60–80.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855.