9-18-2021

An Accelerated Variance-Reduced Conditional Gradient Sliding Algorithm For First-Order And Zeroth-Order Optimization

Xiyuan Wei
Nanjing University of Information Science & Technology

Bin Gu
Nanjing University of Information Science & Technology & Mohamed bin Zayed University of Artificial Intelligence & JD Finance American Cooperation, USA

Heng Huang
JD Finance America Corporation & University of Pittsburgh

An Accelerated Variance-Reduced Conditional Gradient Sliding Algorithm for First-order and Zeroth-order Optimization

Xiyuan Wei XYWEI00@GMAIL.COM
School of Computer & Software
Nanjing University of Information Science & Technology
Nanjing, Jiangsu, 210044, China
Bin Gu BIN.GU@MBZUAI.AC.AE
MBZUAI, United Arab Emirates
JD Finance America Corporation
Heng Huang HENG.HUANG@PITT.EDU
Department of Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, PA, 15261, USA
JD Finance America Corporation

Abstract

The conditional gradient algorithm (also known as the Frank-Wolfe algorithm) has recently regained popularity in the machine learning community due to its projection-free property to solve constrained problems. Although many variants of the conditional gradient algorithm have been proposed to improve performance, they depend on first-order information (gradient) to optimize. Naturally, these algorithms are unable to function properly in the field of increasingly popular zeroth-order optimization, where only zeroth-order information (function value) is available. To fill in this gap, we propose a novel Accelerated variance-Reduced Conditional gradient Sliding (ARCS) algorithm for finite-sum problems, which can use either first-order or zeroth-order information to optimize. To the best of our knowledge, ARCS is the first zeroth-order conditional gradient sliding type algorithms solving convex problems in zeroth-order optimization. In first-order optimization, the convergence results of ARCS substantially outperform previous algorithms in terms of the number of gradient query oracle. Finally we validated the superiority of ARCS by experiments on real-world datasets.

1. Introduction

In this paper, we consider the following constrained finite-sum minimization problem:

$$\min_{x\in\mathcal{C}} \left\{ f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\} \tag{1}$$

where f is (τ-strongly) convex and L-smooth, each f_i is L-smooth and convex. $\mathcal{C} \subset \mathbb{R}^d$ is a convex set. We are particularly interested in the case where the domain $\mathcal{C}$ admits fast linear optimization. Problem (1) summarizes an extensive number of important learning problems, *e.g.*, matrix completion (Zhang et al., 2012), LASSO regression (Tibshirani, 1996), and sparsity constrained classification (Jaggi, 2013). One common approach for solving the constrained problem (1) is the projected gradient algorithm (Iusem, 2003), which conducts a projection onto the constrained set

$\mathcal{C}$ after a gradient step. However, the projection is often expensive to compute for constrained sets, for example, the set of matrices whose nuclear norm is bounded by a positive real number.

The conditional gradient (CG) algorithm (also known as the Frank-Wolfe algorithm (Frank et al., 1956)) and its variants are also natural candidates for solving problem (1). Compared to the projected gradient algorithm, CG type algorithms solve a linear optimization subproblem to bound the solution to the constrained set, which does not conduct projection, and solving the subproblem is much faster than conducting a projection. These algorithms thus have better performance due to the projection-free property, and they are gaining popularity in the machine learning community recently. The key step of CG type algorithms can be summarized as follows.

$$\begin{aligned} v^s &= \arg\max_{x\in\mathcal{C}} \langle -g^s, x\rangle \\ x^s &= (1-\gamma_s)x^{s-1} + \gamma_s v^s \end{aligned} \tag{2}$$

where $s = 1, 2, ...$ denotes the epoch, $\gamma_s \in [0, 1]$ denotes the step size. The first line of (2) calls a linear oracle to solve the linear optimization subproblem and the second line ensures that $x^s \in \mathcal{C}$ due to the convexity of the constrained set. In the conditional gradient (CG) algorithm, g^s is set to be the gradient $\nabla f(x^{s-1})$.

Formally, we denote gradient query complexity of an algorithm to be the number of calls of gradient query oracle to achieve ϵ-accuracy, *i.e.*, to get an output $x \in \mathcal{C}$ such that $f(x) - \min_{y\in\mathcal{C}} f(y) \leq \epsilon$. The CG algorithm has a gradient query complexity of $\mathcal{O}\left(n\epsilon^{-1}\right)$ for convex problems. Lan and Zhou (2016) proposed a novel variant of the CG algorithm named Conditional Gradient Sliding (CGS) algorithm which calls CG recursively in each iteration to solve a quadratic subproblem. CGS has gradient query complexity of $\mathcal{O}\left(n\epsilon^{-1/2}\right)$ and $\mathcal{O}\left(n\log\left(\epsilon^{-1}\right)\right)$ for convex and strongly-convex problems respectively. SCGS, the stochastic version of CGS, which was also proposed by Lan and Zhou (2016), has gradient query complexity of $\mathcal{O}\left(\epsilon^{-2}\right)$ for convex problems. The stochastic version of CG was analysed by Hazan and Luo (2016), which has gradient query complexity of $\mathcal{O}\left(\epsilon^{-3}\right)$ for convex problems. Hazan and Luo (2016) and Yurtsever et al. (2019) respectively combines popular variance-reduction techniques with SCGS and proposed STORC and SPIDER CGS. The linear oracle complexity (number of calls of linear oracle) of all these algorithms above is $\mathcal{O}(\epsilon^{-1})$. It can be seen that CGS type algorithms outperform CG type algorithms in terms of gradient query complexity, thus in this paper we focus on CGS type algorithms.

Although the literature is rich, most CGS type algorithms are first-order algorithms, which take advantage of the gradients to optimize. However, in many complex machine learning problems, the explicit gradient of the problem is expensive to compute or even inaccessible, e.g., problems concerning black-box adversarial attacks (Chen et al., 2017), bandit optimization (Flaxman et al., 2005), reinforcement learning (Choromanski et al., 2018) and metric learning (Kulis et al., 2012). Thus first-order algorithms are not applicable to these problems. Zeroth-order algorithm is a promising substitute since it only uses function value to optimize. But zeroth-order conditional gradient sliding type algorithms for the finite-sum problem are understudied. To the best of our knowledge, only Gao and Huang (2020) studied the zeroth-order version of SPIDER CGS, but it is only analysed for non-convex problems. Thus there have not been analyses on zeroth-order conditional gradient sliding type algorithms for convex problems.

To fill in the gap, we propose an Accelerated variance-Reduced Conditional gradient Sliding (ARCS) algorithm, which leverages variance-reduction techinque and a novel momentum acceleration technique proposed by Lan et al. (2019). Our ARCS algorithm can be used in either first-order

Table 1: Comparison of conditional gradient sliding type algorithms solving *convex* problems. $D_0 = \mathcal{O}([f(\tilde{x}^0) - f(x^*)] + L\|x^0 - x^*\|^2)$. **F** indicates that the result is for the first-order case and **Z** indicates that the result is for the zeroth-order case. Note that our ARCS is the first zeroth-order conditional gradient sliding type algorithm solving convex problems. $\tilde{\mathcal{O}}$ hides a logarithmic factor.

	Algorithm	Oracle Complexity	
		Gradient / Function Query	Linear Oracle
F	CGS (Lan and Zhou, 2016)	$\mathcal{O}\left(\frac{n}{\sqrt{\epsilon}}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
F	SCGS (Lan and Zhou, 2016)	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
F	SPIDER CGS (Yurtsever et al., 2019)	$\tilde{\mathcal{O}}\left(n + \frac{1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
F	STORC (Hazan and Luo, 2016)	$\tilde{\mathcal{O}}\left(n + \frac{1}{\epsilon^{1.5}}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
F	**Ours (first-order)**	$\begin{cases} \tilde{\mathcal{O}}\left(n\right), & \epsilon \geq \frac{3D_0}{n} \\ \tilde{\mathcal{O}}\left(n + \sqrt{\frac{n}{\epsilon}}\right), & \epsilon < \frac{3D_0}{n} \end{cases}$	$\begin{cases} \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right), & \epsilon \geq \frac{3D_0}{n} \\ \tilde{\mathcal{O}}\left(n^2 + \frac{n}{\epsilon}\right), & \epsilon < \frac{3D_0}{n} \end{cases}$
Z	**Ours (zeroth-order)**	$\begin{cases} \tilde{\mathcal{O}}\left(nd\right), & \epsilon \geq \frac{5D_0}{n} \\ \tilde{\mathcal{O}}\left(nd + d\sqrt{n/\epsilon}\right), & \epsilon < \frac{5D_0}{n} \end{cases}$	$\begin{cases} \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right), & \epsilon \geq \frac{5D_0}{n} \\ \tilde{\mathcal{O}}\left(n^2 + \frac{n}{\epsilon}\right), & \epsilon < \frac{5D_0}{n} \end{cases}$

or zeroth-order optimization. In first-order optimization, it outperforms all existing conditional gradient type algorithms with respect to gradient query complexity. In zeroth-order optimization, it is the first conditional gradient sliding type algorithm for convex problems. Since zeroth-order algorithms use function values instead of gradients to optimize, it is natural to consider the number of calls of function query oracle to achieve ϵ-accuracy when assessing the performance of zeroth-order algorithms, which we denote to be the function query complexity.

Besides theoretical analyses, we conduct numerical experiments on real-world datasets, and the results also show the optimality of our ARCS in gradient/function query complexity in both first-order and zeroth-order optimization.

Contributions. The main contributions of this paper are summarized as follows:

- We propose an Accelerated variance-Reduced Conditional gradient Sliding (ARCS) algorithm. Our ARCS algorithm is based on the stochastic conditional gradient sliding (SCGS) algorithm and it leverages the variance-reduction technique and a novel momentum acceleration technique. We give convergence results of ARCS in zeroth-order optimization. To the best of our knowledge, our ARCS algorithm is the first zeroth-order conditional gradient sliding type algorithm addressing the convex and strongly-convex finite-sum problems. Numerical experiments also show its optimality.

- As a by-product, we give convergence results of ARCS in first-order optimization. Both theoretic and numerical results confirm that ARCS have significantly improved gradient query complexities on convex and strongly-convex problems in first-order optimization.

Table 2: Comparison of conditional gradient sliding type algorithms solving *strongly convex* problems. $D_0 = \mathcal{O}([f(\tilde{x}^0) - f(x^*)] + L\|x^0 - x^*\|^2)$. **F** indicates that the result is for the first-order case and **Z** indicates that the result is for the zeroth-order case. Note that our ARCS is the first zeroth-order conditional gradient sliding type algorithm solving strongly-convex problems. $\tilde{\mathcal{O}}$ hides a logarithmic factor.

	Algorithm	Oracle Complexity	
		Gradient / Function Query	Linear Oracle
F	CGS (Lan and Zhou, 2016)	$\tilde{\mathcal{O}}\left(n\sqrt{\frac{L}{\tau}}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
	SCGS (Lan and Zhou, 2016)	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
	STORC (Hazan and Luo, 2016)	$\tilde{\mathcal{O}}\left(n+\frac{L^2}{\tau^2}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$
	Ours (first-order)	$\begin{cases}\tilde{\mathcal{O}}\left(n\right), & \epsilon \geq 5D_0/n \text{ or } n \geq 3L/4\tau \\ \tilde{\mathcal{O}}\left(n+\sqrt{\frac{nL}{\tau}}\right), & \epsilon < 5D_0/n \text{ and } n < 3L/4\tau\end{cases}$	$\begin{cases}\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right), & \epsilon \geq \frac{5D_0}{n} \\ \tilde{\mathcal{O}}\left(n^2+\frac{n}{\epsilon}\right), & \epsilon < \frac{5D_0}{n}\end{cases}$
Z	**Ours (zeroth-order)**	$\begin{cases}\tilde{\mathcal{O}}\left(nd\right), & \epsilon \geq 8D_0/n \text{ or } n \geq 3L/4\tau \\ \tilde{\mathcal{O}}\left(nd+d\sqrt{\frac{nL}{\tau}}\right), & \epsilon < 8D_0/n \text{ and } n < 3L/4\tau\end{cases}$	$\begin{cases}\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right), & \epsilon \geq \frac{8D_0}{n} \\ \tilde{\mathcal{O}}\left(n^2+\frac{n}{\epsilon}\right), & \epsilon < \frac{8D_0}{n}\end{cases}$

2. Related Works

Conditional Gradient Algorithms. Frank et al. (1956) proposed the conditional gradient (CG) algorithm, also known as Frank-Wolfe (FW) algorithm, to avoid projection in solving constrained problems. Motivated by removing the influence of "bad" visited vertices, Wolfe (1970) proposed away-step Frank-Wolfe (AFW) algorithm. Goldfarb et al. (2017) proposed ASFW, the stochastic version of AFW. Lan and Zhou (2016) proposed a variant of CG called conditional gradient sliding (CGS) algorithm which calls CG recursively in each iteration until a good solution is obtained. SCGS, the stochastic version of CGS was also proposed by Lan and Zhou (2016). Hazan and Luo (2016) gave convergence results of the stochastic version of CG, which is called SFW. Also, Hazan and Luo (2016) combined the variance-reduction technique proposed by Johnson and Zhang (2013) with SFW and SCGS to get SVRF and STORC respectively. Yurtsever et al. (2019) combined another variance-reduction technique proposed by Fang et al. (2018) with SCGS to get SPIDER CGS.

Zeroth-Order Optimization. Zeroth-order optimization is a classical technique in the optimization community. Nesterov and Spokoiny (2017) proposed zeroth-order gradient descent (ZO-GD) algorithm. Then Ghadimi and Lan (2013) proposed its stochastic counterpart ZO-SGD. Lian et al. (2016) proposed an asynchronous zeroth-order stochastic gradient (ASZO) algorithm for parallel op-

timization. Gu et al. (2018) further improved the convergence rate of ASZO by combining variance reduction technique with coordinate-wise gradient estimators. Liu et al. (2018) proposed ZO-SVRG based algorithms using three different gradient estimators. Fang et al. (2018) proposed a SPIDER based zeroth-order method named SPIDER-SZO.Ji et al. (2019) further improved ZO SVRG based and SPIDER based algorithms. Chen et al. (2019) proposed zeroth-order adaptive momentum method (ZO-AdaMM). Chen et al. (2020) proposed ZO-Varag which leverages acceleration and variance-reduced technique. Sahu et al. (2019) proposed zeroth-order versions of (stochastic) conditional gradient method. Balasubramanian and Ghadimi (2018) proposed zeroth-order versions of stochastic conditional gradient method and stochastic conditional gradient sliding method. These zeroth-order conditional gradient type algorithms mentioned above did not consider the finite-sum problem (1).

3. Preliminaries

For simplicity, we denote $x^* \stackrel{\text{def}}{=} \arg\min_{x\in\mathcal{C}} f(x)$ to be the optimal solution to the problem (1) and denote $\|\cdot\|$ to be the norm associated with inner product in $\mathbb{R}^d$. First we give formal definitions of some basic concepts.

Definition 1 *For function $f : \mathbb{R}^d \to \mathbb{R}$, we have*

- *f is L-smooth if f has continuous gradients and $\forall\, x, y \in \mathbb{R}^d$, it satisfies $|f(y) - f(x) - \langle\nabla f(x), y - x\rangle| \le \frac{L}{2}||y - x||^2$.*
- *f is convex if $\forall\, x, y \in \mathbb{R}^d$, it satisfies $f(y) \ge f(x) + \langle\nabla f(x), y - x\rangle$.*
- *f is τ-strongly-convex if $f(x) - \frac{\tau}{2}||x||^2$ is convex,* i.e., *$\forall\, x, y \in \mathbb{R}^d$, it satisfies $f(y) \ge f(x) + \langle\nabla f(x), y - x\rangle + \frac{\tau}{2}||y - x||^2$.*

From Definition 1 we know if f is convex, then it is 0-strongly-convex. Next we give assumptions that will be used in our analyses.

3.1 Assumptions

A 2 *f is convex and each $f_i, i = 1, ..., n$ is L-smooth.*

A 3 *f is τ-strongly-convex with $\tau > 0$ and each $f_i, i = 1, ..., n$ is L-smooth.*

A 4 *For any $x, y \in \mathcal{C}$, there exists $D < \infty$ such that $\|x - y\| \le D$.*

Assumption 4 is standard for the convergence analysis of conditional gradient type algorithms (Jaggi, 2013; Lan and Zhou, 2016; Hazan and Luo, 2016). Next we specify the oracles that are used in our algorithms.

3.2 Oracles

We introduce three oracles called in our algorithm.

- Gradient Query Oracle (GQO): GQO returns the gradient of a given component function at point x, which is $\nabla f_i(x)$.

- Function Query Oracle (FQO): FQO returns the value of a given component function at point x, which is $f_i(x)$.
- Linear Oracle (LO): LO sovles the linear programming problem for vector u and returns $\arg\max_{v\in\mathcal{C}}\langle u, v\rangle$.

In this paper, we consider the following two cases:

- First-order Case: We have access to GQO and LO.
- Zeroth-order Case: We have access to FQO and LO.

3.3 Zeroth-order Gradient Estimation

For the zeroth-order case, we only have access to the function query oracle rather than the gradient query oracle. Then we can utilize the difference of the function value at two close points to estimate the gradient. Two gradient estimators are widely used in zeroth-order optimization: the two-point Gaussian random gradient estimator (Nesterov and Spokoiny, 2017) and the coordinate-wise gradient estimator (Lian et al., 2016). Liu et al. (2018) showed that the coordinate-wise gradient estimator has better performance than the two-point Gaussian random gradient estimator. So we only consider the coordinate-wise gradient estimator in this paper, which is defined as follows:

$$\hat{\nabla}_{coord} f(x) = \sum_{i=1}^{d} \frac{f(x+\mu e_i) - f(x-\mu e_i)}{2\mu} e_i \tag{3}$$

where e_i is the i-th vector of the standard basis of $\mathbb{R}^d$ and $\mu > 0$ is a smoothing parameter.

4. Algorithms and Analyses

Lan and Zhou (2016) proposed a novel variant of the conditional gradient algorithm named Conditional Gradient Sliding (CGS) algorithm. CGS calls the linear oracle recursively in each iteration until a good solution is obtained. The idea of CGS can be summarized as follows:

$$\begin{aligned} z^s =& (1-\alpha_s) y^{s-1} + \alpha_s x^{s-1} \\ x^s =& \text{CondG}(\nabla f(z^s), x^{s-1}, 0, \gamma_s, 0, \eta_s) \quad \text{(Algo. 1)} \\ y^s =& (1-\alpha_s) y^{s-1} + \alpha_s x^s \end{aligned} \tag{4}$$

The second line of CGS calls Algorithm 1. In each iteration, the linear oracle is called to produce an output v_t of (10). If the value $V_{g,u,y,\gamma,\tau}(u_t) \le \eta$, then it sets $u^+ = u_t$ and returns. Thus Algorithm 1 outputs a solution u^+ such that

$$\max_{x\in\mathcal{C}} \langle \nabla h(u^+), u^+ - x\rangle \le \eta \tag{5}$$

where h is a quadratic function defined as

$$h(x) \stackrel{\text{def}}{=} \gamma \left[\langle g, x\rangle + \frac{\tau}{2}\|x-y\|^2 \right] + \frac{1}{2}\|x-u\|^2 \tag{6}$$

On the other hand, if $V_{g,u,y,\gamma,\tau}(u_t) > \eta$, then u_t is updated with line search, *i.e.*, $u_{t+1} = (1-\beta_t)u_t + \beta v_t$, where

$$\beta_t = \arg\min_{\beta\in[0,1]} h((1-\beta)u_t + \beta v_t) \tag{7}$$

Denote $u^* = \arg\min_{u\in\mathcal{C}} h(u)$, from the convexity of h, the output u^+ satisfies

$$h(u^+) - h(u^*) \le \langle \nabla h(u^+), u^+ - u^* \rangle \le \eta \tag{8}$$

Then it is clear that Algorithm 1 is in fact the standard conditional gradient algorithm (2) minimizing h. In the CGS algorithm (4), we have $x^s = \mathrm{CondG}(\nabla f(z^s), x^{s-1}, 0, \gamma_s, 0, \eta_s)$, so h in CGS can be rewritten as

$$h_s'(x) \stackrel{\mathrm{def}}{=} \gamma_s \langle \nabla f(z^s), x\rangle + \frac{1}{2}\|x - x_{s-1}\|^2 \tag{9}$$

Note that if $\mathcal{C} = \mathbb{R}^d$, then the minimizer of (9) has a closed form solution and it is in fact an accelerated gradient descent step. We choose the more complicated form (6) since it gives our algorithm better performance when problem (1) is strongly convex ($\tau > 0$). When problem (1) is convex ($\tau = 0$), (6) is identical to (9).

Algorithm 1 CondG Algorithm

Input: $(g, u, y, \gamma, \tau, \eta)$
Define $h(x) \stackrel{\mathrm{def}}{=} \gamma\left[\langle g, x\rangle + \frac{\tau}{2}\|x-y\|^2\right] + \frac{1}{2}\|x-u\|^2$
Set $u_1 = u$.
for $t = 1, 2, \ldots$ **do**
  Let v_t be an optimal solution of the subproblem

$$V_{g,u,y,\gamma,\tau}(u_t) = \max_{x\in\mathcal{C}} \langle \nabla h(u_t), u_t - x\rangle \tag{10}$$

  if $V_{g,u,y,\gamma,\tau}(u_t) \le \eta$ **then**
    Output $u^+ = u_t$.
  else
    Set $u_{t+1} = (1-\beta_t)u_t + \beta_t v_t$ with $\beta_t = \max\left\{0, \min\left\{1, \frac{\langle \nabla h(u_t), u_t - v_t\rangle}{(\gamma\tau+1)\|u_t - v_t\|^2}\right\}\right\}$
  end if
end for

Lan et al. (2019) proposed a VAriance-Reduced Accelerated Gradient (Varag) algorithm for unconstrained finite-sum problems, which leverages the variance-reduction technique and a novel momentum technique. Inspired by Varag, we combined variance-reduction technique and momentum with the conditional gradient sliding algorithm, and proposed our Accelerated variance-Reduced Conditional gradient Sliding (ARCS) algorithm. The detail of ARCS is described in Algorithm 2.

At the beginning of epoch s, ARCS computes a full gradient $\tilde{g}$ at point $\tilde{x}^{s-1}$, which is the solution provided by the preceding epoch. Then the full gradient is used repeatedly in each inner loop to form a gradient blending G_t. This is the classic variance-reduction technique proposed by Johnson and Zhang (2013). Each inner loop maintains three sequences: $\{\underline{x}_t\}, \{x_t\}, \{\bar{x}_t\}$, which is a novel momentum technique proposed by Lan et al. (2019) and plays an important role in the acceleration scheme. The choice of the additional parameters $\{T_s\}, \{p_s\}, \{a_s\}, \{\gamma_s\}, \{\eta_{s,t}\}, \{\theta_t\}$ will

Algorithm 2 Accelerated variance-Reduced Conditional gradient Sliding (ARCS) algorithm

Input: $x_0 \in \mathcal{C}, \{T_s\}, \{\gamma_s\}, \{\alpha_s\}, \{p_s\}, \{\theta_t\}, \{\eta_{s,t}\}$
Set $\tilde{x}^0 = x^0$.
for $s = 1, 2, ...$ **do**
  Set $\tilde{x} = \tilde{x}^{s-1}$ and $\tilde{g} = \begin{cases} \nabla f(\tilde{x}), & \textit{// for first-order case} \\ \hat{\nabla}_{coord} f(\tilde{x}), & \textit{// for zeroth-order case} \end{cases}$
  Set $x_0 = x^{s-1}$, $\bar{x}_0 = \tilde{x}$ and $T = T_s$.
  for $t = 1, ..., T$ **do**
    Pick $i_t \in \{1, ..., n\}$ randomly.
    Set $\underline{x}_t = \frac{[(1+\tau\gamma_s)(1-\alpha_s-p_s)\bar{x}_{t-1} + \alpha_s x_{t-1} + (1+\tau\gamma_s)p_s\tilde{x}]}{(1+\tau\gamma_s(1-\alpha_s))}$
    $G_t = \begin{cases} \nabla f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\tilde{x}) + \tilde{g}, & \textit{// for first-order case} \\ \hat{\nabla}_{coord} f_{i_t}(\underline{x}_t) - \hat{\nabla}_{coord} f_{i_t}(\tilde{x}) + \tilde{g}, & \textit{// for zeroth-order case} \end{cases}$
    $x_t = \mathrm{CondG}(G_t, x_{t-1}, \underline{x}_t, \gamma_s, \tau, \eta_{s,t})$ *// Algorithm 1*
    $\bar{x}_t = (1 - \alpha_s - p_s)\bar{x}_{t-1} + \alpha_s x_t + p_s\tilde{x}$.
  end for
  Set $x^s = x_T$ and $\tilde{x}^s = \sum_{t=1}^{T}(\theta_t \bar{x}_t) / \sum_{t=1}^{T} \theta_t$.
end for

be specified in our convergence analyses for first-order and zeroth-order case, convex and strongly-convex problems respectively. First we provide the convergence results of our ARCS solving *convex* problems. The proof of Theorem 5 is left in the appendix.

Theorem 5 (Convex) *Suppose Assumptions 2 and 4 holds. Denote* $s_0 = \lfloor \log n \rfloor + 1$, *set*

$$T_s = \begin{cases} 2^{s-1}, & s \le s_0 \\ T_{s_0}, & s > s_0 \end{cases},\ \alpha_s = \begin{cases} \frac{1}{2}, & s \le s_0 \\ \frac{2}{s - s_0 + 4}, & s > s_0 \end{cases},\ p_s = \frac{1}{2},\ \eta_{s,t} = \frac{D_0}{sT_sL},\ \theta_t = \begin{cases} \frac{\gamma_s}{\alpha_s}(\alpha_s + p_s), & t \le T_s - 1 \\ \frac{\gamma_s}{\alpha_s}, & t = T_s \end{cases}$$

where D_0 *will be specified below for two cases respectively.*

- *For the first-order case, set* $\gamma_s = \frac{1}{3L\alpha_s}, D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 3L\|x^0 - x^*\|^2$, *we have*

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \le \begin{cases} \dfrac{3D_0(\log S + 2)}{2^{S+1}}, & S \le s_0 \\ \dfrac{48D_0(\log S + 2)}{n(S - s_0 + 4)^2}, & S > s_0 \end{cases}$$

- *For the zeroth-order case, set* $\gamma_s = \frac{1}{5L\alpha_s}, D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 5L\|x^0 - x^*\|^2$, *we have*

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \le \begin{cases} \dfrac{5D_0(\log S + 2)}{2^{S+1}} + D\mu L\sqrt{\dfrac{d^2}{2} + 2d} + \dfrac{\mu^2 Ld}{2}, & S \le s_0 \\ \dfrac{80D_0(\log S + 2)}{n(S - s_0 + 4)^2} + \Delta^\mu, & S > s_0 \end{cases}$$

where $\Delta^\mu = 2(S - s_0 + 4)\mu^2 Ld + 4(S - s_0 + 4)D\mu L\sqrt{\frac{d^2}{2} + 2d}$.

Corollary 6 *With parameters set in Theorem 5, for* convex *problems, we have ($\tilde{\mathcal{O}}$ hides a logarithmic factor)*

- *For the first-order case, the gradient query complexity can be bounded as*

$$N_{GQO} = \begin{cases} \tilde{\mathcal{O}}\left(n\log\frac{D_0}{\epsilon}\right), & \epsilon \geq \tilde{\mathcal{O}}\left(\frac{D_0}{n}\right) \\ \tilde{\mathcal{O}}\left(n\log n + \sqrt{\frac{nD_0}{\epsilon}}\right), & \epsilon < \tilde{\mathcal{O}}\left(\frac{D_0}{n}\right) \end{cases}$$

- *For the zeroth-order case, the function query complexity can be bounded as*

$$N_{FQO} = \begin{cases} \tilde{\mathcal{O}}\left(nd\log\frac{D_0}{\epsilon}\right), & \epsilon \geq \tilde{\mathcal{O}}\left(\frac{D_0}{n}\right) \\ \tilde{\mathcal{O}}\left(nd\log n + d\sqrt{\frac{nD_0}{\epsilon}}\right), & \epsilon < \tilde{\mathcal{O}}\left(\frac{D_0}{n}\right) \end{cases}$$

- *For both cases, the linear oracle complexity can be bounded as*

$$N_{LO} = \begin{cases} \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right), & \epsilon \geq \tilde{\mathcal{O}}\left(\frac{D_0}{n}\right) \\ \tilde{\mathcal{O}}\left(n^2 + \frac{n}{\epsilon}\right), & \epsilon < \tilde{\mathcal{O}}\left(\frac{D_0}{n}\right) \end{cases}$$

From Table 1 it can be seen that the known best algorithms with lowest gradient query complexity for solving convex problems are CGS and STORC, whose results are $\mathcal{O}\left(n\epsilon^{-1/2}\right)$ and $\mathcal{O}\left(n\log\left(\epsilon^{-1}\right) + \epsilon^{-3/2}\right)$ respectively. CGS outperforms STORC when $\epsilon < n^{-1}$ and STORC takes the lead otherwise. But it is easy to verify that the gradient query complexity of ARCS is always lower that of CGS and STORC. The gradient query complexity of ARCS is $\tilde{\mathcal{O}}\left(n\log\left(\epsilon^{-1}\right)\right)$ when $\epsilon \geq 3D_0/n$ and $\tilde{\mathcal{O}}\left(n\log(n) + n^{1/2}\epsilon^{-1/2}\right) = \tilde{\mathcal{O}}\left(n^{1/2}\epsilon^{-1/2}\right)$ otherwise. Thus ARCS outperforms all existing algorithms in terms of gradient query complexity.

However, Theorem 5 (Theorem 7 as well) implies that ARCS has a higher linear oracle complexity than CGS and STORC. To explain this, we make a comparison between ARCS and STORC since they are both accelerated variance-reduced stochastic conditional gradient sliding algorithms. For completeness we include STORC and its key theorems in the appendix. The key differences between ARCS and STORC lie in a) the choice of γ_s and α_s, b) the choice of $\{\underline{x}_t\}$, $\{x_t\}$ and $\{\bar{x}_t\}$, c) minibatch of stochastic gradients.

To be specific, a) the choice of γ_s and α_s contributes most to the difference in convergence results. We have $\alpha\gamma = \mathcal{O}(L^{-1})$ for each inner iteration in both ARCS and STORC. For each epoch (*i.e.*, s is fixed), γ and α in ARCS are constant while in STORC, α diminish with a rate of $\mathcal{O}(t^{-1})$. This adds to $\eta_{s,t}$ a factor of $\mathcal{O}(t^{-1})$ so that $\eta_{s,t}$ can be chosen $\mathcal{O}(t)$ times larger. Thus the linear oracle complexity is lowered down (the linear oracle complexity is proportional to $\eta_{s,t}^{-1}$ from Jaggi 2013). However, this comes with a price. The decrease of α requires a larger minibatch of stochastic gradients in each inner iteration to lower down the variance. Thus STORC has a higher gradient query complexity, which becomes even higher than CGS when $\epsilon < n^{-1}$. b) the choice of $\{\underline{x}_t\}$, $\{x_t\}$ and $\{\bar{x}_t\}$ leverages the acceleration technique and yields accelerated convergence rates for both ARCS and STORC. c) minibatch of stochastic gradients in STORC is required by the choice of γ_s and α_s to lower down the variance of stochastic gradients in the analyses. The points discussed above also work on CGS. In fact, CGS is a deterministic conditional gradient sliding algorithm and it a) benefits from choice of γ_s and α_s as STORC, b) maintains similar acceleration sequences $\{z^s\}$, $\{x^s\}$ and $\{y^s\}$ (see (4)). Next we give convergence results of our ARCS solving *strongly-convex* problems.

Theorem 7 (Strongly-convex) *Suppose Assumptions 3 and 4 hold. Denote* $s_0 = \lfloor \log n \rfloor + 1$, *set*

$$T_s = \begin{cases} 2^{s-1}, & s \le s_0 \\ T_{s_0}, & s > s_0 \end{cases},\ \alpha_s = \begin{cases} \frac{1}{2}, & s \le s_0 \\ \min\left\{\sqrt{\frac{n}{4\varsigma}}, \frac{1}{2}\right\}, & s > s_0 \end{cases}$$

$$p_s = \frac{1}{2},\ \theta_t = \begin{cases} \Gamma_{t-1} - (1-\alpha_s - p_s)\Gamma_t, & t \le T_s - 1 \\ \Gamma_{t-1}, & t = T_s \end{cases},\ \eta_{s,t} = \begin{cases} \frac{D_0}{sT_sL}, & s \le s_0 \\ \frac{\left(\frac{4}{5}\right)^{s-s_0-1} D_0}{snL}, & s > s_0 \text{ and } n \ge \varsigma \\ \frac{\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0-1} D_0}{snL}, & s > s_0 \text{ and } n < \varsigma \end{cases}$$

where ς, Γ_t, D_0 *will be specified below for two cases respectively.*

• *For the first-order case, set* $\gamma_s = \frac{1}{3L\alpha_s}, \varsigma = \frac{3L}{4\tau}, \Gamma_t = (1+\tau\gamma_s)^t, D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 3L\|x^0 - x^*\|^2$. *We have*

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \le \begin{cases} \frac{3D_0(\log S+2)}{2^{S+1}}, & s \le s_0 \\ \left(\frac{4}{5}\right)^{S-s_0} \frac{5D_0(\log S+2)}{n}, & s > s_0 \text{ and } n \ge \varsigma \\ \left(1 + \frac{1}{2}\sqrt{\frac{n\tau}{3L}}\right)^{-(S-s_0)} \frac{5D_0(\log S+2)}{n}, & s > s_0 \text{ and } n < \varsigma \end{cases}$$

• *For the zeroth-order case, set* $\gamma_s = \frac{1}{12dL\alpha_s}, \varsigma = \frac{5L}{4\tau}, \Gamma_t = \left(1 + \frac{\tau\gamma_s}{2}\right)^t, D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 5L\|x^0 - x^*\|^2$. *We have*

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \le \begin{cases} \frac{5D_0(\log S+2)}{2^{S+1}} + \frac{\mu^2L^2d(d+4)}{4\tau} + 2\mu^2Ld, & s \le s_0 \\ \left(\frac{4}{5}\right)^{S-s_0} \frac{8D_0(\log S+2)}{n} + \Delta_1^\mu, & s > s_0,\ n \ge \varsigma \\ \left(1 + \frac{1}{4}\sqrt{\frac{n\tau}{5L}}\right)^{-(S-s_0)} \frac{8D_0(\log S+2)}{n} + \Delta_2^\mu, & s > s_0,\ n < \varsigma \end{cases}$$

where $\Delta_1^\mu = \frac{5\mu^2L^2d(d+4)}{\tau} + 12\mu^2Ld, \Delta_2^\mu = \frac{5\mu^2L^2d(d+4)}{\tau} + 4\mu^2Ld\left(1 + 2\sqrt{\frac{5L}{n\tau}}\right)$.

Corollary 8 *With parameters set in Theorem 7, for* strongly-convex *problems, we have (*$\tilde{\mathcal{O}}$ *hides a logarithmic factor)*

• *For the first-order case, the gradient query complexity can be bounded as*

$$N_{GQO} = \begin{cases} \tilde{\mathcal{O}}\left(n \log\left(\frac{D_0}{\epsilon}\right)\right), & \epsilon \ge \tilde{\mathcal{O}}(\frac{D_0}{n}) \text{ or } n \ge \varsigma \\ \tilde{\mathcal{O}}\left(n\log n + \sqrt{\frac{nL}{\tau}} \log\left(\frac{D_0}{n\epsilon}\right)\right), & \epsilon < \tilde{\mathcal{O}}(\frac{D_0}{n}) \text{ and } n < \varsigma \end{cases}$$

• *For the zeroth-order case, the function query complexity can be bounded as*

$$N_{FQO} = \begin{cases} \tilde{\mathcal{O}}\left(nd \log\left(\frac{D_0}{\epsilon}\right)\right), & \epsilon \ge \tilde{\mathcal{O}}(\frac{D_0}{n}) \text{ or } n \ge \varsigma \\ \tilde{\mathcal{O}}\left(nd\log n + d\sqrt{\frac{nL}{\tau}} \log\left(\frac{D_0}{n\epsilon}\right)\right), & \epsilon < \tilde{\mathcal{O}}(\frac{D_0}{n}) \text{ and } n < \varsigma \end{cases}$$

• *For both cases, the linear oracle complexity can be bounded as*

$$N_{LO} = \begin{cases} \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right), & \epsilon \geq \tilde{\mathcal{O}}(D_0/n) \text{ or } n \geq \varsigma \\ \tilde{\mathcal{O}}\left(n^2 + \frac{n}{\epsilon}\right), & \epsilon < \tilde{\mathcal{O}}(D_0/n) \text{ and } n < \varsigma \end{cases}$$

From Table 2 it can be seen that the known best algorithms with lowest gradient query oracle complexity for solving *strongly-convex* problems are CGS and STORC, whose results are $\mathcal{O}\left(nL^{1/2}\tau^{-1/2}\log\left(\epsilon^{-1}\right)\right)$ and $\mathcal{O}\left(\left(n + L^2\tau^{-2}\right)\log\left(\epsilon^{-1}\right)\right)$. CGS outperforms STORC when $n < L^{3/2}\tau^{-3/2}$ and STORC takes the lead otherwise. But it is easy to verify that the gradient query complexity of ARCS is always lower that of CGS and STORC. The gradient query complexity of ARCS is $\tilde{\mathcal{O}}\left(n\log\left(\epsilon^{-1}\right)\right)$ when $\epsilon \geq 5D_0/n$ or $n \geq 3L/4\tau$ and $\tilde{\mathcal{O}}\left(n\log(n) + n^{1/2}L^{1/2}\tau^{-1/2}\log\left(\epsilon^{-1}\right)\right)$ $= \tilde{\mathcal{O}}\left(n^{1/2}L^{1/2}\tau^{-1/2}\log\left(\epsilon^{-1}\right)\right)$ otherwise. Thus ARCS outperforms all existing algorithms in terms of gradient query complexity. But the linear oracle complexity of ARCS is higher than that of CGS and STORC, which is discussed after Corollary 6.

5. Experiments

In this section, we validate the effectiveness of our ARCS with experiments on different machine learning tasks. We conduct two experiments on ARCS and other compared algorithms listed in Table 1 with five real-world datasets. Specifically, the first experiment is the low-rank matrix completion task, and the second experiment addresses the sparsity-constrained logistic regression problem.

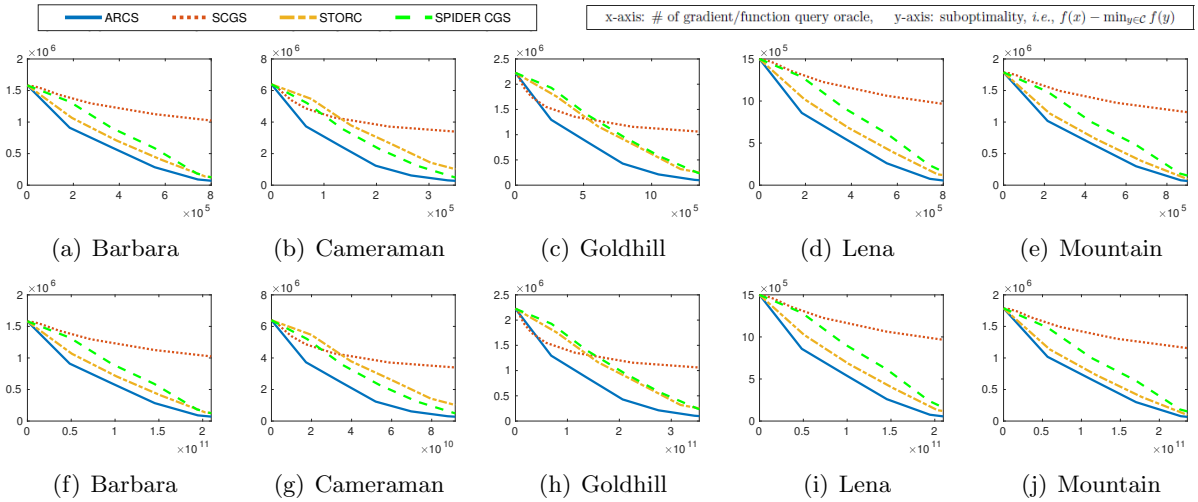
5.1 Low-Rank Matrix Completion Problem

In this experiment, we intend to recover a low rank matrix by solving the following matrix completion problem:

$$\min_{\|X\|_* \leq R} \sum_{(i,j)\in\Omega} (X_{i,j} - Y_{i,j})^2 \tag{11}$$

where $\|\cdot\|_*$ denotes the nuclear norm. $Y \in \mathbb{R}^{d_1 \times d_2}$ is a matrix whose elements were partly observed, and Ω denotes the set of subscripts of observed elements. Following Gu et al. (2019), we use the low-rank matrix completion problem to achieve image recovery such that Y in (11) is the matrix of an incomplete gray-scale image [1], and the solution X is a low rank matrix of the complete image we get. Specifically, we choose five images, which are Barbara (512×512 pixels), Cameraman (256×256 pixels), Goldhill (512×512 pixels), Lena (512×512 pixels) and Mountain (640×480 pixels). To get incomplete images, we eliminate 30% of the pixels in each of them. Note that for the matrix completion problem (11) the zeroth-order coordinate-wise gradient estimator (3) happens to be the true gradient, and the number of function query to construct a coordinate estimator of gradient is $2d$ times of the number of gradient query to construct a true gradient. Thus the figures of results for *zeroth-order* case are exactly the same as that for the *first-order* case, except that the x-axis is slightly different. The parameters are set according to Theorem 7 since the problem is quadratic. For the three variance-reduced algorithms, *i.e.*, ARCS, STORC and SPIDER CGS, we use a mini batch of 256 and for SCGS, we set the mini batch according to (Lan and Zhou, 2016, Algo. 4) since a mini batch of 256 leads to poor performance of SCGS. The results are shown in Figure 1, where

1. The gray-scale images can be found at `https://homepages.cae.wisc.edu/~ece533/images/`



(a) Barbara (b) Cameraman (c) Goldhill (d) Lena (e) Mountain

(f) Barbara (g) Cameraman (h) Goldhill (i) Lena (j) Mountain

Figure 1: Low-rank matrix completion problem. (a)-(e) are results for the *first-order* case and (f)-(j) are results for the *zeroth-order* case. The x-axis represents number of gradient query oracle for (a) - (e) and number of function query oracle for (f) - (j); the y-axis represents suboptimality, i.e., $f(x) - \min_{y\in\mathcal{C}} f(y)$. The curves of the first-order case and the zeroth-order case look the same since the coordinate-wise gradient estimator equals the true gradient.

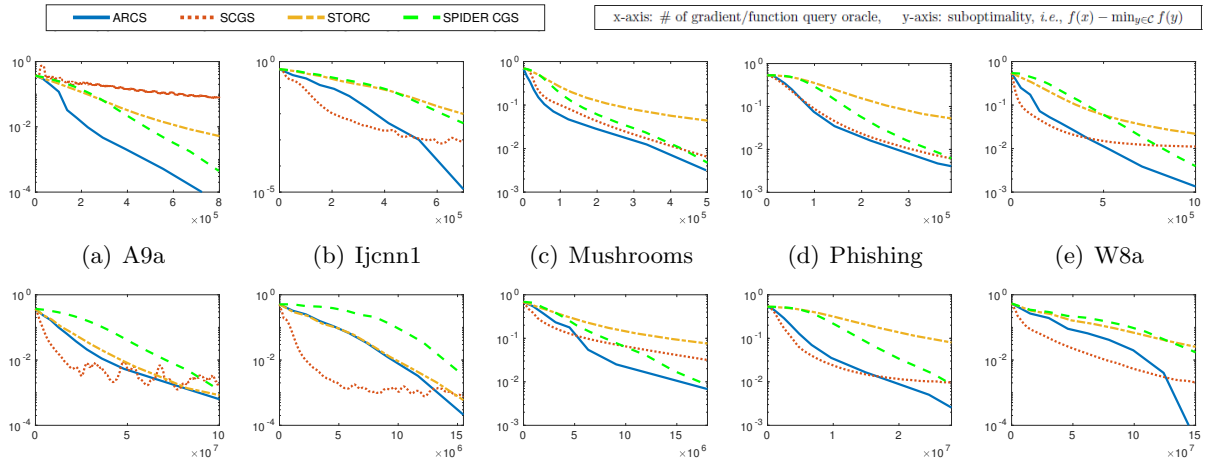
(a)-(e) are results for the *first-order* case and (f)-(j) are results for the *zeroth-order* case. It can be seen that our ARCS outperform all other algorithms compared in terms of gradient/function query complexity.

5.2 Sparsity-Constrained Logistic Regression

In this experiment, we focus on the sparsity-constrained logistic regression:

$$\min_{\|x\|_1\leq r} \frac{1}{n}\sum_{i=1}^{n} -\left(y_i \log\sigma(-x^T a_i) + (1-y_i)\log\sigma(x^T a_i)\right)$$

where $\sigma(z) = 1/\left(1+\exp(-z)\right)$ denotes the sigmoid function, $a_i \in \mathbb{R}^d$ denotes the data and $y_i \in \{0,1\}$ denotes the corresponding label. We conduct the experiment on five LIBSVM (Chang and Lin, 2011) datasets: a9a ($n = 32,561, d = 123$), ijcnn1 ($n = 49,990, d = 22$), mushrooms ($n = 8,124, d = 112$), phishing ($n = 11,055, d = 68$) and w8a ($n = 49,749, d = 300$). We set the parameters according to Theorem 5. For all the four algorithms, we use a mini batch of 256. The results are shown in Figure 2, where (a)-(e) are results for the *first-order* case and (f)-(j) are results for the *zeroth-order* case. For some datasets, our ARCS is slower than SCGS at first but outperforms SCGS later. This corresponds to the gradient query complexity presented in Table 1. For the first-order case, the gradient query complexity of ARCS has a dependence on n and $\epsilon^{-1/2}$, while that of SCGS only has a dependence on ϵ^{-3}. At the beginning, ϵ is relatively big, ϵ^{-3} is relatively small and n is relatively big, thus the gradient query complexity of ARCS is higher than that of SCGS. When ϵ diminishes, the gradient query complexity of ARCS gradually becomes lower than that of SCGS.



(a) A9a (b) Ijcnn1 (c) Mushrooms (d) Phishing (e) W8a

(f) A9a (g) Ijcnn1 (h) Mushrooms (i) Phishing (j) W8a

Figure 2: Sparsity-constrained logistic regression. (a)-(e) are results for the *first-order* case and (f)-(j) are results for the *zeroth-order* case. The x-axis represents number of gradient query oracle for (a) - (e) and number of function query oracle for (f) - (j); the y-axis represents suboptimality, i.e., $f(x) - \min_{y\in\mathcal{C}} f(y)$.

6. Conclusion

In this paper, we proposed an Accelerated variance-Reduced Conditional gradient Sliding (ARCS) algorithm for solving constrained finite-sum problems, which combines the variance-reduction technique and a novel momentum with conditional gradient sliding algorithm. Then We give the convergence results of our ARCS under convex and strongly-convex setting. Our ARCS can be used in either first-order (where gradient query oracle is available) or zeroth-order (where function query oracle is available) optimization. In first-order optimization, it outperforms all existing conditional gradient type algorithms with respect to gradient query complexity. In zeroth-order optimization, it is the first conditional gradient sliding type algorithm for convex problems. Finally we conduct numerical experiments with real-world datasets to show the superiority of our ARCS.

References

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality and saddle-points. *arXiv preprint arXiv:1809.06474*, 2018.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. In *Advances in Neural Information Processing Systems*, pages 7202–7213, 2019.

Yuwen Chen, Antonio Orvieto, and Aurelien Lucchi. An accelerated dfo algorithm for finite-sum convex functions. *arXiv preprint arXiv:2007.03311*, 2020.

Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard E Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. *arXiv preprint arXiv:1804.02395*, 2018.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

Abraham D Flaxman, Adam Tauman Kalai, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Hongchang Gao and Heng Huang. Can stochastic zeroth-order frank-wolfe method converge faster for non-convex problems? In *Thirty-seventh International Conference on Machine Learning (ICML 2020)*, 2020.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Donald Goldfarb, Garud Iyengar, and Chaoxu Zhou. Linear convergence of stochastic frank wolfe variants. *arXiv preprint arXiv:1703.07269*, 2017.

Bin Gu, Zhouyuan Huo, Cheng Deng, and Heng Huang. Faster derivative-free stochastic algorithm for shared memory machines. In *International Conference on Machine Learning*, pages 1812–1821, 2018.

Bin Gu, Wenhan Xian, and Heng Huang. Asynchronous stochastic frank-wolfe algorithms for nonconvex optimization. In *28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, 2019.

Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

Alfredo N Iusem. On the convergence properties of the projected gradient method for convex optimization. *Computational & Applied Mathematics*, 22(1):37–52, 2003.

Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International Conference on Machine Learning*, pages 3100–3109, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

Brian Kulis et al. Metric learning: A survey. *Foundations and trends in machine learning*, 5(4): 287–364, 2012.

Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

Guanghui Lan, Zhize Li, and Yi Zhou. A unified variance-reduced accelerated gradient method for convex optimization. In *Advances in Neural Information Processing Systems*, pages 10462–10472, 2019.

Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, and Ji Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. In *Advances in Neural Information Processing Systems*, pages 3054–3062, 2016.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3727–3737, 2018.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. Towards gradient free and projection free stochastic optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3468–3477, 2019.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

P Wolfe. Convergence theory in nonlinear programming. In *Integer and Nonlinear Programming*, pages 1–36. North-Holland publishing Company-Amsterdam-London, 1970.

Alp Yurtsever, Suvrit Sra, and Volkan Cevher. Conditional gradient methods via stochastic path-integrated differential estimator. In *International Conference on Machine Learning*, pages 7282–7291. PMLR, 2019.

Xinhua Zhang, Dale Schuurmans, and Yao-liang Yu. Accelerated training for matrix-norm regularization: A boosting approach. In *Advances in Neural Information Processing Systems*, pages 2906–2914, 2012.

Appendix A. Fundamental Lemmas

For simplicity, we denote

$$\begin{aligned} l_f(z,x) &:= f(z) + \langle \nabla f(z), x - z\rangle \\ \delta_t &:= G_t - \nabla f(\underline{x}_t) \\ x_{t-1}^+ &:= \frac{1}{1+\tau\gamma_s}(x_{t-1} + \tau\gamma_s \underline{x}_t) \end{aligned} \tag{12}$$

With the above notations, we have

$$\begin{aligned} \bar{x}_t - \underline{x}_t =& (1-\alpha_s-p_s)\bar{x}_{t-1} + \alpha_s x_t + p_s\tilde{x} - \underline{x}_t \\ \overset{①}{=}& \alpha_s x_t + \frac{1}{1+\tau\gamma_s}\left[(1+\tau\gamma_s(1-\alpha_s))\,\underline{x}_t - \alpha_s x_{t-1}\right] - \underline{x}_t = \alpha_s(x_t - x_{t-1}^+) \end{aligned} \tag{13}$$

where ① comes from the definition of $\underline{x}_t$ in Algorithm 2.

Lemma 9 *For any $x \in \mathcal{C}$, we have*

$$\begin{aligned} &\gamma_s\left[l_f(\underline{x}_t, x_t) - l_f(\underline{x}_t, x)\right] \\ \le& \frac{\gamma_s\tau}{2}\|x - x_t\|^2 + \frac{1}{2}\|x - x_{t-1}\|^2 - \frac{1+\gamma_s\tau}{2}\|x - x_t\|^2 - \frac{1+\gamma_s\tau}{2}\|x_t - x_{t-1}^+\|^2 - \gamma_s\langle \delta_t, x_t - x\rangle + \eta_{s,t} \end{aligned} \tag{14}$$

Proof From Algorithm 2, we have $\langle \nabla h(x_t), x_t - x\rangle \le \eta_{s,t}$ for any $x \in \mathcal{C}$. Observe that h is $(1+\tau\gamma_s)$-strongly convex, then we have

$$h(x_t) - h(x) + \frac{1+\tau\gamma_s}{2}\|x - x_t\|^2 \le \langle \nabla h(x_t), x_t - x\rangle \le \eta_{s,t},\ \forall\, x \in C \tag{15}$$

which is

$$\begin{aligned} &\gamma_s\left[\langle G_t, x_t\rangle + \frac{\tau}{2}\|x_t - \underline{x}_t\|^2\right] + \frac{1}{2}\|x_t - x_{t-1}\|^2 \\ \le& \gamma_s\left[\langle G_t, x\rangle + \frac{\tau}{2}\|x - \underline{x}_t\|^2\right] + \frac{1}{2}\|x - x_{t-1}\|^2 - \frac{1+\tau\gamma_s}{2}\|x - x_t\|^2 + \eta_{s,t},\ \forall\, x \in C \end{aligned} \tag{16}$$

Rearranging the terms, we get

$$\begin{aligned} &\gamma_s\left[\langle G_t, x_t - x\rangle + \frac{\tau}{2}\|x_t - \underline{x}_t\|^2\right] + \frac{1}{2}\|x_t - x_{t-1}\|^2 \\ \le& \frac{\gamma_s\tau}{2}\|x - x_t\|^2 + \frac{1}{2}\|x - x_{t-1}\|^2 - \frac{1+\tau\gamma_s}{2}\|x - x_t\|^2 + \eta_{s,t} \end{aligned} \tag{17}$$

With the notation of $l_f(\cdot,\cdot)$, we have

$$\langle G_t, x_t - x\rangle = \langle \nabla f(\underline{x}_t), x_t - x\rangle + \langle \delta_t, x_t - x\rangle = l_f(\underline{x}_t, x_t) - l_f(\underline{x}_t, x) + \langle \delta_t, x_t - x\rangle \tag{18}$$

Also we have

$$\begin{aligned} &\frac{\gamma_s\tau}{2}\|x_t - \underline{x}_t\|^2 + \frac{1}{2}\|x_t - x_{t-1}\|^2 \\ =& \frac{\gamma_s\tau}{2}\|x_t\|^2 - \langle x_t, \gamma_s\tau\underline{x}_t\rangle + \frac{\gamma_s\tau}{2}\|\underline{x}_t\|^2 + \frac{1}{2}\|x_t\|^2 - \langle x_t, x_{t-1}\rangle + \frac{1}{2}\|x_{t-1}\|^2 \\ =& \frac{1+\gamma_s\tau}{2}\|x_t\|^2 - (1+\gamma_s\tau)\langle x_t, \frac{1}{1+\gamma_s\tau}(\gamma_s\tau\underline{x}_t + x_{t-1})\rangle + \underbrace{\frac{\gamma_s\tau}{2}\|\underline{x}_t\|^2 + \frac{1}{2}\|x_{t-1}\|^2}_{Q_1} \end{aligned} \tag{19}$$

For the term Q_1, we have

$$\begin{aligned}2(1+\gamma_s\tau)Q_1 =&(1+\gamma_s\tau)\left(\gamma_s\tau\|\underline{x}_t\|^2+\|x_{t-1}\|^2\right)\\ =&\gamma_s\tau\|\underline{x}_t\|^2+\gamma_s\tau\|x_{t-1}\|^2+\left(\gamma_s^2\tau^2\|\underline{x}_t\|^2+\|x_{t-1}\|^2\right)\\ =&\gamma_s\tau\|\underline{x}_t\|^2+\gamma_s\tau\|x_{t-1}\|^2-2\gamma_s\tau\langle\underline{x}_t,x_{t-1}\rangle+\|\gamma_s\tau\underline{x}_t+x_{t-1}\|^2\end{aligned}\tag{20}$$

With the notation of x_{t-1}^+, we have

$$2(1+\gamma_s\tau)Q_1\geq\|\gamma_s\tau\underline{x}_t+x_{t-1}\|^2=(1+\gamma_s\tau)^2\|x_{t-1}^+\|^2\tag{21}$$

which is

$$Q_1\geq\frac{1+\gamma_s\tau}{2}\|x_{t-1}^+\|^2\tag{22}$$

Plugging (22) into (19), we get

$$\begin{aligned}&\frac{\gamma_s\tau}{2}\|x_t-\underline{x}_t\|^2+\frac{1}{2}\|x_t-x_{t-1}\|^2\\ \geq&\frac{1+\gamma_s\tau}{2}\|x_t\|^2-(1+\gamma_s\tau)\langle x_t,\frac{1}{1+\gamma_s\tau}(\gamma_s\tau\underline{x}_t+x_{t-1})\rangle+\frac{1+\gamma_s\tau}{2}\|x_{t-1}^+\|^2=\frac{1+\gamma_s\tau}{2}\|x_t-x_{t-1}^+\|^2\end{aligned}\tag{23}$$

Plugging (18) and (23) into (17) and rearranging the terms, we get

$$\begin{aligned}&\gamma_s\left[l_f(\underline{x}_t,x_t)-l_f(\underline{x}_t,x)\right]\\ \leq&\frac{\gamma_s\tau}{2}\|x-x_t\|^2+\frac{1}{2}\|x-x_{t-1}\|^2-\frac{1+\tau\gamma_s}{2}\|x-x_t\|^2+\eta_{s,t}-\frac{1+\gamma_s\tau}{2}\|x_t-x_{t-1}^+\|^2-\gamma_s\langle\delta_t,x_t-x\rangle\end{aligned}\tag{24}$$

Then we complete the proof. ■

Lemma 10 *Suppose f is τ-strongly ($\tau\geq 0$) convex and each $f_{i\in[n]}$ is L-smooth. Conditioning on $x_1,...,x_{t-1}$*

- *For the first-order case, assume that $\alpha_s\in[0,1],p_s\in[0,1]$ and $\gamma_s>0$ satisfy*

$$1+\tau\gamma_s-L\alpha_s\gamma_s>0,\ 1-\alpha_s-p_s\geq 0,\qquad p_s-\frac{L\alpha_s\gamma_s}{1+\tau\gamma_s-L\alpha_s\gamma_s}>0$$

Then we have

$$\begin{aligned}&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t)-f(x)\right]+\frac{1+\tau\gamma_s}{2}\mathbb{E}\left[\|x-x_t\|^2\right]\\ \leq&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1})-f(x)\right]+\frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x})-f(x)\right]+\frac{1}{2}\|x-x_{t-1}\|^2+\eta_{s,t}\end{aligned}$$

- *For the zeroth-order case, assume that $\alpha_s\in[0,1],p_s\in[0,1]$ and $\gamma_s>0$ satisfy*

$$1+\tau\gamma_s-L\alpha_s\gamma_s>0,\ 1-\alpha_s-p_s\geq 0,\qquad p_s-\frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s-L\alpha_s\gamma_s}>0$$

Then we have

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t)-f(x)\right]+\frac{1+\tau\gamma_s}{2}\mathbb{E}\left[\|x-x_t\|^2\right]\\
\leq&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1})-f(x)\right]+\frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x})-f(x)\right]+\frac{1}{2}\|x-x_{t-1}\|^2+\eta_{s,t}\\
&-\gamma_s\langle\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t),x_t-x\rangle+\frac{6\gamma_s^2\mu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}
\end{aligned}$$

Proof Since f is τ-strongly convex and L-smooth, then we have

$$\begin{aligned}
f(\bar{x}_t)\leq& l_f(\underline{x}_t,\bar{x}_t)+\frac{L}{2}\|\bar{x}_t-\underline{x}_t\|^2\\
\overset{①}{=}&(1-\alpha_s-p_s)l_f(\underline{x}_t,\bar{x}_{t-1})+\alpha_s l_f(\underline{x}_t,x_t)+p_s l_f(\underline{x}_t,\tilde{x})+\frac{L\alpha_s^2}{2}\|x_t-x_{t-1}^+\|^2
\end{aligned}\tag{25}$$

where ① comes from the definition of $\bar{x}_t$ in Algorithm 2. Plugging Lemma 9 into (25), we get

$$\begin{aligned}
f(\bar{x}_t)\leq&(1-\alpha_s-p_s)l_f(\underline{x}_t,\bar{x}_{t-1})\\
&+\alpha_s\left[l_f(\underline{x}_t,x)+\frac{\tau}{2}\|x-\underline{x}_t\|^2+\frac{1}{2\gamma_s}\|x-x_{t-1}\|^2-\frac{1+\tau\gamma_s}{2\gamma_s}\|x-x_t\|^2+\frac{\eta_{s,t}}{\gamma_s}\right]\\
&+p_s l_f(\underline{x}_t,\tilde{x})-\frac{\alpha_s}{2\gamma_s}\left(1+\tau\gamma_s-L\alpha_s\gamma_s\right)\|x_t-x_{t-1}^+\|^2-\alpha_s\langle\delta_t,x_t-x\rangle\\
\overset{①}{\leq}&(1-\alpha_s-p_s)f(\bar{x}_{t-1})+\alpha_s\left[f(x)+\frac{1}{2\gamma_s}\|x-x_{t-1}\|^2-\frac{1+\tau\gamma_s}{2\gamma_s}\|x-x_t\|^2+\frac{\eta_{s,t}}{\gamma_s}\right]\\
&+p_s l_f(\underline{x}_t,\tilde{x})-\frac{\alpha_s}{2\gamma_s}\left(1+\tau\gamma_s-L\alpha_s\gamma_s\right)\|x_t-x_{t-1}^+\|^2-\alpha_s\langle\delta_t,x_t-x_{t-1}^+\rangle-\alpha_s\langle\delta_t,x_{t-1}^+-x\rangle
\end{aligned}\tag{26}$$

where ① comes from the fact that f is τ-strongly convex. Now we give proof to the first-order and zeroth-order case respectively.

First-order Case: Using the fact that $b\langle u,v\rangle-a\|v\|^2/2\leq b^2\|u\|^2/(2a)$, we have from (26)

$$\begin{aligned}
f(\bar{x}_t)\leq&(1-\alpha_s-p_s)f(\bar{x}_{t-1})+\alpha_s\left[f(x)+\frac{1}{2\gamma_s}\|x-x_{t-1}\|^2-\frac{1+\tau\gamma_s}{2\gamma_s}\|x-x_t\|^2+\frac{\eta_{s,t}}{\gamma_s}\right]\\
&\underbrace{+p_s l_f(\underline{x}_t,\tilde{x})+\frac{\alpha_s\gamma_s\|\delta_t\|^2}{2(1+\tau\gamma_s-L\alpha_s\gamma_s)}-\alpha_s\langle\delta_t,x_{t-1}^+-x\rangle}_{Q_2}
\end{aligned}\tag{27}$$

Using Lemma 18, we can bound Q_2 as

$$\begin{aligned}
&\mathbb{E}\left[p_s l_f(\underline{x}_t,\tilde{x})+\frac{\alpha_s\gamma_s\|\delta_t\|^2}{2(1+\tau\gamma_s-L\alpha_s\gamma_s)}-\alpha_s\langle\delta_t,x_{t-1}^+-x\rangle\right]=\mathbb{E}\left[p_s l_f(\underline{x}_t,\tilde{x})+\frac{\alpha_s\gamma_s\|\delta_t\|^2}{2(1+\tau\gamma_s-L\alpha_s\gamma_s)}\right]\\
\overset{①}{\leq}&p_s l_f(\underline{x}_t,\tilde{x})+\frac{L\alpha_s\gamma_s}{1+\tau\gamma_s-L\alpha_s\gamma_s}\left(f(\tilde{x})-l_f(\underline{x}_t,\tilde{x})\right)\\
=&\left(p_s-\frac{L\alpha_s\gamma_s}{2(1+\tau\gamma_s-L\alpha_s\gamma_s)}\right)l_f(\underline{x}_t,\tilde{x})+\frac{L\alpha_s\gamma_s}{1+\tau\gamma_s-L\alpha_s\gamma_s}f(\tilde{x})\overset{②}{\leq}p_s f(\tilde{x})
\end{aligned}\tag{28}$$

where ① comes from Lemma 18 and ② comes from the assumption that $p_s - \frac{L\alpha_s\gamma_s}{1+\tau\gamma_s-L\alpha_s\gamma_s} > p_s - \frac{2L\alpha_s\gamma_s}{1+\tau\gamma_s-L\alpha_s\gamma_s} > 0$ and the convexity of f. Plugging (28) into (27), we get

$$\begin{aligned}&\mathbb{E}\left[f(\bar{x}_t)+\frac{\alpha_s(1+\tau\gamma_s)}{2\gamma_s}\|x-x_t\|^2\right]\\ \leq&(1-\alpha_s-p_s)f(\bar{x}_{t-1})+\alpha_s f(x)+p_s f(\tilde{x})+\frac{\alpha_s}{2\gamma_s}\|x-x_{t-1}\|^2+\frac{\alpha_s}{\gamma_s}\eta_{s,t}\end{aligned} \tag{29}$$

Subtracting both sides with $f(x)$ and then multiplying both sides with $\frac{\gamma_s}{\alpha_s}$, we get

$$\begin{aligned}&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t)-f(x)\right]+\frac{1+\tau\gamma_s}{2}\mathbb{E}\left[\|x-x_t\|^2\right]\\ \leq&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1})-f(x)\right]+\frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x})-f(x)\right]+\frac{1}{2}\|x-x_{t-1}\|^2+\eta_{s,t}\end{aligned} \tag{30}$$

Then we get the desired result for the first-order case. Next we give proof to the zeroth-order case.

Zeroth-order Case: We can rewrite (26) as

$$\begin{aligned}f(\bar{x}_t)\leq&(1-\alpha_s-p_s)f(\bar{x}_{t-1})+\alpha_s\left[f(x)+\frac{1}{2\gamma_s}\|x-x_{t-1}\|^2-\frac{1+\tau\gamma_s}{2\gamma_s}\|x-x_t\|^2+\frac{\eta_{s,t}}{\gamma_s}\right]\\ &+p_s l_f(\underline{x}_t,\tilde{x})-\frac{\alpha_s}{2\gamma_s}\left(1+\tau\gamma_s-L\alpha_s\gamma_s\right)\|x_t-x_{t-1}^+\|^2-\alpha_s\langle\delta_t-\mathbb{E}\left[\delta_t\right],x_t-x_{t-1}^+\rangle\\ &-\alpha_s\langle\mathbb{E}\left[\delta_t\right],x_t-x_{t-1}^+\rangle-\alpha_s\langle\delta_t,x_{t-1}^+-x\rangle\\ \overset{①}{\leq}&(1-\alpha_s-p_s)f(\bar{x}_{t-1})+\alpha_s\left[f(x)+\frac{1}{2\gamma_s}\|x-x_{t-1}\|^2-\frac{1+\tau\gamma_s}{2\gamma_s}\|x-x_t\|^2+\frac{\eta_{s,t}}{\gamma_s}\right]\\ &+p_s l_f(\underline{x}_t,\tilde{x})+\frac{\alpha_s\gamma_s\|\delta_t-\mathbb{E}\left[\delta_t\right]\|^2}{2(1+\tau\gamma_s-L\alpha_s\gamma_s)}-\alpha_s\langle\mathbb{E}\left[\delta_t\right],x_t-x_{t-1}^+\rangle-\alpha_s\langle\delta_t,x_{t-1}^+-x\rangle\end{aligned} \tag{31}$$

where ① comes from the fact that $b\langle u,v\rangle-a\|v\|^2/2\leq b^2\|u\|^2/(2a)$. Using Lemma 18, we have

$$\begin{aligned}&\mathbb{E}\left[p_s l_f(\underline{x}_t,\tilde{x})+\frac{\alpha_s\gamma_s\|\delta_t-\mathbb{E}\left[\delta_t\right]\|^2}{2(1+\tau\gamma_s-L\alpha_s\gamma_s)}\right]\\ \overset{①}{\leq}&p_s l_f(\underline{x}_t,\tilde{x})+\frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s-L\alpha_s\gamma_s}\left(f(\tilde{x})-l_f(\underline{x}_t,\tilde{x})\right)+\frac{6\alpha_s\gamma_s\mu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}\\ =&\left(p_s-\frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s-L\alpha_s\gamma_s}\right)l_f(\underline{x}_t,\tilde{x})+\frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s-L\alpha_s\gamma_s}f(\tilde{x})+\frac{6\alpha_s\gamma_s\mu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}\\ \overset{②}{\leq}&p_s f(\tilde{x})+\frac{6\alpha_s\gamma_s\mu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}\end{aligned} \tag{32}$$

where ① comes from Lemma 18 and ② comes from the assumption that $p_s-\frac{4\alpha_s\gamma_s dL}{1+\tau\gamma_s-L\alpha_s\gamma_s}>0$ and the convexity of f. Also we have

$$\mathbb{E}\left[-\alpha_s\langle\mathbb{E}\left[\delta_t\right],x_t-x_{t-1}^+\rangle-\alpha_s\langle\delta_t,x_{t-1}^+-x\rangle\right]=-\alpha_s\langle\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t),x_t-x\rangle \tag{33}$$

which comes from Lemma 18. Plugging (32), (33) into (31), we get

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{x}_t)+\frac{\alpha_s(1+\tau\gamma_s)}{2\gamma_s}\|x-x_t\|^2\right]\\
\leq&(1-\alpha_s-p_s)f(\bar{x}_{t-1})+\alpha_s f(x)+p_s f(\tilde{x})+\frac{\alpha_s}{2\gamma_s}\|x-x_{t-1}\|^2+\frac{\alpha_s}{\gamma_s}\eta_{s,t} \qquad (34)\\
&-\alpha_s\langle\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t),x_t-x\rangle+\frac{6\alpha_s\gamma_s\mu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}
\end{aligned}$$

Subtracting both sides with $f(x)$ and then multiplying both sides with $\frac{\gamma_s}{\alpha_s}$, we get

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t)-f(x)\right]+\frac{1+\tau\gamma_s}{2}\mathbb{E}\left[\|x-x_t\|^2\right]\\
\leq&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1})-f(x)\right]+\frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x})-f(x)\right]+\frac{1}{2}\|x-x_{t-1}\|^2+\eta_{s,t} \qquad (35)\\
&-\gamma_s\langle\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t),x_t-x\rangle+\frac{6\gamma_s^2\mu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}
\end{aligned}$$

Then we get the desired result for the zeroth-order case. Then we complete the proof. ■

Appendix B. Proof of Theorem 5

Lemma 11 *Suppose Assumption 2 holds. Denote* $\mathcal{L}_s=\frac{\gamma_s}{\alpha_s}+(T_s-1)\frac{\gamma_s(\alpha_s+p_s)}{\alpha_s}$, $\mathcal{R}_s=\frac{\gamma_s}{\alpha_s}(1-\alpha_s)+(T_s-1)\frac{\gamma_s p_s}{\alpha_s}$. *Set* $\theta_t=\begin{cases}\frac{\gamma_s}{\alpha_s}(\alpha_s+p_s), & t\leq T_s-1\\ \frac{\gamma_s}{\alpha_s}, & t=T_s\end{cases}$

- *For the first-order case, assume that* $\alpha_s\in[0,1],p_s\in[0,1]$ *and* $\gamma_s>0$ *satisfy*

$$1+\tau\gamma_s-L\alpha_s\gamma_s>0,\ 1-\alpha_s-p_s\geq 0,\qquad p_s-\frac{L\alpha_s\gamma_s}{1+\tau\gamma_s-L\alpha_s\gamma_s}>0$$

Then we have

$$\mathcal{L}_s\mathbb{E}\left[f(\tilde{x}^s)-f(x)\right]\leq\mathcal{R}_s\mathbb{E}\left[f(\tilde{x}^{s-1})-f(x)\right]+\mathbb{E}\left[\frac{1}{2}\|x-x^{s-1}\|^2-\frac{1}{2}\|x-x^s\|^2\right]+\sum_{t=1}^{T_s}\eta_{s,t}$$

- *For the zeroth-order case, assume that* $\alpha_s\in[0,1],p_s\in[0,1]$ *and* $\gamma_s>0$ *satisfy*

$$1+\tau\gamma_s-L\alpha_s\gamma_s>0,\ 1-\alpha_s-p_s\geq 0,\qquad p_s-\frac{4\alpha_s\gamma_s L}{1+\tau\gamma_s-L\alpha_s\gamma_s}>0$$

Then we have

$$\begin{aligned}
\mathcal{L}_s\mathbb{E}\left[f(\tilde{x}^s)-f(x)\right]\leq&\mathcal{R}_s\mathbb{E}\left[f(\tilde{x}^{s-1})-f(x)\right]+\mathbb{E}\left[\frac{1}{2}\|x-x^{s-1}\|^2-\frac{1}{2}\|x-x^s\|^2\right]+\sum_{t=1}^{T_s}\eta_{s,t}\\
&+\gamma_s T_s D\mu L\sqrt{d}+T_s\frac{6\gamma_s^2\mu^2L^2d}{1-L\alpha_s\gamma_s}
\end{aligned}$$

Proof Define

$$\Delta_t = \begin{cases} 0, & \text{for the first-order case} \\ -\gamma_s\langle \hat{\nabla}_{coord} f(\underline{x}_t) - \nabla f(\underline{x}_t), x_t - x\rangle + \frac{6\gamma_s^2\mu^2 L^2 d}{1+\tau\gamma_s - L\alpha_s\gamma_s}, & \text{for the zeroth-order case} \end{cases}$$

From Lemma 10, we have

$$\begin{aligned} \frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] \leq & \frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\mathbb{E}\left[f(\bar{x}_{t-1}) - f(x)\right] + \frac{\gamma_s p_s}{\alpha_s}\mathbb{E}\left[f(\tilde{x}) - f(x)\right] \\ & + \mathbb{E}\left[\frac{1}{2}\|x - x_{t-1}\|^2 - \frac{1}{2}\|x - x_t\|^2\right] + \eta_{s,t} + \Delta_t \end{aligned} \tag{36}$$

Summing the above inequality over $t = 1, ..., T_s$, with the definition of θ_t, we have

$$\begin{aligned} & \sum_{t=1}^{T_s}\theta_t\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] \\ & \leq \left[\frac{\gamma_s}{\alpha_s}(1-\alpha_s) + (T_s-1)\frac{\gamma_s p_s}{\alpha_s}\right]\mathbb{E}\left[f(\tilde{x}) - f(x)\right] + \mathbb{E}\left[\frac{1}{2}\|x - x_0\|^2 - \frac{1}{2}\|x - x_{T_s}\|^2\right] + \sum_{t=1}^{T_s}\eta_{s,t} + \sum_{t=1}^{T_s}\Delta_t \end{aligned} \tag{37}$$

Note that $\tilde{x}^s = \sum_{t=1}^{T}(\theta_t\bar{x}_t)/\sum_{t=1}^{T}\theta_t, x^s = x_{T_s}, x^{s-1} = x_0, \tilde{x}^{s-1} = \tilde{x}$, the convexity of f and Jensen's inequality, we have

$$\begin{aligned} & \sum_{t=1}^{T_s}\theta_t\mathbb{E}\left[f(\tilde{x}^s) - f(x)\right] \\ & \leq \left[\frac{\gamma_s}{\alpha_s}(1-\alpha_s) + (T_s-1)\frac{\gamma_s p_s}{\alpha_s}\right]\mathbb{E}\left[f(\tilde{x}^{s-1}) - f(x)\right] + \mathbb{E}\left[\frac{1}{2}\|x - x^{s-1}\|^2 - \frac{1}{2}\|x - x^s\|^2\right] \\ & \quad + \sum_{t=1}^{T_s}\eta_{s,t} + \sum_{t=1}^{T_s}\Delta_t \end{aligned} \tag{38}$$

Denote $\mathcal{L}_s = \frac{\gamma_s}{\alpha_s} + (T_s-1)\frac{\gamma_s(\alpha_s+p_s)}{\alpha_s}, \mathcal{R}_s = \frac{\gamma_s}{\alpha_s}(1-\alpha_s) + (T_s-1)\frac{\gamma_s p_s}{\alpha_s}$, we have

$$\mathcal{L}_s\mathbb{E}\left[f(\tilde{x}^s) - f(x)\right] \leq \mathcal{R}_s\mathbb{E}\left[f(\tilde{x}^{s-1}) - f(x)\right] + \mathbb{E}\left[\frac{1}{2}\|x - x^{s-1}\|^2 - \frac{1}{2}\|x - x^s\|^2\right] + \sum_{t=1}^{T_s}\eta_{s,t} + \sum_{t=1}^{T_s}\Delta_t \tag{39}$$

First-order Case: Using the definition of Δ_t and (39), we have

$$\mathcal{L}_s\mathbb{E}\left[f(\tilde{x}^s) - f(x)\right] \leq \mathcal{R}_s\mathbb{E}\left[f(\tilde{x}^{s-1}) - f(x)\right] + \mathbb{E}\left[\frac{1}{2}\|x - x^{s-1}\|^2 - \frac{1}{2}\|x - x^s\|^2\right] + \sum_{t=1}^{T_s}\eta_{s,t} \tag{40}$$

Then we get the desired result for first-order case.

Zeroth-order Case: Using the definition of Δ_t and (39), we have

$$\begin{aligned}\mathcal{L}_s\mathbb{E}\left[f(\tilde{x}^s)-f(x)\right] \leq & \mathcal{R}_s\mathbb{E}\left[f(\tilde{x}^{s-1})-f(x)\right]+\mathbb{E}\left[\frac{1}{2}\|x-x^{s-1}\|^2-\frac{1}{2}\|x-x^s\|^2\right]+\sum_{t=1}^{T_s}\eta_{s,t} \\ & \underbrace{-\gamma_s\sum_{t=1}^{T_s}\mathbb{E}\left[\langle\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t),x_t-x\rangle\right]}_{Q_3}+T_s\frac{6\gamma_s^2\mu^2L^2d}{1-L\alpha_s\gamma_s}\end{aligned} \tag{41}$$

Next we have

$$\begin{aligned}Q_3 = & -\gamma_s\sum_{t=1}^{T_s}\mathbb{E}\left[\langle\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t),\ x_t-x\rangle\right] \\ \overset{①}{\leq} & \gamma_s\sum_{t=1}^{T_s}\mathbb{E}\left[\|\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t)\|\cdot\|x_t-x\|\right]\overset{②}{\leq}\gamma_sT_sD\mu L\sqrt{d}\end{aligned} \tag{42}$$

where ① comes from Cauchy-Schwartz inequality and ② comes from Lemma 16 and Assumption 2. Plugging (42) into (41), we get

$$\begin{aligned}\mathcal{L}_s\mathbb{E}\left[f(\tilde{x}^s)-f(x)\right] \leq & \mathcal{R}_s\mathbb{E}\left[f(\tilde{x}^{s-1})-f(x)\right]+\mathbb{E}\left[\frac{1}{2}\|x-x^{s-1}\|^2-\frac{1}{2}\|x-x^s\|^2\right]+\sum_{t=1}^{T_s}\eta_{s,t} \\ & +\gamma_sT_sD\mu L\sqrt{d}+T_s\frac{6\gamma_s^2\mu^2L^2d}{1-L\alpha_s\gamma_s}\end{aligned} \tag{43}$$

Then we get the desired result for zeroth-order case. Then we complete the proof. ■

Proof [of Theorem 5] We give proof to the first-order case and zeroth-order case respectively.

First-order Case: Lemma 11 implies

$$\mathcal{L}_s\mathbb{E}\left[f(\tilde{x}^s)-f(x)\right]\leq\mathcal{R}_s\mathbb{E}\left[f(\tilde{x}^{s-1})-f(x)\right]+\mathbb{E}\left[\frac{1}{2}\|x-x^{s-1}\|^2-\frac{1}{2}\|x-x^s\|^2\right]+\sum_{t=1}^{T_s}\eta_{s,t} \tag{44}$$

Summing the above inequality over $s=1,...,S$, and set $x=x^*=\arg\min_{x\in\mathcal{C}}f(x)$, we have

$$\begin{aligned}& \mathcal{L}_S\mathbb{E}\left[f(\tilde{x}^S)-f(x^*)\right]+\sum_{s=1}^{S-1}\left(\mathcal{L}_s-\mathcal{R}_{s+1}\right)\mathbb{E}\left[f(\tilde{x})-f(x^*)\right] \\ \leq & \mathcal{R}_1\mathbb{E}\left[f(\tilde{x}^0)-f(x^*)\right]+\mathbb{E}\left[\frac{1}{2}\|x^*-x^0\|^2-\frac{1}{2}\|x^*-x^S\|^2\right]+\sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t}\end{aligned} \tag{45}$$

Next we prove $\mathcal{L}_s-\mathcal{R}_{s+1}\leq 0$ for all $s=1,...,S-1$. When $s<s_0$, we have $\alpha_{s+1}=\alpha_s=\frac{1}{2},\gamma_{s+1}=\gamma_s,p_{s+1}=p_s=\frac{1}{2},T_{s+1}=2T_s$, thus

$$\begin{aligned}\mathcal{L}_s-\mathcal{R}_{s+1} = & \frac{\gamma_s}{\alpha_s}+(T_s-1)\frac{\gamma_s(\alpha_s+p_s)}{\alpha_s}-\left[\frac{\gamma_{s+1}}{\alpha_{s+1}}(1-\alpha_{s+1})+(T_{s+1}-1)\frac{\gamma_{s+1}p_{s+1}}{\alpha_{s+1}}\right] \\ = & \frac{\gamma_s}{\alpha_s}\left[1+(T_s-1)(\alpha_s+p_s)-(1-\alpha_s)-(2T_s-1)p_s\right]=\frac{\gamma_s}{\alpha_s}\left[T_s(\alpha_s-p_s)\right]=0\end{aligned} \tag{46}$$

When $s \geq s_0$, we have $\alpha_s = \frac{2}{s-s_0+4}, \gamma_s = \frac{1}{3L\alpha_s}, p_{s+1} = p_s = \frac{1}{2}, T_{s+1} = T_s$, thus

$$\begin{aligned}\mathcal{L}_s - \mathcal{R}_{s+1} =& \frac{\gamma_s}{\alpha_s} + (T_s - 1)\frac{\gamma_s(\alpha_s + p_s)}{\alpha_s} - \left[\frac{\gamma_{s+1}}{\alpha_{s+1}}(1-\alpha_{s+1}) + (T_{s+1}-1)\frac{\gamma_{s+1}p_{s+1}}{\alpha_{s+1}}\right] \\ =& \frac{\gamma_s}{\alpha_s} - \frac{\gamma_{s+1}}{\alpha_{s+1}}(1-\alpha_{s+1}) + (T_{s_0}-1)\left[\frac{\gamma_s(\alpha_s+p_s)}{\alpha_s} - \frac{\gamma_{s+1}p_{s+1}}{\alpha_{s+1}}\right] \\ =& \frac{1}{12L} + \frac{(T_{s_0}-1)(2(s-s_0+4)-1)}{24L} \geq 0\end{aligned} \tag{47}$$

Thus $\mathcal{L}_s - \mathcal{R}_{s+1} \geq 0$ for $s = 1, ..., S-1$. Note that $\mathcal{R}_1 = \frac{2}{3L}$. Plugging this inequality into (45), we get

$$\begin{aligned}\mathcal{L}_S \mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq& \mathcal{L}_S \mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] + \sum_{s=1}^{S-1}\left(\mathcal{L}_s - \mathcal{R}_{s+1}\right)\mathbb{E}\left[f(\tilde{x}) - f(x^*)\right] \\ \leq& \mathcal{R}_1 \mathbb{E}\left[f(\tilde{x}^0) - f(x^*)\right] + \mathbb{E}\left[\frac{1}{2}\|x^* - x^0\|^2 - \frac{1}{2}\|x^* - x^S\|^2\right] + \sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} \\ \leq& \frac{2}{3L}\left[f(\tilde{x}^0) - f(x^*)\right] + \frac{1}{2}\|x^* - x^0\|^2 + \sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} \overset{①}{\leq} \frac{D_0}{6L} + \sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t}\end{aligned} \tag{48}$$

where ① comes from the definition of D_0 that $D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 3L\|x^0 - x^*\|^2$.

- If $S \leq s_0$, then $\mathcal{L}_s = \frac{2^{S+1}}{3L}$, we have

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \frac{D_0}{2^{S+2}} + \frac{3L}{2^{S+1}}\sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} \tag{49}$$

  With the choice of $\eta_{s,t}$, we have

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} \overset{①}{\leq} \sum_{s=1}^{S}\frac{D_0}{sL} \leq \frac{D_0(\log S + 1)}{L} \tag{50}$$

  where ① comes from $\sum_{k=1}^{n}\frac{1}{k} \leq \log n + 1$. Plugging this inequality into (49) we get

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \frac{3D_0(\log S + 2)}{2^{S+1}} \tag{51}$$

- If $S > s_0$, we have

$$\begin{aligned}\mathcal{L}_S =& \frac{1}{3L\alpha_S^2}\left[1 + (T_S - 1)(\alpha_S + \frac{1}{2})\right] \\ =& \frac{(S-s_0+4)(T_{s_0}-1)}{6L} + \frac{(S-s_0+4)^2(T_{s_0}+1)}{24L} \overset{①}{\geq} \frac{(S-s_0+4)^2 n}{48L}\end{aligned} \tag{52}$$

  where ① comes from $T_{s_0} = 2^{\lfloor \log_2 n \rfloor + 1 - 1} \geq n/2$. Then we have

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \frac{8D_0}{n(S-s_0+4)^2} + \frac{48L}{n(S-s_0+4)^2}\sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} \tag{53}$$

With the choice of $\eta_{s,t}$, we have

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} \le \sum_{s=1}^{S}\frac{D_0}{sL} \le \frac{D_0(\log S+1)}{L} \tag{54}$$

Plugging this inequality into (53) we get

$$\mathbb{E}\left[f(\tilde{x}^S)-f(x^*)\right] \le \frac{48D_0(\log S+2)}{n(S-s_0+4)^2} \tag{55}$$

Then we get the desired result for the first-order case.

Zeroth-order Case: From Lemma 11, we get

$$\begin{aligned}\mathcal{L}_s\mathbb{E}\left[f(\tilde{x}^s)-f(x)\right] \le& \mathcal{R}_s\mathbb{E}\left[f(\tilde{x}^{s-1})-f(x)\right]+\mathbb{E}\left[\frac{1}{2}\|x-x^{s-1}\|^2-\frac{1}{2}\|x-x^s\|^2\right]+\sum_{t=1}^{T_s}\eta_{s,t}\\ &+\gamma_s T_s D\mu L\sqrt{d}+T_s\frac{6\gamma_s^2\mu^2L^2d}{1-L\alpha_s\gamma_s}\end{aligned} \tag{56}$$

Summing the above inequality over $s=1,...,S$ and set $x=x^*=\arg\min_{x\in\mathcal{C}} f(x)$, we have

$$\begin{aligned}&\mathcal{L}_S\mathbb{E}\left[f(\tilde{x}^S)-f(x^*)\right]+\sum_{s=1}^{S-1}\left(\mathcal{L}_s-\mathcal{R}_{s+1}\right)\mathbb{E}\left[f(\tilde{x})-f(x^*)\right]\\ \le&\mathcal{R}_1\mathbb{E}\left[f(\tilde{x}^0)-f(x^*)\right]+\mathbb{E}\left[\frac{1}{2}\|x^*-x^0\|^2-\frac{1}{2}\|x^*-x^S\|^2\right]+\sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t}\\ &+\sum_{s=1}^{S}\gamma_s T_s D\mu L\sqrt{d}+\sum_{s=1}^{S}T_s\frac{6\gamma_s^2\mu^2L^2d}{1-L\alpha_s\gamma_s}\end{aligned} \tag{57}$$

Next we prove $\mathcal{L}_s-\mathcal{R}_{s+1}\le 0$ for all $s=1,...,S-1$. When $s\le s_0$, we have

$$\mathcal{L}_s-\mathcal{R}_{s+1}=\frac{\gamma_s}{\alpha_s}\left[1+(T_s-1)(\alpha_s+p_s)-(1-\alpha_s)(2T_s-1)p_s\right]=\frac{\gamma_s}{\alpha_s}\left[T_s(\alpha_s-p_s)\right]=0 \tag{58}$$

When $s>s_0$, $\frac{\gamma_s}{\alpha_s}=\frac{1}{5L\alpha_s^2}=\frac{(s-s_0+4)^2}{20L}$, we have

$$\begin{aligned}\mathcal{L}_s-\mathcal{R}_{s+1}=&\frac{\gamma_s}{\alpha_s}+(T_s-1)\frac{\gamma_s(\alpha_s+p_s)}{\alpha_s}-\left[\frac{\gamma_{s+1}}{\alpha_{s+1}}(1-\alpha_{s+1})+(T_{s+1}-1)\frac{\gamma_{s+1}p_{s+1}}{\alpha_{s+1}}\right]\\ =&\frac{\gamma_s}{\alpha_s}-\frac{\gamma_{s+1}}{\alpha_{s+1}}(1-\alpha_{s+1})+(T_{s_0}-1)\left[\frac{\gamma_s(\alpha_s+p_s)}{\alpha_s}-\frac{\gamma_{s+1}p_{s+1}}{\alpha_{s+1}}\right]\\ =&\frac{1}{20L}+\frac{(T_{s_0}-1)(2(s-s_0+4)-1)}{40L}\ge 0\end{aligned} \tag{59}$$

Thus $\mathcal{L}_s - \mathcal{R}_{s+1} \geq 0$ for $s = 1, ..., S-1$. Plugging this inequality into (57), we get

$$\begin{aligned}
&\mathcal{L}_S \mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \mathcal{L}_S \mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] + \sum_{s=1}^{S-1} \left(\mathcal{L}_s - \mathcal{R}_{s+1}\right) \mathbb{E}\left[f(\tilde{x}) - f(x^*)\right] \\
\leq &\mathcal{R}_1 \mathbb{E}\left[f(\tilde{x}^0) - f(x^*)\right] + \mathbb{E}\left[\frac{1}{2}\|x^* - x^0\|^2 - \frac{1}{2}\|x^* - x^S\|^2\right] + \sum_{s=1}^{S}\sum_{t=1}^{T_s} \eta_{s,t} + \sum_{s=1}^{S} \gamma_s T_s D\mu L\sqrt{d} + \sum_{s=1}^{S} T_s \frac{6\gamma_s^2 \mu^2 L^2 d}{1 - L\alpha_s\gamma_s} \\
\leq &\frac{2}{5L}\left[f(\tilde{x}^0) - f(x^*)\right] + \frac{1}{2}\|x^* - x^0\|^2 + \sum_{s=1}^{S}\sum_{t=1}^{T_s} \eta_{s,t} + \sum_{s=1}^{S} \gamma_s T_s D\mu L\sqrt{d} + \sum_{s=1}^{S} T_s \frac{6\gamma_s^2 \mu^2 L^2 d}{1 - \frac{1}{5}} \\
\overset{①}{\leq} &\frac{D_0}{10L} + \sum_{s=1}^{S}\sum_{t=1}^{T_s} \eta_{s,t} + \sum_{s=1}^{S} \gamma_s T_s D\mu L\sqrt{d} + \sum_{s=1}^{S} T_s \frac{\gamma_s}{\alpha_s} \frac{3\mu^2 L d}{2} \\
= &\frac{D_0}{10L} + \sum_{s=1}^{S}\sum_{t=1}^{T_s} \eta_{s,t} + \sum_{s=1}^{S} \gamma_s T_s D\mu L\sqrt{d} + \sum_{s=1}^{S} T_s \frac{\gamma_s}{\alpha_s} \frac{3\mu^2 L d}{2}
\end{aligned} \tag{60}$$

where ① comes from the definition of D_0 that $D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 5L\|x^0 - x^*\|^2$.

- If $S \leq s_0$, then $\mathcal{L}_S = \frac{2^{S+1}}{5L}, \alpha_S = \frac{1}{2}, \gamma_S = \frac{2}{5L}$. We have

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \\
\leq &\frac{D_0}{2^{S+2}} + \frac{5L}{2^{S+1}} \sum_{s=1}^{S}\sum_{t=1}^{T_s} \eta_{s,t} + D\mu L\sqrt{d} + 3\mu^2 L d
\end{aligned} \tag{61}$$

  With the choice of $\eta_{s,t}$, we have

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s} \eta_{s,t} \leq \sum_{s=1}^{S} \frac{D_0}{sL} \leq \frac{D_0(\log S + 1)}{L} \tag{62}$$

  Plugging this inequality into (61) we get

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \frac{5D_0(\log S + 2)}{2^{S+1}} + D\mu L\sqrt{d} + 3\mu^2 L d \tag{63}$$

- If $S > s_0$, then $\mathcal{L}_s - \mathcal{R}_s = \gamma_s T_s > 0$ for $s > s_0$, and $\sum_{s=1}^{s_0} 2^{s-1} = 2^{s_0} - 1 \leq 2n$. We have

$$\begin{aligned}
\mathcal{L}_S = &\frac{1}{5L\alpha_S^2}\left[1 + (T_S - 1)(\alpha_S + \frac{1}{2})\right] \\
= &\frac{(S - s_0 + 4)(T_{s_0} - 1)}{10L} + \frac{(S - s_0 + 4)^2 (T_{s_0} + 1)}{40L} \overset{①}{\geq} \frac{(S - s_0 + 4)^2 n}{80L}
\end{aligned} \tag{64}$$

  where ① comes from $T_{s_0} = 2^{\lfloor \log_2 n \rfloor + 1 - 1} \geq n/2$. And

$$\sum_{s=1}^{S} T_s \frac{\gamma_s}{\alpha_s} = \sum_{s=1}^{s_0} \frac{2^{s+1}}{5L} + \sum_{s=s_0+1}^{S} \frac{(s - s_0 + 4)^2}{20L} 2^{s_0 - 1} \overset{①}{\leq} \frac{8n}{5L} + \frac{n(S - s_0 + 4)^3}{20L} \leq \frac{n(S - s_0 + 4)^3}{10L} \tag{65}$$

where ① comes from $\sum_{i=1}^{n} i^2 \leq n^3$. And

$$\sum_{s=1}^{S} \gamma_s T_s \leq \frac{1}{2} \sum_{s=1}^{S} T_s \frac{\gamma_s}{\alpha_s} \leq \frac{n(S-s_0+4)^3}{20L} \tag{66}$$

Thus

$$\begin{aligned}\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq & \frac{8D_0}{n(S-s_0+4)^2} + \frac{80L}{n(S-s_0+4)^2} \sum_{s=1}^{S} \sum_{t=1}^{T_s} \eta_{s,t} + 8(S-s_0+4)D\mu L\sqrt{d} \\ & + 12(S-s_0+4)\mu^2 L d\end{aligned} \tag{67}$$

With the choice of $\eta_{s,t}$, we have

$$\sum_{s=1}^{S} \sum_{t=1}^{T_s} \eta_{s,t} \leq \sum_{s=1}^{S} \frac{D_0}{sL} \leq \frac{D_0(\log S + 1)}{L} \tag{68}$$

Plugging this inequality into (67) we get

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \frac{80D_0(\log S + 2)}{n(S-s_0+4)^2} + 8(S-s_0+4)D\mu L\sqrt{d} + 12(S-s_0+4)\mu^2 L d \tag{69}$$

Then we get the desired result for the zeroth-order case. Then we complete the proof. ■

Appendix C. Proof of Theorem 7

Theorem 7 is a direct result of Theorem 13, Theorem 14 and Theorem 15. First we give a refined version of Lemma 10.

Lemma 12 *Suppose Assumption 3 holds. Conditioning on $x_1, ..., x_{t-1}$*

- *For the first-order case, assume that $\alpha_s \in [0,1], p_s \in [0,1]$ and $\gamma_s > 0$ satisfy*

$$1 + \tau\gamma_s - L\alpha_s\gamma_s > 0,\ 1 - \alpha_s - p_s \geq 0, \qquad p_s - \frac{L\alpha_s\gamma_s}{1 + \tau\gamma_s - L\alpha_s\gamma_s} > 0$$

Then we have

$$\begin{aligned}&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1+\tau\gamma_s}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \\ \leq & \frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1}) - f(x)\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x - x_{t-1}\|^2 + \eta_{s,t}\end{aligned}$$

- *For the zeroth-order case, assume that $\alpha_s \in [0,1], p_s \in [0,1]$ and $\gamma_s > 0$ satisfy*

$$1 + \tau\gamma_s - L\alpha_s\gamma_s > 0,\ 1 - \alpha_s - p_s \geq 0, \qquad p_s - \frac{4\alpha_s\gamma_s dL}{1 + \tau\gamma_s - L\alpha_s\gamma_s} > 0$$

Then we have

$$\begin{aligned}&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1+(1-c)\tau\gamma_s}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \\ \leq & \frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1}) - f(x)\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x - x_{t-1}\|^2 + \eta_{s,t} + \frac{\gamma_s\mu^2L^2d}{2c\tau} + \frac{6\gamma_s^2\mu^2L^2d}{1+\tau\gamma_s - L\alpha_s\gamma_s}\end{aligned}$$

Proof For the first-order case, the result is the same as that in Lemma 10. Now we give proof to the result of the zeroth-order case. From Lemma 10 we have

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t)-f(x)\right]+\frac{1+\tau\gamma_s}{2}\mathbb{E}\left[\|x-x_t\|^2\right]\\
\leq&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1})-f(x)\right]+\frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x})-f(x)\right]+\frac{1}{2}\|x-x_{t-1}\|^2+\eta_{s,t}\\
&-\gamma_s\langle\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t),x_t-x\rangle+\frac{6\gamma_s^2\mu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}
\end{aligned}\tag{70}$$

From $b\langle u,v\rangle-\frac{a}{2}\|v\|^2\leq\frac{b^2}{2a}\|u\|^2$ we have for $c>0$

$$-\gamma_s\langle\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t),x_t-x\rangle-\frac{c\tau\gamma_s}{2}\|x_t-x\|^2\leq\frac{\gamma_s}{2c\tau}\|\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t)\|^2\tag{71}$$

Plugging (71) into (70), we get

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t)-f(x)\right]+\frac{1+(1-c)\tau\gamma_s}{2}\mathbb{E}\left[\|x-x_t\|^2\right]\\
\leq&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1})-f(x)\right]+\frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x})-f(x)\right]+\frac{1}{2}\|x-x_{t-1}\|^2+\eta_{s,t}\\
&+\frac{\gamma_s}{2c\tau}\mathbb{E}\left[\|\hat{\nabla}_{coord}f(\underline{x}_t)-\nabla f(\underline{x}_t)\|^2\right]+\frac{6\gamma_s^2\mu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}\\
\overset{①}{\leq}&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1})-f(x)\right]+\frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x})-f(x)\right]+\frac{1}{2}\|x-x_{t-1}\|^2+\eta_{s,t}\\
&+\frac{\gamma_s\mu^2L^2d}{2c\tau}+\frac{6\gamma_s^2\mu^2L^2d}{1+\tau\gamma_s-L\alpha_s\gamma_s}
\end{aligned}\tag{72}$$

where ① comes from Lemma 16. Then we complete the proof. ■

Theorem 13 *Suppose Assumption 3 holds. Denote* $s_0=\lfloor\log n\rfloor+1$. *Suppose* $s\leq s_0$, *set* $\{T_s\},\{\alpha_s\},\{p_s\},\{\eta_{s,t}\},\{\theta_t\}$ *as*

$$T_s=2^{s-1},\ \alpha_s=\frac{1}{2},\ p_s=\frac{1}{2},\ \eta_{s,t}=\frac{D_0}{sT_sL},\ \theta_t=\begin{cases}\Gamma_{t-1}-(1-\alpha_s-p_s)\Gamma_t, & t\leq T_s-1\\ \Gamma_{t-1}, & t=T_s\end{cases}$$

where D_0,Γ_t *will be specified below for two cases respectively.*

• *For the first-order case, set* $\gamma_s=\frac{1}{3L\alpha_s},\Gamma_t=(1+\tau\gamma_s)^t,D_0=4(f(\tilde{x}^0)-f(x^*))+3L\|x^0-x^*\|^2$, *we have*

$$\mathbb{E}\left[f(\tilde{x}^S)-f(x^*)\right]\leq\frac{3D_0(\log S+2)}{2^{S+1}}$$

• *For the zeroth-order case, set* $\gamma_s=\frac{1}{5L\alpha_s},\Gamma_t=\left(1+\frac{\tau\gamma_s}{2}\right)^t,D_0=4(f(\tilde{x}^0)-f(x^*))+5L\|x^0-x^*\|^2$, *we have*

$$\mathbb{E}\left[f(\tilde{x}^S)-f(x^*)\right]\leq\ \frac{5D_0(\log S+2)}{2^{S+1}}+\frac{\mu^2L^2d}{2\tau}+3\mu^2Ld$$

Proof We give proof to the first-order case and zeroth-order case respectively.

First-order Case: We have $\alpha_s = p_s = \frac{1}{2}, \gamma_s = \frac{2}{3L}, T_s = 2^{s-1}$. Summing up Lemma 12 for $t = 1, ..., T_s$, we get

$$\begin{aligned}&\sum_{t=1}^{T_s} \frac{\gamma_s}{\alpha_s} \mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1}{2}\mathbb{E}\left[\|x_{T_s} - x\|^2\right] + \sum_{t=1}^{T_s} \frac{\tau\gamma_s}{2}\|x_t - x\|^2 \\ \leq & T_s \frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x_0 - x\|^2 + \sum_{t=1}^{T_s} \eta_{s,t}\end{aligned} \tag{73}$$

From the definition of $\tilde{x}^s, \tilde{x}^{s-1}, x^s, x^{s-1}$, the fact that $\frac{\gamma_s}{\alpha_s} = \frac{4}{3L}$, the convexity of f and Jensen's inequality, we have

$$\begin{aligned}\frac{4T_s}{3L}\mathbb{E}\left[f(\tilde{x}^s) - f(x)\right] + \frac{1}{2}\mathbb{E}\left[\|x^s - x\|^2\right] \leq & \frac{4T_s}{6L}\left[f(\tilde{x}^{s-1}) - f(x)\right] + \frac{1}{2}\|x^{s-1} - x\|^2 + \sum_{t=1}^{T_s} \eta_{s,t} \\ = & \frac{4T_{s-1}}{3L}\left[f(\tilde{x}^{s-1}) - f(x)\right] + \frac{1}{2}\|x^{s-1} - x\|^2 + \sum_{t=1}^{T_s} \eta_{s,t}\end{aligned} \tag{74}$$

Summing up (74) for $s = 1, ..., S$, we get

$$\frac{4T_S}{3L}\mathbb{E}\left[f(\tilde{x}^S) - f(x)\right] + \frac{1}{2}\mathbb{E}\left[\|x^S - x\|^2\right] \leq \frac{2}{3L}\left[f(\tilde{x}^0) - f(x)\right] + \frac{1}{2}\|x^0 - x\|^2 + \sum_{s=1}^{S}\sum_{t=1}^{T_s} \eta_{s,t} \tag{75}$$

Set $x = x^*$, we get

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] + \frac{3L}{8T_S}\mathbb{E}\left[\|x^S - x^*\|^2\right] \leq \frac{f(\tilde{x}^0) - f(x^*)}{2T_S} + \frac{3L}{8T_S}\|x^0 - x^*\|^2 + \frac{3L}{4T_S}\sum_{s=1}^{S}\sum_{t=1}^{T_s} \eta_{s,t} \tag{76}$$

With the choice of $\eta_{s,t}$, we have

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s} \eta_{s,t} \leq \sum_{s=1}^{S} \frac{D_0}{sL} \leq \frac{D_0(\log S + 1)}{L} \tag{77}$$

Plugging this inequality into (76), with the definition of D_0 and $T_s = 2^{s-1}$, we get

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] + \frac{3L}{8T_s}\mathbb{E}\left[\|x^S - x^*\|^2\right] \leq \frac{3D_0(\log S + 2)}{2^{S+1}} \tag{78}$$

Then we get the desired result for the first-order case.

Zeroth-order Case: We have $\alpha_s = p_s = \frac{1}{2}, \gamma_s = \frac{2}{5L}, T_s = 2^{s-1}$. Summing up Lemma 12 with $c = 1$ for $t = 1, ..., T_s$, we get

$$\begin{aligned}&\sum_{t=1}^{T_s} \frac{\gamma_s}{\alpha_s} \mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1}{2}\|x_{T_s} - x\|^2 \\ \leq & T_s \frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x_0 - x\|^2 + \sum_{t=1}^{T_s} \eta_{s,t} + T_s\frac{\gamma_s\mu^2L^2d}{2\tau} + T_s\frac{6\mu^2d}{5}\end{aligned} \tag{79}$$

From the definition of $\tilde{x}^s, \tilde{x}^{s-1}, x^s, x^{s-1}$, the fact that $\frac{\gamma_s}{\alpha_s} = \frac{4}{5L}$, the convexity of f and Jensen's inequality, we have

$$\begin{aligned}
&\frac{4T_s}{5L}\mathbb{E}\left[f(\tilde{x}^s) - f(x)\right] + \frac{1}{2}\mathbb{E}\left[\|x^s - x\|^2\right] \leq \sum_{t=1}^{T_s}\frac{4}{5L}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1}{2}\|x_{T_s} - x\|^2 \\
\leq &\frac{2T_s}{5L}\left[f(\tilde{x}^{s-1}) - f(x)\right] + \frac{1}{2}\|x^{s-1} - x\|^2 + \sum_{t=1}^{T_s}\eta_{s,t} + T_s\frac{\mu^2 Ld}{5\tau} + T_s\frac{6\mu^2 d}{5} \\
= &\frac{4T_{s-1}}{5L}\left[f(\tilde{x}^{s-1}) - f(x)\right] + \frac{1}{2}\|x^{s-1} - x\|^2 + \sum_{t=1}^{T_s}\eta_{s,t} + T_s\frac{\mu^2 Ld}{5\tau} + T_s\frac{6\mu^2 d}{5}
\end{aligned} \tag{80}$$

Summing up (80) for $s = 1, ..., S$, we get

$$\begin{aligned}
&\frac{4T_S}{5L}\mathbb{E}\left[f(\tilde{x}^S) - f(x)\right] + \frac{1}{2}\mathbb{E}\left[\|x^S - x\|^2\right] \\
\leq &\frac{4T_0}{5L}\left[f(\tilde{x}^0) - f(x)\right] + \frac{1}{2}\|x^0 - x\|^2 + \sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} + \sum_{s=1}^{S}T_s\frac{\mu^2 Ld}{5\tau} + \sum_{s=1}^{S}T_s\frac{6\mu^2 d}{5}
\end{aligned} \tag{81}$$

Note that $T_s = 2^{s-1}$. Setting $x = x^*$, we have

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] + \frac{5L}{8T_S}\mathbb{E}\left[\|x^S - x^*\|^2\right] \\
\leq &\frac{1}{2^S}\left[f(\tilde{x}^0) - f(x^*)\right] + \frac{5L}{2^{S+2}}\|x^0 - x^*\|^2 + \frac{5L}{2^{S+1}}\sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} + \frac{\mu^2 L^2 d}{2\tau} + 3\mu^2 Ld \\
\overset{①}{\leq} &\frac{D_0}{2^{S+2}} + \frac{5L}{2^{S+1}}\sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} + \frac{\mu^2 L^2 d}{2\tau} + 3\mu^2 Ld
\end{aligned} \tag{82}$$

where ① comes from the definition of D_0. With the choice of $\eta_{s,t}$, we have

$$\sum_{s=1}^{S}\sum_{t=1}^{T_s}\eta_{s,t} \leq \sum_{s=1}^{S}\frac{D_0}{sL} \leq \frac{D_0(\log S + 1)}{L} \tag{83}$$

Plugging this inequality into (82) we get

$$\begin{aligned}
\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq &\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] + \frac{5L}{8T_S}\mathbb{E}\left[\|x^S - x^*\|^2\right] \\
\leq &\frac{5D_0(\log S + 2)}{2^{S+1}} + \frac{\mu^2 L^2 d}{2\tau} + 3\mu^2 Ld
\end{aligned} \tag{84}$$

Then we get the desired result for the zeroth-order case. Then we complete the proof. ■

Theorem 14 *Suppose Assumption 3 holds. Denote* $s_0 = \lfloor \log n \rfloor + 1$. *Suppose* $s > s_0$, *set* $\{T_s\}, \{\alpha_s\}, \{p_s\}, \{\eta_{s,t}\}, \{\theta_t\}$ *as*

$$T_s = T_{s_0} = 2^{s_0 - 1},\ \alpha_s = \frac{1}{2},\ p_s = \frac{1}{2},\ \eta_{s,t} = \frac{\left(\frac{4}{5}\right)^{s-s_0-1}D_0}{snL},\ \theta_t = \begin{cases}\Gamma_{t-1} - (1 - \alpha_s - p_s)\Gamma_t, & t \leq T_s - 1 \\ \Gamma_{t-1}, & t = T_s\end{cases}$$

where D_0, Γ_t will be specified below for two cases respectively.

• *For the first-order case, set $\gamma_s = \frac{1}{3L\alpha_s}, \Gamma_t = (1+\tau\gamma_s)^t, D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 3L\|x^0 - x^*\|^2$ for $s > s_0$. Suppose $n \geq \frac{3L}{4\tau}$, we have*

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \left(\frac{4}{5}\right)^{S-s_0} \frac{5D_0(\log S + 2)}{n}$$

• *For the zeroth-order case, set $\gamma_s = \frac{1}{5L\alpha_s}, \Gamma_t = \left(1 + \frac{\tau\gamma_s}{2}\right)^t, D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 5L\|x^0 - x^*\|^2$ for $s > s_0$. Suppose $n \geq \frac{5L}{4\tau}$, we have*

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \left(\frac{4}{5}\right)^{S-s_0} \frac{8D_0(\log S + 2)}{n} + \frac{5\mu^2 L^2 d}{\tau} + 18\mu^2 L d$$

Proof We give proof to the first-order case and zeroth-order case respectively.

First-order Case: We have $\alpha_s = p_s = \frac{1}{2}, \gamma_s = \frac{2}{3L}$. Note that for the first-order case, we have $f = f$. From Lemma 12 we have

$$\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1+\tau\gamma_s}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \leq \frac{\gamma_s}{2\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x - x_{t-1}\|^2 + \eta_{s,t} \tag{85}$$

Multiplying both sides of (85) with $\theta_t = \Gamma_{t-1} = (1+\tau\gamma_s)^{t-1}$, we get

$$\frac{\gamma_s}{\alpha_s}\theta_t\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_t}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \leq \frac{\gamma_s}{2\alpha_s}\theta_t\left[f(\tilde{x}) - f(x)\right] + \frac{\Gamma_{t-1}}{2}\|x - x_{t-1}\|^2 + \theta_t\eta_{s,t} \tag{86}$$

Summing up (86) for $t = 1, ..., T_s$, we get

$$\frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_{T_s}}{2}\mathbb{E}\left[\|x_{T_s} - x\|^2\right] \leq \frac{\gamma_s}{2\alpha_s}\sum_{t=1}^{T_s}\theta_t\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x_0 - x\|^2 + \sum_{t=1}^{T_s}\theta_t\eta_{s,t} \tag{87}$$

Since we have

$$\Gamma_{T_s} = (1+\tau\gamma_s)^{T_s} = (1+\tau\gamma_s)^{T_{s_0}} \geq 1 + \tau\gamma_s T_{s_0} \geq 1 + \frac{\tau n}{3L} \overset{①}{\geq} \frac{5}{4} \tag{88}$$

where ① comes from the assumption that $n \geq \frac{3L}{4\tau}$. Then we get

$$\begin{aligned}
&\frac{5}{4}\left(\frac{\gamma_s}{2\alpha_s}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1}{2}\mathbb{E}\left[\|x_{T_s} - x\|^2\right]\right) \leq \frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_{T_s}}{2}\mathbb{E}\left[\|x_{T_s} - x\|^2\right] \\
&\leq \frac{\gamma_s}{2\alpha_s}\sum_{t=1}^{T_s}\theta_t\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x_0 - x\|^2 + \sum_{t=1}^{T_s}\theta_t\eta_{s,t}
\end{aligned} \tag{89}$$

From the definition of $\tilde{x}^s, \tilde{x}^{s-1}, x^s, x^{s-1}$, the fact that $\frac{\gamma_s}{\alpha_s} = \frac{4}{3L}$, the convexity of f and Jensen's inequality, we have

$$\begin{aligned}
&\frac{5}{4}\left(\frac{2}{3L}\mathbb{E}\left[f(\tilde{x}^s) - f(x)\right] + \frac{1}{2\sum_{t=1}^{T_s}\theta_t}\mathbb{E}\left[\|x^s - x\|^2\right]\right) \\
\leq &\frac{2}{3L}\left[f(\tilde{x}^{s-1}) - f(x)\right] + \frac{1}{2\sum_{t=1}^{T_s}\theta_t}\|x^{s-1} - x\|^2 + \frac{\sum_{t=1}^{T_s}\theta_t\eta_{s,t}}{\sum_{t=1}^{T_s}\theta_t}
\end{aligned} \tag{90}$$

Applying the inequality recursively for $s \geq s_0$, we get

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^S) - f(x)\right] + \frac{3L}{4\sum_{t=1}^{T_s}\theta_t}\mathbb{E}\left[\|x^S - x\|^2\right] \\
\leq &\left(\frac{4}{5}\right)^{S-s_0}\left([f(\tilde{x}^{s_0}) - f(x)] + \frac{3L}{4\sum_{t=1}^{T_s}\theta_t}\|x^{s_0} - x\|^2\right) + \sum_{k=s_0+1}^{S}\left(\frac{4}{5}\right)^{S+1-k}\frac{3L\sum_{t=1}^{T_s}\theta_t\eta_{k,t}}{2\sum_{t=1}^{T_s}\theta_t} \qquad (91)\\
\overset{①}{\leq} &\left(\frac{4}{5}\right)^{S-s_0}\left([f(\tilde{x}^{s_0}) - f(x)] + \frac{3L}{4T_{s_0}}\|x^{s_0} - x\|^2\right) + \left(\frac{4}{5}\right)^{S-s_0}\frac{3D_0(\log S+1)}{2n}D_0D_0
\end{aligned}$$

where ① comes from the choice of $\eta_{s,t}$ that

$$\sum_{k=s_0+1}^{S}\left(\frac{4}{5}\right)^{S+1-k}\frac{3L\sum_{t=1}^{T_s}\theta_t\eta_{k,t}}{2\sum_{t=1}^{T_s}\theta_t} = \sum_{k=s_0+1}^{S}\left(\frac{4}{5}\right)^{S-s_0}\frac{3}{2kn}D_0 \leq \left(\frac{4}{5}\right)^{S-s_0}\frac{3D_0(\log S+1)}{2n}D_0$$

and $\sum_{t=1}^{T_s}\theta_t \geq T_s = T_{s_0}$. From (75) we have

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^{s_0}) - f(x^*)\right] + \frac{3L}{4T_{s_0}}\mathbb{E}\|x^{s_0} - x^*\|^2 \\
\leq &2\left(\mathbb{E}\left[f(\tilde{x}^{s_0}) - f(x^*)\right] + \frac{3L}{8T_{s_0}}\mathbb{E}\|x^{s_0} - x^*\|^2\right) \leq \frac{3D_0(\log s_0+2)}{2^{s_0}}
\end{aligned} \qquad (92)$$

Plugging (92) into (91), setting $x = x^*$, we get

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \\
\leq &\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] + \frac{3L}{4\sum_{t=1}^{T_s}\theta_t}\mathbb{E}\left[\|x^S - x^*\|^2\right] \\
\leq &\left(\frac{4}{5}\right)^{S-s_0}\left([f(\tilde{x}^{s_0}) - f(x^*)] + \frac{3L}{4T_{s_0}}\|x^{s_0} - x^*\|^2\right) + \left(\frac{4}{5}\right)^{S-s_0}\frac{3D_0(\log S+1)}{2n}D_0 \qquad (93)\\
\leq &\left(\frac{4}{5}\right)^{S-s_0}\left(\frac{3D_0(\log s_0+2)}{2^{s_0}} + \frac{3D_0(\log S+1)}{2n}D_0\right) \\
\overset{①}{\leq} &\left(\frac{4}{5}\right)^{S-s_0}\left(\frac{3D_0(\log s_0+2)}{n} + \frac{3D_0(\log S+1)}{2n}D_0\right) \leq \left(\frac{4}{5}\right)^{S-s_0}\frac{5D_0(\log S+2)}{n}
\end{aligned}$$

where ① comes from the fact that $2^{s_0} \geq n$. Then we get the desired result for the first-order case.

Zeroth-order Case: We have $\alpha_s = p_s = \frac{1}{2}, \gamma_s = \frac{2}{5L}, T_s = T_{s_0} = 2^{s_0-1}$. Setting $c = \frac{1}{2}$ in Lemma 12, we have

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1+\frac{\tau\gamma_s}{2}}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \\
\leq &\frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x - x_{t-1}\|^2 + \eta_{s,t} + \frac{\gamma_s\mu^2L^2d}{\tau} + \frac{6\gamma_s^2\mu^2L^2d}{1+\tau\gamma_s - L\alpha_s\gamma_s} \qquad (94)\\
\leq &\frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x - x_{t-1}\|^2 + \eta_{s,t} + \frac{\gamma_s\mu^2L^2d}{\tau} + \frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2Ld}{2}
\end{aligned}$$

Multiplying both sides of (94) with $\theta_t = \Gamma_{t-1} = \left(1 + \frac{\tau\gamma_s}{2}\right)^{t-1}$, we get

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\theta_t \mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_t}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \\
\leq &\frac{\gamma_s}{2\alpha_s}\theta_t\left[f(\tilde{x}) - f(x)\right] + \frac{\Gamma_{t-1}}{2}\|x - x_{t-1}\|^2 + \theta_t\eta_{s,t} + \theta_t\frac{\gamma_s\mu^2L^2d}{\tau} + \theta_t\frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2Ld}{2}
\end{aligned} \tag{95}$$

Summing up (95) for $t = 1, ..., T_s$, we get

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\theta_t \mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_{T_s}}{2}\mathbb{E}\left[\|x_{T_s} - x\|^2\right] \\
\leq &\frac{\gamma_s}{2\alpha_s}\sum_{t=1}^{T_s}\theta_t\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x_0 - x\|^2 + \sum_{t=1}^{T_s}\theta_t\eta_{s,t} + \sum_{t=1}^{T_s}\frac{\theta_t\gamma_s\mu^2L^2d}{\tau} + \sum_{t=1}^{T_s}\theta_t\frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2Ld}{2}
\end{aligned} \tag{96}$$

Since we have

$$\Gamma_{T_s} = (1 + \frac{\tau\gamma_s}{2})^{T_s} = (1 + \frac{\tau\gamma_s}{2})^{T_{s_0}} \geq 1 + \frac{\tau\gamma_s}{2}T_{s_0} \geq 1 + \frac{\tau\gamma_s}{2}\cdot\frac{n}{2} = 1 + \frac{\tau n}{5L} \overset{①}{\geq} \frac{5}{4} \tag{97}$$

where ① comes from the assumption that $n \geq \frac{5L}{4\tau}$. Then we get

$$\begin{aligned}
&\frac{5}{4}\left(\frac{\gamma_s}{2\alpha_s}\sum_{t=1}^{T_s}\theta_t \mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1}{2}\mathbb{E}\left[\|x_{T_s} - x\|^2\right]\right) \\
\leq &\frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\theta_t \mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_{T_s}}{2}\mathbb{E}\left[\|x_{T_s} - x\|^2\right] \\
\leq &\frac{\gamma_s}{2\alpha_s}\sum_{t=1}^{T_s}\theta_t\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x_0 - x\|^2 + \sum_{t=1}^{T_s}\theta_t\eta_{s,t} + \sum_{t=1}^{T_s}\frac{\theta_t\gamma_s\mu^2L^2d}{\tau} + \sum_{t=1}^{T_s}\theta_t\frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2Ld}{2}
\end{aligned} \tag{98}$$

From the definition of $\tilde{x}^s, \tilde{x}^{s-1}, x^s, x^{s-1}$, the fact that $\frac{\gamma_s}{\alpha_s} = \frac{4}{5L}$, the convexity of f and Jensen's inequality, we have

$$\begin{aligned}
&\frac{5}{4}\left(\frac{2}{5L}\mathbb{E}\left[f(\tilde{x}^s) - f(x)\right] + \frac{1}{2\sum_{t=1}^{T_s}\theta_t}\mathbb{E}\left[\|x^s - x\|^2\right]\right) \\
\leq &\frac{2}{5L}\left[f(\tilde{x}^{s-1}) - f(x)\right] + \frac{1}{2\sum_{t=1}^{T_s}\theta_t}\|x^{s-1} - x\|^2 + \frac{\sum_{t=1}^{T_s}\theta_t\eta_{s,t}}{\sum_{t=1}^{T_s}\theta_t} + \frac{2\mu^2Ld}{5\tau} + \frac{6\mu^2d}{5}
\end{aligned} \tag{99}$$

Applying the inequality recursively for $s \geq s_0$, we get

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^S) - f(x)\right] + \frac{5L}{\sum_{t=1}^{T_s}\theta_t}\mathbb{E}\left[\|x^S - x\|^2\right] \\
&\leq \left(\frac{4}{5}\right)^{S-s_0}\left([f(\tilde{x}^{s_0}) - f(x)] + \frac{5L}{4\sum_{t=1}^{T_s}\theta_t}\|x^{s_0} - x\|^2\right) \\
&\quad + \sum_{k=s_0+1}^{S}\left(\frac{4}{5}\right)^{S+1-k}\left(\frac{5L\sum_{t=1}^{T_s}\theta_t\eta_{k,t}}{2\sum_{t=1}^{T_s}\theta_t} + \frac{\mu^2L^2d}{\tau} + 3\mu^2Ld\right) \qquad (100) \\
&\overset{①}{\leq} \left(\frac{4}{5}\right)^{S-s_0}\left([f(\tilde{x}^{s_0}) - f(x)] + \frac{5L}{4T_{s_0}}\|x^{s_0} - x\|^2\right) \\
&\quad + \left(\frac{4}{5}\right)^{S-s_0}\frac{5(\log S+1)}{2n}D_0 + \frac{4\mu^2L^2d}{\tau} + 12\mu^2Ld
\end{aligned}$$

where ① comes from the choice of $\eta_{s,t}$ that

$$\sum_{k=s_0+1}^{S}\left(\frac{4}{5}\right)^{S+1-k}\frac{5L\sum_{t=1}^{T_s}\theta_t\eta_{k,t}}{2\sum_{t=1}^{T_s}\theta_t} = \sum_{k=s_0+1}^{S}\left(\frac{4}{5}\right)^{S-s_0}\frac{5}{2kn}D_0 \leq \left(\frac{4}{5}\right)^{S-s_0}\frac{5(\log S+1)}{2n}D_0$$

and

$$\sum_{k=s_0+1}^{S}\left(\frac{4}{5}\right)^{S+1-k} \leq \frac{4}{5}\cdot\frac{1}{1-\frac{4}{5}} = 4$$

and $\sum_{t=1}^{T_s}\theta_t \geq T_s = T_{s_0}$. From (84) we have

$$\begin{aligned}
\mathbb{E}\left[f(\tilde{x}_0^s) - f(x^*)\right] + \frac{5L}{4T_{s_0}}\mathbb{E}\left[\|x_0^s - x^*\|^2\right] &\leq 2\left(\mathbb{E}\left[f(\tilde{x}_0^s) - f(x^*)\right] + \frac{5L}{8T_{s_0}}\mathbb{E}\left[\|x_0^s - x^*\|^2\right]\right) \\
&\leq \frac{5D_0(\log s_0 + 2)}{2^{s_0}} + \frac{\mu^2L^2d}{\tau} + 6\mu^2Ld
\end{aligned} \qquad (101)$$

Plugging (101) into (100), setting $x = x^*$, we get

$$\begin{aligned}
\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] &\leq \left(\frac{4}{5}\right)^{S-s_0}\left(\frac{5D_0(\log s_0+2)}{2^{s_0}} + \frac{\mu^2L^2d}{\tau} + 6\mu^2Ld\right) \\
&\quad + \left(\frac{4}{5}\right)^{S-s_0}\frac{5(\log S+1)}{2n}D_0 + \frac{4\mu^2L^2d(d+4)}{\tau} + 12\mu^2Ld \qquad (102) \\
&\overset{①}{\leq} \left(\frac{4}{5}\right)^{S-s_0}\frac{8D_0(\log S+2)}{n} + \frac{5\mu^2L^2d}{\tau} + 18\mu^2Ld
\end{aligned}$$

where ① comes from the fact that $2^{s_0} \geq n$. Then we get the desired result for the zeroth-order case. Then we complete the proof. ■

Theorem 15 *Suppose Assumption 3 holds. Denote* $s_0 = \lfloor \log n \rfloor + 1$. *Suppose* $s > s_0$, *set* $\{T_s\}, \{p_s\}, \{\eta_{s,t}\}, \{\theta_t\}$ *as*

$$T_s = T_{s_0} = 2^{s_0-1},\ p_s = \frac{1}{2},\ \eta_{s,t} = \frac{\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{s-s_0-1} D_0}{snL},\ \theta_t = \begin{cases} \Gamma_{t-1} - (1-\alpha_s-p_s)\Gamma_t, & t \le T_s - 1 \\ \Gamma_{t-1}, & t = T_s \end{cases}$$

where D_0, Γ_t *will be specified below for two cases respectively.*

• *For the first-order case, set* $\alpha_s = \sqrt{\frac{n\tau}{3L}}, \gamma_s = \frac{1}{3L\alpha_s}, \Gamma_t = (1+\tau\gamma_s)^t, D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 3L\|x^0 - x^*\|^2$ *for* $s > s_0$. *Suppose* $n < \frac{3L}{4\tau}$, *we have*

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \le \left(1 + \frac{1}{2}\sqrt{\frac{n\tau}{3L}}\right)^{-(S-s_0)} \frac{5D_0(\log S + 2)}{n}$$

• *For the zeroth-order case, set* $\alpha_s = \sqrt{\frac{n\tau}{5L}}, \gamma_s = \frac{1}{5L\alpha_s}, \Gamma_t = \left(1 + \frac{\tau\gamma_s}{2}\right)^t, D_0 = 4(f(\tilde{x}^0) - f(x^*)) + 5L\|x^0 - x^*\|^2$ *for* $s > s_0$. *Suppose* $n < \frac{5L}{4\tau}$, *we have*

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \le \left(1 + \frac{1}{4}\sqrt{\frac{n\tau}{5L}}\right)^{-(S-s_0)} \frac{8D_0(\log S + 2)}{n} + \left(\frac{\mu^2L^2d}{\tau} + 6\mu^2Ld\right)\left(1 + 8\sqrt{\frac{5L}{n\tau}}\right)$$

Proof We give proof to the first-order case and zeroth-order case respectively.

First-order Case: We have $\alpha_s = \sqrt{\frac{n\tau}{3L}}, p_s = \frac{1}{2}, \gamma_s = \frac{1}{\sqrt{3n\tau L}}, T_s = T_{s_0} = 2^{s_0-1}$. Note that for the first-order case, we have $f = f$. From Lemma 12, we have

$$\begin{aligned} &\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1+\tau\gamma_s}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \\ \le &\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1}) - f(x)\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x - x_{t-1}\|^2 + \eta_{s,t} \end{aligned} \tag{103}$$

Multiplying both sides of the above inequality with $\Gamma_{t-1} = (1+\tau\gamma_s)^{t-1}$, we get

$$\begin{aligned} &\frac{\gamma_s}{\alpha_s}\Gamma_{t-1}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_t}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \\ \le &\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\Gamma_{t-1}\left[f(\bar{x}_{t-1}) - f(x)\right] + \frac{\gamma_s p_s}{\alpha_s}\Gamma_{t-1}\left[f(\tilde{x}) - f(x)\right] + \frac{\Gamma_{t-1}}{2}\|x - x_{t-1}\|^2 + \Gamma_{t-1}\eta_{s,t} \end{aligned} \tag{104}$$

Summing up the above inequality for $t = 1, ..., T_s$, using the definition of θ_t, we get

$$\begin{aligned} &\frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_{T_s}}{2}\mathbb{E}\left[\|x_{T_s} - x\|^2\right] \\ \le &\frac{\gamma_s}{\alpha_s}\left[1 - \alpha_s - p_s + p_s\sum_{t=1}^{T_s}\Gamma_{t-1}\right]\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x_0 - x\|^2 + \sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{s,t} \end{aligned} \tag{105}$$

From the definition of $\tilde{x}^s, \tilde{x}^{s-1}, x^s, x^{s-1}$, the convexity of f and Jensen's inequality, we have

$$\begin{aligned} &\frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\left[f(\tilde{x}^s) - f(x)\right] + \frac{\Gamma_{T_s}}{2}\mathbb{E}\left[\|x^s - x\|^2\right] \\ \le &\frac{\gamma_s}{\alpha_s}\left[1 - \alpha_s - p_s + p_s\sum_{t=1}^{T_s}\Gamma_{t-1}\right]\left[f(\tilde{x}^{s-1}) - f(x)\right] + \frac{1}{2}\|x^{s-1} - x\|^2 + \sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{s,t} \end{aligned} \tag{106}$$

From the definition of θ_t, we have

$$\begin{aligned}\sum_{t=1}^{T_{s_0}}\theta_t =&\Gamma_{T_{s_0}-1}+\sum_{t=1}^{T_{s_0}-1}\left(\Gamma_{t-1}-(1-\alpha_s-p_s)\Gamma_t\right)\\ =&\Gamma_{T_{s_0}}(1-\alpha_s-p_s)+\sum_{t=1}^{T_{s_0}}\left(\Gamma_{t-1}-(1-\alpha_s-p_s)\Gamma_t\right)\\ =&\Gamma_{T_{s_0}}(1-\alpha_s-p_s)+\left[1-(1-\alpha_s-p_s)(1+\tau\gamma_s)\right]\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\end{aligned}\tag{107}$$

Since $T_{s_0}=2^{s_0-1}\le n$, we have

$$\alpha_s=\sqrt{\frac{n\tau}{3L}}\ge\sqrt{\frac{T_{s_0}\tau}{3L}}=\tau\sqrt{\frac{T_{s_0}n}{3n\tau L}}\ge\tau\gamma_s T_{s_0}\tag{108}$$

Then we have

$$\begin{aligned}1-(1-\alpha_s-p_s)(1+\tau\gamma_s)=&(1+\tau\gamma_s)(\alpha_s-\tau\gamma_s+p_s)+\tau^2\gamma_s^2\\ \overset{①}{\ge}&(1+\tau\gamma_s)(\tau\gamma_s T_{s_0}-\tau\gamma_s+p_s)\\ =&p_s(1+\tau\gamma_s)(2(T_{s_0}-1)\tau\gamma_s+1)\\ \overset{②}{\ge}&p_s(1+\tau\gamma_s)^{T_{s_0}}=p_s\Gamma_{T_s}\end{aligned}\tag{109}$$

where ① comes from (108) and ② comes from the fact that $(1+a)^b\le 1+2ab$, for $b\ge 1, ab\in[0,1]$. Plugging (109) into (107), we get

$$\sum_{t=1}^{T_{s_0}}\theta_t\ge\Gamma_{T_{s_0}}\left[1-\alpha_s-p_s+p_s\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\right]\tag{110}$$

Plugging (110) into (106), we get

$$\begin{aligned}&\Gamma_{T_{s_0}}\left(\frac{\gamma_s}{\alpha_s}\left[1-\alpha_s-p_s+p_s\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\right]\mathbb{E}\left[f(\tilde{x}^s)-f(x)\right]+\frac{1}{2}\mathbb{E}\left[\|x^s-x\|^2\right]\right)\\ \le&\frac{\gamma_s}{\alpha_s}\left[1-\alpha_s-p_s+p_s\sum_{t=1}^{T_s}\Gamma_{t-1}\right]\left[f(\tilde{x}^{s-1})-f(x)\right]+\frac{1}{2}\|x^{s-1}-x\|^2+\sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{s,t}\end{aligned}\tag{111}$$

Since we have

$$\frac{\gamma_s}{\alpha_s}\left[1-\alpha_s-p_s+p_s\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\right]\ge\frac{\gamma_s p_s}{\alpha_s}\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\ge\frac{\gamma_s p_s}{\alpha_s}T_{s_0}\tag{112}$$

Dividing both sides of the above inequality with $\Gamma_{T_{s_0}}\frac{\gamma_s}{\alpha_s}\left[1-\alpha_s-p_s+p_s\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\right]$, we have

$$\begin{aligned}&\mathbb{E}\left[f(\tilde{x}^s)-f(x)\right]+\frac{\alpha_s}{\gamma_s T_{s_0}}\mathbb{E}\left[\|x^s-x\|^2\right]\\ \le&\left(\frac{1}{\Gamma_{T_{s_0}}}\right)\left(\left[f(\tilde{x}^s)-f(x)\right]+\frac{\alpha_s}{\gamma_s T_{s_0}}\left[\|x^s-x\|^2\right]\right)+\left(\frac{1}{\Gamma_{T_{s_0}}}\right)\frac{2\alpha_s\sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{s,t}}{\gamma_s\sum_{t=1}^{T_s}\Gamma_{t-1}}\end{aligned}\tag{113}$$

Summing up the above inequality for $s > s_0$, setting $x = x^*$, we get

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] + \frac{\alpha_S}{\gamma_S T_{s_0}}\mathbb{E}\left[\|x^S - x^*\|^2\right] \\
\leq &\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S-s_0}\left([f(\tilde{x}^{s_0}) - f(x^*)] + \frac{\alpha_{s_0}}{\gamma_{s_0}T_{s_0}}\left[\|x^{s_0} - x^*\|^2\right]\right) + \sum_{k=s_0+1}^{S}\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S+1-k}\frac{2\alpha_s\sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{k,t}}{\gamma_s\sum_{t=1}^{T_s}\Gamma_{t-1}} \\
\overset{①}{\leq} &\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S-s_0}\left([f(\tilde{x}^{s_0}) - f(x^*)] + \frac{3L}{4T_{s_0}}\left[\|x^{s_0} - x^*\|^2\right]\right) + \sum_{k=s_0+1}^{S}\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S+1-k}\frac{2\alpha_s\sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{k,t}}{\gamma_s\sum_{t=1}^{T_s}\Gamma_{t-1}}
\end{aligned} \tag{114}$$

where ① comes from the fact that $\frac{\alpha_s}{\gamma_s} = n\tau \leq \frac{3L}{4}$. From Theorem 13 we have

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^{s_0}) - f(x^*)\right] + \frac{3L}{4T_{s_0}}\mathbb{E}\left[\|x^{s_0} - x^*\|^2\right] \\
\leq &2\left(\mathbb{E}\left[f(\tilde{x}^{s_0}) - f(x^*)\right] + \frac{3L}{8T_{s_0}}\mathbb{E}\left[\|x^{s_0} - x^*\|^2\right]\right) \overset{①}{\leq} \frac{3D_0(\log s_0 + 2)}{2^{s_0}}
\end{aligned} \tag{115}$$

where ① comes from (78). With the choice of $\eta_{s,t}$, we have

$$\begin{aligned}
\sum_{k=s_0+1}^{S}\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S+1-k}\frac{2\alpha_s\sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{k,t}}{\gamma_s\sum_{t=1}^{T_s}\Gamma_{t-1}} &= \sum_{k=s_0+1}^{S}\left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S-s_0}\frac{2}{kn}D_0 \\
&\leq \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S-s_0}\frac{2(\log S + 1)}{n}D_0
\end{aligned} \tag{116}$$

Plugging (116), (115) into (114), we get

$$\begin{aligned}
\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] &\leq \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S-s_0}\frac{3D_0(\log s_0 + 2)}{2^{s_0}} + \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S-s_0}\frac{2(\log S + 1)}{n}D_0 \\
&\overset{①}{\leq} \left(\frac{1}{\Gamma_{T_{s_0}}}\right)^{S-s_0}\frac{5D_0(\log S + 2)}{n}
\end{aligned} \tag{117}$$

where ① comes from $2^{s_0} \geq n$. From the definition of $\Gamma_{T_{s_0}}$, we have

$$\Gamma_{T_{s_0}} = (1 + \tau\gamma_s)^{T_{s_0}} \geq 1 + \tau\gamma_s T_{s_0} \geq 1 + \frac{\tau\gamma_s n}{2} = 1 + \frac{1}{2}\sqrt{\frac{n\tau}{3L}} \tag{118}$$

Plugging (118) into (117), we get

$$\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \left(1 + \frac{1}{2}\sqrt{\frac{n\tau}{3L}}\right)^{-(S-s_0)}\frac{5D_0(\log S + 2)}{n} \tag{119}$$

Then we get the desired result for the first-order case.

Zeroth-order Case: We have $\alpha_s = \sqrt{\frac{n\tau}{5L}}, p_s = \frac{1}{2}, \gamma_s = \frac{1}{\sqrt{5n\tau L}}, T_s = T_{s_0} = 2^{s_0-1}$. Setting $c = \frac{1}{2}$ in Lemma 12, we have

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{1+\frac{\tau\gamma_s}{2}}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \\
\leq&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1}) - f(x)\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x - x_{t-1}\|^2 + \eta_{s,t} \\
&+ \frac{\gamma_s\mu^2L^2d}{\tau} + \frac{6\gamma_s^2\mu^2L^2d}{1+\tau\gamma_s - L\alpha_s\gamma_s} \\
\leq&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\left[f(\bar{x}_{t-1}) - f(x)\right] + \frac{\gamma_s p_s}{\alpha_s}\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x - x_{t-1}\|^2 + \eta_{s,t} \\
&+ \frac{\gamma_s\mu^2L^2d}{\tau} + \frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2Ld}{2}
\end{aligned} \tag{120}$$

Multiplying both sides of the inequality with $\Gamma_{t-1} = \left(1+\frac{\tau\gamma_s}{2}\right)^{t-1}$, we get

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\Gamma_{t-1}\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_t}{2}\mathbb{E}\left[\|x - x_t\|^2\right] \\
\leq&\frac{\gamma_s(1-\alpha_s-p_s)}{\alpha_s}\Gamma_{t-1}\left[f(\bar{x}_{t-1}) - f(x)\right] + \frac{\gamma_s p_s}{\alpha_s}\Gamma_{t-1}\left[f(\tilde{x}) - f(x)\right] + \frac{\Gamma_{t-1}}{2}\|x - x_{t-1}\|^2 + \Gamma_{t-1}\eta_{s,t} \\
&+ \Gamma_{t-1}\frac{\gamma_s\mu^2L^2d}{\tau} + \Gamma_{t-1}\frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2Ld}{2}
\end{aligned} \tag{121}$$

Summing up the inequality for $t = 1, ..., T_s$, using the definition of θ_t, we get

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\left[f(\bar{x}_t) - f(x)\right] + \frac{\Gamma_{T_s}}{2}\mathbb{E}\left[\|x_{T_s} - x\|^2\right] \\
\leq&\frac{\gamma_s}{\alpha_s}\left(1-\alpha_s-p_s+p_s\sum_{t=1}^{T_s}\Gamma_{t-1}\right)\left[f(\tilde{x}) - f(x)\right] + \frac{1}{2}\|x_0 - x\|^2 + \sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{s,t} \\
&+ \sum_{t=1}^{T_s}\Gamma_{t-1}\frac{\gamma_s\mu^2L^2d}{\tau} + \sum_{t=1}^{T_s}\Gamma_{t-1}\frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2Ld}{2}
\end{aligned} \tag{122}$$

From the definition of $\tilde{x}^s, \tilde{x}^{s-1}, x^s, x^{s-1}$ and the convexity of f, we have

$$\begin{aligned}
&\frac{\gamma_s}{\alpha_s}\sum_{t=1}^{T_s}\theta_t\mathbb{E}\left[f(\tilde{x}^s) - f(x)\right] + \frac{\Gamma_{T_s}}{2}\mathbb{E}\left[\|x^s - x\|^2\right] \\
\leq&\frac{\gamma_s}{\alpha_s}\left(1-\alpha_s-p_s+p_s\sum_{t=1}^{T_s}\Gamma_{t-1}\right)\left[f(\tilde{x}^{s-1}) - f(x)\right] + \frac{1}{2}\|x^{s-1} - x\|^2 + \sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{s,t} \\
&+ \sum_{t=1}^{T_s}\Gamma_{t-1}\frac{\gamma_s\mu^2L^2d}{\tau} + \sum_{t=1}^{T_s}\Gamma_{t-1}\frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2Ld}{2}
\end{aligned} \tag{123}$$

From the definition of θ_t, we have

$$\begin{aligned}\sum_{t=1}^{T_{s_0}}\theta_t =&\Gamma_{T_{s_0}-1}+\sum_{t=1}^{T_{s_0}-1}\left(\Gamma_{t-1}-(1-\alpha_s-p_s)\Gamma_t\right)\\ =&\Gamma_{T_{s_0}}(1-\alpha_s-p_s)+\sum_{t=1}^{T_{s_0}}\left(\Gamma_{t-1}-(1-\alpha_s-p_s)\Gamma_t\right)\\ =&\Gamma_{T_{s_0}}(1-\alpha_s-p_s)+\left[1-(1-\alpha_s-p_s)\left(1+\frac{\tau\gamma_s}{2}\right)\right]\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\end{aligned}\tag{124}$$

Since $T_{s_0}=2^{s_0-1}\le n$, we have

$$\alpha_s=\sqrt{\frac{n\tau}{5L}}\ge\sqrt{\frac{T_{s_0}\tau}{5L}}=\frac{1}{\sqrt{5n\tau L}}\cdot\tau\sqrt{T_{s_0}n}\ge\tau\gamma_s T_{s_0}\ge\frac{\tau\gamma_s T_{s_0}}{2}\tag{125}$$

Then we have

$$\begin{aligned}1-(1-\alpha_s-p_s)\left(1+\frac{\tau\gamma_s}{2}\right)=&\left(1+\frac{\tau\gamma_s}{2}\right)\left(\alpha_s+p_s-\frac{\tau\gamma_s}{2}\right)+\frac{\tau^2\gamma_s^2}{4}\\ \overset{①}{\ge}&\left(1+\frac{\tau\gamma_s}{2}\right)\left(\frac{\tau\gamma_s}{2}T_{s_0}+p_s-\frac{\tau\gamma_s}{2}\right)\\ =&p_s\left(1+\frac{\tau\gamma_s}{2}\right)\left(1+2(T_{s_0}-1)\frac{\tau\gamma_s}{2}\right)\\ \overset{②}{\ge}&p_s\left(1+\frac{\tau\gamma_s}{2}\right)^{T_{s_0}}=p_s\Gamma_{T_{s_0}}\end{aligned}\tag{126}$$

where ① comes from (125) and ② comes from the fact that $(1+a)^b\le 1+2ab$, for $b\ge 1, ab\in[0,1]$. Plugging (126) into (124), we get

$$\sum_{t=1}^{T_{s_0}}\theta_t\ge\Gamma_{T_{s_0}}\left[1-\alpha_s-p_s+p_s\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\right]\tag{127}$$

Then plugging (127) into (123), setting $x=x^*$, we get

$$\begin{aligned}&\Gamma_{T_s}\left(\frac{\gamma_s}{\alpha_s}\left[1-\alpha_s-p_s+p_s\sum_{t=1}^{T_s}\Gamma_{t-1}\right]\mathbb{E}\left[f(\tilde{x}^s)-f(x^*)\right]+\frac{1}{2}\mathbb{E}\left[\|x^s-x^*\|^2\right]\right)\\ \le&\frac{\gamma_s}{\alpha_s}\left[1-\alpha_s-p_s+p_s\sum_{t=1}^{T_s}\Gamma_{t-1}\right]\left[f(\tilde{x}^{s-1})-f(x^*)\right]+\frac{1}{2}\|x^{s-1}-x^*\|^2+\sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{s,t}\\ &+\sum_{t=1}^{T_s}\Gamma_{t-1}\frac{\gamma_s\mu^2L^2d}{\tau}+\sum_{t=1}^{T_s}\Gamma_{t-1}\frac{\gamma_s}{\alpha_s}\cdot\frac{3\mu^2Ld}{2}\end{aligned}\tag{128}$$

Denote $\alpha = \alpha_s = \sqrt{\frac{n\tau}{5L}}, p = p_s = \frac{1}{2}, \gamma = \gamma_s = \frac{1}{\sqrt{5n\tau L}}$ Rearranging the terms and summing up the above inequality for $s > s_0$, we get

$$\begin{aligned}
&\frac{\gamma}{\alpha}\left[1-\alpha-p+p\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\right]\mathbb{E}\left[f(\tilde{x}^S)-f(x^*)\right]+\frac{1}{2}\mathbb{E}\left[\|x^S-x^*\|^2\right]\\
&\le\left(\frac{1}{\Gamma_{T_s}}\right)^{S-s_0}\left(\frac{\gamma}{\alpha}\left[1-\alpha-p+p\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\right][f(\tilde{x}^{s_0})-f(x^*)]+\frac{1}{2}\|x^{s_0}-x^*\|^2\right)\\
&\quad+\sum_{k=s_0+1}^{S}\left(\frac{1}{\Gamma_{T_s}}\right)^{S+1-k}\left(\sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{k,t}+\sum_{t=1}^{T_s}\Gamma_{t-1}\frac{\gamma\mu^2L^2d}{\tau}+\sum_{t=1}^{T_s}\Gamma_{t-1}\frac{\gamma}{\alpha}\cdot\frac{3\mu^2Ld}{2}\right)
\end{aligned}\tag{129}$$

Since we have

$$\frac{\gamma}{\alpha}\left[1-\alpha-p+p\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\right]\ge\frac{\gamma p}{\alpha}\sum_{t=1}^{T_{s_0}}\Gamma_{t-1}\ge\frac{\gamma pT_{s_0}}{\alpha}\tag{130}$$

Plugging (130) into (129), we get

$$\begin{aligned}
\mathbb{E}\left[f(\tilde{x}^S)-f(x^*)\right]&\le\left(\frac{1}{\Gamma_{T_s}}\right)^{S-s_0}\left[[f(\tilde{x}^{s_0})-f(x^*)]+\frac{\alpha}{\gamma T_{s_0}}\|x^{s_0}-x^*\|^2\right]\\
&\quad+\sum_{k=s_0+1}^{S}\left(\frac{1}{\Gamma_{T_s}}\right)^{S+1-k}\left[\frac{2\alpha\sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{k,t}}{\gamma\sum_{t=1}^{T_s}\Gamma_{t-1}}+\frac{2\alpha\mu^2L^2d}{\tau}+3\mu^2Ld\right]\\
&\overset{①}{\le}\left(\frac{1}{\Gamma_{T_s}}\right)^{S-s_0}\left[[f(\tilde{x}^{s_0})-f(x^*)]+\frac{\alpha}{\gamma T_{s_0}}\|x^{s_0}-x^*\|^2\right]\\
&\quad+\left(\frac{1}{\Gamma_{T_s}}\right)^{S-s_0}\frac{3(\log S+1)}{n}D_0+\frac{1}{\Gamma_{T_{s_0}}-1}\left[\frac{2\alpha\mu^2L^2d}{\tau}+3\mu^2Ld\right]
\end{aligned}\tag{131}$$

where ① comes from the choice of $\eta_{s,t}$ that

$$\sum_{k=s_0+1}^{S}\left(\frac{1}{\Gamma_{T_s}}\right)^{S+1-k}\frac{2\alpha\sum_{t=1}^{T_s}\Gamma_{t-1}\eta_{k,t}}{\gamma\sum_{t=1}^{T_s}\Gamma_{t-1}}=\sum_{k=s_0+1}^{S}\left(\frac{1}{\Gamma_{T_s}}\right)^{S-s_0}\frac{3}{kn}D_0\le\left(\frac{1}{\Gamma_{T_s}}\right)^{S-s_0}\frac{3(\log S+1)}{n}D_0$$

and

$$\sum_{k=s_0+1}^{S}\left(\frac{1}{\Gamma_{T_s}}\right)^{S+1-k}\le\frac{1}{\Gamma_{T_{s_0}}-1}$$

From (84) we know

$$\mathbb{E}\left[f(\tilde{x}^{s_0})-f(x^*)\right]+\frac{5L}{8T_{s_0}}\mathbb{E}\left[\|x^{s_0}-x^*\|^2\right]\le\frac{5D_0(\log s_0+2)}{2^{s_0+1}}+\frac{\mu^2L^2d}{2\tau}+3\mu^2Ld\tag{132}$$

From the definition of α, γ and the assumption that $n < \frac{5L}{4\tau}$, we have $\frac{\alpha}{\gamma} = 5L\alpha^2 = n\tau \leq \frac{5L}{4}$. Thus we get

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^{s_0}) - f(x^*)\right] + \frac{\alpha}{\gamma T_{s_0}}\mathbb{E}\left[\|x^{s_0} - x^*\|^2\right] \\
\leq& 2\left(\mathbb{E}\left[f(\tilde{x}^{s_0}) - f(x^*)\right] + \frac{5L}{8T_s}\mathbb{E}\left[\|x^{s_0} - x^*\|^2\right]\right) \\
\leq& \frac{5D_0(\log s_0 + 2)}{2^{s_0}} + \frac{\mu^2 L^2 d}{\tau} + 6\mu^2 L d \overset{①}{\leq} \frac{5D_0(\log s_0 + 2)}{n} + \frac{\mu^2 L^2 d}{\tau} + 6\mu^2 L d
\end{aligned} \tag{133}$$

where ① holds since $2^{s_0} \geq n$. Plugging (133) into (131), we get

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \\
\leq& \left(\frac{1}{\Gamma_{T_s}}\right)^{S-s_0}\left[\frac{8D_0(\log S + 2)}{n} + \frac{\mu^2 L^2 d}{\tau} + 6\mu^2 L d\right] + \frac{1}{\Gamma_{T_{s_0}} - 1}\left[\frac{2\alpha\mu^2 L^2 d}{\tau} + 3\mu^2 L d\right]
\end{aligned} \tag{134}$$

From the definition of $\Gamma_{T_{s_0}}$, we have

$$\Gamma_{T_{s_0}} = \left(1 + \frac{\tau\gamma}{2}\right)^{T_{s_0}} \geq 1 + \frac{\tau\gamma}{2}T_{s_0} \geq 1 + \frac{\tau\gamma n}{4} \overset{①}{=} 1 + \frac{1}{4}\sqrt{\frac{n\tau}{5L}} \tag{135}$$

where ① comes from the definition of γ. Then we have

$$\frac{1}{\Gamma_{T_{s_0}} - 1} \leq 4\sqrt{\frac{5L}{n\tau}}, \text{ and } \left(\frac{1}{\Gamma_{T_s}}\right)^{S-s_0} \leq \left(1 + \frac{1}{4}\sqrt{\frac{n\tau}{5L}}\right)^{-(S-s_0)} \tag{136}$$

Plugging (136) into (134), using the definition of α, γ, we get

$$\begin{aligned}
&\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \\
\leq& \left(1 + \frac{1}{4}\sqrt{\frac{n\tau}{5L}}\right)^{-(S-s_0)}\frac{8D_0(\log S + 2)}{n} + \left(\frac{\mu^2 L^2 d}{\tau} + 6\mu^2 L d\right)\left(1 + 8\sqrt{\frac{5L}{n\tau}}\right)
\end{aligned} \tag{137}$$

Then we get the desired result for the zeroth-order case. Then we complete the proof. ■

Appendix D. Auxillary Lemmas

Lemma 16 (Coordinate-wise Gradient Estimator) *For all $x \in \mathcal{C}$, we have*

$$\|\hat{\nabla}_{coord} f(x) - \nabla f(x)\|^2 \leq \mu^2 L^2 d$$

Proof See [Ji et al. (2019), Appendix, Lemma 3]. ■

Lemma 17 *Suppose each $f_{i\in[n]}$ is L-smooth, for any $x, y \in \mathcal{C}$, we have*

$$\mathbb{E}\left[\|\nabla f_i(x) - \nabla f_i(y)\|^2\right] \leq 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y\rangle\right)$$

Proof Denote $\phi_i(x) = f_i(x) - f(y) - \langle \nabla f(y), x - y\rangle$. It is easy to verify that ϕ_i is also L-smooth. Clearly $\nabla\phi_i(y) = 0$ and hence $\min_{x\in\mathcal{C}} \phi_i(x) = \phi_i(y) = 0$. Then for $\alpha \in \mathbb{R}$, we have

$$\begin{aligned}\phi_i(y) &\le \min_{\alpha} \{\phi_i(x - \alpha\nabla\phi_i(x))\} \\ &\overset{①}{\le} \min_{\alpha} \left\{\phi_i(x) - \alpha\|\nabla\phi_i(x)\|^2 + \frac{L\alpha^2}{2}\|\nabla\phi_i(x)\|^2\right\} = \phi_i(x) - \frac{1}{2L}\|\nabla\phi_i(x)\|^2\end{aligned} \tag{138}$$

where ① comes from the smoothness of ϕ_i. Rearranging the terms and using the definition of ϕ_i we get

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \le 2L\left(f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y\rangle\right) \tag{139}$$

Taking expectation with respect to i, we get

$$\mathbb{E}\left[\|\nabla f_i(x) - \nabla f_i(y)\|^2\right] \le 2L\left(f(x) - f(y) - \langle \nabla f(y), x - y\rangle\right) \tag{140}$$

Then we complete the proof. ■

Lemma 18 *Suppose each $f_{i\in[n]}$ is L-smooth. Conditioning on $x_1, ..., x_{t-1}$*

- *For the first-order case, we have*

$$\mathbb{E}\left[\delta_t\right] = 0$$

and

$$\mathbb{E}\left[\|\delta_t - \mathbb{E}\left[\delta_t\right]\|^2\right] \le 2L\left[f(\tilde{x}) - f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\right]$$

- *For the zeroth-order case, we have*

$$\mathbb{E}\left[\delta_t\right] = \hat{\nabla}_{coord} f(\underline{x}_t) - \nabla f(\underline{x}_t) \neq 0$$

and

$$\mathbb{E}\left[\|\delta_t - \mathbb{E}\left[\delta_t\right]\|^2\right] \le 8L\left(f(\tilde{x}) - f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\right) + 12\mu^2 L^2 d$$

where the expectation is taken with respect to all variables.

Proof The part for the first-order case is proved in [Lan et al. (2019), Lemma 3]. Now we give a proof to the zeroth-order case. For the zeroth-order case, we have

$$\begin{aligned}\mathbb{E}\left[\delta_t\right] &= \mathbb{E}\left[\hat{\nabla}_{coord} f_{i_t}(\underline{x}_t) - \hat{\nabla}_{coord} f_{i_t}(\tilde{x}) + \tilde{g} - \nabla f(\underline{x}_t)\right] \\ &= \mathbb{E}\left[\hat{\nabla}_{coord} f_{i_t}(\underline{x}_t) - \nabla f(\underline{x}_t)\right] = \hat{\nabla}_{coord} f(\underline{x}_t) - \nabla f(\underline{x}_t)\end{aligned} \tag{141}$$

Then we prove the upper bound of $\mathbb{E}\left[\|\delta_t - \mathbb{E}\left[\delta_t\right]\|^2\right]$. We have

$$\begin{aligned}
&\mathbb{E}\left[\|\delta_t - \mathbb{E}\left[\delta_t\right]\|^2\right] \overset{①}{\leq} \mathbb{E}\left[\|\delta_t\|^2\right] \\
=&\mathbb{E}\Big[\|\left(\nabla f_{i_t}\left(\underline{x}_t\right) - \nabla f_{i_t}(\tilde{x}) - \left[\nabla f(\underline{x}_t) - \nabla f(\tilde{x})\right]\right) + \left(\hat{\nabla}_{coord} f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\underline{x}_t)\right) \\
&- \left(\hat{\nabla}_{coord} f_{i_t}(\tilde{x}) - \nabla f_{i_t}(\tilde{x})\right) + \left(\hat{\nabla}_{coord} f(\tilde{x}) - \nabla f(\tilde{x})\right)\|^2\Big] \\
\overset{②}{\leq}&4\mathbb{E}\Big[\|\nabla f_{i_t}\left(\underline{x}_t\right) - \nabla f_{i_t}(\tilde{x}) - \left[\nabla f(\underline{x}_t) - \nabla f(\tilde{x})\right]\|^2 + \|\hat{\nabla}_{coord} f_{i_t}(\underline{x}_t) - \nabla f_{i_t}(\underline{x}_t)\|^2 \\
&+\|\hat{\nabla}_{coord} f_{i_t}(\tilde{x}) - \nabla f_{i_t}(\tilde{x})\|^2 + \|\hat{\nabla}_{coord} f(\tilde{x}) - \nabla f(\tilde{x})\|^2\Big] \\
\overset{③}{\leq}&4\mathbb{E}\left[\|\nabla f_{i_t}\left(\underline{x}_t\right) - \nabla f_{i_t}(\tilde{x})\|^2\right] + 12\mu^2 L^2 d \\
\overset{④}{\leq}&8L\left(f(\tilde{x}) - f(\underline{x}_t) - \langle \nabla f(\underline{x}_t), \tilde{x} - \underline{x}_t\rangle\right) + 12\mu^2 L^2 d
\end{aligned} \tag{142}$$

where ① comes from $\mathbb{E}\left[\|x - \mathbb{E}\left[x\right]\|^2\right] = \mathbb{E}\left[\|x\|^2\right] - \mathbb{E}\left[x\right]^2 \leq \mathbb{E}\left[\|x\|^2\right]$, ② comes from the Cauchy-Schwarz inequality, ③ comes from $\mathbb{E}\left[\|x - \mathbb{E}\left[x\right]\|^2\right] \leq \mathbb{E}\left[\|x\|^2\right]$ and Lemma 16, ④ comes from Lemma 17. Then we complete the proof. ■

Appendix E. The STORC Algorithm

In this section. we include the STORC algorithm proposed by Hazan and Luo (2016) and its key theorems for completeness.

Algorithm E.3 STOchastic variance-Reduced Conditional gradient sliding (STORC)

Input: $x_0 \in \mathcal{C}, \{T_s\}, \{\gamma_{s,t}\}, \{\alpha_{s,t}\}, \{\eta_{s,t}\}$
Set $\tilde{x}^0 = x^0$.
for $s = 1, 2, ...$ **do**
  Set $x_0 = \bar{x}_0 = \tilde{x} = \tilde{x}^{s-1}$ and $\tilde{g} = \nabla f(\tilde{x})$
  Set $T = T_s$.
  for $t = 1, ..., T$ **do**
    Pick $\mathcal{I}_t \subset \{1, ..., n\}$ randomly with $|\mathcal{I}_t| = m_{s,t}$
    Set $\underline{x}_t = (1 - \alpha_{s,t})\bar{x}_{t-1} + \alpha_{s,t} x_{t-1}$
    Set $G_t = \frac{1}{m_{s,t}} \sum_{i \in \mathcal{I}_t} \left[\nabla f_i(\underline{x}_t) - \nabla f_i(\tilde{x}) + \tilde{g}\right]$
    $x_t = \mathrm{CondG}(G_t, x_{t-1}, 0, \gamma_{s,t}, 0, \eta_{s,t})$ *// Algorithm 1*
    $\bar{x}_t = (1 - \alpha_{s,t})\bar{x}_{t-1} + \alpha_{s,t} x_t$.
  end for
  Set $\tilde{x}^s = \bar{x}_t$.
end for

Theorem 19 (2 of Hazan and Luo (2016)) *With the following parameters (where D_s is defined later below):*

$$\alpha_{s,t} = \frac{2}{t+1}, \qquad \gamma_{s,t} = \frac{t}{3L}, \qquad \eta_{s,t} = \frac{2D_s^2}{3T_s}$$

Algorithm E.3 ensures $\mathbb{E}\left[f(\tilde{x}^S) - f(x^*)\right] \leq \frac{LD^2}{2^{S+1}}$ *if any of the following three cases holds:*

(a) $\nabla f(x^*) = 0$ *and* $D_s = D$, $T_s = \lceil 2^{s/2+2} \rceil$, $m_{s,t} = 900T_s$.

(b) f *is* G*-Lipschitz and* $D_s = D$, $T_s = \lceil 2^{s/2+2} \rceil$, $m_{s,t} = 700T_s + \frac{24T_sG(t+1)}{LD}$.

(c) f *is* τ*-strongly convex and* $D_s = \frac{LD^2}{\tau 2^{s-1}}$, $T_s = \lceil \sqrt{\frac{32L}{\tau}} \rceil$, $m_{s,t} = \frac{5600T_sL}{\tau}$.

From the following proof (especially (146)), we can see clearly how the decrease of $\alpha_{s,t}$ helps lower down the linear oracle complexity and raise the gradient query complexity.

Lemma 20 (3 of Hazan and Luo (2016)) *Suppose* $0 \leq D_s \leq D$ *is such that* $\mathbb{E}\left[\|\bar{x}_0 - x^*\|^2\right] \leq D_s^2$. *For any* t, *we have* $\mathbb{E}\left[f(\bar{x}_t) - f(x^*)\right] \leq \frac{8LD_s^2}{t(t+1)}$ *if* $\mathbb{E}\left[\|G_k - \nabla f(\underline{x}_k)\|^2\right] \leq \frac{L^2D_s^2}{T_s(k+1)^2}$ *for all* $k \leq t$.

Proof Since f is L-smooth, then we have

$$
\begin{aligned}
& f(\bar{x}_t) \leq l_f(\underline{x}_t, \bar{x}_t) + \frac{L}{2}\|\bar{x}_t - \underline{x}_t\|^2 \\
\overset{①}{=}& (1-\alpha_{s,t}) l_f(\underline{x}_t, \bar{x}_{t-1}) + \alpha_{s,t} l_f(\underline{x}_t, x^*) + \alpha_{s,t}\langle \nabla f(\underline{x}_t), x_t - x^* \rangle + \frac{L\alpha_{s,t}^2}{2}\|x_t - x_{t-1}\|^2 \\
\overset{②}{\leq}& (1-\alpha_{s,t}) f(\bar{x}_{t-1}) + \alpha_{s,t} f(x^*) + \alpha_{s,t}\langle \nabla f(\underline{x}_t), x_t - x^* \rangle + \frac{L\alpha_{s,t}^2}{2}\|x_t - x_{t-1}\|^2 \\
=& (1-\alpha_{s,t}) f(\bar{x}_{t-1}) + \alpha_{s,t} f(x^*) + \alpha_{s,t}\langle G_t, x_t - x^* \rangle + \frac{L\alpha_{s,t}^2}{2}\|x_t - x_{t-1}\|^2 + \alpha_{s,t}\langle \delta_t, x^* - x_t \rangle \\
\overset{③}{\leq}& (1-\alpha_{s,t}) f(\bar{x}_{t-1}) + \alpha_{s,t} f(x^*) + \frac{\alpha_{s,t}}{\gamma_{s,t}}\eta_{s,t} - \frac{\alpha_{s,t}}{\gamma_{s,t}}\langle x_t - x_{t-1}, x_t - x^* \rangle \\
& + \frac{L\alpha_{s,t}^2}{2}\|x_t - x_{t-1}\|^2 + \alpha_{s,t}\langle \delta_t, x^* - x_t \rangle \\
=& (1-\alpha_{s,t}) f(\bar{x}_{t-1}) + \alpha_{s,t} f(x^*) + \frac{\alpha_{s,t}}{\gamma_{s,t}}\eta_{s,t} + \frac{\alpha_{s,t}}{2\gamma_{s,t}}\left(\|x_{t-1} - x^*\|^2 - \|x_t - x^*\|^2\right) \\
& + \frac{\alpha_{s,t}}{2}\left[\left(L\alpha_{s,t} - \frac{1}{\gamma_{s,t}}\right)\|x_t - x_{t-1}\|^2 + 2\langle \delta_t, x_{t-1} - x_t \rangle + 2\langle \delta_t, x^* - x_{t-1} \rangle\right] \\
\overset{④}{\leq}& (1-\alpha_{s,t}) f(\bar{x}_{t-1}) + \alpha_{s,t} f(x^*) + \frac{\alpha_{s,t}}{\gamma_{s,t}}\eta_{s,t} + \frac{\alpha_{s,t}}{2\gamma_{s,t}}\left(\|x_{t-1} - x^*\|^2 - \|x_t - x^*\|^2\right) \\
& + \frac{\alpha_{s,t}}{2}\left[\frac{\gamma_{s,t}\|\delta_t\|^2}{1 - L\alpha_{s,t}\gamma_{s,t}} + 2\langle \delta_t, x^* - x_{t-1} \rangle\right]
\end{aligned}
\tag{143}
$$

where ① comes from the definition of $\underline{x}_t$ and x_t, ② comes from the convexity of f, ③ comes from Line 10 of Algorithm E.3, ④ comes from the fact that $b\langle u, v\rangle - a\|v\|^2/2 \leq b^2\|u\|^2/(2a)$. Note that $\mathbb{E}\left[\langle \delta_t, x^* - x_{t-1} \rangle\right] = 0$. So with the condition $\mathbb{E}\left[\|\delta_t\|^2\|\right] \leq \frac{L^2D_s^2}{T_s(k+1)^2} \overset{\text{def}}{=} \sigma_t^2$ we arrive at

$$
\begin{aligned}
& \mathbb{E}\left[f(\bar{x}_t) - f(x^*)\right] \\
\leq& (1-\alpha_{s,t})\mathbb{E}\left[f(\bar{x}_{t-1}) - f(x^*)\right] \\
& + \alpha_{s,t}\left[\frac{1}{\gamma_{s,t}}\eta_{s,t} + \frac{1}{2\gamma_{s,t}}\left(\mathbb{E}\left[\|x_{t-1} - x^*\|^2\right] - \mathbb{E}\left[\|x_t - x^*\|^2\right]\right) + \frac{\gamma_{s,t}\sigma_t^2}{2\left(1 - L\alpha_{s,t}\gamma_{s,t}\right)}\right]
\end{aligned}
\tag{144}
$$

Now we define $\Gamma_t = \Gamma_{t-1}(1-\alpha_{s,t})$ when $t > 1$ and $\Gamma_1 = 1$. By induction, one can verify $\Gamma_t = \frac{2}{t(t+1)}$ and the following:

$$\begin{aligned}&\mathbb{E}\left[f(\bar{x}_t) - f(x^*)\right]\\ \leq&\Gamma_t \sum_{k=1}^{t} \frac{\alpha_{s,k}}{\Gamma_k}\left[\frac{1}{\gamma_{s,k}}\eta_{s,k} + \frac{1}{2\gamma_{s,k}}\left(\mathbb{E}\left[\|x_{k-1}-x^*\|^2\right] - \mathbb{E}\left[\|x_k - x^*\|^2\right]\right) + \frac{\gamma_{s,k}\sigma_k^2}{2\left(1-L\alpha_{s,k}\gamma_{s,k}\right)}\right]\end{aligned} \tag{145}$$

which is at most

$$\begin{aligned}&\Gamma_t \sum_{k=1}^{t} \frac{\alpha_{s,k}}{\Gamma_k}\left[\frac{1}{\gamma_{s,k}}\eta_{s,k} + \frac{\gamma_{s,k}\sigma_k^2}{2\left(1-L\alpha_{s,k}\gamma_{s,k}\right)}\right]\\ &+ \frac{\Gamma_t}{2}\left[\frac{\alpha_{s,1}}{\gamma_{s,1}\Gamma_1}\mathbb{E}\left[\|x_0 - x^*\|^2\right] + \sum_{k=2}^{t}\left(\frac{\alpha_{s,k}}{\gamma_{s,k}\Gamma_k} - \frac{\alpha_{s,k-1}}{\gamma_{s,k-1}\Gamma_{k-1}}\right)\mathbb{E}\left[\|x_{k-1}-x^*\|^2\right]\right]\end{aligned} \tag{146}$$

Finally plugging in the parameters $\alpha_{s,k}, \gamma_{s,k}, \eta_{s,k}, \Gamma_k$ and the bound $\mathbb{E}\left[\|\bar{x}_0 - x^*\|^2\right] \leq D_s^2$ concludes the proof:

$$\mathbb{E}\left[f(\bar{x}_t) - f(x^*)\right] \leq \frac{2}{t(t+1)}\sum_{k=1}^{t} k\left[\frac{2LD_s^2}{T_s k} + \frac{LD_s^2}{2T_s(k+1)}\right] + \frac{3LD_s^2}{t(t+1)} \leq \frac{8LD_s^2}{t(t+1)} \tag{147}$$

■

In (146), the factor before $\eta_{s,k}$ is $\frac{1}{\gamma_{s,k}}$, which is $\mathcal{O}\left(\frac{1}{k}\right)$. Thus $\eta_{s,t}$ can be chosen $\mathcal{O}(t)$ larger, which leads to lower linear oracle complexity. However, the factor before the variance σ_k^2 is $\gamma_{s,k}$, which is $\mathcal{O}(k)$. Thus σ_k^2 has to be $\mathcal{O}(k)$ smaller. From (Hazan and Luo, 2016) we know σ_t^2 is proportional to $\frac{1}{m_{s,t}}$. Thus $m_{s,t}$ has to be chosen $\mathcal{O}(t)$ larger, which leads to higher gradient complexity.