ALEXR: An Optimal Single-Loop Algorithm for Convex Finite-Sum Coupled Compositional Stochastic Optimization

Bokun Wang* Tianbao Yang*

Abstract

This paper revisits a class of convex Finite-Sum Coupled Compositional Stochastic Optimization (cFCCO) problems with many applications, including group distributionally robust optimization (GDRO), learning with imbalanced data, reinforcement learning, and learning to rank. To better solve these problems, we introduce an efficient single-loop primal-dual block-coordinate proximal algorithm, dubbed ALEXR. This algorithm leverages **block-coordinate stochastic mirror ascent** updates for the dual variable and stochastic proximal gradient descent updates for the primal variable. We establish the convergence rates of ALEXR in both convex and strongly convex cases under smoothness and non-smoothness conditions of involved functions, which not only improve the best rates in previous works on smooth cFCCO problems but also expand the realm of cFCCO for solving more challenging non-smooth problems such as the dual form of GDRO. Finally, we present lower complexity bounds to demonstrate that the convergence rates of ALEXR are optimal among first-order block-coordinate stochastic algorithms for the considered class of cFCCO problems.

1 Introduction

In this paper, we focus on the following class of convex finite-sum coupled compositional optimization (cFCCO) problems:

$$\min_{x\in\mathcal{X}} F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(g_i(x)) + r(x), \quad \text{where } g_i(x) = \mathbf{E}_{\zeta_i\sim\mathbb{P}_i}[g_i(x;\zeta_i)], \tag{1.1}$$

where $\mathcal{X}\subset\mathbb{R}^d$ is a convex closed set, $g_i:\mathbb{R}^d\to\mathbb{R}^m$ is convex while $f_i:\mathbb{R}^m\to\mathbb{R}$ and $r:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$ are closed proper convex. The problem (1.1) is more challenging than empirical risk minimization in machine learning and conventional two-level stochastic compositional optimization (SCO) [1, 2] due to some unique challenges. Firstly, computing the gradient of each term $f_i(g_i(x))$ poses difficulties because the inner function g_i is in an expectation form. Therefore, existing algorithms based on stochastic gradient descent do not apply to (1.1). Second, the FCCO problems involve a substantial number of inner functions g_i coupled with the outer summation index i, making FCCO distinct from previous SCO problems with a single inner function [1, 2].

The problem (1.1) is closely related to the empirical X-risk minimization introduced in [3] to formulate many objectives in machine learning [4, 5, 6, 7, 8]. Several existing algorithms have been proposed to solve FCCO problems with provable convergence guarantees [4, 5, 9, 10, 11, 12].

*Department of Computer Science and Engineering, Texas A&M University, College Station, TX.
Correspondence to: bokun-wang@tamu.edu, tianbao-yang@tamu.edu

However, most existing results are devoted to non-convex FCCO where the functions f_i and g_i are non-convex, yielding slow convergence to stationary points. Despite the global convergence guarantees established for convex problems in [5, 10], these results do not simultaneously achieve three crucial desiderata: (i) **optimal rate** in terms of the accuracy level ϵ of the objective gap or the distance to optimal solution; (ii) **parallel speed-up** through both inner and outer mini-batches; (iii) **single-loop** algorithmic design. In particular, the analysis of the SOX algorithm in [5] either lacks parallel speed-up in terms of the inner batch size or only achieves a sub-optimal rate in terms of ϵ for convex problems. The double-loop algorithm MSVR [10], while achieving the optimal rate in terms of ϵ for the objective gap of a convex objective, exhibits only partial parallel speed-up with the inner mini-batch size (see Table 1). Furthermore, their convergence results do not hold when either f_i or g_i is a non-smooth function, limiting their applicability to broader problems.

The overarching goal of this paper is to design an algorithm to attain the three nice properties mentioned above. Under the non-decreasing monotonicity and Lipschitz continuity of f_i (see Section 4.2), we can reformulate (1.1) into a convex-concave min-max problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} L(x,y) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big[\Big\langle y^{(i)}, g_i(x)\Big\rangle - f_i^*(y^{(i)})\Big] + r(x), \tag{1.2}$$

where f_i^* the convex conjugate of f_i, $y^{(i)} \in \mathcal{Y}_i \subseteq \mathbb{R}_+^m$ is the i-th block of y, $\mathcal{Y}_i$ is convex and compact, and $\mathcal{Y} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n \subseteq \mathbb{R}_+^{nm}$. To solve the above problem, we propose a primal-dual block-coordinate proximal algorithm named ALEXR[1] to efficiently solve (1.2). This is motivated by state-of-the-art primal-dual algorithms for empirical risk minimization with linear models [13, 14] and for convex-concave min-max optimization problems [15]. However, (1.2) possesses unique characteristics that make it non-trivial to extend existing algorithms and analysis to solving (1.2): (i) the objective in (1.2) is not necessarily bilinear as in [13, 14]; (ii) it is prohibitive to access the stochastic gradient for all dual variables $\{y_1, \ldots, y_n\}$, which is different from [15] assuming the stochastic gradient for y is given at each iteration. The key steps of ALEXR are the following block-coordinate stochastic mirror ascent update for the dual variable and stochastic proximal gradient descent update for the primal variable:

$$\begin{aligned}
\tilde{g}_t^{(i)} &= g_i(x_t; \mathcal{B}_t^{(i)}) + \theta(g_i(x_t; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}; \mathcal{B}_t^{(i)})), \quad \forall i \in \mathcal{S}_t \\
y_{t+1}^{(i)} &= \begin{cases} \arg\max_{y^{(i)}\in\mathcal{Y}_i} \left\{ y^{(i)}\tilde{g}_t^{(i)} - f_i^*(y^{(i)}) - \tau U_{\psi_i}(y^{(i)}, y_t^{(i)}) \right\}, & \text{if } i \in \mathcal{S}_t \\ y_t^{(i)} & \text{o.w.} \end{cases} \\
x_{t+1} &= \arg\min_{x\in\mathcal{X}} \left\{ \left\langle \frac{1}{S}\sum_{i\in\mathcal{S}_t} [\nabla g_i(x_t; \tilde{\mathcal{B}}_t^{(i)})]^\top y_{t+1}^{(i)}, x \right\rangle + r(x) + \frac{\eta}{2}\|x - x_t\|_2^2 \right\},
\end{aligned} \tag{1.3}$$

where $\theta \in (0,1]$, $\mathcal{S}_t \subset \{1,2,\ldots,n\}$ refers to the outer mini-batch, $\mathcal{B}_t^i$ and $\tilde{\mathcal{B}}_t^i$ are two independent inner mini-batches sampled from $\mathbb{P}_i$ for each $i \in \mathcal{S}_t$, ψ_i is a convex distance-generating function, the prox-function associated with a distance-generating function ψ_i is defined as $U_{\psi_i}(u,v) = \psi_i(u) - \psi_i(v) - \psi_i'(v)(u-v)$ for $u, v \in \mathbb{R}^m$, $\psi_i'(v) \in \partial\psi_i(v)$. ALEXR has some interesting connections with existing algorithms for FCCO and convex-concave problems, including SOX [5], MSVR [10] and SAPD [15], which will be discussed in subsection 5.1. Let $S = |\mathcal{S}_t|$ be the outer mini-batch size and $B = |\mathcal{B}_t^i| = |\tilde{\mathcal{B}}_t^i|$ be the inner mini-batch size.

Our contributions can be summarized as follows:

[1]Instead of naming our algorithm based on the techniques used, we name it based on what problems it can address. In particular, ALEXR means **A**lgorithms for **L**earning with **E**mpirical **X**-**R**isks.

• We introduce a single-loop primal-dual block-coordinate algorithm called ALEXR to tackle (1.1), which requires only $O(1)$ oracles and $O(d)$ computational cost per iteration.
• For cFCCO problems with μ-strongly convex r and smooth f_i, g_i, our ALEXR requires $T = O\left(\frac{n}{S} + \frac{1}{\mu} + \frac{\sqrt{n}}{\sqrt{S\mu}} + \frac{\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon} + \frac{n\sigma_0^2}{BS\epsilon}\right)$ iterations to achieve the ϵ level of distance gap[2], where $\sigma_0^2, \sigma_1^2, \delta^2$ are variances. For non-strongly convex cFCCO problems with smooth g_i and possibly non-smooth f_i, ALEXR requires $T = O\left(\frac{\sqrt{n}}{\sqrt{S}\epsilon} + \frac{\sigma_1^2}{B\epsilon^2} + \frac{\delta^2}{S\epsilon^2} + \frac{n\sigma_0^2}{BS\epsilon^2}\right)$ iterations to achieve the ϵ level of objective gap. In both convex and strongly convex cases, the convergence rates of our ALEXR improve upon the rates in previous works [5, 10] on cFCCO problems with smooth f_i, g_i (see Table 1 for a detailed comparison). Besides, we also provide the convergence rates of ALEXR for cFCCO problems with non-smooth g_i in both convex and strongly convex cases.
• For cFCCO problems with the non-smooth outer function f_i, we present a lower complexity bound to show that an abstract first-order update scheme with S oracles per iteration (covering our ALEXR and previous algorithms SOX [5], MSVR [10] with $B = 1$ as special cases) requires at least $T = \Omega\left(\frac{n\sigma_0^2}{S\epsilon^2}\right)$ iterations to achieve the ϵ level of objective gap. For cFCCO problems with the smooth outer function f_i and strongly convex $r(x)$, we also that any algorithm in the abstract first-order update scheme requires at least $T = \Omega\left(\frac{1}{\mu\epsilon} \vee \frac{n\sigma_0^2}{S\epsilon}\right)$ iterations to achieve the ϵ level of the distance gap to the optimum. Thus, the convergence rate of ALEXR is optimal among first-order stochastic algorithms for cFCCO problems, in terms of n and ϵ.

2 Applications

In this section, we present several motivating examples of the cFCCO problem in (1.1) and its special case, where the distribution $\mathbb{P}_i$ is independent of i, denoted as $\mathbb{P}_i \equiv \mathbb{P}$ for all i in the set $[n]$.

2.1 Group Distributionally Robust Optimization

Machine learning models are typically trained through the process of empirical risk minimization (ERM), which often results in high average accuracy on similarly distributed test data. However, models with high *average* accuracy may perform poorly on some rare sub-populations. The Group Distributionally Robust Optimization (GDRO) framework was proposed to tackle this problem [16]. Suppose that there are n predefined groups and the data distribution of the i-th group is $\mathbb{P}_i$. The ϕ-divergence (Csiszár divergence) penalized GDRO problem can be formulated as

$$\min_{w\in\mathcal{W}} \mathcal{L}(w) := \max_{q\in\Delta_n} \left\{\sum_{i=1}^n \left(q^{(i)} R_i(w) - \frac{\lambda}{n}\phi(nq_i)\right)\right\} + r(w), \ R_i(w) := \mathbf{E}_{z\sim\mathbb{P}_i}[\ell(w;z)], \tag{2.1}$$

where w is the model parameter, $R_i(w)$ is the expected loss of the i-th group, domain $\mathcal{W} \subset \mathbb{R}^d$ is convex compact, penalty $\lambda \geq 0$, generator $\phi : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$, $\phi(1) = 0$, and Δ_n is the probability simplex in $\mathbb{R}^n$. Several prior works [16, 17] discarded the ϕ-divergence penalty ($\lambda = 0$ in (2.1)) and consider the problem $\min_{w\in\mathcal{W}} \max_{i\in[n]} R_i(w)$, which minimizes the risk of the *worst* group. However, the model trained through worst-group risk minimization may be vacuous if the worst group is an outlier. Moreover, the sizes of groups may follow a long-tailed distribution such that multiple rare groups exist. To resolve these issues, we choose $\lambda > 0$ and consider the penalized GDRO problem with CVaR divergence $\phi = \mathbb{I}_{[0,\alpha^{-1}]}$ or χ^2-divergence $\phi(t) = \frac{1}{2}(t-1)^2$. The challenges of directly

[2] See Table 1 for the definition of distance gap.

solving (2.1) using stochastic min-max algorithms lie in estimating the stochastic gradient of q and controlling its variance when $n \gg 1$ is large [17]. To address these issues, we transfer the above problem into the following dual form by the duality relationship [18]:

$$\min_{w\in\mathcal{W}, c\in[\underline{c},\overline{c}]} F(w,c), \;\; F(w,c) = \frac{\lambda}{n}\sum_{i=1}^{n}\phi^*\left(\frac{R_i(w)-c}{\lambda}\right) + c + r(w), \tag{2.2}$$

where $R_i(w)$ is convex and ϕ^* is monotonically non-decreasing, e.g., $\phi^*(u) = \frac{1}{\alpha}(u)_+$ for CVaR divergence and $\phi^*(u) = \frac{1}{4}(u+2)_+^2 - 1$ for χ^2-divergence, $(\cdot)_+ \coloneqq \max(\cdot, 0)$. Indeed, for any $(w_{\text{out}}, c_{\text{out}})$ that satisfies $\mathbf{E}[F(w_{\text{out}}, c_{\text{out}}) - \min_{w,c} F(w,c)] \leq \epsilon$ and an optimal solution $w_* \in \arg\min_{w\in\mathcal{W}} \mathcal{L}(w)$ to (2.1), we have $\mathcal{L}(w_{\text{out}}) - \mathcal{L}(w_*) = \min_c F(w_{\text{out}}, c) - \min_c F(w_*, c) \leq F(w_{\text{out}}, c_{\text{out}}) - \min_{w,c} F(w,c)$ (see Section A.1.2 in Levy et al. [18]). Thus, an approximate solution to the dual formulation (2.2) also leads to an approximate solution to the original problem (2.1). The dual formulation in (2.2) is recognized as a difficult open problem in Sagawa et al. [16] due to the biased stochastic estimator (refer to footnote 4 in their paper). In this work, we can solve the problem in (2.2) by viewing it as a cFCCO problem with a convex outer function $f_i(\cdot) = \lambda\phi^*(\cdot)$ and an inner function $g_i(x) = (R_i(w) - c)/\lambda$ that is jointly convex to $x = (w, c)$. In Section 7, we provide the convergence rates and per-iteration computational costs of our new algorithm ALEXR for solving (2.2), in comparison to existing algorithms specifically designed for addressing the GDRO problem. Our algorithm only requires sampling $O(1)$ groups and $O(1)$ samples and does not involve handling expensive dual projection onto (constrained) $(n-1)$-dimensional simplex, yet enjoy competitive performance compared with stochastic min-max algorithms for solving (2.1) directly.

2.2 Partial AUC Maximization with Restricted True Positive Rate

The Area Under the ROC Curve (AUC) is acknowledged as a more informative metric than accuracy for assessing the performance of binary classifiers in the context of imbalanced data [19]. In scenarios influenced by diagnostic or monetary considerations, the primary objective may be to maximize the partial AUC (pAUC) with a specified lower bound α for the true positive rate (TPR). As shown in [6, 3], a surrogate objective for maximizing pAUC with restricted TPR is formulated as

$$\min_{w\in\mathbb{R}^d} \frac{1}{n_+ n_-} \sum_{a_i\in\mathcal{S}_+^{\uparrow}[1, n_+(1-\alpha)]} \sum_{a_j\in\mathcal{S}_-} L(w; a_i, a_j), \tag{2.3}$$

Here $\mathcal{S}_+, \mathcal{S}_-$ are the sets of positive/negative data, w refers to the model and $L(w; a_i, a_j) = \ell(h_w(a_j) - h_w(a_i))$ represents a continuous pairwise surrogate loss, where $h_w(a_i)$ denotes the prediction score for data a_i. Additionally, $\mathcal{S}_+^{\uparrow}[1,k]$ the bottom-k positive data based on the prediction scores. In particular, ℓ is a convex and monotonically non-decreasing function, ensuring the consistency of the surrogate objective [20]. Following Lemma 7 of Zhu et al. [6], pAUC maximization with restricted TPR$\geq \alpha$ is equivalent to

$$\max_{w\in\mathbb{R}^d} \max_{\substack{y\in\Delta_{n_+} \\ y^{(i)}\leq \frac{1}{n_+(1-\alpha)}}} \sum_{a_i\in\mathcal{S}_+} y^{(i)} L(w; a_i, \mathcal{S}_-), \quad L(w; a_i, \mathcal{S}_-) \coloneqq \frac{1}{n_-}\sum_{a_j\in\mathcal{S}_-} L(w; a_i, a_j), \tag{2.4}$$

which can be transformed into its dual form:

$$\min_{w\in\mathbb{R}^d, s\in\mathbb{R}} \frac{1}{n_+(1-\alpha)} \sum_{a_i\in\mathcal{S}_+} \left(\frac{1}{n_-}\sum_{a_j\in\mathcal{S}_-} L(w; a_i, a_j) - s\right)_+ + s, \tag{2.5}$$

where $(\cdot)_+ \coloneqq \max(\cdot, 0)$ is monotonically non-decreasing and convex and $\frac{1}{n_-}\sum_{a_j\in\mathcal{S}_-} L(w; a_i, a_j) - s$ is jointly convex to (w, s). Thus, the problem in (2.5) is a cFCCO problem.

2.3 Other Applications

In addition to GDRO and pAUC, there are many other intriguing applications of the cFCCO problem.

• *Robust Logistic Regression:* Consider a collection of data-label pairs, denoted as $(a_i, b_i)_{i=1}^n$. We formulate the robust logistic regression problem as $\min_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \log(1+\exp b_i \mathbf{E}[\mathcal{A}(a_i)^\top x \mid a_i]) + r(x)$. In this formulation, $\mathcal{A}(a_i)$ represents the perturbed data generated from an underlying distribution $\mathbb{P}_i$. This problem aligns with the structure of (1.1), where the functions $f_i(\cdot)$ are convex and monotonically non-decreasing given by $f_i(\cdot) = \log(1+\exp(b_i\cdot))$, and $g_i(x) = \mathbf{E}_{A(a_i)\sim\mathbb{P}_i}[\mathcal{A}(a_i)^\top x]$.

• *Bellman Residual Minimization:* The task of approximating the value function, denoted as $V^\pi(s)$, for each state s under policy π using a linear mapping can be expressed as $\min_{x\in\mathcal{X}} \sum_{s=1}^S (\phi_s^\top x - \sum_{s'} \mathbf{P}^\pi_{s,s'}[r_{s,s'} + \gamma\cdot\phi_{s'}^\top x])^2$. In this formulation, ϕ_s and $\phi_{s'}$ are feature vectors representing states s and s', respectively. Additionally, $r_{s,s'}$ represents the random reward obtained during the transition from state s to s', $\gamma < 1$ is the discount factor, π denotes the policy, and $\mathbf{P}^\pi_{s,s'}$ represents the probability of transitioning from state s to s' under policy π. This problem can be formulated as (1.1), where the functions $f_s(\cdot)$ are convex and given by $f_s(\cdot) = \frac{1}{S}(\cdot)^2$, and the affine function $g_s(x) = \phi_s^\top x - \sum_{s'} \mathbf{P}^\pi_{s,s'}[r_{s,s'} + \gamma\cdot\phi_{s'}^\top x]$.

• *Bipartite Ranking for Classification or Retrieval:* Imbalanced data classification is usually tackled in the context of the bipartite ranking problem. There is often a desire to penalize those positive examples with lower scores. One approach is the p-norm push, introduced by Rudin et al. [21]. It formulates the problem as $\min_{x\in\mathcal{X}} \frac{1}{n_+}\sum_{a_i\in\mathcal{D}_+}\left(\frac{1}{n_-}\sum_{a_j\in\mathcal{D}_-}\ell(s_x(a_j)-s_x(a_i))\right)^p + r(x)$, $p\geq 1$. Here, $\mathcal{D}_+$ and $\mathcal{D}_-$ represent positive and negative data sets. The function $s_x(a)$ denotes the ranking score of data point a, which is determined by a linear model parameterized by x. The loss function ℓ is non-negative, convex, and monotonically non-decreasing, for instance, $\ell(\cdot) = \exp(\cdot)$. The p-norm push method is in the structure of (1.1), where the functions $f_i(\cdot)$ are convex and monotonically non-decreasing and given by $f_i(\cdot) = (\cdot)^p$, and the convex function $g_i(x) = \frac{1}{n_+}\sum_{a_j\in\mathcal{D}_+}\ell(s_x(a_j)-s_x(a_i))$. One popular approach for retrieval problems is maximizing the precision or recall at top k positions (prec/rec@k). Yang [3] has formulated the problem as $\min_{x\in\mathcal{X}} \frac{1}{n_+}\sum_{a_i\in\mathcal{D}_+}\ell_1(\sum_{a_j\in\mathcal{D}_+\cup\mathcal{D}_-}\ell_2(s_x(a_j) - s_x(a_i) - k)) + r(x)$, where ℓ_1, ℓ_2 are monotonically non-decreasing convex surrogate losses of the zero-one loss. Hence, maximizing precision or recall at top k positions with a convex model $s_x(a)$ is covered by (1.1).

• *Multi-Task GDRO:* The Group Distributionally Robust Optimization (GDRO) problem can be extended to the multi-task setting. Consider a scenario with n tasks and m groups. We represent the data distribution for the i-th task and the j-th group as $\mathbb{P}_{i,j}$. Additionally, let $\ell(x;z)$ be the loss function associated with parameter x on data point z. The Multi-Task GDRO, with a regularization term r, is formulated as $\min_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \max_{j\in[m]} \mathbf{E}[\ell(x;z_{ij})] + r(x)$. In this formulation, the functions $f_i(\cdot)$ are defined as $f_i(g_i) = \max_{j\in[m]}(g_{ij})$, and $g_{ij}(x) = \mathbf{E}[\ell(x;z_{ij})]$, where $g_i(x) = [g_{i1}(x), \ldots, g_{im}(x)]$. Alternatively, we may consider the smooth $f_i(g_i) = \log\sum_{j\in[m]}\exp(g_{ij})$. This problem fits within the structure of (1.1) and is particularly relevant when dealing with a scenario featuring a substantial number n of tasks, such as identity prediction in human faces, with a limited number m of groups (e.g. lightning conditions).

3 Related Work

Problem (1.1) or its min-max reformulation (1.2) is closely related to several widely studied problems.

Table 1: Comparison of **iteration complexities** and per-iteration #oracles for achieving ϵ-optimal solution of (1.1) with **smooth** g_i in terms of some optimality gap, where "Dist." denotes the distance gap $\mathbf{E}\frac{\mu}{2}\|x_{\text{out}} - x_*\|_2^2 \leq \epsilon$ in the strongly convex case, "Obj." denotes the objective gap $\mathbf{E}[F(x_{\text{out}}) - F(x_*)] \leq \epsilon$, and "Gap" denotes the duality gap $\mathbf{E}[\text{Gap}(\text{x}_{\text{out}}, \text{y}_{\text{out}})] = \mathbf{E}\max_{x,y}\{L(x_{\text{out}}, y) - L(x, y_{\text{out}})\} \leq \epsilon$, "W-Gap" denotes a weak duality gap $\max_{x,y}\mathbf{E}\{L(x_{\text{out}}, y) - L(x, y_{\text{out}})\} \leq \epsilon$. Here x_{out} (probably also y_{out}) is the output of each algorithm. We hide other constant quantities except for n, variances $\sigma_0^2, \sigma_1^2, \delta^2$, modulus of strong convexity μ, and batch sizes B, S. "-" means that the result is missing. Besides, $\tilde{O}$ hides poly $\log(1/\epsilon)$ factors.

Algorithm	#Oracles[(1)]	Strongly Convex r Smooth f_i		Convex r		Single-Loop?
		Complexity	Metric	Complexity	Metric	
ASC-PG [(2)][2]	$O(n)/O(1)$	$O\left(\frac{1}{\mu\epsilon^{1.25}}\right)$	Dist.	$O\left(\frac{1}{\epsilon^{3.5}}\right)$[(3)]	Obj.	✓
SAPD [(4)][15]	$O(n)/O(1)$	$\tilde{O}\left(\frac{1}{\mu} + \frac{1}{\sqrt{n\mu}} + \frac{\sigma_0^2}{\epsilon} + \frac{\sigma_1^2+\delta^2}{\mu\epsilon}\right)$	Dist.	$O\left(\frac{1}{\epsilon} + \frac{\sigma_0^2+\sigma_1^2+\delta^2}{\epsilon^2}\right)$	Gap	✓
SSD [22]	$O(n)/O(1)$	$O\left(\frac{1}{\sqrt{\epsilon}} + \frac{\sigma_0^2}{\epsilon} + \frac{\sigma_1^2+\delta^2}{\mu\epsilon}\right)$[(5)]	Dist.	$O\left(\frac{1}{\sqrt{\epsilon}} + \frac{\sigma_0^2+\sigma_1^2+\delta^2}{\epsilon^2}\right)$[(3)] $O\left(\frac{1}{\epsilon} + \frac{\sigma_0^2+\sigma_1^2+\delta^2}{\epsilon^2}\right)$	Obj.	✓
BSGD [(2)][23]	$O\left(\frac{1}{\epsilon}\right)/O\left(\frac{1}{\epsilon}\right)$[(6)]	$O\left(\frac{1}{\mu\epsilon}\right)$	Obj.[(7)]	$O\left(\frac{1}{\epsilon^2}\right)$	Obj.	✓
MSVR [(2)][10]	$O(1)/O(1)$	$O\left(\frac{n}{\mu\sqrt{BS}\epsilon}\right)$	Obj.[(7)]	$O\left(\frac{n}{\sqrt{BS}\epsilon^2}\right)$[(3)]	Obj.	✗
SOX-Boost [(2)][5]	$O(1)/O(1)$	$O\left(\frac{1}{\mu\min\{B,S\}\epsilon} + \frac{n\sigma_0^2}{\mu^2 BS\epsilon}\right)$	Obj.	$O\left(\frac{1}{\min\{B,S\}\epsilon^2} + \frac{n\sigma_0^2}{BS\epsilon^3}\right)$[(3)]	Obj.	✗
SOX [5]	$O(1)/O(1)$	$\tilde{O}(\frac{n}{S\mu\epsilon})$	Dist.	$O\left(\frac{n}{S\epsilon^2}\right)$	W-Gap	✓
ALEXR (**This Work**)	$O(1)/O(1)$	$\tilde{O}\left(\frac{1}{\mu} + \frac{\sqrt{n}}{\sqrt{S\mu}} + \frac{n\sigma_0^2}{BS\epsilon} + \frac{\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon}\right)$	Dist.	$O\left(\frac{\sqrt{n}}{\sqrt{S}\epsilon} + \frac{n\sigma_0^2}{BS\epsilon^2} + \frac{\sigma_1^2}{B\epsilon^2} + \frac{\delta^2}{S\epsilon^2}\right)$[(8)]	Obj.	✓

[(1)] Representing the number of zeroth-order oracles for $g_i(x)$ / the number of first-order oracles for $\nabla g_i(x)$ in each iteration.
[(2)] Under the assumption that $F(x)$ is convex, which is slightly weaker than the layerwise convexity assumption stated in Section 4.2.
[(3)] Requiring smoothness of f_i.
[(4)] For general convex-concave $L(x, y)$. It does not need that $L(x, y)$ is linear in y or $\mathcal{Y}$ is block-separable over $i \in [n]$.
[(5)] The $O\left(1/\sqrt{\epsilon}\right)$ term can be improved to $\tilde{O}(1/\sqrt{\mu})$ by the restarting technique, which makes SSD a double-loop algorithm.
[(6)] Requiring smoothness of f_i. If f_i is non-smooth, the number of oracles per iteration increases to $O\left(\frac{1}{\epsilon^2}\right)$
[(7)] In strongly convex stochastic optimization problem, converting the $O(1/(\mu\epsilon))$ convergence rate in terms of "Obj." can be converted to the optimal $O(1/(\mu\epsilon))$ rate [24, 25, 26] for first-order methods in terms of "Dist." by the strong convexity of F.
[(8)] This result requires the distance-generating function ψ_i to be smooth.

3.1 History of FCCO

The FCCO problem was first introduced in [4] for optimizing average precision (AP) to address the large batch size issue of previous stochastic algorithms for AP maximization. Later, it was used for solving a wide range of problems in the field of machine learning, including optimizing listwise losses for learning to rank [7], optimizing surrogate losses of partial areas under the curves for imbalanced data classification [6], and optimizing global contrastive losses for contrastive self-supervised learning [8, 27].

Qi et al. [4] proposed an algorithm SOAP and analyzed its convergence for solving a non-convex surrogate loss of AP in the form of FCCO. Wang et al. [9] adopted the momentum update to accelerate the convergence rate of SOAP for AP maximization, improving the iteration complexity from $O\left(\frac{n}{\epsilon^5}\right)$ to $O\left(\frac{n}{S\epsilon^4}\right)$ for finding an ϵ-stationary point. The work by Wang and Yang [5] was the first to study the general form of FCCO and proposed SOX to further improve the rate by enjoying the parallel speed-up of using inner and outer mini-batches. For non-convex and smooth problems, SOX has a convergence rate of $O\left(\frac{n}{BS\epsilon^4}\right)$ for finding an ϵ-stationary solution. Jiang et al. [10] proposed a new variance reduction technique (MSVR) for tracking and estimating multiple inner functions g_i by accessing only a constant number of samples per iteration. For non-convex

Table 2: Comparison of **iteration complexities** and per-iteration #oracles for achieving ϵ-optimal solution of (1.1) with **non-smooth** g_i in terms of some optimality gap, where Dist. denotes the distance gap $\mathbf{E}\frac{\mu}{2}\|x_{\text{out}} - x_*\|_2^2 \leq \epsilon$ in the strongly convex case, Obj. denotes the objective gap $\mathbf{E}[F(x_{\text{out}}) - F(x_*)] \leq \epsilon$. Here x_{out} is the output of each algorithm. We hide other constant quantities except for n, variances $\sigma_0^2, \sigma_1^2, \delta^2$, modulus of strong convexity μ, and batch sizes B, S. "-" means that the result is missing. Besides, $\tilde{O}$ hides $\text{poly}\log(1/\epsilon)$ factors.

Algorithm	#Oracles[(1)]	Strongly Convex r Smooth f_i		Convex r		Single-Loop?
		Complexity	Metric	Complexity	Metric	
SCGD [(2)][1]	$O(n)/O(1)$	$O\left(\frac{1}{\mu^{4.5}\epsilon^{1.5}}\right)$	Dist.	$O\left(\frac{1}{\epsilon^4}\right)$	Obj.	✓
nSSD [22]	$O(n)/O(1)$	-	-	$O\left(\frac{1+\sigma_0^2+\sigma_1^2+\delta^2}{\epsilon^2}\right)$	Obj.	✓
ALEXR (**This Work**)	$O(1)/O(1)$	$\tilde{O}\left(\frac{\sqrt{n}}{\sqrt{S\mu}} + \frac{n\sigma_0^2}{BS\epsilon} + \frac{\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon} + \frac{1}{\mu\epsilon}\right)$	Dist.	$O\left(\frac{\sqrt{n}}{\sqrt{S}\epsilon} + \frac{n\sigma_0^2}{BS\epsilon^2} + \frac{\sigma_1^2}{B\epsilon^2} + \frac{\delta^2}{S\epsilon^2} + \frac{1}{\epsilon^2}\right)$[(3)]	Obj.	✓

[(1)] Representing the number of zeroth-order oracles for $g_i(x)$ / the number of first-order oracles for $\nabla g_i(x)$ in each iteration.
[(2)] Under the assumption that $F(x)$ is convex, which is slightly weaker than the layerwise convexity assumption stated in Section 4.2.
[(3)] This result requires the distance-generating function ψ_i to be smooth.

problems, they improve the complexity to $O\left(\frac{n}{\sqrt{B}S\epsilon^3}\right)$ for finding an ϵ-stationary point.

To establish the convergence for convex FCCO problems, the two studies [5, 10] have used the restarting trick to boost the convergence rate for finding an ϵ-optimal solution of strongly convex and convex problems. In particular, restarted SOX (named SOX-boost) suffers rates of $O\left(\frac{n}{BS\mu^2\epsilon}\right)$ for strongly convex problems and $O\left(\frac{n}{BS\epsilon^3}\right)$ for convex problems, and restarted MSVR suffers rates of $O\left(\frac{n}{S\sqrt{B}\mu\epsilon}\right)$ for strongly convex problems and $O\left(\frac{n}{S\sqrt{B}\epsilon^2}\right)$ for convex problems, where $\mu > 0$ is the strong convexity parameter. It is notable that SOX-boost achieves full mini-batch speedup but non-optimal rates, while MSVR enjoys the optimal rates in terms of μ and ϵ but has only partial speedup with B. Both algorithms only require $O(1)$ oracles per iteration. Under the slightly stronger assumption (f is convex and monotonically non-decreasing while g is convex), Wang and Yang [5] follow the technique in [22] to reformulate the FCCO as the saddle point problem in (1.2). Subsequently, they propose a randomized block-coordinate variant of SCGD [1]. This variant, similar to SOX-boost, demands merely $O(1)$ oracles per iteration and $T = O\left(\frac{n}{S\epsilon^2}\right)$ iterations to find a $(\bar{x}, \bar{y})$ satisfying $\max_{x,y} \mathbf{E}[L(\bar{x}, y) - L(x, \bar{y})] \leq \epsilon$ for the convex FCCO problem. However, it is important to note that achieving $\max_{x,y} \mathbf{E}[L(\bar{x}, y) - L(x, \bar{y})] \leq \epsilon$ does not guarantee $\mathbf{E}[F(\bar{x}) - F(x_*)] \leq \epsilon$, as demonstrated in Section 3.1.1 of Alacaoglu et al. [14]. Consequently, the convergence criterion in terms of $\max_{x,y} \mathbf{E}[L(\bar{x}, y) - L(x, \bar{y})]$ is considered to be relatively weak and is not of primary interest. It is worth mentioning that the previous results [4, 9, 5, 10] restrict to those problems in which both f_i and g_i are smooth. In this work, we also consider non-smooth f_i and non-smooth g_i.

Advantage of Single-Loop Algorithms: It is worth noting that some algorithms such as MSVR and SOX-boost contain nested inner loops to solve some subproblems inexactly with high accuracy. However, the termination criterion of each inner loop depends on problem-specific unknown constants. Thus, *single-loop* algorithms are easier to implement than multi-loop counterparts.

3.2 Convex Stochastic Compositional Problem

The stochastic compositional optimization (SCO) problem takes the form of $\min_{x\in\mathbb{R}^d} F(x)$, with $F(x) = \mathbf{E}_\xi\left[f(\mathbf{E}_\zeta\left[g(x;\zeta)\right];\xi)\right]$, where ζ and ξ are mutually independent. In this context, when F is μ-convex ($\mu \geq 0$) and f is smooth, Wang et al. [1] have introduced a stochastic method named SCGD, which achieves a convergence rate of $O\left(\frac{1}{\epsilon^4}\right)$ for convex problems, ensuring that $\mathbf{E}[F(\bar{x}) - F(x_*)] \leq \epsilon$. For strongly convex problems, it requires $O\left(\frac{1}{\mu^2\epsilon^{1.5}}\right)$ iterations to reach $\frac{\mu}{2}\mathbf{E}\left\|\bar{x} - x_*\right\|_2^2 \leq \epsilon$. Further exploiting the smoothness of function g, Wang et al. [2] proposed ASC-PG, which improves the convergence rate to $O(1/\epsilon^{3.5})$ for convex problems and $O\left(\frac{1}{\mu\epsilon^{1.25}}\right)$ for strongly convex problems. Lian et al. [28] consider the finite-sum SCO problem $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(\frac{1}{m}\sum_{j=1}^m g_j(x))$ and utilizes the technique from SVRG [29] to obtain linear convergence for strongly convex F. Similar to SVRG, the algorithm presented in [28] follows a double-loop structure and requires full gradient evaluations at the start of each outer-loop iteration. Notably, the algorithms designed for the SCO problem can also be applied to tackle the FCCO problem, as discussed in Appendix G.1. However, it is important to underscore that all these algorithms require $O(n)$ oracles per iteration, which can become computationally demanding when n takes on large values.

Building upon a slightly stronger assumption than that composition F is convex, specifically that f is convex and monotonically non-decreasing while g is convex, Zhang and Lan [22] reformulate SCO problem as a convex-concave min-max-max problem and propose the stochastic sequential dual (SSD) algorithm to obtain the optimal $O\left(\frac{1}{\sqrt{\epsilon}} + \frac{\sigma^2}{\epsilon^2}\right)$ rate in terms of $\mathbf{E}[F(\bar{x}) - F(x_*)]$ in the convex case and $O\left(\frac{1}{\sqrt{\epsilon}} + \frac{\sigma^2}{\mu\epsilon}\right)$ rate in terms of $\frac{\mu}{2}\mathbf{E}\left\|\bar{x} - x_*\right\|_2^2$ in the strongly convex case. Moreover, the primal-dual algorithm in [22] has a primal-only implementation when f is smooth similar to SCGD by properly choosing the distance-generating function for its dual mirror step.

Hu et al. [23] consider the conditional stochastic optimization (CSO) problem $\min_x F(x)$, $F(x) = \mathbf{E}_\xi\left[f(\mathbf{E}_{\zeta|\xi}\left[g_\xi(x;\zeta)\right];\xi)\right]$. Compared to the SCO problem, the inner function g and the distribution of the inner random vector ζ depend on the outer random vector ξ. For convex and smooth F, SGD with biased oracles (BSGD) in [23] requires $O(\epsilon^{-2})$ iterations and $B = O(\epsilon^{-1})$ large inner batch size per iteration to find an $\bar{x}$ s.t. $\mathbf{E}[F(\bar{x}) - F(x_*)] \leq \epsilon$ for convex problems, and $O\left(\frac{1}{\mu\epsilon}\right)$ iterations and $B = O(\epsilon^{-1})$ large inner batch size per iteration to find an $\bar{x}$ s.t. $\mathbf{E}[F(\bar{x}) - F(x_*)] \leq \epsilon$ for strongly convex problems.

3.3 Convex-Concave Saddle Point (SP) Problem

The saddle point (SP) problem $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} L(x,y)$ that is μ_x-convex in x and μ_y-concave in y ($\mu_x, \mu_y \geq 0$) has been thoroughly studied. We refer to the SP problem with $\mu_x, \mu_y > 0$ as a strongly-convex-strongly-concave (SCSC) problem while those with $\mu_x, \mu_y = 0$ as a convex-concave (CC) problem. A saddle point (x_*, y_*), if it exists, satisfies the condition $L(x_*, y) \leq L(x_*, y_*) \leq L(x, y_*)$, $\forall (x,y) \in \mathcal{X}\times\mathcal{Y}$. Besides, the saddle point (x_*, y_*) is unique in an SCSC problem. To assess the optimality of any point $(\bar{x}, \bar{y}) \in \mathcal{X}\times\mathcal{Y}$, we can employ the concept of the duality gap, defined as $\mathrm{Gap}(\bar{x}, \bar{y}) \coloneqq \max_{x,y}\{L(\bar{x}, y) - L(x, \bar{y})\}$, and for SCSC problems, we can also use the Euclidean distance to the saddle point, given by $D(\bar{x}, \bar{y}) \coloneqq \frac{\mu_x}{2}\left\|\bar{x} - x_*\right\|_2^2 + \frac{\mu_y}{2}\left\|\bar{y} - y_*\right\|_2^2$. The SP problem is closely related to the more general monotone variational inequalities (VI), which involve finding a point $z_* = (x_*, y_*)$ such that $\langle \Phi(z_*), z - z_*\rangle \geq 0$, $\Phi(z) = (\partial_x L(x,y), -\partial_y L(x,y))$, $\forall z \in \mathcal{X}\times\mathcal{Y}$. The convergence rate is quantified by measuring the number of iterations required to find an ϵ-approximate saddle point $(\bar{x}, \bar{y})$ or an ϵ-approximate solution to the VI, satisfying one of the

following conditions:

$$\mathrm{Gap}(\bar{x}, \bar{y}) \le \epsilon, \quad \text{or} \quad D(\bar{x}, \bar{y}) \le \epsilon, \quad \text{or} \quad \max_{z \in \mathcal{X} \times \mathcal{Y}} \langle \Phi(\bar{z}), \bar{z} - z \rangle \le \epsilon.$$

Notably, extragradient methods (EG), initially introduced in [30], have proven to achieve the optimal convergence $O(1/\epsilon)$ rate among first-order approaches for solving deterministic monotone Lipschitz Variational Inequalities (VIs) in both Euclidean and non-Euclidean spaces [31, 32, 33]. Moreover, EG can be viewed as approximations of the implicit proximal point (PP) method [34]. All of these methods share a convergence rate of $\tilde{O}\left(\frac{1}{\mu_x} \vee \frac{1}{\mu_y}\right)$[3], where $\tilde{O}(\cdot)$ hides a $\mathrm{poly}\log\left(\frac{1}{\epsilon}\right)$ term when applied to problems with smooth and SCSC objective functions [35, 36]. For deterministic problems that are both smooth and SCSC, the convergence rate of $\tilde{O}\left(\frac{1}{\sqrt{\mu_x}} \vee \frac{1}{\sqrt{\mu_y}} + \frac{1}{\sqrt{\mu_x \mu_y}}\right)$ presented in [37] is known to be optimal for first-order algorithms, as demonstrated by lower bounds established in [38]. Regarding CC problems, a lower bound result [39] indicates that the $O\left(\frac{1}{\epsilon}\right)$ convergence rate achieved by extragradient methods [31, 32, 40] is indeed optimal. Furthermore, the primal-dual hybrid gradient (PDHG) method [41, 42] and some more recent works [43, 44, 45, 46] have concentrated on the bilinear problem with $L(x, y) = x^\top A y$. Hamedani and Aybat [47] have extended this focus to problems where $L(x, y)$ is convex in x and linear in y.

Accessing exact oracles such as $\nabla_x L$ and $\nabla_y L$ may not be feasible in many real-world scenarios. Instead, the available resources provide only unbiased stochastic estimators, denoted as $\tilde{\nabla}_x L$ and $\tilde{\nabla}_y L$, with variances bounded by σ^2. This limitation has prompted the development of numerous algorithms tailored for addressing the stochastic saddle point problem (SPP) and the more general stochastic variational inequalities (SVIs). For instance, the stochastic mirror descent (SMD) method [48] achieves the optimal convergence rate of $O\left(\frac{1}{\epsilon^2}\right)$ for non-Lipschitz SVIs. For Lipschitz monotone SVIs, the stochastic mirror-prox (SMP) method [49] attains the optimal rate of $O\left(\frac{1}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right)$. For SCSC and non-smooth SP problems, Yan et al. [50] establish the $\tilde{O}\left(\frac{1}{\epsilon} + \frac{1}{\mu_x \epsilon} \vee \frac{1}{\mu_y \epsilon}\right)$ rate with probability $1 - p$. Hsieh et al. [51] propose a single-call stochastic extragradient (SSEG) method that achieves a rate of $O\left(\frac{1}{\epsilon} + \frac{\sigma^2}{\mu_x \epsilon} \vee \frac{\sigma^2}{\mu_y \epsilon}\right)$ for Lipschitz and strongly monotone SVIs. More recently, several works have devised stochastic algorithms for both the SSP and SVI problems, achieving (near-)optimal deterministic and stochastic convergence rates simultaneously. Zhang et al. [15] introduce the SAPD algorithm, which reaches a convergence rate of $\tilde{O}\left(\frac{1}{\mu_x} \vee \frac{1}{\mu_y} + \frac{1}{\sqrt{\mu_x \mu_y}} + \frac{\sigma^2}{\mu_x \epsilon} \vee \frac{\sigma^2}{\mu_y \epsilon}\right)$ for the SCSC problem and $O\left(\frac{1}{\epsilon} + \frac{\sigma^2}{\epsilon^2}\right)$ for the CC problem. Du et al. [52] further close the gap for the SCSC problem by improving the rate to $\tilde{O}\left(\frac{1}{\sqrt{\mu_x}} \vee \frac{1}{\sqrt{\mu_y}} + \frac{1}{\sqrt{\mu_x \mu_y}} + \frac{\sigma^2}{\mu_x \epsilon} \vee \frac{\sigma^2}{\mu_y \epsilon}\right)$.

3.4 Coordinate Methods for the Block-Separable Deterministic SP Problem

A special class of bilinearly-coupled SP problem is in the form $\min_x \max_y L(x, y) \coloneqq \frac{1}{n} \sum_{i=1}^n (\langle y^{(i)}, Ax \rangle - \phi_i(y^{(i)})) + r(x)$, where $L(x, y)$ is block-separable w.r.t. the dual variable y. One illustrative example is the primal-dual reformulation of the (regularized) empirical risk minimization (ERM) problem, denoted as $\min_x F(x)$, where $F(x)$ is defined as $F(x) \coloneqq \frac{1}{n} \sum_{i=1}^n \ell(a_i^\top x) + r(x)$. This reformulation applies to data-label pairs $(a_i, b_i)_{i=1}^n$ in the context of a linear model. Particularly in scenarios with a significantly large value of n, the computational overhead of computing $\nabla_y L(x, y)$ and updating y can become prohibitively expensive. In such cases, randomized coordinate methods offer a viable solution by reducing the per-iteration oracle cost from $O(n)$ to $O(1)$. The SPDC

[3] $a \wedge b$ denotes $\min\{a, b\}$ and $a \vee b$ denotes $\max\{a, b\}$.

method [13] leads to $\tilde{O}\left(n+\sqrt{\frac{n}{\mu_x\mu_y}}\right)$ convergence rate to make $\mathbf{E}[D(\bar{x},\bar{y})]\le\epsilon$ for the SCSC problem and $\tilde{O}\left(n+\frac{\sqrt{n}}{\epsilon}\right)$ rate to make $\mathbf{E}[F(\bar{x})-F(x_*)]\le\epsilon$ for the CC problem. Recently, Alacaoglu et al. [14] extended the PURE-CD originally proposed in [53] to incorporate importance sampling and exploit the potential sparsity in A. For the CC problem with dense A, PURE-CD not only achieves an improved rate of $O\left(n+\frac{\sqrt{n}}{\epsilon}\right)$ to guarantee $\mathbf{E}[F(\bar{x})-F(x_*)]\le\epsilon$ but also attains a rate of $\tilde{O}\left(\frac{n}{\epsilon}\right)$ to ensure $\mathbf{E}[\mathrm{Gap}(\bar{x},\bar{y})]\le\epsilon$. It is worth noting that $\mathbf{E}[\mathrm{Gap}(\bar{x},\bar{y})]\le\epsilon$ serves as a sufficient but not necessary condition for $\mathbf{E}[F(\bar{x})-F(x_*)]\le\epsilon$.

In addition to addressing the bilinearly-coupled block-separable saddle point (SP) problem, Hamedani et al. [54] have extended their focus to the more general Convex-Concave (CC) problem, defined as $L(x,y)=\Phi(x,y)-\phi(y)+\sum_{i=1}^m h_i(x^{(i)})$. Their work establishes a convergence rate of $O\left(\frac{m}{\epsilon}\right)$ for a randomized block-coordinate primal-dual method, ensuring that $\mathbf{E}[\mathrm{Gap}(\bar{x},\bar{y})]\le\epsilon$. Furthermore, Jalilzadeh et al. [55] have delved into scenarios where $L(x,y)$ exhibits block-separability to both x and y. In this context, $L(x,y)$ is defined as $L(x,y)=\Phi(x,y)-\sum_{i=1}^n\phi_i(y^{(i)})+\sum_{j=1}^m h_j(x^{(j)})$. They introduce a doubly-randomized block-coordinate method to address such problems. It is worth emphasizing that all the works mentioned in this section [13, 53, 14, 55] rely on the assumption of having access to the exact $\nabla_x\Phi(x,y)$ and $\nabla_y\Phi(x,y)$. In contrast, our work addresses the more challenging problem where only stochastic oracles are available.

4 Preliminaries

In this section, we present the necessary notations, definitions, and assumptions.

4.1 Notations and Definitions

The following notations are used throughout this paper.

- For a vector $y\in\mathbb{R}^{nm}$, we use $y^{(i)}\in\mathbb{R}^m$ to represent the i-th coordinate (block) of y, i.e., $y=(y^{(1)},\dots,y^{(n)})^\top$.
- f_i^* denotes the convex conjugate of f_i.
- The prox-function associated with a distance-generating function (d.-g.f.) $\psi_i:\mathbb{R}^m\to\mathbb{R}$ is defined as $U_{\psi_i}(u,v)\coloneqq\psi_i(u)-\psi_i(v)-\langle\psi_i'(v),u-v\rangle$ for $u,v\in\mathbb{R}^m$, where $\psi_i'(v)\in\partial\psi_i(v)$ is a subgradient of ψ_i. Besides, we define $U_\psi(y_1,y_2)\coloneqq\sum_{i=1}^n U_{\psi_i}(y_1^{(i)},y_2^{(i)})$ for $y_1,y_2\in\mathbb{R}^{nm}$.
- For a function $g(x)=\mathbf{E}_{\zeta\sim\mathbb{P}}[g(x;\zeta)]$, we define the stochastic estimator based on the mini-batch $\mathcal{B}$ as $g_i(x;\mathcal{B})\coloneqq\frac{1}{|\mathcal{B}|}\sum_{\zeta\in\mathcal{B}}g(x;\zeta)$.
- For $a,b\in\mathbb{R}$, we define $a\vee b\coloneqq\max(a,b)$ and $a\wedge b\coloneqq\min(a,b)$.
- $a\asymp b$ means that there exists $c,C>0$ such that $cb\le a\le Cb$.
- For a set $\mathcal{X}$, we define its diameter w.r.t. a d.-g.f. ψ as $D_{\psi,\mathcal{X}}\coloneqq[\max_{x\in\mathcal{X}}\psi(x)-\min_{x\in\mathcal{X}}\psi(x)]^{1/2}$. If $\psi(\cdot)=\frac{1}{2}\|\cdot\|_2^2$, we simply denote $D_{\psi,\mathcal{X}}$ as $D_{\mathcal{X}}$.
- Let $\mathcal{X}$ be a normed vector space with $\|\cdot\|_2$. For each $i\in[n]$, let $\mathcal{Y}_i\subset\mathbb{R}^m$ be a normed vector space with a general norm $\|\cdot\|$. The norm of the dual space $\mathcal{Y}_i^*\subset\mathbb{R}^m$ is defined as $\|\cdot\|_*\coloneqq\sup_{\|v\|\le1}\langle\cdot,v\rangle$.
- For any linear operator $T_i:\mathcal{X}\to\mathcal{Y}_i^*$, we define the operator norm of T_i as $\|T_i\|_{\mathrm{op}}\coloneqq\sup_{x\in\mathcal{X}}\left\{\frac{\|T_ix\|_*}{\|x\|_2}\right\}$ and the operator norm of its adjoint operator $T_i^*:\mathcal{Y}_i\to\mathcal{X}$ is defined as $\|T_i^*\|_{\mathrm{op}}\coloneqq\sup_{y^{(i)}\in\mathcal{Y}_i}\left\{\frac{\|T_i^*y^{(i)}\|_2}{\|y^{(i)}\|}\right\}$.
- We say $g_i:\mathcal{X}\to\mathbb{R}^m$ is L_g-smooth if it is differentiable on $\mathcal{X}$ and there exists $L_g>0$ such that $\|g_i(x)-g_i(x')-\nabla g_i(x')(x-x')\|_*\le\frac{L_g}{2}\|x-x'\|_2^2$, for any $x,x'\in\mathcal{X}$.

- We say that $f_i : \mathbb{R}^m \to \mathbb{R}$ is L_f-smooth if it is differentiable on its domain and there exists $L_f > 0$ such that $|f_i(u) - f_i(u') - \langle \nabla f_i(u'), u - u' \rangle| \leq \frac{L_f}{2} \|u - u'\|_*^2$, for any $u, u' \in \mathcal{Y}_i^*$.

4.2 Assumptions

Throughout the paper, we make the following assumptions that are standard in the literature [1, 22].

Assumption 1. $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex and closed set. Besides, r is μ-convex on $\mathcal{X}$, $\mu \geq 0$.

Assumption 2. $f_i : \mathbb{R}^m \to \mathbb{R}$ is proper convex and lower-semicontinuous. Besides, there exists $C_f > 0$ such that $|f_i(u) - f_i(u')| \leq C_f \|u - u'\|_*$ for any $u, u' \in \mathcal{Y}_i^*$.

Assumption 2 implies that $\|y^{(i)}\| \leq C_f \ \forall y^{(i)} \in \mathrm{dom} f_i^*, \ \forall i \in [n]$. Thus, (1.1) is equivalent to (1.2) with a closed and proper f_i^* and a convex and compact $\mathcal{Y} = \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n$.

Assumption 3. If g_i is not affine, we assume that f_i is monotonically non-decreasing for each coordinate of its input.

The assumption above ensures that the dual domain $\mathcal{Y}$ in (1.2) satisfies $\mathcal{Y}_i \subseteq \mathbb{R}_+^m$ for each $i \in [n]$ such that the min-max problem in (1.2) is convex-concave. Note that the outer functions f_i of all examples in Section 2 satisfy Assumption 3.

Assumption 4. g_i is convex and Lipschitz continuous, i.e., $\|g_i(x) - g_i(x')\|_* \leq C_g \|x - x'\|_2$ for some $C_g > 0$ and any $x, x' \in \mathcal{X}$.

While the smoothness conditions of f_i and g_i are not obligatory in this work, incorporating them leads to better convergence bounds. Lastly, we assume that the variances of stochastic estimators are bounded.

Assumption 5. Assume that $\mathbf{E}_{\zeta_i} \|g_i(x) - g_i(x; \zeta_i)\|_*^2 \leq \sigma_0^2 < \infty$, $\mathbf{E}_{\zeta_i} \|[g_i'(x)]^\top - [g_i'(x; \zeta_i)]^\top\|_{\mathrm{op}}^2 \leq \sigma_1^2 < \infty$ for any $g_i'(x) \in \partial g_i(x)$, $x \in \mathcal{X}$, and $\zeta_i \sim \mathbb{P}_i$. Besides, $\mathbf{E}_\iota \left\| [g_\iota'(x)]^\top y^{(\iota)} - \frac{1}{n} \sum_{i=1}^n [g_i'(x)]^\top y^{(i)} \right\|_2^2 \leq \delta^2 \leq C_f^2 C_g^2$ for any $g_i'(x) \in \partial g_i(x)$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$.

5 A Primal-Dual Block-Coordinate Stochastic Algorithm for cFCCO

First, we describe the proposed algorithm, which is named ALEXR (refer to Algorithm 1), designed for solving (1.1). We also establish connections between our algorithm and several existing methods. Subsequently, we delve into the main technical challenges in our convergence analysis, emphasizing the differences from previous works.

Each iteration of ALEXR consists of two main steps. The first step involves a block-coordinate stochastic proximal mirror ascent update of the selected dual variables from a random block $\mathcal{S}_t$ out of $\{1, 2, \ldots, n\}$, which occurs between Line 3 and Line 9 in Algorithm 1. It is notable that we use proximal mapping to tackle $f_i^*(y^{(i)})$ and we use an extrapolated stochastic gradient $\tilde{g}_t^{(i)}$ of the linear term $y^{(i)} g_i(x_t)$ in terms of $y^{(i)}$. The second step, involving a stochastic proximal gradient descent update of the primal variable, occurs between Line 10 and Line 11 in Algorithm 1 to compute the next x, where G_t is a (sub)gradient estimator of the coupling term $\frac{1}{n} \sum_i y_{t+1}^{(i)} g_i(x_t)$ using an independent mini-batch $\tilde{\mathcal{B}}_t^{(i)}$.

It is crucial to carefully select the distance-generating function ψ_i for the proximal mirror ascent step to satisfy the following necessary condition and the proximal mapping can be efficiently computed without requiring inner loops for an inexact solution.

Algorithm 1 ALEXR

Initialize $x_0 \in \mathcal{X}$, $y_0 \in \mathcal{Y}$
for $t = 0, 1, \ldots, T-1$ **do**
  Sample a batch $\mathcal{S}_t \subset \{1, \ldots, n\}$, $|\mathcal{S}_t| = S$
  for each $i \in \mathcal{S}_t$ **do**
    Sample independent size-B mini-batches $\mathcal{B}_t^{(i)}$, $\tilde{\mathcal{B}}_t^{(i)}$ from $\mathbb{P}_i$
    Compute $\tilde{g}_t^{(i)} = g_i(x_t; \mathcal{B}_t^{(i)}) + \theta(g_i(x_t; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}; \mathcal{B}_t^{(i)}))$
    Update $y_{t+1}^{(i)} = \arg\max_{y^{(i)} \in \mathcal{Y}_i} \left\{ y^{(i)} \tilde{g}_t^{(i)} - f_i^*(y^{(i)}) - \tau U_{\psi_i}(y^{(i)}, y_t^{(i)}) \right\}$
  end for
  For each $i \notin \mathcal{S}_t$, $y_{t+1}^{(i)} = y_t^{(i)}$
  Compute $g_i'(x_t; \tilde{\mathcal{B}}_t^{(i)}) \in \partial g_i(x_t; \tilde{\mathcal{B}}_t^{(i)})$ and $G_t = \frac{1}{S} \sum_{i \in \mathcal{S}_t} [g_i'(x_t; \tilde{\mathcal{B}}_t^{(i)})]^\top y_{t+1}^{(i)}$
  Update $x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \langle G_t, x \rangle + r(x) + \frac{\eta}{2} \|x - x_t\|_2^2 \right\}$
end for

Assumption 6. Distance-generating function ψ_i is μ_ψ-strongly convex on $\mathcal{Y}_i$ w.r.t. $\|\cdot\|$.

Next, we give some general recipes and specific examples of ψ_i for applications of (1.1) considered in subsection 2.

• When f_i is smooth on its domain, we can select $\psi_i = f_i^*$. By the first-order optimality condition, it is not difficult to show that (see Lemma 11 in Appendix A):

$$y_{t+1}^{(i)} = \nabla f_i(u_{t+1}^{(i)}), \quad u_{t+1}^{(i)} = \frac{\tilde{g}_t^{(i)} + \tau u_t^{(i)}}{1 + \tau}, \forall i \in \mathcal{S}_t \tag{5.1}$$

Then, ALEXR has a primal-only implementation similar to SOX and MSVR. This applies to the Bellman residual minimization/p-norm push problem with $f_i(\cdot) = (\cdot)^2$, pre/rec@k maximization with a smooth surrogate loss ℓ_1, the GDRO problem with χ^2-divergence, as well as the multi-task GDRO problem with smooth $f_i(g_i) = \log \sum_{j \in [m]} \exp(g_{ij})$.

• When f_i is non-smooth, we need to choose a strongly convex function for ψ_i depending on the problems. For example, for multi-task GDRO, we can use the entropy function $\psi_i(y^{(i)}) = \sum_{j=1}^m y^{(i,j)} \log(y^{(i,j)})$, which is 1-strongly convex to $\|\cdot\|_1$.

• When r is non-strongly convex, to derive a better rate in Section 6.2, we also require that ψ_i is smooth. If $f_i(\cdot)$ is strongly convex, we can set $\psi_i = f_i^*$; Otherwise, we need a smooth and strongly convex ψ_i. For example, we can choose $\psi_i(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ for the multi-task group DRO problems, where the proximal mapping can be solved by efficient projection onto the probability simplex [56]. For p-norm push with $f_i(\cdot) = (\cdot)^3$, we can also choose $\psi_i(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, where the proximal mapping of both cases has a closed-form solution (see Section 6.9 of Beck [57]). Moreover, we can choose $\psi_i(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ for any structured non-smooth[4] functions f_i. For example, the outer function $f_i(\cdot) = \frac{1}{\alpha}(\cdot)_+$ in GDRO with CVaR divergence is structured non-smooth, where the proximal mapping of f_i^* with $\psi_i(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ can be efficiently computed by the projection onto a closed interval.

[4]The definition of structured non-smooth functions is from Zhang and Lan [22]: We call a function g structured non-smooth if there is some convex closed set Π and convex closed and proper function g^* such that $g(y) = \max_{\pi \in \Pi} \langle \pi, y \rangle - g^*(\pi), \forall y \in Y$. Additionally, the proximal mapping of g^* with $\frac{1}{2}\|\cdot\|_2^2$ as the prox-function can be efficiently computable.

5.1 Relations with Existing Algorithms

Our algorithm ALEXR exhibits certain similarities to SOX [5], MSVR [10] and SAPD [15] with remarkable differences for adressing their limitations for solving (1.1) or (1.2).

Relationship with SOX. By setting $\theta = 0$, $\psi_i = f_i^*$ in ALEXR, the dual update and the gradient estimator become similar to that used in SOX [5]. In particular, the update of $u_{t+1}^{(i)}$ in (5.1) becomes the moving average estimator, i.e., $u_{t+1}^{(i)} = (1-\gamma)u_t^{(i)} + \gamma g_i(x_t; \mathcal{B}_t^{(i)})$, where $\gamma = 1/(1+\tau)$. Hence, the updates of ALEXR with $\theta = 0$, $\psi_i = f_i^*$ reduces to SOX without a momentum update[5]. SOX without the momentum update is analyzed in [5] for solving convex FCCO, but only achieves an iteration complexity of $O\left(\frac{n}{S\epsilon^2}\right)$ for the weak duality gap. However, a convergence guarantee in terms of the objective gap for SOX on non-strongly convex problems remains absent.

Relationship with MSVR. ALEXR with $\psi_i = f_i^*$ is closely related to MSVR [10] but has a subtle difference that gives ALEXR an advantage over MSVR. In particular, the update of $u_{t+1}^{(i)}$ in (5.1) can be written as

$$u_{t+1}^{(i)} = (1-\gamma)u_t^{(i)} + \gamma g_i(x_t; \mathcal{B}_t^{(i)}) + \gamma\theta(g_i(x_t; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}; \mathcal{B}_t^{(i)})), \forall i \in \mathcal{S}_t$$

where $\gamma = 1/(1+\tau) < 1$. This estimator is similar to the one used in MSVR except that the scaling factor before the correction term $(g_i(x_t; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}; \mathcal{B}_t^{(i)}))$ is $\beta = \frac{n-S}{S(1-\gamma)} + 1 - \gamma$, which could be much larger than 1 when $S \ll n$ and $\gamma \ll 1$. In contrast, the scaling factor in ALEXR is $\gamma\theta \leq 1$. Notably, several existing works have reported better empirical performance using a scaling factor less than one [11, 58], which is consistent with our setting and theory. Another difference between ALEXR and MSVR is that ALEXR does not use the variance-reduction technique (e.g., STORM [59]) for computing the gradient estimator of the primal variable. For FCCO problems under the PL condition, MSVR employs the STORM gradient estimator of the primal variable to accelerate the convergence. Then, the convergence rate of MSVR on convex FCCO problems is derived by adding a small quadratic regularization term and using the restarting trick. In our work, we find that the STORM gradient estimator is unnecessary for convex FCCO problems: It demands more memory and computational costs, albeit resulting in a worse convergence rate compared to that of ALEXR.

Relationship with SAPD. Both SAPD [15] and ALEXR employ the extrapolated estimator in (6) for the dual step. The difference is that SAPD [15] updates all coordinates $i \in [n]$ while ALEXR only updating those sampled coordinates $i \in \mathcal{S}_t$, $\mathcal{S}_t \subset [n]$. This design characterizes ALEXR as a randomized block-coordinate variant of SAPD. However, it introduces several novel challenges in the convergence analysis, not present in the analysis of SAPD:

• Firstly, one might initially intend to follow the proof of SAPD in the convex case to bound the gap $\mathbf{E}(L(x_{t+1}, y) - L(x, y_{t+1})) = \mathbf{E}[\frac{1}{n}(y^{(i)}g_i(x_{t+1}) - y_{t+1}^{(i)}g_i(x)) - \frac{1}{n}\sum_{i=1}^n(f_i^*(y^{(i)}) - f_i^*(y_{t+1}^{(i)})) + r(x_{t+1}) - r(x)]$ for the t-th iteration. However, the single-iteration analysis of ALEXR only yields a bound containing $\frac{1}{S}\sum_{i\in\mathcal{S}_t} f_i^*(y^{(i)})$ instead of $\frac{1}{n}\sum_{i=1}^n f_i^*(y^{(i)})$ due to the coordinate update of y. Unfortunately, we cannot easily get around this by taking conditional expectation, as the desired $\frac{1}{S}\mathbf{E}[\sum_{i\in\mathcal{S}_t} f_i^*(y^{(i)})] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[f_i^*(y^{(i)})]$ is valid only when the chosen $y^{(i)}$ does not depend on $\mathcal{S}_t$. To achieve convergence results in terms of $\mathbf{E}[F(x_{\text{out}}) - F(x_*)] \leq \mathbf{E}\max_y[L(x_{\text{out}}, y) - L(x_*, y_{\text{out}})]$ for the output $x_{\text{out}} = \frac{1}{T}\sum_{t=0}^{T-1} x_t$, the optimal y involved is $y^{(i)} \in \arg\max_v\{v^\top g_i(x_{\text{out}}) - f_i^*(v)\}$ for each $i \in [n]$, which actually *depends* on $\mathcal{S}_t$ such that the proof does not directly go through. To address this challenge, we need to introduce a virtual sequence $\{\bar{y}_t\}$ where each $\bar{y}_t$ is obtained by updating all coordinates of y_t. Nevertheless, this will make the analysis much more involved.

[5] Another difference is that SOX computes the gradient estimator by $G_t = y_t^{(i)}\nabla g_i(x_t, \mathcal{B}_t^{(i)})$ instead of $G_t = y_{t+1}^{(i)}\nabla g_i(x_t, \tilde{\mathcal{B}}_t^{(i)})$. However, this is not the fundamental difference as SOX can also use the latter one.

• Furthermore, ALEXR offers more options for distance-generating functions ψ_i other than $\psi_i(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ in SAPD for the dual step, enhancing its flexibility to a broader range of problems. In addition, we also provide convergence results for non-smooth problems, which are lacking in [15].

6 Convergence Analysis

In this section, we first present the convergence analysis of ALEXR for a class of strongly convex problems, where r is μ-strongly convex and f_i is smooth (i.e. f_i^* is strongly convex). Following that, we shift our focus to the case that r is only convex ($\mu = 0$ in Assumption 1) and f_i can be smooth or non-smooth.

Proposition 1. *There exists $\rho \geq 0$ such that $U_{f_i^*}(u,v) \geq \rho U_{\psi_i}(u,v)$. $\forall u, v \in \mathcal{Y}_i$. When $\rho > 0$ and ψ is μ_ψ-strongly convex w.r.t. $\|\cdot\|$, we have f_i is L_f-smooth w.r.t. $\|\cdot\|_*$, where $L_f = \frac{1}{\mu_\psi \rho}$.*

For instance, when choosing $\psi_i = f_i^*$, we have $\rho = 1$, and when setting $\psi_i(\cdot) = \frac{1}{2}\|\cdot\|^2$, ρ becomes $\frac{1}{L_f}$—the inverse of the smoothness constant of f_i. Besides, we have $U_{f_i^*}(u,v) \geq 0$ regardless of the smoothness of f_i.

For the convenience of analysis, we define the virtual sequence $\{\bar{y}_t\}_{t\geq 0}$ as follows.

$$\bar{y}_{t+1}^{(i)} = \arg\max_{y^{(i)}\in\mathcal{Y}_i} \left\{ y^{(i)}\tilde{g}_t^{(i)} - f_i^*(y^{(i)}) - \tau U_{\psi_i}(y^{(i)}, y_t^{(i)}) \right\}, \quad \tilde{g}_t = (\tilde{g}_t^{(1)}, \ldots, \tilde{g}_t^{(n)})^\top, \quad \forall i \in [n].$$

The reason for introducing this virtual sequence is to decouple the dependence between y_{t+1} and $\mathcal{S}_t$. Note that only those coordinates $i \in \mathcal{S}_t$ of $\tilde{g}_t$ are computed in the t-th iteration of Algorithm 1. Lemma 2 describes the progress made in the t-th iteration of ALEXR.

Lemma 2. *Under Assumptions 1 and 6, the following holds for any $x \in \mathcal{X}, y \in \mathcal{Y}$ after the t-th iteration of Algorithm 1.*

$$\begin{aligned} &L(x_{t+1}, y) - L(x, \bar{y}_{t+1}) \qquad (6.1)\\ &\leq \frac{\tau}{n}U_\psi(y, y_t) - \frac{\tau+\rho}{n}U_\psi(y, \bar{y}_{t+1}) - \frac{\tau}{n}U_\psi(\bar{y}_{t+1}, y_t) + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - \tilde{g}_t^{(i)},\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + \frac{\eta}{2}\|x - x_t\|_2^2 \\ &\quad - \frac{\eta+\mu}{2}\|x - x_{t+1}\|_2^2 - \frac{\eta}{2}\|x_{t+1} - x_t\|_2^2 + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - g_i(x),\, \bar{y}_{t+1}^{(i)} \right\rangle - \langle G_t,\, x_{t+1} - x\rangle. \end{aligned}$$

6.1 Strongly Convex and Smooth Case

We first consider the scenario $\mu > 0, \rho > 0$, where r is μ-strongly convex and f_i is $\frac{1}{\mu_\psi\rho}$-smooth. In this case, $L(x,y)$ in (1.2) is strongly-convex-strongly-concave and a unique saddle point (x_*, y_*) exists where $x_* = \arg\min_{x\in\mathcal{X}} F(x)$. Note that both x_* and y_* are independent of the randomness in the algorithm's execution. We define that $\mathcal{G}_t$ is the σ-algebra generated by $\{\mathcal{B}_0, \mathcal{S}_0, \ldots, \mathcal{B}_{t-1}, \mathcal{S}_{t-1}, \mathcal{B}_t\}$ and $\mathcal{F}_t$ is the σ-algebra generated by $\{\mathcal{B}_0, \mathcal{S}_0, \ldots, \mathcal{B}_{t-1}, \mathcal{S}_{t-1}, \mathcal{B}_t, \mathcal{S}_t\}$. Note that $\mathcal{G}_t \subset \mathcal{F}_t$ and y_{t+1} is $\mathcal{F}_t$-measurable. For any $i \in [n]$, we have

$$\mathbf{E}[U_{\psi_i}(y_*^{(i)}, y_{t+1}^{(i)}) \mid \mathcal{G}_t] = \frac{S}{n}U_{\psi_i}(y_*^{(i)}, \bar{y}_{t+1}^{(i)}) + \frac{n-S}{n}U_{\psi_i}(y_*^{(i)}, y_t^{(i)}).$$

Thus, the expectation of the $\frac{\tau}{n}U_\psi(y, y_t) - \frac{\tau+\rho}{n}U_\psi(y, \bar{y}_{t+1})$ terms in (6.1) with $y = y_*$ can form a contraction as

$$\mathbf{E}\left[\frac{\tau}{n}U_\psi(y_*, y_t) - \frac{\tau+\rho}{n}U_\psi(y_*, \bar{y}_{t+1})\right] = \frac{\tau + \rho\left(1 - \frac{S}{n}\right)}{S}\mathbf{E}[U_\psi(y_*, y_t)] - \frac{\tau+\rho}{S}\mathbf{E}[U_\psi(y_*, y_{t+1})]. \qquad (6.2)$$

Based on Lemma 2, (6.2), and other intermediate results in Appendix C, we can derive the following results for the strongly convex case.

Theorem 3. *Suppose that Assumptions 1, 2, 3, 4, 5, 6 hold. Moveover, r is μ-strongly convex with $\mu > 0$ while ρ in Proposition 1 satisfies that $\rho > 0$, i.e., f_i is L_f-smooth, $L_f := \frac{1}{\mu_\psi \rho}$.*

- *If g_i is L_g-smooth, Algorithm 1 with $\eta = \frac{\mu\theta}{1-\theta}$, $\tau = \frac{S}{n(1-\theta)}$, and a specific $\theta < 1$ can make $\frac{\mu}{2}\mathbf{E}\left\|x_T - x_*\right\|_2^2 \le \epsilon$ after $T = \tilde{O}\left(\frac{n}{S} + \frac{C_g\sqrt{nL_f}}{\sqrt{S\mu}} + \frac{L_g C_f}{\mu} + \frac{nL_f\sigma_0^2}{BS\epsilon} + \frac{C_f^2\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon}\right)$ iterations.*
- *If g_i is non-smooth, Algorithm 1 under the same settings of η, τ and $\theta < 1$ can make $\frac{\mu}{2}\mathbf{E}\left\|x_T - x_*\right\|_2^2 \le \epsilon$ after $T = \tilde{O}\left(\frac{n}{S} + \frac{C_g\sqrt{nL_f}}{\sqrt{S\mu}} + \frac{C_f^2 C_g^2}{\mu\epsilon} + \frac{nL_f\sigma_0^2}{BS\epsilon} + \frac{C_f^2\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon}\right)$ iterations.*

Remark 4. We would like to highlight several observations:
(1) When $n = 1$, Problem (1.2) with smooth f_i, g_i becomes a standard μ-strongly-convex-ρ-strongly-concave and smooth stochastic min-max optimization problem. Our ALEXR algorithm reverts to SAPD [15] and achieves the same $\tilde{O}\left(\frac{1}{\mu} + \frac{1}{\sqrt{\mu\rho}} + \frac{\sigma_0^2+\sigma_1^2}{\mu\epsilon}\right)$ convergence rate;

(2) When f_i is the identity mapping and $n = 1$ (i.e., $\sigma_0 = 0$[6] and $\delta = 0$), Problem (1.1) with a smooth g_i degenerates into the classical strongly convex and smooth stochastic optimization problem and our ALEXR with $\tau = 0$ reverts to (proximal) stochastic gradient descent (SGD) and nearly matches its $O\left(\frac{\sigma_1^2}{\mu\epsilon}\right)$ lower iteration complexity bound [25];

(3) When g_i is smooth, ALEXR needs $T = \tilde{O}\left(\frac{n\sigma_0^2}{BS\epsilon} + \frac{\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon}\right)$ iterations to find an ϵ-accurate solution $\bar{x}$ (i.e., $\mathbf{E}\frac{\mu}{2}\left\|\bar{x} - x_*\right\|_2^2 \le \epsilon$), which improves upon the $T = O\left(\frac{n}{\mu\sqrt{BS\epsilon}}\right)$ iterations needed by MSVR [10].
(4) When g_i is non-smooth, the complexity has a worse term $\tilde{O}(\frac{1}{\mu\epsilon})$ compared to that $\tilde{O}(\frac{1}{\mu})$ for the smooth g_i highlighted in blue in the theorem. It persists even when the variances are zero, and hence offers no parallel speedup in this term for mini-batches $\mathcal{B}$. Such a result is similar to those of (stochastic) subgradient methods.

6.2 Convex Case

Directly converting the result in the strongly convex case to that in the non-strongly case leads to an unsatisfactory $\tilde{O}\left(\frac{n\sigma_0^2}{BS\epsilon^2} + \frac{\sigma_1^2}{B\epsilon^3} + \frac{\delta^2}{S\epsilon^3}\right)$ convergence rate (See Appendix C.3). To address this issue, we provide a better result matching the optimal rate by restricting the distance-generating function ψ_i to be L_ψ-smooth and μ_ψ-strongly convex w.r.t. $\|\cdot\|$.
To derive a bound of the objective gap $\mathbf{E}[F(\bar{x}_T) - F(x_*)]$ for the time-average iterate $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$, we will plug $x = x_*$ and $y^{(i)}(\bar{x}_T) \in \arg\max_{v\in\mathcal{Y}_i}\{v^\top g_i(\bar{x}_T) - f_i^*(v)\} \in \partial f_i(g_i(\bar{x}_T))$ for all $i \in [n]$ into Lemma 2. It is important to note that the sum $\sum_{t=0}^{T-1}\left(\frac{\tau}{n}U_\psi(y, y_t) - \frac{\tau}{n}U_\psi(y, \bar{y}_{t+1})\right)$ in Lemma 2 does not form a telescoping sum. Additionally, the technique outlined in (6.2) does not address this issue because y also depends on $\mathcal{S}_t$. Instead, we handle this by employing the lemma below, which draws inspiration from Lemma A.2 in Alacaoglu et al. [14] but extends it to mini-batch sampling and a general smooth and strongly convex distance-generating function ψ_i.

[6]Here $\sigma_0 = 0$ because we have $f_i' \equiv 1$ for the identity mapping $f_i(g_i(x)) = g_i(x)$ such that there is no need to compute the stochastic estimator of $g_i(x)$ because $\nabla f_i(g_i(x)) = \nabla g_i(x)$.

Lemma 5. *Under Assumption 6, the following holds for Algorithm 1 with L_ψ-smooth distance-generating function ψ_i and any $\lambda_1 > 0$ satisfies that*

$$\mathbf{E}\left[\frac{\tau}{n}\big(U_\psi(y,y_t)-U_\psi(y,\bar{y}_{t+1})\big)-\frac{\tau}{n}U_\psi(\bar{y}_{t+1},y_t)\right] \tag{6.3}$$
$$\le \mathbf{E}\left[\frac{\tau}{S}(U_\psi(y,y_t)-U_\psi(y,y_{t+1}))+\frac{\tau\lambda_1}{S}(U_\psi(y,\hat{y}_t)-U_\psi(y,\hat{y}_{t+1}))\right]-\frac{\tau}{n}\left(1-\frac{L_\psi^2}{\lambda_1\mu_\psi^2 S}\right)\mathbf{E}\big[U_\psi(\bar{y}_{t+1},y_t)\big],$$

where the sequence $\{\hat{y}_t\}_t, \hat{y}_t \in \mathcal{Y}$ is virtual.

Based on Lemma 2, Lemma 5 and other intermediate results in Appendix D, we are ready to present the main theorem for the convergence of ALEXR in the convex case. To facilitate the discussion, we introduce the following quantity:

$$\Omega_{\mathcal{Y}}^0 = \mathbf{E}[\sum_{i=1}^n U_{\psi_i}(y^{(i)}(\bar{x}_T), y_0^{(i)})] \le \sum_{i=1}^n D_{\psi_i,\mathcal{Y}_i}^2 = O(n).$$

Theorem 6. *Suppose Assumptions 1, 2, 3, 4, 5, 6 hold. Besides, Algorithm 1 selects a smooth distance-generating function ψ_i. We denote that $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$.*

• *If g_i is L_g-smooth, Algorithm 1 with $\theta = 1$, $\eta = O\left(L_g C_f \vee \frac{\sqrt{n}C_g}{\sqrt{S}} \vee \frac{\delta^2}{S\epsilon} \vee \frac{C_f^2\sigma_1^2}{B\epsilon}\right)$, $\tau = O\left(\frac{\sqrt{S}C_g}{\mu_\psi\sqrt{n}} \vee \frac{\sigma_0^2}{\mu_\psi B\epsilon}\right)$ can make $\mathbf{E}[F(\bar{x}_T) - F(x_*)] \le \epsilon$ after*

$$T = O\left(\frac{L_g C_f D_{\mathcal{X}}^2}{\epsilon} + \frac{\sqrt{n}C_g D_{\mathcal{X}}^2}{\sqrt{S}\epsilon} + \frac{C_g(1+L_\psi^2/(S\mu_\psi^2))\Omega_{\mathcal{Y}}^0}{\mu_\psi\sqrt{nS}\epsilon} + \frac{D_{\mathcal{X}}^2\delta^2}{S\epsilon^2} + \frac{C_f^2\sigma_1^2 D_{\mathcal{X}}^2}{B\epsilon^2} + \frac{\sigma_0^2(1+L_\psi^2/(S\mu_\psi^2))\Omega_{\mathcal{Y}}^0}{\mu_\psi BS\epsilon^2}\right).$$

• *If g_i is non-smooth, Algorithm 1 with $\theta = 1$, $\eta = O\left(\frac{\sqrt{n}C_g}{\sqrt{S}} \vee \frac{\delta^2}{S\epsilon} \vee \frac{C_f^2\sigma_1^2}{B\epsilon} \vee \frac{C_f^2 C_g^2}{\epsilon}\right)$, $\tau = O\left(\frac{\sqrt{S}C_g}{\mu_\psi\sqrt{n}} \vee \frac{\sigma_0^2}{\mu_\psi B\epsilon}\right)$ can make $\mathbf{E}[F(\bar{x}_T) - F(x_*)] \le \epsilon$ after*

$$T = O\left(\frac{\sqrt{n}C_g D_{\mathcal{X}}^2}{\sqrt{S}\epsilon} + \frac{C_g(1+L_\psi^2/(S\mu_\psi^2))\Omega_{\mathcal{Y}}^0}{\mu_\psi\sqrt{nS}\epsilon} + \frac{C_f^2 C_g^2 D_{\mathcal{X}}^2}{\epsilon^2} + \frac{D_{\mathcal{X}}^2\delta^2}{S\epsilon^2} + \frac{C_f^2\sigma_1^2 D_{\mathcal{X}}^2}{B\epsilon^2} + \frac{\sigma_0^2(1+L_\psi^2/(S\mu_\psi^2))\Omega_{\mathcal{Y}}^0}{\mu_\psi BS\epsilon^2}\right).$$

Remark: We discuss the results in the worst case where $\Omega_{\mathcal{Y}}^0 = n\max_i D_{\psi_i,\mathcal{Y}_i}^2$, where $D_{\psi_i,\mathcal{Y}_i} \coloneqq [\max_{v\in\mathcal{Y}_i}\psi_i(v) - \min_{v\in\mathcal{Y}_i}\psi_i(v)]^{1/2}$. When g_i is smooth, the leading term in the iteration complexity of ALEXR is $O\left(\frac{n\sigma_0^2}{BS\epsilon^2} + \frac{\delta^2}{S\epsilon^2} + \frac{\sigma_1^2}{B\epsilon^2}\right)$. This result outperforms the $\tilde{O}\left(\frac{n\sigma_0^2}{BS\epsilon^2} + \frac{\delta^2}{S\epsilon^3} + \frac{\sigma_1^2}{B\epsilon^3}\right)$ rate derived from the strongly convex result, as well as the $O\left(\frac{n}{\sqrt{BS}\epsilon^2}\right)$ rate of MSVR [10]. Nevertheless, the requirement for ψ_i to be smooth prevents us from selecting $\psi_i = f_i^*$ except when f_i is strongly convex. However, it is worth noting that Theorem 6 remains applicable to our motivating examples, which include GDRO with CVaR divergence, pAUC maximization with restricted TPR, Bellman residual minimization, p-norm push, and multi-task GDRO. In these problems, the dual mirror step of our proposed ALEXR can be efficiently solved with a smooth ψ_i, e.g. $\psi_i(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, as indicated in the paragraph below Assumption 6. Moreover, when the optimal dual variable $y(\bar{x}_T)$ at $\bar{x}_T$ has some sparsity structure, then $\Omega_{\mathcal{Y}}^0$ could be much smaller than $n\max_i D_{\psi_i,\mathcal{Y}_i}$, the results above indicate much better complexity of ALEXR when n is large. An example is considered in the next section.

7 Application to GDRO with ϕ-divergence

We discuss two examples of the GDRO problem with ϕ-divergence: CVaR divergence with a hyper-parameter $\alpha \in (0,1)$ and χ^2-divergence with a hyper-parameter $\lambda > 0$. We compare ALEXR to the following baselines:

• SMD [48, 17]: It can be applied to the GDRO problem in (2.1) with CVaR divergence, where the dual mirror step with the entropy d.-g.f. can be efficiently solved by projection onto the permutahedron [60]. Moreover, SMD can also be applied to the worst-group DRO problem [16] (i.e., $\lambda = 0$ in (2.1) or $\alpha = \frac{1}{n}$ in CVaR). The iteration complexity of SMD is $T = O\left(\frac{\log n}{\epsilon^2}\right)$, which is independent of α (Theorem 1 in [17]). Besides, it requires $O(n \log n)$ computational cost for performing the dual projection and $O(n)$ oracles in each iteration. Note that SMD cannot be applied to the GDRO problem in (2.1) with χ^2-divergence due to the non-linear penalty term.

• OOA [16]: This algorithm can be viewed as a variant of the SMD algorithm with the dual gradient estimator $[0, \ldots, n\ell(w_t; z_t^{(i_t)}), \ldots, 0]^\top$ such that it only requires $O(1)$ oracles per iteration. But the dual projection cost in each iteration is still $O(n \log n)$. The iteration complexity of SMD is $T = O\left(\frac{n^2 \log n}{\epsilon^2}\right)$, which is also independent of α. OOA is not applicable to the GDRO problem in (2.1) with χ^2-divergence either.

It comes to our attention that the NOL algorithm [17] designed for the worst-group DRO problem ($\lambda = 0$ in (2.1) or $\alpha = \frac{1}{n}$ in CVaR) can achieve $T = O(\frac{n \log n}{\epsilon^2})$ iteration complexity in high probability with per-iteration $O(1)$ oracles. However, this result cannot be extended to the GDRO problem with CVaR or χ^2-divergence, since their proof technique relies on powerful tools for multi-armed bandits. Besides, Soma et al. [61] also consider the GDRO problem with CVaR divergence but their convergence analysis suffers from dependency issues, as pointed out in Zhang et al. [17]. Recently, Hu et al. [62] studied non-smooth weakly convex FCCO problems and proposed an algorithm SONX, which can be applied to solving GDRO with CVaR divergence. However, their algorithm does not leverage the convexity of the inner function and hence suffers from a worse complexity of $O\left(\frac{n}{S\sqrt{B}\epsilon^6}\right)$.

7.1 GDRO with CVaR divergence

GDRO with CVaR divergence can be formulated as (1.1) with $f_i(\cdot) = \alpha^{-1}(\cdot)_+$, $\alpha \in (0,1)$ and $g_i(w,c) = R_i(w) - c$ such that $C_f = \frac{1}{\alpha}$ and $C_g = C_R + 1$, where C_R is the Lipschitz constant of R_i. The dual update (7) of ALEXR with $\psi_i(\cdot) = \frac{1}{2}(\cdot)^2$ has the closed-form expression $y_{t+1}^{(i)} = \begin{cases} \text{Proj}_{[0,\alpha^{-1}]}\left[y_t^{(i)} + (1/\tau)\tilde{g}_t^{(i)}\right], & i \in \mathcal{S}_t \\ y_t^{(i)}, & i \notin \mathcal{S}_t \end{cases}$. According to Theorem 6, we can derive the following result.

Corollary 7. *Suppose that R_i is convex and Lipschitz continuous. For the GDRO problem (2.1) with CVaR divergence, the* ALEXR *algorithm with $y_0^{(i)} = 0$ requires $T = O\left(\frac{\sqrt{n}}{\sqrt{S}\epsilon} + \frac{1}{\alpha^2\epsilon^2} + \frac{\delta^2}{S\epsilon^2} + \frac{\sigma_1^2}{\alpha^2 B\epsilon^2} + \frac{\sigma_0^2\Omega_{\mathcal{Y}}^0}{BS\epsilon^2}\right)$ iterations to return an ϵ-approximate solution w_{out} that satisfies $\mathbf{E}[\mathcal{L}(w_{out}) - \mathcal{L}(w_*)] \leq \epsilon$, where $\Omega_{\mathcal{Y}}^0 \coloneqq \mathbf{E}[\sum_{i=1}^n U_{\psi_i}(\tilde{y}^{(i)}, 0)]$ and $\tilde{y}^{(i)} \in \partial f_i(g_i(w_{\text{out}}, c_{\text{out}}))$.*

Remark 8. The worst-case estimate of the $\Omega_{\mathcal{Y}}^0$ term is $\Omega_{\mathcal{Y}}^0 = \mathbf{E}[\sum_{i=1}^n U_{\psi_i}(\tilde{y}^{(i)}, 0)] \leq \frac{nC_f^2}{2} = \frac{n}{2\alpha^2}$ when $\psi_i = \frac{1}{2}(\cdot)^2$. However, it could be much smaller than $\frac{n}{2\alpha^2}$ in practice since $\tilde{y}^{(i)} = 0$ for those coordinates i that satisfy $R_i(w_{\text{out}}) \leq c_{\text{out}}$, i.e., the ALEXR algorithm can benefit from the "sparsity" of $\tilde{y}^{(i)} \in \partial f_i(g_i(w_{\text{out}}, c_{\text{out}}))$. In particular, when $(w_{\text{out}}, c_{\text{out}})$ is close to the optimal solution, then roughly about αn number of groups such that $[R_i(w_{\text{out}}) - c_{\text{out}}]_+ > 0$. As a result, $\Omega_{\mathcal{Y}}^0 = \mathbf{E}[\sum_{i=1}^n U_{\psi_i}(\tilde{y}^{(i)}, 0)] \approx$

Table 3: Comparison of iteration complexities, dual projection cost, and per-iteration #oracles for achieving ϵ-optimal solution of the GDRO problem in (2.1) in terms of $\mathbf{E}[\mathcal{L}(w_{\text{out}}) - \mathcal{L}(w_*)] \le \epsilon$. Here x_{out} is the output of each algorithm. We hide other constant quantities except for n, variances $\sigma_0^2, \sigma_1^2, \delta^2$, and batch sizes B, S. Besides, $\tilde{O}$ hides $\text{poly}\log(1/\epsilon)$ factors.

ϕ-Divergence	Algorithm	#Oracles[(1)]	Dual Proj.	Iteration Complexity	
CVaR	SMD [17]	$O(n)/O(1)$	$O(n\log n)$	$O\left(\frac{\log n}{\epsilon^2}\right)$	
	OOA [16]	$O(1)/O(1)$	$O(n\log n)$	$O\left(\frac{n^2\log n}{\epsilon^2}\right)$	
	ALEXR	$O(1)/O(1)$	$O(1)$	$O\left(\frac{\sqrt{n}}{\alpha^2\sqrt{S}\epsilon} + \frac{1}{\alpha^2\epsilon^2} + \frac{\delta^2}{S\epsilon^2} + \frac{\sigma_1^2}{\alpha^2 B\epsilon^2} + \frac{\sigma_0^2\Omega_{\mathcal{Y}}^0}{BS\epsilon^2}\right)$[(2)]	
χ^2	ALEXR	$O(1)/O(1)$	$O(1)$	Convex r	Strongly Convex r
				$O\left(\frac{\sqrt{n}}{\lambda\sqrt{S}\epsilon} + \frac{1}{\lambda^2\epsilon^2} + \frac{\delta^2}{S\epsilon^2} + \frac{\sigma_1^2}{B\epsilon^2} + \frac{\sigma_0^2\Omega_{\mathcal{Y}}^0}{BS\epsilon^2}\right)$[(3)]	$\tilde{O}\left(\frac{\sqrt{n}}{\lambda\sqrt{S\mu}} + \frac{1}{\mu\lambda^2\epsilon} + \frac{n\sigma_0^2}{BS\epsilon} + \frac{\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon}\right)$[(4)]

[(1)] Representing #zeroth-order oracles for $g_i(x)$/#first-order oracles for $\nabla g_i(x)$ in each iteration.
[(2)] The worst-case estimate of $\Omega_{\mathcal{Y}}^0$ is $\frac{n}{2\alpha^2}$, but it could be much smaller than $\frac{n}{2\alpha^2}$ in practice, as explained in Remark 8.
[(3)] The worst-case estimate of $\Omega_{\mathcal{Y}}^0$ is $\frac{nC_f^2}{2}$, but it could be much smaller than $\frac{nC_f^2}{2}$ in practice.
[(4)] In terms of the distance gap.

$\frac{n\alpha C_f^2}{2} = \frac{n}{2\alpha}$. Then, the complexity may become $T = O\left(\frac{\sqrt{n}}{\sqrt{S}\epsilon} + \frac{1}{\alpha^2\epsilon^2} + \frac{\delta^2}{S\epsilon^2} + \frac{\sigma_1^2}{\alpha^2 B\epsilon^2} + \frac{\sigma_0^2 n}{\alpha BS\epsilon^2}\right)$.

7.2 GDRO with χ^2- divergence

GDRO with χ^2- divergence can be formulated as (1.1) with $f_i(\cdot) = \lambda\left(\frac{1}{4}(\cdot+2)_+^2 - 1\right)$ and $g_i(w,c) = (R_i(w) - c)/\lambda$ such that $C_f = \frac{\max\{B_R - \underline{c}, B_R + \bar{c}\}}{2}$ [7] and $C_g = \frac{C_R + 1}{\lambda}$, where $B_R := \max_{w\in\mathcal{W}} |R_i(w)|$. In this case, the proximal mapping of $f_i^*(y^{(i)}) = \frac{\lambda}{2}(y^{(i)}/\lambda - 1)^2$ with $\psi_i(\cdot) = \frac{1}{2}(\cdot)^2$ can also be efficiently solved. We can also consider the GDRO problem with a convex regularization term $r(x)$. With a strongly convex regularizer, we can choose either $\psi_i = f_i^*$ or $\psi_i(\cdot) = \frac{1}{2}(\cdot)^2$. When $\psi_i = f_i^*$, the dual update (7) of ALEXR has the closed-form expression $y_{t+1}^{(i)} = f_i'(u_{t+1}^{(i)})$, $u_{t+1}^{(i)} = \begin{cases} \frac{\tau}{1+\tau}u_t^{(i)} + \frac{1}{1+\tau}\tilde{g}_t^{(i)}, & i \in \mathcal{S}_t \\ u_t^{(i)}, & i \notin \mathcal{S}_t \end{cases}$. When $\psi_i(\cdot) = \frac{1}{2}(\cdot)^2$, the dual update (7) of ALEXR has the closed-form expression $y_{t+1}^{(i)} = \begin{cases} \frac{\lambda}{1+\lambda}(y_t^{(i)} + (1/\tau)\tilde{g}_t^{(i)} + 1), & i \in \mathcal{S}_t \\ y_t^{(i)}, & i \notin \mathcal{S}_t \end{cases}$. Theorem 3 and Theorem 6 imply the following convergence result.

Corollary 9. *Suppose that $R_i(w)$ is uniformly bounded and Lipschitz continuous. For the GDRO problem (2.1) with χ^2-divergence and a convex regularization term $r(x)$, the* ALEXR *algorithm with $y_0^{(i)} = 0$ requires $T = \tilde{O}\left(\frac{\sqrt{n}}{\lambda\sqrt{S\mu}} + \frac{1}{\mu\lambda^2\epsilon} + \frac{\sigma_0^2\Omega_{\mathcal{Y}}^0}{BS\epsilon} + \frac{\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon}\right)$ iterations to return an ϵ-approximate solution w_{out} that satisfies $\mathbf{E}[\mathcal{L}(w_{out}) - \mathcal{L}(w_*)] \le \epsilon$, where $\Omega_{\mathcal{Y}}^0 := \mathbf{E}[\sum_{i=1}^n U_{\psi_i}(\tilde{y}^{(i)}, 0)]$ and $\tilde{y}^{(i)} \in \partial f_i(g_i(w_{\text{out}}, c_{\text{out}}))$. If r is μ-strongly convex, it takes $T = \tilde{O}\left(\frac{\sqrt{n}}{\lambda\sqrt{S\mu}} + \frac{1}{\mu\lambda^2\epsilon} + \frac{n\sigma_0^2}{BS\epsilon} + \frac{\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon}\right)$ iterations to find an w_{out} such that $\frac{\mu}{2}\|w_{\text{out}} - w_*\|_2^2 \le \epsilon$.*

[7] A valid choice of $\underline{c}, \bar{c}$ is $\underline{c} = -\lambda$, $\bar{c} = B_R$ (see Appendix A.3 in Levy et al. [18]).

7.3 Comparison with Baselines

In Table 3, we compare our ALEXR to the baseline algorithms OOA and SMD. It is notable that although SMD has a better iteration complexity for CVaR divergence, it requires $O(n)$ oracles at each iteration. In contrast, ALEXR and OOA only require $O(1)$ oracles in each iteration. On the GDRO problem with CVaR divergence, the iteration complexity of ALEXR is better than OOA when $\Omega_{\mathcal{Y}}^0 = o(n^2 \log n)$ or the variance σ_0^2 is small. In the worst case, we have $\Omega_{\mathcal{Y}}^0 = O(n/\alpha^2)$, then ALEXR has a better complexity than OOA when $\frac{1}{\alpha} = o(\sqrt{n \log n})$. In practice, we have $\Omega_{\mathcal{Y}}^0 = O(n/\alpha)$, then ALEXR has a better complexity than OOA when $\frac{1}{\alpha} = o(n \log n)$. In addition, OOA cannot enjoy the parallel speedup with respect to the inner batch size B due to its scaled dual gradient estimator. Moreover, we also provide the iteration complexity of ALEXR on this the GDRO problem with χ^2-divergence, with or without a regularization term.

8 Lower Complexity Bounds

The proposed ALEXR and previous algorithms SOX [5], MSVR [10] are all special instantiations of an abstract first-order block-coordinate stochastic update scheme (Algorithm 2) with $O(1)$ oracles and $O(d)$ computation cost per iteration. For an affine subspace $\mathfrak{S} \subset \mathbb{R}^d$, we denote the linear span of $\{s^{(i)} \mid s \in \mathfrak{S}\}$ as $\mathfrak{S}^{(i)}$ for each $i \in [d]$. The "+" in Algorithm 2 refers to the Minkowski addition.

Algorithm 2 Abstract First-Order Block-Coordinate Stochastic Update Scheme

Initialize affine subspaces $\mathfrak{X}_0$, $\mathfrak{Y}_0$, $\mathfrak{g}_0$, $\mathfrak{G}_0$
for $t = 0, 1, \ldots, T-1$ **do**
  Sample a batch $\mathcal{S}_t \subset \{1, \ldots, n\}$, $|\mathcal{S}_t| = S$
  for each $i \in \mathcal{S}_t$ **do**
    Sample independent size-B mini-batches $\mathcal{B}_t^{(i)}, \tilde{\mathcal{B}}_t^{(i)}$ from $\mathbb{P}_i$
    $\mathfrak{g}_{t+1}^{(i)} = \mathfrak{g}_t^{(i)} + \operatorname{span}\{g_i(\hat{x}; \mathcal{B}_t^{(i)}) \mid \hat{x} \in \mathfrak{X}_t\}$
    $\mathfrak{Y}_{t+1}^{(i)} = \mathfrak{Y}_t^{(i)} + \operatorname{span}\left\{\arg\max_{y^{(i)}} \left\{y^{(i)} \hat{g}^{(i)} - f_i^*(y^{(i)}) - \tau U_{\psi_i}(y^{(i)}, \hat{y}^{(i)})\right\} \mid \hat{g}^{(i)} \in \mathfrak{g}_{t+1}^{(i)}, \hat{y}^{(i)} \in \mathfrak{Y}_t^{(i)}\right\}$
  end for
  For each $i \notin \mathcal{S}_t$, $\mathfrak{g}_{t+1}^{(i)} = \mathfrak{g}_t^{(i)}, \mathfrak{Y}_{t+1}^{(i)} = \mathfrak{Y}_t^{(i)}$
  $\mathfrak{G}_{t+1} = \mathfrak{G}_t + \operatorname{span}\left\{\frac{1}{S} \sum_{i \in \mathcal{S}_t} \hat{y}^{(i)} \nabla g_i(\hat{x}; \hat{\mathcal{B}}_t^{(i)}) \mid \hat{x} \in \mathfrak{X}_t, \hat{y} \in \mathfrak{Y}_{t+1}\right\}$
  $\mathfrak{X}_{t+1} = \mathfrak{X}_t + \operatorname{span}\left\{\arg\min_x \left\{\langle \hat{G}, x\rangle + r(x) + \frac{\eta}{2}\|x - \hat{x}\|_2^2\right\} \mid \hat{x} \in \mathfrak{X}_t, \hat{G} \in \mathfrak{G}_{t+1}\right\}$
end for

To obtain the best possible iteration complexity of this abstract scheme on cFCCO problems, we consider a special instance of problem (1.1) that is separable over the coordinates i and $\mathbb{P}_i = \mathbb{P}$.

$$\min_{x \in [-D, D]^n} F(x), \quad F(x) = \frac{1}{n}\left(\sum_{i=1}^n f(g_i(x)) + \frac{\alpha}{2}\|x\|^2\right), \tag{8.1}$$
$$g_i(x) = \mathbf{E}_{\zeta \sim \mathbb{P}}[g_i(x; \zeta)], \; g_i(x; \zeta) = x^{(i)} + \zeta,$$

where the additive noise ζ follows

$$\zeta = \begin{cases} -\nu & \text{w.p. } 1-p, \\ \nu(1-p)/p & \text{w.p. } p. \end{cases}, \quad \text{where } p := \frac{\nu^2}{\sigma^2} \in (0, 1).$$

Based on the abstract scheme (Algorithm 2) and the "hard" problems in Appendix E, we can derive the following lower complexity bounds.

Theorem 10. *Consider the abstract scheme (Algorithm 2) with inner mini-batch size $B = 1$ and initialization $\mathfrak{X}_0^{(i)} = \{0\}$, $\mathfrak{Y}_0^{(i)} = \{0\}$, $\mathfrak{g}_0^{(i)} = \varnothing$, $\mathfrak{G}_0^{(i)} = \varnothing$.*

- *There exists a cFCCO problem (1.1) with smooth f_i and μ-strongly convex r such that any algorithm in the abstract scheme requires at least $T = \Omega\left(\frac{1}{\mu\epsilon}\right)$ iterations to find an $\bar{x}$ that satisfies $\mathbf{E}\left[\frac{\mu}{2}\|\bar{x} - x_*\|_2^2\right] \leq \epsilon$. Moreover, there exists another cFCCO problem (1.1) with smooth f_i and μ-strongly convex r such that any algorithm in the abstract scheme requires at least $T = \Omega\left(\frac{n\sigma_0^2}{S\epsilon}\right)$ iterations to find an $\bar{x}$ that satisfies $\mathbf{E}\left[\frac{\mu}{2}\|\bar{x} - x_*\|_2^2\right] \leq \epsilon$ or $\mathbf{E}[F(\bar{x}) - F(x_*)] \leq \epsilon$.*
- *There exists a cFCCO problem (1.1) with non-smooth f_i such that any algorithm in the abstract scheme requires at least $T = \Omega\left(\frac{n\sigma_0^2}{S\epsilon^2}\right)$ iterations to find an $\bar{x}$ that satisfies $\mathbf{E}[F(\bar{x}) - F(x_*)] \leq \epsilon$.*

Compared with the upper bound results in Section 6, this theorem demonstrates that ALEXR is optimal among first-order block-coordinate stochastic algorithms for cFCCO problems.

9 Numerical Experiments

In this section, we show experimental results on the Group Distributionally Robust Optimization (GDRO) and Partial AUC Maximization with restricted TPR. More details of the experiments and additional results can be found in Appendix F.

9.1 Experiments on GDRO with the CVaR divergence

First, we numerically compare our proposed ALEXR and baseline methods on the GDRO problem in (2.1) with the CVaR divergence for the binary classification task, where the objective function $\mathcal{L}(w)$ is the average risk on the top-αn worst groups, i.e., $\mathcal{L}(w) = \frac{1}{\lfloor \alpha n \rfloor} \sum_{i=1}^{\lfloor \alpha n \rfloor} R_{[\![i]\!]}(w)$ and $[\![i]\!]$ refers to the i-th worst group in descending order. We consider the linear model w and the logistic loss $\ell(w; z)$.

Baselines. We compare our ALEXR with previous algorithms on the FCCO problem and, more specifically, the GDRO problem: BSGD [23], SOX [5], SONX [11], OOA [16], and SGD with up-weighting (SGD-UW) [63, 64]. Note that the MSVR algorithm [10] is not applicable since the outer function f_i in this problem is non-smooth. Instead, we consider the non-smooth variant SONX of the MSVR algorithm. To show the benefit of GDRO, we also include SGD which is based on empirical risk minimization (ERM) and neglects the group information. Besides, OOA needs a projection onto an $(n-1)$-dimensional capped simplex $\{y \in \mathbb{R}^n \mid \sum_{i=1}^n y^{(i)} = 1, 0 \leq y^{(i)} \leq \frac{1}{\alpha n}\}$ in each iteration.

Datasets. We perform experiments on two datasets: a tabular dataset Adult [65] and an image dataset CelebA [66]. The Adult dataset contains 48,842 data points, where we use 22,792 data points for training, 9,769 points for validation, and 16,281 points for testing [67]. We construct 83 groups for the Adult dataset according to some categorical features such as race and gender. The original CelebA dataset is composed of 162,770 celebrity images for training, 198,670 images for validation, and 199,620 images for testing. We construct 160 groups for the CelebA dataset. On the Adult dataset, the task is to predict the income, whereas on the CelebA dataset, the goal is to determine whether the individual in the image possesses blonde hair. More details of the datasets and the preprocessing steps can be found in Appendix F.1.1.

Results. We report the loss curves on the validation dataset for FCCO algorithms that share the same objective function (2.1) in Figure 1. Then, we report test accuracy for all algorithms in Table 4 on the worst-(αn) groups under 4 different values of α. First, we notice that the vanilla SGD performs poorly on the worst-(αn) groups' data. While the up-weighting trick offers some improvement for

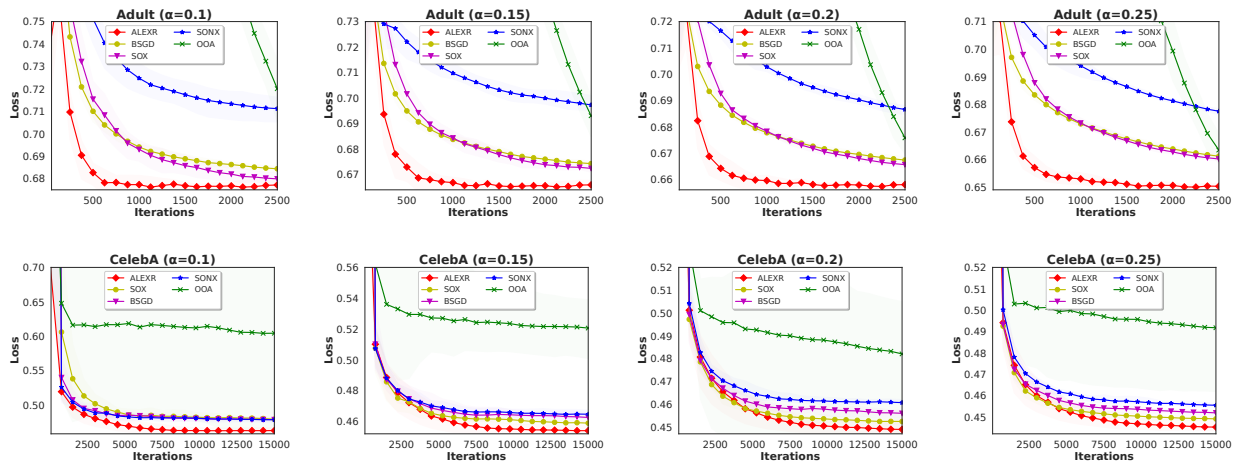


Figure 1: Losses evaluated on the validation dataset during training with different $\alpha \in \{0.1, 0.15, 0.2, 0.25\}$.

Table 4: Comparison of test accuracy (%) on the worst-(αn) groups with $\alpha \in \{0.1, 0.15, 0.2, 0.25\}$. The best accuracy is highlighted in black while the second-best one is highlighted in gray.

Methods	Adult				CelebA			
	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.2$	$\alpha = 0.25$	$\alpha = 0.1$	$\alpha = 0.15$	$\alpha = 0.2$	$\alpha = 0.25$
SGD	0.71±0.20	1.87±0.25	4.14±0.26	7.35±0.27	2.75±0.08	4.89±0.10	6.61±0.07	8.02±0.08
SGD-UW	23.70±1.01	26.26±1.06	31.84±0.71	36.77±0.65	73.70±0.13	74.18±0.13	74.79±0.12	75.28±0.11
OOA	51.46±2.21	54.12±2.04	56.08±1.98	57.45±1.73	66.40±6.37	73.43± 0.79	75.62±0.01	74.90±0.02
BSGD	55.81±0.70	**58.58±0.61**	59.48±0.47	60.48±0.45	75.30±0.27	76.16±0.12	**77.00±0.09**	77.53±0.07
SOX	56.34 ±1.15	58.36±0.44	**60.39±0.36**	61.45±0.25	75.04±0.20	76.10±0.30	76.85±0.11	**77.59±0.25**
SONX	47.78±1.06	49.49±0.95	51.78±0.66	54.42±0.65	75.34±0.28	76.17±0.09	76.99±0.06	77.47±0.09
ALEXR	**56.58±0.69**	58.52±0.71	60.23±0.50	**61.76±0.36**	**75.79±0.05**	**76.29±0.07**	76.80±0.12	77.28±0.10

SGD, its effectiveness still falls short of Group DRO algorithms. Among GDRO algorithms, our proposed ALEXR, exhibits faster convergence compared to baseline methods. Additionally, ALEXR also achieves superior testing performance in most cases.

9.2 Partial AUC Maximization with Restricted TPR

Next, we compare the proposed ALEXR and existing baselines on the pAUC maximization problem with restricted TPR in (2.3) and its equivalent forms (2.4), (2.5). In our experiments, we consider linear prediction model w and two different lower bounds α of TPR: 0.5 and 0.75.

Baselines. Previous FCCO algorithms BSGD, SOX and SONX can be applied to (2.5) and OOA can be applied to (2.4). We also include SGD with over-sampling to minimize the cross-entropy (CE) loss. In Zhu et al. [6], they propose an algorithm called SOTA for the weakly convex pAUC problem with TPR and FPR restrictions. We modify the SOTA algorithm for the convex pAUC problem with only restricted TPR.

Datasets. We perform experiments on four datasets: Covtype, Cardiomegaly, Lung-mass, and Higgs. The Covtype and Higgs datasets are from the LibSVM repository[8]. We create the imbalanced datasets by randomly removing 99.5% positive data from the Covtype dataset and 99.9% positive

[8] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

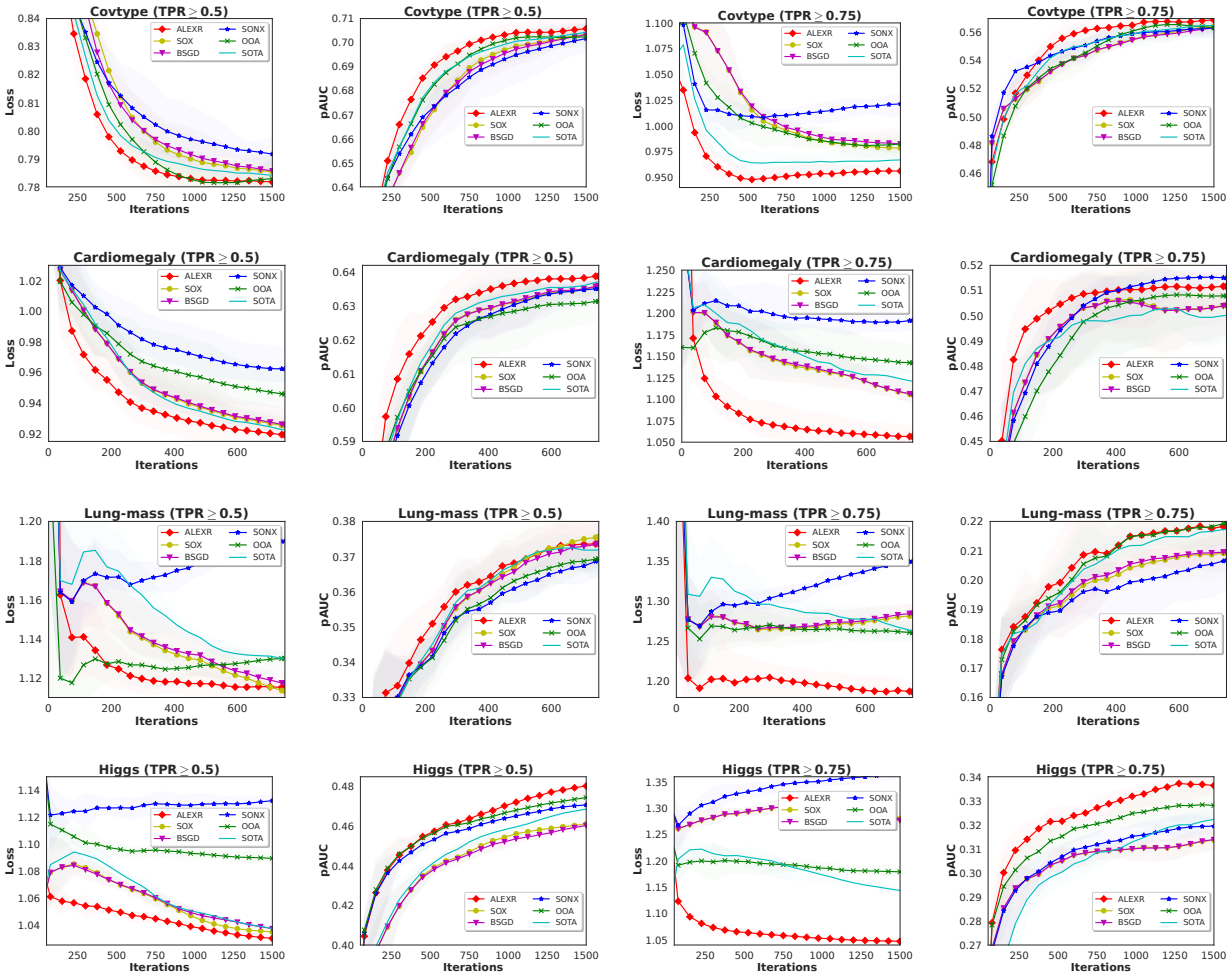


Figure 2: Loss and partial AUC evaluated on the validation set during training under TPR≥ 0.5 and TPR≥ 0.75.

data from the Higgs dataset. Cardiomegaly and Lung-mass are two imbalanced datasets that share the same collection of Chest X-ray images and different label annotations from the MedMNIST repository [68]. Details of the dataset and the preprocessing steps can be found in Appendix F.2.1.

Table 5: Comparison of partial AUC on the test dataset. The best result is highlighted in black.

Methods	Covtype		Cardiomegaly		Lung-mass		Higgs	
	TPR≥ 0.5	TPR≥ 0.75	TPR≥ 0.5	TPR≥ 0.75	TPR≥ 0.5	TPR≥ 0.75	TPR≥ 0.5	TPR≥ 0.75
CE	0.707±0.001	0.581±0.004	0.524±0.001	0.378±0.002	0.292±0.009	0.143±0.009	0.482±0.001	0.318±0.003
OOA	0.716±0.002	0.594±0.003	**0.613±0.001**	0.477±0.005	0.323±0.008	0.181±0.012	0.482±0.002	0.316±0.005
BSGD	0.726±0.003	0.597±0.005	0.606±0.001	**0.481±0.003**	0.329±0.004	0.170±0.004	0.482±0.001	0.320±0.004
SOX	0.726±0.002	0.597±0.005	0.607±0.000	**0.481±0.003**	0.329±0.003	0.170±0.004	0.481±0.002	0.319±0.004
SONX	0.723±0.003	0.597±0.004	0.603±0.002	0.474±0.003	0.318±0.002	0.165±0.005	0.481±0.001	0.318±0.002
SOTA	0.726±0.004	0.600±0.007	0.611±0.002	**0.481±0.001**	0.332±0.008	0.183±0.008	0.483±0.002	0.321±0.003
ALEXR	**0.727±0.003**	**0.605±0.005**	**0.613±0.002**	0.477±0.004	**0.333±0.005**	**0.185±0.014**	**0.485±0.000**	**0.322±0.003**

Results. For each algorithm, we present the objective function values in (2.3) and the partial AUC values evaluated on the validation dataset throughout the training process, as depicted in Figure 2. Additionally, we also compare the final partial AUC values on the test dataset for all algorithms, summarized in Table 5. The results suggest that optimizing the surrogate loss in Equation (2.3) outperforms optimizing the Cross-Entropy (CE) loss for maximizing the partial AUC with restricted TPR. Among algorithms tailored for optimizing (2.3), our proposed algorithm, named ALEXR, demonstrates overall superior performance when compared to previous algorithms.

10 Conclusion

In this paper, we delve into a class of convex finite-sum compositional stochastic optimization (cFCCO) problems, as represented by (1.1), by leveraging the min-max reformulation in (1.2). To tackle this problem, we propose a single-loop primal-dual block-coordinate proximal algorithm called ALEXR. Our proposed ALEXR achieves improved convergence rates compared to previous works on both convex and strongly convex problems. Furthermore, we present lower complexity bounds to show that the convergence rate of ALEXR stands as optimal among first-order block-coordinate stochastic methods for cFCCO problems. We demonstrate that ALEXR has applications in a broad spectrum of problems, including the Group Distributionally Robust Optimization (GDRO) problem and partial AUC maximization problem with restricted TPR. Numerical experiments demonstrate the promising performance of ALEXR on the GDRO and pAUC problems.

Acknowledgements

We are deeply grateful to Guanghui Lan for his invaluable feedback on this paper. We are also thankful to Stephen J. Wright for bringing the analysis of the duality gap for PureCD to our attention. We thank Guanghui Wang for the initial discussion of the problem.

References

[1] Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

[2] Mengdi Wang, Ji Liu, and Ethan X Fang. Accelerating stochastic composition optimization. *Journal of Machine Learning Research*, 18:1–23, 2017.

[3] Tianbao Yang. Algorithmic foundations of empirical x-risk minimization. *arXiv preprint arXiv:2206.00439*, 2022.

[4] Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] Bokun Wang and Tianbao Yang. Finite-sum coupled compositional stochastic optimization: Theory and applications. In *International Conference on Machine Learning*, pages 23292–23317. PMLR, 2022.

[6] Dixian Zhu, Gang Li, Bokun Wang, Xiaodong Wu, and Tianbao Yang. When AUC meets DRO: optimizing partial AUC for deep learning with non-convex convergence guarantee. In *International Conference on Machine Learning*, 2022.

[7] Zi-Hao Qiu, Quanqi Hu, Yongjian Zhong, Lijun Zhang, and Tianbao Yang. Large-scale stochastic optimization of ndcg surrogates for deep learning with provable convergence. In *International Conference on Machine Learning*, pages 18122–18152. PMLR, 2022.

[8] Zhuoning Yuan, Yuexin Wu, Zi-Hao Qiu, Xianzhi Du, Lijun Zhang, Denny Zhou, and Tianbao Yang. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In *International Conference on Machine Learning*, pages 25760–25782. PMLR, 2022.

[9] Guanghui Wang, Ming Yang, Lijun Zhang, and Tianbao Yang. Momentum accelerates the convergence of stochastic auprc maximization. *arXiv preprint arXiv:2107.01173*, 2021.

[10] Wei Jiang, Gang Li, Yibo Wang, Lijun Zhang, and Tianbao Yang. Multi-block-single-probe variance reduced estimator for coupled compositional optimization. *arXiv preprint arXiv:2207.08540*, 2022.

[11] Quanqi Hu, Dixian Zhu, and Tianbao Yang. Non-smooth weakly-convex finite-sum coupled compositional optimization. In *NeurIPS*, 2023.

[12] Lie He and Shiva Prasad Kasiviswanathan. Debiasing conditional stochastic optimization. *arXiv preprint arXiv:2304.10613*, 2023.

[13] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In *ICML*, pages 353–361, 2015.

[14] Ahmet Alacaoglu, Volkan Cevher, and Stephen J Wright. On the complexity of a practical primal-dual coordinate method. *arXiv preprint arXiv:2201.07684*, 2022.

[15] Xuan Zhang, Necdet Serhat Aybat, and Mert Gürbüzbalaban. Robust accelerated primal-dual methods for computing saddle points. *arXiv preprint arXiv:2111.12743*, 2021.

[16] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[17] Lijun Zhang, Peng Zhao, Tianbao Yang, and Zhi-Hua Zhou. Stochastic approximation approaches to group distributionally robust optimization. *arXiv preprint arXiv:2302.09267*, 2023.

[18] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

[19] Tianbao Yang and Yiming Ying. Auc maximization in the era of big data and ai: A survey. *ACM Computing Surveys*, 55(8):1–37, 2022.

[20] W. Gao and Z.-H. Zhou. On the consistency of AUC pairwise optimization. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, pages 939–945, 2015.

[21] C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10(Oct):2233–2271, 2009.

[22] Zhe Zhang and Guanghui Lan. Optimal algorithms for convex nested stochastic composite optimization. *arXiv preprint arXiv:2011.10076*, 2020.

[23] Yifan Hu, Siqi Zhang, Xin Chen, and Niao He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[24] Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Advances in neural information processing systems*, 27, 2014.

[25] Phuong Ha Nguyen, Lam Nguyen, and Marten van Dijk. Tight dimension independent lower bound on the expected convergence rate for diminishing step sizes in sgd. *Advances in Neural Information Processing Systems*, 32, 2019.

[26] Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

[27] Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, and Tianbao Yang. Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2023.

[28] Xiangru Lian, Mengdi Wang, and Ji Liu. Finite-sum composition optimization via variance reduced gradient descent. In *Artificial Intelligence and Statistics*, pages 1159–1167. PMLR, 2017.

[29] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

[30] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[31] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[32] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

[33] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

[34] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[35] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[36] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915. PMLR, 2019.

[37] Dmitry Kovalev and Alexander Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:14691–14703, 2022.

[38] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019.

[39] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1-2):1–35, 2021.

[40] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

[41] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

[42] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

[43] Kiran K Thekumparampil, Niao He, and Sewoong Oh. Lifted primal-dual method for bilinearly coupled smooth minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 4281–4308. PMLR, 2022.

[44] Dmitry Kovalev, Alexander Gasnikov, and Peter Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *Advances in Neural Information Processing Systems*, 35:21725–21737, 2022.

[45] Guangzeng Xie, Yuze Han, and Zhihua Zhang. Dippa: An improved method for bilinear saddle point problems. *arXiv preprint arXiv:2103.08270*, 2021.

[46] Chris Junchi Li, Huizhuo Yuan, Gauthier Gidel, Quanquan Gu, and Michael I. Jordan. Nesterov meets optimism: Rate-optimal separable minimax optimization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 20351–20383. PMLR, 2023.

[47] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2, 2018.

[48] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[49] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[50] Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[51] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *arXiv preprint arXiv:1908.08465*, 2019.

[52] Simon S Du, Gauthier Gidel, Michael I Jordan, and Chris Junchi Li. Optimal extragradient-based bilinearly-coupled saddle-point optimization. *arXiv preprint arXiv:2206.08573*, 2022.

[53] Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. Random extrapolation for primal-dual coordinate descent. In *International conference on machine learning*, pages 191–201. PMLR, 2020.

[54] Erfan Yazdandoost Hamedani, Afrooz Jalilzadeh, and Necdet S Aybat. Randomized primal-dual methods with adaptive step sizes. In *International Conference on Artificial Intelligence and Statistics*, pages 11185–11212. PMLR, 2023.

[55] Afrooz Jalilzadeh, Erfan Yazdandoost Hamedani, and Necdet S Aybat. A doubly-randomized block-coordinate primal-dual method for large-scale saddle point problems. *arXiv preprint arXiv:1907.03886*, 2019.

[56] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the L1-ball for learning in high dimensions. In *International Conference on Machine Learning*, 2008.

[57] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[58] Quanqi Hu, Zi-Hao Qiu, Zhishuai Guo, Lijun Zhang, and Tianbao Yang. Blockwise stochastic variance-reduced methods with parallel speedup for multi-block bilevel optimization. In *International Conference on Machine Learning*, 2023.

[59] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 15236–15245, 2019.

[60] Cong Han Lim and Stephen J Wright. Efficient bregman projections onto the permutahedron and related polytopes. In *Artificial Intelligence and Statistics*, pages 1205–1213. PMLR, 2016.

[61] Tasuku Soma, Khashayar Gatmiry, and Stefanie Jegelka. Optimal algorithms for group distributionally robust optimization and beyond. *arXiv preprint arXiv:2212.13669*, 2022.

[62] Quanqi Hu, Dixian Zhu, and Tianbao Yang. Non-smooth weakly-convex finite-sum coupled compositional optimization. In *Advances in Neural Information Processing Systems*, volume abs/2310.03234, 2023.

[63] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 467–482. Springer, 2016.

[64] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[65] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[66] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[67] John C Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods*, pages 185–208, 1999.

[68] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[69] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

[70] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

[71] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

Appendix

A Basic Lemmas

Lemma 11. *Suppose that* $y_0^{(i)} = f_i'(u_0^{(i)}) \in \partial f_i(u_0^{(i)})$ *for some* $u_0^{(i)} \in \mathbb{R}$ *and*

$$u_{t+1}^{(i)} = \begin{cases} \frac{\tau}{1+\tau} u_t^{(i)} + \frac{1}{1+\tau} \tilde{g}_t^{(i)}, & i \in \mathcal{S}_t \\ u_t^{(i)}, & i \notin \mathcal{S}_t \end{cases}. \tag{A.1}$$

Algorithm 1 with $\psi_i = f_i^*$ *satisfies that* $y_t^{(i)} = f_i'(u_t^{(i)}) \in \partial f_i(u_t^{(i)})$ *for all* $i \in \{1, \dots, n\}$ *and* $t \geq 0$.

Proof. We prove it by induction. The base case follows from the premise. Assume that $y_t^{(i)} = f_i'(u_t^{(i)}) \in \partial f_i(u_t^{(i)})$. We discuss two cases.

- Case I ($i \notin \mathcal{S}_t$): Note that $y_{t+1}^{(i)} = y_t^{(i)}$ and $u_{t+1}^{(i)} = u_t^{(i)}$. Thus, $y_{t+1}^{(i)} = f_i'(u_{t+1}^{(i)}) \in \partial f_i(u_{t+1}^{(i)})$.
- Case II ($i \in \mathcal{S}_t$): This part resembles Lemma 2 in Zhang and Lan [22]. Based on the update rule and the premise $y_t^{(i)} \in \partial f_i(u_t^{(i)})$ ($\Leftrightarrow u_t^{(i)} \in \partial f_i^*(y_t^{(i)})$), we have

$$\begin{aligned} y_{t+1}^{(i)} &= \arg\max_{y^{(i)}} \left\{ y^{(i)} \tilde{g}_t^{(i)} - f_i^*(y^{(i)}) - \tau \left(f_i^*(y^{(i)}) - (f_i^*)'(y_t^{(i)}) \cdot y^{(i)} \right) \right\} \\ &= \arg\max_{y^{(i)}} \left\{ \left(\frac{1}{1+\tau} \tilde{g}_t^{(i)} + \frac{\tau}{1+\tau} u_t^{(i)} \right) \cdot y^{(i)} - f_i^*(y^{(i)}) \right\} \\ &\in \partial f_i \left(\frac{1}{1+\tau} \tilde{g}_t^{(i)} + \frac{\tau}{1+\tau} u_t^{(i)} \right) = \partial f_i(u_{t+1}^{(i)}). \end{aligned}$$

□

Lemma 12 (Corollary 2 in [49])**.** *Consider an adapted sequence* $\{\Delta_t, \mathcal{F}_t\}_{t \geq 0}$ *where* (Δ_t) *is a martingale difference sequence. Define a sequence* $\{\hat{\pi}_t\}_t$*:*

$$\hat{\pi}_0 = 0, \quad \hat{\pi}_{t+1} = \arg\min_v \{ \langle -\Delta_t, v \rangle + \alpha U_\psi(v, \hat{\pi}_t) \},$$

where we also assume that ψ *is* μ_ψ*-strongly convex w.r.t.* $\|\cdot\|$ ($\mu_\psi > 0$)*. For any* v *(that possibly depends on* Δ_t*) we have*

$$\mathbf{E}\left[\langle \Delta_t, v \rangle \right] \leq \mathbf{E}\left[\alpha U_\psi(v, \hat{\pi}_t) - \alpha U_\psi(v, \hat{\pi}_{t+1}) \right] + \frac{1}{2\alpha\mu_\psi} \mathbf{E} \|\Delta_t\|_*^2 .$$

Proof. Use the three-point inequality:

$$\langle -\Delta_t, \hat{\pi}_{t+1} - v \rangle \leq \alpha U_\psi(v, \hat{\pi}_t) - \alpha U_\psi(v, \hat{\pi}_{t+1}) - \alpha U_\psi(\hat{\pi}_{t+1}, \hat{\pi}_t).$$

Add $\langle -\Delta_t, \hat{\pi}_t - \hat{\pi}_{t+1} \rangle$ to both sides and use Young's inequality.

$$\begin{aligned} \langle -\Delta_t, \hat{\pi}_t - v \rangle &\leq \alpha U_\psi(v, \hat{\pi}_t) - \alpha U_\psi(v, \hat{\pi}_{t+1}) - \alpha U_\psi(\hat{\pi}_{t+1}, \hat{\pi}_t) + \langle \Delta_t, \hat{\pi}_{t+1} - \hat{\pi}_t \rangle \\ &\leq \alpha U_\psi(v, \hat{\pi}_t) - \alpha U_\psi(v, \hat{\pi}_{t+1}) - \alpha U_\psi(\hat{\pi}_{t+1}, \hat{\pi}_t) + \frac{\alpha\mu_\psi}{2} \|\hat{\pi}_{t+1} - \hat{\pi}_t\|^2 + \frac{1}{2\alpha\mu_\psi} \|\Delta_t\|_*^2 . \end{aligned}$$

If ψ is μ_ψ-strongly convex, we have $-U_\psi(\hat{\pi}_{t+1}, \hat{\pi}_t) \leq -\frac{\mu_\psi}{2} \|\hat{\pi}_{t+1} - \hat{\pi}_t\|^2$. Lastly, $\mathbf{E}_t[\Delta_t, \hat{\pi}_t] = 0$. □

Lemma 13 (Lemma 4 in [49] and Lemma 7 in [22]). *Let $\Pi \subset \mathbb{R}^m$ be a non-empty closed and convex domain and let function $u(\pi)$ be μ-strongly convex on Π w.r.t. $\|\cdot\|$. For a $\hat{\pi}$ generated via a prox-mapping with the argument $g+\delta$, $\hat{\pi} \leftarrow \arg\min_{\pi\in\Pi}\{\langle \pi, g+\delta-u'(\underline{\pi})\rangle + u(\pi)\}$ for some $\underline{\pi} \in \Pi$, where δ denotes a noise term with $\mathbf{E}[\delta]=0$ and $\mathbf{E}[\|\delta\|_*^2] \le \sigma_0^2$. Then, for $\bar{\pi}$ generated via a prox-mapping with the argument g, $\bar{\pi} \leftarrow \arg\min_{\pi\in\Pi}\{\langle \pi, g-u'(\underline{\pi})\rangle + u(\pi)\}$, we have*

$$\|\hat{\pi}-\bar{\pi}\| \le \|\delta\|_*/\mu, \tag{A.2}$$

$$|\mathbf{E}\langle \hat{\pi}, \delta\rangle| \le \sigma_0^2/\mu. \tag{A.3}$$

For completeness, we present the proof of the lemma above. We do not claim any novelty here.

Proof. By the optimality condition of prox-mapping, we have

$$\left\langle u'(\hat{\pi})-u'(\underline{\pi})+g+\delta, \hat{\pi}-\pi\right\rangle \le 0, \quad \forall \pi\in\Pi, \tag{A.4}$$

$$\left\langle u'(\bar{\pi})-u'(\underline{\pi})+g, \bar{\pi}-\pi\right\rangle \le 0, \quad \forall \pi\in\Pi. \tag{A.5}$$

Choose $\pi=\bar{\pi}$ in (A.4) and $\pi=\hat{\pi}$ in (A.5). By combining (A.4) and (A.5), we have

$$\|\delta\|_*\|\hat{\pi}-\bar{\pi}\| \ge \langle \delta, \hat{\pi}-\bar{\pi}\rangle \ge \left\langle u'(\hat{\pi})-u'(\bar{\pi}), \hat{\pi}-\bar{\pi}\right\rangle.$$

Since u is μ-strongly convex, we have $\langle u'(\hat{\pi})-u'(\bar{\pi}), \hat{\pi}-\bar{\pi}\rangle \ge \mu\|\hat{\pi}-\bar{\pi}\|^2$. Thus, $\|\hat{\pi}-\bar{\pi}\| \le \|\delta\|_*/\mu$. Moreover, the triangle inequality leads to $|\mathbf{E}\langle \hat{\pi}, \delta\rangle| \le |\mathbf{E}\langle \hat{\pi}-\bar{\pi}, \delta\rangle| + |\mathbf{E}\langle \bar{\pi}, \delta\rangle|$. Note that $\mathbf{E}\langle \bar{\pi}, \delta\rangle = 0$. Moreover, Cauchy-Schwartz inequality and (A.2) leads to

$$|\mathbf{E}\langle \hat{\pi}, \delta\rangle| \le |\mathbf{E}\langle \hat{\pi}-\bar{\pi}, \delta\rangle| \le \mathbf{E}[\|\hat{\pi}-\bar{\pi}\|\,\|\delta\|_*] \le \mathbf{E}\|\delta\|_*^2/\mu \le \sigma_0^2/\mu.$$

□

Next, we present a basic inequality about the mirror proximal update. Similar results have been widely used in the literature, e.g., Lemma 3.8 in Lan [69] and Lemma 7.1 in Hamedani and Aybat [70].

Lemma 14. *Suppose that the function $\phi: \mathcal{X} \to \mathbb{R}$ is on a convex closed domain $\mathcal{X}$ and ϕ is μ-convex ($\mu \ge 0$) with respect to a prox-function $U_\psi(x,y) \coloneqq \psi(x)-\psi(y)-\langle \psi'(y), x-y\rangle$ for any $x, y \in \mathcal{X}$ with a generating function ψ, i.e., $\phi(x) \ge \phi(y)+\langle \phi'(y), x-y\rangle + \mu U_\psi(x,y)$, $\forall x, y \in \mathcal{X}$. For $\hat{x} = \arg\min_{x\in\mathcal{X}}\{\phi(x)+\eta U_\psi(\underline{x}, x)\}$, we have*

$$\phi(\hat{x})-\phi(x) \le \eta U_\psi(x,\underline{x}) - (\eta+\mu)U_\psi(x,\hat{x}) - \eta U_\psi(\hat{x},\underline{x}), \quad \forall x\in\mathcal{X}. \tag{A.6}$$

Proof. By the definition of the prox-function $U_\psi(x,y)$, we have

$$\begin{aligned}
&U_\psi(x,\underline{x}) - U_\psi(x,\hat{x}) - U_\psi(\hat{x},\underline{x}) \\
&= \psi(x)-\psi(\underline{x})-\left\langle \psi'(\underline{x}), x-\underline{x}\right\rangle - \psi(x)+\psi(\hat{x})+\left\langle \psi'(\hat{x}), x-\hat{x}\right\rangle - \psi(\hat{x})+\psi(\underline{x})+\left\langle \psi'(\underline{x}), \hat{x}-\underline{x}\right\rangle \\
&= \left\langle \psi'(\hat{x})-\psi'(\underline{x}), x-\hat{x}\right\rangle.
\end{aligned}$$

By the strong convexity of ϕ with respect to ψ, we have $\phi(x)-\phi(\hat{x}) \ge \langle \phi'(\hat{x}), x-\hat{x}\rangle + \mu U_\psi(x,\hat{x})$. The optimality condition of the prox-mapping implies that $\langle \phi'(\hat{x})+\eta(\psi'(\hat{x})-\psi'(\underline{x})), x-\hat{x}\rangle \ge 0$ for any $x\in\mathcal{X}$. Thus, we obtain $\langle \phi'(\hat{x}), x-\hat{x}\rangle \ge \eta\langle \psi'(\hat{x})-\psi'(\underline{x}), x-\hat{x}\rangle$ such that

$$\begin{aligned}
\phi(x)-\phi(\hat{x}) &\ge \left\langle \phi'(\hat{x}), x-\hat{x}\right\rangle + \mu U_\psi(\hat{x},x) \\
&\ge \eta\left\langle \psi'(\underline{x})-\psi'(\hat{x}), x-\hat{x}\right\rangle + \mu U_\psi(\hat{x},x) \ge -\eta U_\psi(x,\underline{x}) + (\eta+\mu)U_\psi(x,\hat{x}) + U_\psi(\hat{x},\underline{x}).
\end{aligned}$$

□

B Proof of Lemma 2

Proof. According to Lemma 14, the primal update rule implies that

$$-\langle G_t,\, x - x_{t+1}\rangle + r(x_{t+1}) - r(x) \le \frac{\eta}{2}\|x - x_t\|_2^2 - \frac{\eta+\mu}{2}\|x - x_{t+1}\|_2^2 - \frac{\eta}{2}\|x_{t+1} - x_t\|_2^2. \tag{B.1}$$

Similarly, for all $i \in [n]$ the dual update rule implies that

$$\left\langle \tilde{g}_t^{(i)},\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + f_i^*(\bar{y}_{t+1}^{(i)}) - f_i^*(y^{(i)}) \le \tau U_{\psi_i}(y^{(i)}, y_t^{(i)}) - (\tau+\rho)U_{\psi_i}(y^{(i)}, \bar{y}_{t+1}^{(i)}) - \tau U_{\psi_i}(\bar{y}_{t+1}^{(i)}, y_t^{(i)}).$$

Average this equation over $i = 1, \dots, n$.

$$\frac{1}{n}\sum_{i=1}^n \left\langle \tilde{g}_t^{(i)},\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + \frac{1}{n}\sum_{i=1}^n f_i^*(\bar{y}_{t+1}^{(i)}) - \frac{1}{n}\sum_{i=1}^n f_i^*(y^{(i)}) \le \frac{\tau}{n}U_\psi(y, y_t) - \frac{\tau+\rho}{n}U_\psi(y, \bar{y}_{t+1}) - \frac{\tau}{n}U_\psi(\bar{y}_{t+1}, y_t). \tag{B.2}$$

By the definition of $L(x, y)$ in (1.2), we have

$$\begin{aligned}
&L(x_{t+1}, y) - L(x, \bar{y}_{t+1}) \\
&= \frac{1}{n}\sum_{i=1}^n \left\langle y^{(i)},\, g_i(x_{t+1}) \right\rangle - \frac{1}{n}\sum_{i=1}^n f_i^*(y^{(i)}) + r(x_{t+1}) - \frac{1}{n}\sum_{i=1}^n \left\langle \bar{y}_{t+1}^{(i)},\, g_i(x) \right\rangle + \frac{1}{n}\sum_{i=1}^n f_i^*(\bar{y}_{t+1}^{(i)}) - r(x) \\
&= \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + \frac{1}{n}\sum_{i=1}^n f_i^*(\bar{y}_{t+1}^{(i)}) - \frac{1}{n}\sum_{i=1}^n f_i^*(y^{(i)}) + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - g_i(x),\, \bar{y}_{t+1}^{(i)} \right\rangle \\
&\quad + r(x_{t+1}) - r(x).
\end{aligned}$$

Combine the equation above with (B.1) and (B.2).

$$\begin{aligned}
&L(x_{t+1}, y) - L(x, \bar{y}_{t+1}) \\
&\le \frac{\tau}{n}U_\psi(y, y_t) - \frac{\tau+\rho}{n}U_\psi(y, \bar{y}_{t+1}) - \frac{\tau}{n}U_\psi(\bar{y}_{t+1}, y_t) + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - \tilde{g}_t^{(i)},\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + \frac{\eta}{2}\|x - x_t\|_2^2 \\
&\quad - \frac{\eta+\mu}{2}\|x - x_{t+1}\|_2^2 - \frac{\eta}{2}\|x_{t+1} - x_t\|_2^2 + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - g_i(x),\, \bar{y}_{t+1}^{(i)} \right\rangle - \langle G_t,\, x_{t+1} - x\rangle.
\end{aligned}$$

□

C Convergence Analysis of ALEXR in the Strongly Convex Case

In this section, we present several lemmas that upper-bound different terms in (6.1) with $x = x_*$, $y = y_*$, where we define $x_* = \min_{x\in\mathcal{X}} F(x)$, $y_* = \arg\max_{y\in\mathcal{Y}} L(x_*, y)$.

C.1 Supporting Lemmas

Lemma 15. *Under Assumptions 4, 5, 6, (C.1) holds for Algorithm 1 with $\theta < 1$ and any $\lambda_2, \lambda_3 > 0$.*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\langle g_i(x_{t+1}) - \tilde{g}_t^{(i)},\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle \\
&\le \Gamma_{t+1} - \theta\Gamma_t + \frac{C_g^2\|x_{t+1} - x_t\|_2^2}{2\lambda_2} + \frac{\theta C_g^2\|x_t - x_{t-1}\|_2^2}{2\lambda_3} + \frac{(\lambda_2 + \lambda_3\theta)U_\psi(\bar{y}_{t+1}, y_t)}{\mu_\psi n} + \frac{2(1+2\theta)\sigma_0^2}{B\mu_\psi(\rho+\tau)},
\end{aligned} \tag{C.1}$$

where $\Gamma_t \coloneqq \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_{t-1}),\, y_*^{(i)} - y_t^{(i)} \right\rangle$.

Proof. The $\frac{1}{n}\sum_{i=1}^n \mathbf{E}(g_i(x_{t+1}) - \tilde{g}_t^{(i)})(y_*^{(i)} - \bar{y}_{t+1}^{(i)})$ term can be decomposed as

$$\begin{aligned}
\diamond &= \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - \tilde{g}_t^{(i)},\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle \qquad \text{(C.2)}\\
&= \underbrace{\frac{1+\theta}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}),\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle}_{\text{I}} + \underbrace{\frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}),\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle - \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t),\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle}_{\text{II}}\\
&\quad + \underbrace{\frac{\theta}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}) - g_i(x_t),\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle}_{\text{III}} + \underbrace{\frac{\theta}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1};\mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle}_{\text{IV}}.
\end{aligned}$$

Taking conditional expectations of terms I and IV leads to $\mathbf{E}\left[\left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}),\, y_*^{(i)} \right\rangle \mid \mathcal{F}_{t-1}\right] = 0$ and $\mathbf{E}\left[\left\langle g_i(x_{t-1}) - g_i(x_{t-1};\mathcal{B}_t^{(i)}),\, y_*^{(i)} \right\rangle \mid \mathcal{F}_{t-1}\right] = 0$. Define $\dot{y}_{t+1}^{(i)} \coloneqq \arg\max_{v\in\mathcal{Y}_i}\{v^\top \bar{g}_t^{(i)} - f_i^*(v) - \tau U_{\psi_i}(v, y_t^{(i)})\}$ and $\bar{g}_t^{(i)} \coloneqq g_i(x_t) + \theta(g_i(x_t) - g_i(x_{t-1}))$, $\forall i \in [n]$. Note that $\dot{y}_{t+1}^{(i)}$ is independent of $\mathcal{B}_t^{(i)}$ such that $\mathbf{E}\left[\left\langle g_i(x_t;\mathcal{B}_t^{(i)}) - g_i(x_t),\, \dot{y}_{t+1}^{(i)} \right\rangle \mid \mathcal{F}_{t-1}\right] = 0$.

$$\begin{aligned}
&\mathbf{E}\left[\left\langle g_i(x_t;\mathcal{B}_t^{(i)}) - g_i(x_t),\, \bar{y}_{t+1}^{(i)} \right\rangle\right]\\
&= \mathbf{E}\left[\left\langle g_i(x_t;\mathcal{B}_t^{(i)}) - g_i(x_t),\, \bar{y}_{t+1}^{(i)} - \dot{y}_{t+1}^{(i)} \right\rangle\right] \le \mathbf{E}\left\|g_i(x_t;\mathcal{B}_t^{(i)}) - g_i(x_t)\right\|_* \left\|\bar{y}_{t+1}^{(i)} - \dot{y}_{t+1}^{(i)}\right\|\\
&\overset{\text{Lemma 13}}{\le} \frac{1}{\mu_\psi(\rho+\tau)}\mathbf{E}\left\|g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)})\right\|_* \left\|(1+\theta)(g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)})) - \theta(g_i(x_{t-1}) - g_i(x_{t-1};\mathcal{B}_t^{(i)}))\right\|_*\\
&= \frac{(1+\theta)\mathbf{E}\left\|g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)})\right\|_*^2}{\mu_\psi(\rho+\tau)} + \frac{\theta\mathbf{E}\left\|g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)})\right\|_* \left\|g_i(x_{t-1}) - g_i(x_{t-1};\mathcal{B}_t^{(i)})\right\|_*}{\mu_\psi(\rho+\tau)}\\
&\le \frac{(1+\theta)\mathbf{E}\left\|g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)})\right\|_*^2}{\mu_\psi(\rho+\tau)} + \frac{0.5\theta\mathbf{E}\left\|g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)})\right\|_*^2 + 0.5\theta\left\|g_i(x_{t-1}) - g_i(x_{t-1};\mathcal{B}_t^{(i)})\right\|_*^2}{\mu_\psi(\rho+\tau)}\\
&\le \frac{(1+2\theta)\sigma_0^2}{B\mu_\psi(\rho+\tau)},
\end{aligned}$$

$$\mathbf{E}\left[\left\langle g_i(x_{t-1};\mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, \bar{y}_{t+1}^{(i)} \right\rangle\right] = \mathbf{E}\left[\left\langle g_i(x_{t-1};\mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, \bar{y}_{t+1}^{(i)} - \dot{y}_{t+1}^{(i)} \right\rangle\right] \le \frac{(1+2\theta)\sigma_0^2}{B\mu_\psi(\rho+\tau)}.$$

Define $\Gamma_t \coloneqq \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_{t-1}),\, y_*^{(i)} - y_t^{(i)} \right\rangle$. II + III in (C.2) can be rewritten as

$$\begin{aligned}
\text{II} + \text{III} &= \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}),\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle - \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t),\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + \frac{\theta}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}) - g_i(x_t),\, y_*^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle\\
&= \Gamma_{t+1} - \theta\Gamma_t + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - g_i(x_t),\, y_{t+1}^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + \frac{\theta}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}) - g_i(x_t),\, y_t^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle\\
&\le \Gamma_{t+1} - \theta\Gamma_t + \frac{1}{n}\sum_{i=1}^n \left\|g_i(x_{t+1}) - g_i(x_t)\right\|_* \left\|y_{t+1}^{(i)} - \bar{y}_{t+1}^{(i)}\right\| + \frac{\theta}{n}\sum_{i=1}^n \left\|g_i(x_{t-1}) - g_i(x_t)\right\|_* \left\|y_t^{(i)} - \bar{y}_{t+1}^{(i)}\right\|\\
&\le \Gamma_{t+1} - \theta\Gamma_t + \frac{C_g^2\left\|x_{t+1} - x_t\right\|_2^2}{2\lambda_2} + \frac{\theta C_g^2\left\|x_t - x_{t-1}\right\|_2^2}{2\lambda_3} + \frac{(\lambda_2 + \lambda_3\theta)U_\psi(\bar{y}_{t+1}, y_t)}{\mu_\psi n}.
\end{aligned}$$

□

Lemma 16. *Suppose that g_i is L_g-smooth and Assumptions 1, 2, 3, 4, 5 hold. Then, the following holds for Algorithm 1.*

$$\frac{1}{n}\mathbf{E}\sum_{i=1}^{n}\left\langle g_i(x_{t+1})-g_i(x_*),\,\bar{y}_{t+1}^{(i)}\right\rangle-\mathbf{E}\left\langle G_t,\,x_{t+1}-x_*\right\rangle\le\frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\eta+\mu}+\frac{L_gC_f}{2}\left\|x_{t+1}-x_t\right\|_2^2. \tag{C.3}$$

Proof. We define $\Delta_t\coloneqq\frac{1}{S}\sum_{i\in\mathcal{S}_t}[\nabla g_i(x_t;\tilde{\mathcal{B}}_t^{(i)})]^\top y_{t+1}^{(i)}-\frac{1}{n}\sum_{i=1}^{n}[\nabla g_i(x_t)]^\top\bar{y}_{t+1}^{(i)}$.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\langle g_i(x_{t+1})-g_i(x_*),\,\bar{y}_{t+1}^{(i)}\right\rangle-\left\langle G_t,\,x_{t+1}-x_*\right\rangle\\
&=\frac{1}{n}\sum_{i=1}^{n}\left\langle g_i(x_{t+1})-g_i(x_t),\,\bar{y}_{t+1}^{(i)}\right\rangle+\frac{1}{n}\sum_{i=1}^{n}\left\langle g_i(x_t)-g_i(x_*),\,\bar{y}_{t+1}^{(i)}\right\rangle+\left\langle\frac{1}{n}\sum_{i=1}^{n}[\nabla g_i(x_t)]^\top\bar{y}_{t+1}^{(i)}+\Delta_t,\,x_*-x_{t+1}\right\rangle\\
&\overset{\substack{g_i\text{ convex}\\ \mathcal{Y}_i\subseteq\mathbb{R}_+^m}}{\le}\frac{1}{n}\sum_{i=1}^{n}\left\langle g_i(x_{t+1})-g_i(x_t),\,\bar{y}_{t+1}^{(i)}\right\rangle+\left\langle\frac{1}{n}\sum_{i=1}^{n}[\nabla g_i(x_t)]^\top\bar{y}_{t+1}^{(i)},\,x_t-x_*\right\rangle+\left\langle\frac{1}{n}\sum_{i=1}^{n}[\nabla g_i(x_t)]^\top\bar{y}_{t+1}^{(i)}+\Delta_t,\,x_*-x_{t+1}\right\rangle\\
&=\frac{1}{n}\sum_{i=1}^{n}\left\langle\bar{y}_{t+1}^{(i)},\,g_i(x_{t+1})-g_i(x_t)\right\rangle+\left\langle\frac{1}{n}\sum_{i=1}^{n}[\nabla g_i(x_t)]^\top\bar{y}_{t+1}^{(i)},\,x_t-x_{t+1}\right\rangle+\left\langle\Delta_t,\,x_*-x_{t+1}\right\rangle,
\end{aligned}\tag{C.4}$$

We bound the first two terms above by the Lipschitz continuity of f_i and ∇g_i.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\langle\bar{y}_{t+1}^{(i)},\,g_i(x_{t+1})-g_i(x_t)\right\rangle+\left\langle\frac{1}{n}\sum_{i=1}^{n}[\nabla g_i(x_t)]^\top\bar{y}_{t+1}^{(i)},\,x_t-x_{t+1}\right\rangle\\
&=\frac{1}{n}\sum_{i=1}^{n}\left\langle\bar{y}_{t+1}^{(i)},\,g_i(x_{t+1})-g_i(x_t)-\nabla g_i(x_t)(x_{t+1}-x_t)\right\rangle\\
&\le\frac{1}{n}\sum_{i=1}^{n}\left\|\bar{y}_{t+1}^{(i)}\right\|\left\|g_i(x_{t+1})-g_i(x_t)-\nabla g_i(x_t)(x_{t+1}-x_t)\right\|_*\le\frac{C_f}{n}\sum_{i=1}^{n}\left\|g_i(x_{t+1})-g_i(x_t)-\nabla g_i(x_t)(x_{t+1}-x_t)\right\|_*.
\end{aligned}$$

Due to the L_g-smoothness of g_i, we have

$$\left\|g_i(x_{t+1})-g_i(x_t)-\nabla g_i(x_t)(x_{t+1}-x_t)\right\|_*\le\frac{L_g}{2}\left\|x_{t+1}-x_t\right\|_2^2.$$

Thus, the first two terms in (C.4) can be upper bounded by

$$\frac{1}{n}\sum_{i=1}^{n}\left\langle\bar{y}_{t+1}^{(i)},\,g_i(x_{t+1})-g_i(x_t)\right\rangle+\left\langle\frac{1}{n}\sum_{i=1}^{n}[\nabla g_i(x_t)]^\top\bar{y}_{t+1}^{(i)},\,x_t-x_{t+1}\right\rangle\le\frac{L_gC_f}{2}\left\|x_{t+1}-x_t\right\|_2^2. \tag{C.5}$$

Besides, we have that $\mathbf{E}[\langle\Delta_t,\,x_*\rangle\mid\mathcal{F}_{t-1}]=0$. By the Lipschitz continuity of f_i and the definition of the operator norm, we have $\left\|\left([\nabla g_i(x_t)]^\top-[\nabla g_i(x_t;\tilde{B}_t^{(i)})]^\top\right)\bar{y}_{t+1}^{(i)}\right\|_2\le\left\|[\nabla g_i(x_t)]^\top-[\nabla g_i(x_t;\tilde{B}_t^{(i)})]^\top\right\|_{\mathrm{op}}\left\|\bar{y}_{t+1}^{(i)}\right\|\le C_f\left\|[\nabla g_i(x_t)]^\top-[\nabla g_i(x_t;\tilde{B}_t^{(i)})]^\top\right\|_{\mathrm{op}}$. According to Lemma 13 and Assumption 5, we can derive that

$$-\mathbf{E}[\langle x_{t+1},\,\Delta_t\rangle]\le\frac{\mathbf{E}\left\|\Delta_t\right\|_2^2}{\mu+\eta}\le\frac{1}{\mu+\eta}\left(\frac{\delta^2}{S}+\mathbf{E}\left\|\frac{1}{S}\sum_{i\in\mathcal{S}_t}\left([\nabla g_i(x_t)]^\top-[\nabla g_i(x_t;\tilde{B}_t^{(i)})]^\top\right)\bar{y}_{t+1}^{(i)}\right\|_2^2\right)\le\frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\mu+\eta}. \tag{C.6}$$

Then, combining (C.4), (C.5) and (C.6) leads to

$$\frac{1}{n}\mathbf{E}\sum_{i=1}^{n}\left\langle g_i(x_{t+1})-g_i(x_*),\,\bar{y}_{t+1}^{(i)}\right\rangle-\mathbf{E}\left\langle G_t,\,x_{t+1}-x\right\rangle\le\frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\mu+\eta}+\frac{L_gC_f}{2}\left\|x_{t+1}-x_t\right\|_2^2.$$

□

Lemma 17. *Suppose that g_i is non-smooth and Assumptions 1, 2, 3, 4, 5 hold. The following holds for Algorithm 1.*

$$\frac{1}{n}\mathbf{E}\sum_{i=1}^{n}\left\langle g_i(x_{t+1})-g_i(x_*),\,\bar{y}_{t+1}^{(i)}\right\rangle-\mathbf{E}\left\langle G_t,\,x_{t+1}-x\right\rangle\leq\frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}+4C_f^2C_g^2}{\mu+\eta}+\frac{\eta+\mu}{4}\left\|x_{t+1}-x_t\right\|_2^2. \quad \text{(C.7)}$$

Proof. Note that (C.4) and (C.6) still hold. Since g_i is non-smooth, we need to bound the left-hand side of (C.5) in a different way. Based on the definition of the operator norm and the Lipschitz continuity of g_i, we have $\left\|g_i'(x_t)(x_t-x_{t+1})\right\|_*\leq\left\|g_i'(x_t)\right\|_{\text{op}}\left\|x_t-x_{t+1}\right\|_2\leq C_g\left\|x_t-x_{t+1}\right\|_2$ such that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\langle\bar{y}_{t+1}^{(i)},\,g_i(x_{t+1})-g_i(x_t)\right\rangle+\left\langle\frac{1}{n}\sum_{i=1}^{n}[g_i'(x_t)]^\top\bar{y}_{t+1}^{(i)},\,x_t-x_{t+1}\right\rangle\\
&=\frac{1}{n}\sum_{i=1}^{n}\left\langle\bar{y}_{t+1}^{(i)},\,g_i(x_{t+1})-g_i(x_t)\right\rangle+\frac{1}{n}\sum_{i=1}^{n}\left\langle\bar{y}_{t+1}^{(i)},\,g_i'(x_t)(x_t-x_{t+1})\right\rangle\\
&\leq\frac{1}{n}\sum_{i=1}^{n}\left\|\bar{y}_{t+1}^{(i)}\right\|\left\|g_i(x_{t+1})-g_i(x_t)\right\|_*+\frac{1}{n}\sum_{i=1}^{n}\left\|\bar{y}_{t+1}^{(i)}\right\|\left\|g_i'(x_t)(x_t-x_{t+1})\right\|_*\\
&\leq 2C_fC_g\left\|x_{t+1}-x_t\right\|_2\leq\frac{4C_f^2C_g^2}{\eta+\mu}+\frac{\eta+\mu}{4}\left\|x_{t+1}-x_t\right\|_2^2, \qquad \text{(C.8)}
\end{aligned}$$

where $g_i'(x_t)\in\partial g_i(x_t)$. Merge (C.4), (C.6), and (C.8).

$$\frac{1}{n}\mathbf{E}\sum_{i=1}^{n}\left\langle g_i(x_{t+1})-g_i(x_*),\,\bar{y}_{t+1}^{(i)}\right\rangle-\mathbf{E}\left\langle G_t,\,x_{t+1}-x\right\rangle\leq\frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}+4C_f^2C_g^2}{\mu+\eta}+0.25(\eta+\mu)\left\|x_{t+1}-x_t\right\|_2^2.$$

□

C.2 Proof of Theorem 3

Proof. If g_i is smooth, we combine (6.1), (6.2), (C.1), and (C.3).

$$\begin{aligned}
&\mathbf{E}[L(x_{t+1},y_*)-L(x_*,\bar{y}_{t+1})]\\
&\leq\frac{\tau+\rho\left(1-\frac{S}{n}\right)}{S}\mathbf{E}[U_\psi(y_*,y_t)]-\frac{\tau+\rho}{S}\mathbf{E}[U_\psi(y_*,y_{t+1})]+\frac{\eta}{2}\mathbf{E}\left\|x_*-x_t\right\|_2^2-\frac{\eta+\mu}{2}\mathbf{E}\left\|x_*-x_{t+1}\right\|_2^2\\
&\quad-\left(\frac{\tau}{n}-\frac{\lambda_2+\lambda_3\theta}{\mu_\psi n}\right)\mathbf{E}\left[U_\psi(\bar{y}_{t+1},y_t)\right]-\left(\frac{\eta}{2}-\frac{C_g^2}{2\lambda_2}-\frac{L_gC_f}{2}\right)\mathbf{E}\left\|x_{t+1}-x_t\right\|_2^2+\frac{\theta C_g^2}{2\lambda_3}\mathbf{E}\left\|x_t-x_{t-1}\right\|_2^2\\
&\quad+\mathbf{E}[\Gamma_{t+1}-\theta\Gamma_t]+\frac{2(1+2\theta)\sigma_0^2}{B\mu_\psi(\rho+\tau)}+\frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\eta+\mu}. \qquad \text{(C.9)}
\end{aligned}$$

Define $\Upsilon_t^x\coloneqq\frac{1}{2}\mathbf{E}\left\|x_*-x_t\right\|_2^2$ and $\Upsilon_t^y=\frac{1}{S}\mathbf{E}U_\psi(y_*,y_t)$. Note that $L(x_{t+1},y_*)-L(x_*,\bar{y}_{t+1})\geq 0$. Multiply both sides of (C.9) by θ^{-t} and do telescoping sum from $t=0$ to $T-1$. Add $\eta\theta^{-T}\Upsilon_T^x$ to both sides.

$$\begin{aligned}
\eta\theta^{-T}\Upsilon_T^x&\leq\sum_{t=0}^{T-1}\theta^{-t}\left(\left(\eta\Upsilon_t^x+\left(\tau+\rho\left(1-\frac{S}{n}\right)\right)\Upsilon_t^y-\theta\mathbf{E}\Gamma_t\right)-\left((\eta+\mu)\Upsilon_{t+1}^x+(\tau+\rho)\Upsilon_{t+1}^y-\mathbf{E}\Gamma_{t+1}\right)\right)\\
&\quad+\eta\theta^{-T}\Upsilon_T^x+\left(\frac{2(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)}+\frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\eta+\mu}\right)\sum_{t=0}^{T-1}\theta^{-t}-\sum_{t=0}^{T-1}\theta^{-t}\left(\frac{\tau}{n}-\frac{(\lambda_2+\lambda_3\theta)}{\mu_\psi n}\right)\mathbf{E}[U_\psi(\bar{y}_{t+1},y_t)]\\
&\quad-\sum_{t=0}^{T-1}\theta^{-t}\left(\frac{\eta}{2}-\frac{L_gC_f}{2}-\frac{C_g^2}{2\lambda_2}-\frac{C_g^2}{2\lambda_3}\right)\mathbf{E}\left\|x_{t+1}-x_t\right\|_2^2.
\end{aligned}$$

Let $\eta \geq \frac{\mu\theta}{1-\theta}$ such that $\theta \leq \frac{\eta}{\eta+\mu}$ and $\tau \geq \frac{\rho S}{n(1-\theta)}$ such that $\theta \leq \frac{\tau+\rho\left(1-\frac{S}{n}\right)}{\tau+\rho}$. Then,

$$\begin{aligned}
&\sum_{t=0}^{T-1} \theta^{-t}\left(\left(\eta\Upsilon_t^x + \left(\tau+\rho\left(1-\frac{S}{n}\right)\right)\Upsilon_t^y - \theta\mathbf{E}\Gamma_t\right) - \left((\eta+\mu)\Upsilon_{t+1}^x + (\tau+\rho)\Upsilon_{t+1}^y - \mathbf{E}\Gamma_{t+1}\right)\right) \\
&= \eta\Upsilon_0^x + \left(\tau+\rho\left(1-\frac{S}{n}\right)\right)\Upsilon_0^y - \theta\mathbf{E}\Gamma_0 - \theta^{-T+1}\left((\eta+\mu)\Upsilon_T^x + (\tau+\rho)\Upsilon_T^y - \mathbf{E}\Gamma_T\right).
\end{aligned}$$

By setting $x_{-1} = x_0$, we have $\Gamma_0 = 0$. Besides, we have $-\Gamma_T \leq \frac{1}{n}\sum_{i=1}^n \|g_i(x_T) - g_i(x_{T-1})\|_* \left\|y_*^{(i)} - y_t^{(i)}\right\| \leq \frac{C_g}{n}\|x_T - x_{T-1}\|_2 \|y_* - y_T\|$. Thus,

$$\begin{aligned}
\eta\theta^{-T}\Upsilon_T^x \leq\ & \eta\Upsilon_0^x + \left(\tau+\rho\left(1-\frac{S}{n}\right)\right)\Upsilon_0^y - \theta^{-T+1}\left((\eta+\mu)\Upsilon_T^x + (\tau+\rho)\Upsilon_T^y - \frac{\eta}{\theta}\Upsilon_T^x - \frac{C_g}{n}\|x_T - x_{T-1}\|_2\|y_* - y_T\|\right) \\
&- \sum_{t=1}^{T-1}\theta^{-t+1}\underbrace{\left(\left((\eta+\mu)\Upsilon_{t+1}^x + (\tau+\rho)\Upsilon_{t+1}^y - \mathbf{E}\Gamma_{t+1}\right) - \left(\frac{\eta}{\theta}\Upsilon_t^x + \left(\tau+\rho\left(1-\frac{S}{n}\right)\right)/\theta\Upsilon_t^y - \mathbf{E}\Gamma_t\right)\right)}_{\heartsuit} \\
&+ \left(\frac{2(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}\right)\sum_{t=0}^{T-1}\theta^{-t} - \sum_{t=0}^{T-1}\theta^{-t}\underbrace{\left(\frac{\tau}{n} - \frac{(\lambda_2+\lambda_3\theta)}{\mu_\psi n}\right)}_{\heartsuit}\mathbf{E}[U_\psi(\bar{y}_{t+1}, y_t)] \\
&- \sum_{t=0}^{T-1}\theta^{-t}\underbrace{\left(\frac{\eta}{2} - \frac{L_g C_f}{2} - \frac{C_g^2}{2\lambda_2} - \frac{C_g^2}{2\lambda_3}\right)}_{\heartsuit}\mathbf{E}\|x_{t+1} - x_t\|_2^2.
\end{aligned} \tag{C.10}$$

Note that $\eta+\mu-\frac{\eta}{\theta} \geq 0 \Leftrightarrow \theta \geq \frac{\eta}{\eta+\mu}$ such that $(\eta+\mu)\Upsilon_T^x - \frac{\eta}{\theta}\Upsilon_T^x \geq 0$ and $\frac{C_g}{n}\|x_T - x_{T-1}\|_2\|y - y_T\| \leq \frac{C_g^2}{2\lambda_2}\|x_T - x_{T-1}\|_2^2 + \frac{\lambda_2}{2\mu_\psi n^2}U_\psi(y_*, y_T)$. To make the $\heartsuit$ terms in (C.10) be non-negative, we choose $\lambda_2 \asymp \frac{C_g\sqrt{S\rho\mu_\psi}}{\sqrt{n\mu}}$, $\lambda_3 \asymp \frac{C_g\sqrt{S\rho\mu_\psi}}{\sqrt{n\mu}}$ while ensuring that

$$1/\tau \leq O\left(\frac{\sqrt{n\mu\mu_\psi}}{C_g\sqrt{S\rho}}\right), \quad 1/\eta \leq O\left(\frac{\sqrt{S\rho\mu_\psi}}{C_g\sqrt{n\mu}} \wedge \frac{1}{L_g C_f}\right). \tag{C.11}$$

Notice that $\tau + \left(1-\frac{S}{n}\right) \leq \theta(\tau+\rho)$ and $(\tau+\rho)(1-\theta) = \frac{\rho S + \rho n(1-\theta)}{n(1-\theta)}(1-\theta) = \rho\left(\frac{S}{n} + (1-\theta)\right)$.

$$\begin{aligned}
\mu\Upsilon_T^x &\leq \mu\theta^T\Upsilon_0^x + \frac{\left(\tau+\rho\left(1-\frac{S}{n}\right)\right)(1-\theta)}{\theta}\theta^T\Upsilon_0^y + \left(\frac{2(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}\right) \\
&\leq \mu\theta^T\Upsilon_0^x + (\tau+\rho)(1-\theta)\theta^T\Upsilon_0^y + \left(\frac{2(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}\right) \\
&= \mu\theta^T\Upsilon_0^x + \rho\left(\frac{S}{n} + (1-\theta)\right)\theta^T\Upsilon_0^y + \left(\frac{2(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}\right).
\end{aligned}$$

We select $\eta = \frac{\mu\theta}{1-\theta}$, $\tau = \frac{\rho S}{n(1-\theta)}$, and

$$\theta = O\left(1 - \frac{S}{n} \wedge \frac{\mu}{L_g C_f} \wedge \sqrt{\frac{\mu\rho\mu_\psi S}{C_g^2 n}} \wedge \frac{\mu_\psi B\rho S\epsilon}{\sigma_0^2 n} \wedge \frac{B\mu\epsilon}{C_f^2\sigma_1^2} \wedge \frac{S\mu\epsilon}{\delta^2}\right)$$

to make (C.11) hold and

$$\frac{2(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)}+\frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\eta+\mu}\le\frac{2(1+2\theta)(1-\theta)\sigma_0^2 n}{\mu_\psi B\rho S}+\frac{(1-\theta)\left(\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}\right)}{\mu}\le\epsilon.$$

Since $L_f \coloneqq \frac{1}{\mu_\psi\rho}$, the number of iterations needed by Algorithm 1 to make $\mu\Upsilon_T^x\le\epsilon$ is

$$T=\tilde{O}\left(\frac{n}{S}+\frac{L_gC_f}{\mu}+\frac{C_g\sqrt{nL_f}}{\sqrt{S\mu}}+\frac{nL_f\sigma_0^2}{BS\epsilon}+\frac{C_f^2\sigma_1^2}{\mu B\epsilon}+\frac{\delta^2}{\mu S\epsilon}\right),$$

where $\tilde{O}(\cdot)$ hides the polylog$(1/\epsilon)$ factor. In the case that g_i is non-smooth, we utilize (C.7) instead of (C.3). Correspondingly, we need to replace the blue term $\frac{L_gC_f}{2}$ in (C.9) by $0.25(\eta+\mu)$. Additionally, there is a $\frac{4C_f^2C_g^2}{\eta+\mu}$ term on the right-hand side of (C.9). Following similar steps, we can get the iteration complexity to make $\mu\Upsilon_T^x\le\epsilon$ is

$$T=\tilde{O}\left(\frac{n}{S}+\frac{C_g\sqrt{nL_f}}{\sqrt{S\mu}}+\frac{nL_f\sigma_0^2}{BS\epsilon}+\frac{C_f^2\sigma_1^2}{\mu B\epsilon}+\frac{\delta^2}{\mu S\epsilon}+\frac{C_f^2C_g^2}{\mu\epsilon}\right).$$

□

C.3 A Direct Conversion to Non-strongly Convex Results

We can directly convert the results from Theorem 3 for the strongly convex and smooth case to the convex and smooth case using a commonly employed regularization technique. For a non-strongly convex $F(x)$ in (1.1), we can construct the strongly convex $\hat{F}(x)\coloneqq F(x)+\frac{\epsilon}{2}\|x\|_2^2$ and define that $\hat{x}_*=\arg\min_{x\in\mathcal{X}}\hat{F}(x)$. We then apply Algorithm 1 to solve the problem $\min_{x\in\mathcal{X}}\hat{F}(x)$, and the output is denoted as x_{out}. Leveraging the smoothness of f_i, we can convert the iteration complexity from Theorem 3, originally for making $\frac{\mu}{2}\mathbf{E}\|x_{\text{out}}-\hat{x}_*\|_2^2\le\epsilon$, to that for $\hat{F}(x_{\text{out}})-\hat{F}(\hat{x}_*)\le\epsilon$. The objective gap of the original F can be upper-bounded as

$$\begin{aligned}F(x_{\text{out}})-F(x_*)&\le\hat{F}(x_{\text{out}})-F(x_*)=(\hat{F}(x_{\text{out}})-\hat{F}(\hat{x}_*))+\hat{F}(\hat{x}_*)-F(x_*)\\&\le(\hat{F}(x_{\text{out}})-\hat{F}(\hat{x}_*))+\hat{F}(\hat{x}_*)-\hat{F}(x_*)+\frac{\epsilon}{2}\|x_*\|_2^2\le\hat{F}(x_{\text{out}})-\hat{F}(\hat{x}_*)+\frac{\epsilon}{2}\|x_*\|_2^2.\end{aligned}\tag{C.12}$$

Therefore, running our algorithm on the strongly convexified function $\hat{F}$ results in an ϵ-accurate solution for the original convex problem as long as $\|x_*\|$ is bounded.

Theorem 18. *Suppose that f_i is smooth and Assumptions 1, 2, 3, 4, 5, 6 hold. Moveover, ρ in Proposition 1 satisfies that $\rho>0$, i.e., f_i is L_f-smooth, $L_f\coloneqq\frac{1}{\mu_\psi\rho}$.*

- *If g_i is L_g-smooth, Algorithm 1 can find an x_{out} such that $\mathbf{E}[F(x_{\text{out}})-F(x_*)]\le\epsilon$ after $T=\tilde{O}\left(\frac{n}{S}+\frac{L_gC_f}{\epsilon}+\frac{C_g\sqrt{nL_f}}{\sqrt{S\epsilon}}+\frac{nC_g^2L_f^2\sigma_0^2}{BS\epsilon^2}+\frac{C_f^2C_g^2L_f\sigma_1^2}{B\epsilon^3}+\frac{C_g^2L_f\delta^2}{S\epsilon^3}\right)$ iterations.*
- *If g_i is non-smooth, Algorithm 1 can find an x_{out} such that $\mathbf{E}[F(x_{\text{out}})-F(x_*)]\le\epsilon$ after $T=\tilde{O}\left(\frac{n}{S}+\frac{C_g\sqrt{nL_f}}{\sqrt{S\epsilon}}+\frac{nC_g^2L_f^2\sigma_0^2}{BS\epsilon^2}+\frac{C_f^2C_g^2L_f\sigma_1^2}{B\epsilon^3}+\frac{C_g^2L_f\delta^2}{S\epsilon^3}+\frac{L_fC_f^2C_g^4}{\epsilon^3}\right)$ iterations.*

Proof. According to (C.12), the proof can be completed by converting the distance gap result to the objective gap result in the strongly convex case. When g_i is smooth, multiply both sides of (C.9) by

θ^{-t} and do telescoping sum from $t=0$ to $T-1$. Add $\eta\theta^{-T}\Upsilon_T^x$ to both sides.

$$
\begin{aligned}
&\eta\theta^{-T}\Upsilon_T^x + \sum_{t=0}^{T-1}\theta^{-t}\mathbf{E}[L(x_t,y_*) - L(x_*,\bar{y}_t)] \\
&\le \sum_{t=0}^{T-1}\theta^{-t}\left(\left(\eta\Upsilon_t^x + \left(\tau+\rho\left(1-\frac{S}{n}\right)\right)\Upsilon_t^y - \theta\mathbf{E}\Gamma_t\right) - \left((\eta+\mu)\Upsilon_{t+1}^x + (\tau+\rho)\Upsilon_{t+1}^y - \mathbf{E}\Gamma_{t+1}\right)\right) \\
&\qquad + \eta\theta^{-T}\Upsilon_T^x + \left(\frac{2(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\eta+\mu}\right)\sum_{t=0}^{T-1}\theta^{-t} - \sum_{t=0}^{T-1}\theta^{-t}\left(\frac{\tau}{n} - \frac{(\lambda_2+\lambda_3\theta)}{\mu_\psi n}\right)\mathbf{E}[U_\psi(\bar{y}_{t+1},y_t)] \\
&\qquad - \sum_{t=0}^{T-1}\theta^{-t}\left(\frac{\eta}{2} - \frac{L_g C_f}{2} - \frac{C_g^2}{2\lambda_2} - \frac{C_g^2}{2\lambda_3}\right)\mathbf{E}\left\|x_{t+1}-x_t\right\|_2^2.
\end{aligned}
$$

Following similar steps as in the proof of Theorem 3, we can arrive at

$$
\begin{aligned}
&\frac{\mu}{2}\mathbf{E}\left\|x_T - x^*\right\|_2^2 + \frac{\theta^T\sum_{t=0}^{T-1}\theta^{-t}}{\eta}\mathbf{E}[L(\bar{x}_T,y_*) - L(x_*,\bar{\bar{y}}_T)] \\
&\overset{\eta=\frac{\mu\theta}{1-\theta}}{=} \frac{\mu}{2}\mathbf{E}\left\|x_T - x^*\right\|_2^2 + \frac{1-\theta^T}{\mu}\mathbf{E}[L(\bar{x}_T,y_*) - L(x_*,\bar{\bar{y}}_T)] \\
&\le \mu\theta^T\Upsilon_0^x + \frac{2\rho\theta^T S}{n}\Upsilon_0^y + \left(\frac{(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\eta+\mu}\right),
\end{aligned}
\tag{C.13}
$$

where $\bar{x}_T = \sum_{t=0}^{T-1}\frac{\theta^{-t}}{\sum_{t=0}^{T-1}\theta^{-t}}x_t$ and $\bar{\bar{y}}_T = \sum_{t=0}^{T-1}\frac{\theta^{-t}}{\sum_{t=0}^{T-1}\theta^{-t}}\bar{y}_t$. Note that $1-\theta^T \ge \frac{1}{2}$ when $T \ge \frac{\log(2)}{1-\theta}$ due to $\exp(-u) \ge 1-u$ for any $u\in\mathbb{R}$. Recall that f_i is $\frac{1}{\rho\mu_\psi}$-smooth. For $\tilde{y}_T^{(i)} = \arg\max_{v\in\mathcal{Y}_i}\{vg_i(\bar{x}_T) - f_i^*(v)\}$, we have $\tilde{y}_T^{(i)} = f_i'(g_i(\bar{x}_T)) \Leftrightarrow g_i(\bar{x}_T) \in \partial f_i^*(\tilde{y}_T^{(i)})$ and

$$
\begin{aligned}
F(\bar{x}_T) - F(x_*) &\le L(\bar{x}_T,\tilde{y}_T) - L(x_*,\bar{\bar{y}}_T) \\
&= L(\bar{x}_T,\tilde{y}_T) - L(\bar{x}_T,y_*) + L(\bar{x}_T,y_*) - L(x_*,\bar{\bar{y}}_T) \\
&= \frac{1}{n}\sum_{i=1}^n\left(\left\langle\tilde{y}_T^{(i)} - y_*^{(i)},\, g_i(\bar{x}_T)\right\rangle + f_i^*(y_*^{(i)}) - f_i^*(\tilde{y}_T^{(i)})\right) + L(\bar{x}_T,y_*) - L(x_*,\bar{\bar{y}}_T) \\
&= \frac{1}{n}\sum_{i=1}^n U_{f_i^*}(\tilde{y}_T^{(i)},y_*^{(i)}) + L(\bar{x}_T,y_*) - L(x_*,\bar{\bar{y}}_T) \qquad \triangleright\ g_i(\bar{x}_T)\in\partial f_i^*(\tilde{y}_T^{(i)}) \\
&= \frac{1}{n}\sum_{i=1}^n U_{f_i}(g_i(\bar{x}_T),g_i(x_*)) + L(\bar{x}_T,y_*) - L(x_*,\bar{\bar{y}}_T) \\
&\le \frac{1}{2\rho\mu_\psi}\left\|g_i(\bar{x}_T) - g_i(x_*)\right\|_*^2 + L(\bar{x}_T,y_*) - L(x_*,\bar{\bar{y}}_T) \\
&\le \frac{C_g^2}{2\rho\mu_\psi}\left\|\bar{x}_T - x_*\right\|_2^2 + L(\bar{x}_T,y_*) - L(x_*,\bar{\bar{y}}_T).
\end{aligned}
\tag{C.14}
$$

Next, we turn to bound the distance between the average iterate $\bar{x}_T$ and the optimum x_*.

$$\begin{aligned}
&\mathbf{E}\frac{\mu}{2}\left\|\bar{x}_T - x_*\right\|_2^2 \le \sum_{t=0}^{T-1}\frac{\theta^{-t}}{\sum_{t=0}^{T-1}\theta^{-t}}\frac{\mu}{2}\left\|x_t - x^*\right\|_2^2 \\
&\le \left(\mu\Upsilon_0^x + \frac{2\rho S}{n}\Upsilon_0^y\right)\sum_{t=0}^{T-1}\frac{\theta^{-t}}{\sum_{t=0}^{T-1}\theta^{-t}}\theta^t + \left(\frac{(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}\right) \\
&= \left(\mu\Upsilon_0^x + \frac{2\rho S}{n}\Upsilon_0^y\right)\frac{\theta^{-1}-1}{\theta^{-T}-1} + \left(\frac{(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}\right) \\
&\le \theta^{(T-1)}\left(\mu\Upsilon_0^x + \frac{2\rho S}{n}\Upsilon_0^y\right) + \left(\frac{(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}\right). \qquad \text{(C.15)}
\end{aligned}$$

Then, we can upper bound $\mathbf{E}[F(\bar{x}_T) - F(x_*)]$ by plug (C.13) and (C.15) into (C.14).

$$\mathbf{E}[F(\bar{x}_T) - F(x_*)] \le \left(\frac{C_g^2}{\mu\rho\mu_\psi} + 2\mu\theta\right)\left(\theta^{(T-1)}\left(\mu\Upsilon_0^x + \frac{2\rho S}{n}\Upsilon_0^y\right) + \left(\frac{(1+2\theta)\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}\right)\right).$$

□

D Convergence Analysis of ALEXR in the Convex Case

D.1 Proof of Lemma 5

Proof. When $\rho = 0$, we decompose the $\triangle$ term in (6.1) as

$$\begin{aligned}
&\frac{\tau}{n}U_\psi(y, y_t) - \frac{\tau}{n}U_\psi(y, \bar{y}_{t+1}) - \frac{\tau}{n}U_\psi(\bar{y}_{t+1}, y_t) \qquad \text{(D.1)} \\
&= \frac{\tau}{S}U_\psi(y, y_t) - \frac{\tau}{S}U_\psi(y, y_{t+1}) - \frac{\tau}{n}U_\psi(\bar{y}_{t+1}, y_t) + \left(\frac{\tau}{S}U_\psi(y, y_{t+1}) - \frac{\tau}{n}U_\psi(y, \bar{y}_{t+1}) + \frac{(S-n)\tau}{nS}U_\psi(y, y_t)\right).
\end{aligned}$$

We rewrite the last three terms above as follows.

$$\begin{aligned}
&\frac{\tau}{S}U_\psi(y, y_{t+1}) - \frac{\tau}{n}U_\psi(y, \bar{y}_{t+1}) + \frac{(S-n)\tau}{nS}U_\psi(y, y_t) \\
&= \frac{\tau}{S}\sum_{i=1}^n\left(\psi_i(y^{(i)}) - \psi_i(y_{t+1}^{(i)}) - \left\langle\nabla\psi_i(y_{t+1}^{(i)}),\, y^{(i)} - y_{t+1}^{(i)}\right\rangle\right) - \frac{\tau}{n}\sum_{i=1}^n\left(\psi_i(y^{(i)}) - \psi_i(\bar{y}_{t+1}^{(i)}) - \left\langle\nabla\psi_i(\bar{y}_{t+1}^{(i)}),\, y^{(i)} - \bar{y}_{t+1}^{(i)}\right\rangle\right) \\
&\quad + \frac{(S-n)\tau}{nS}\sum_{i=1}^n\left(\psi_i(y^{(i)}) - \psi_i(y_t^{(i)}) - \left\langle\nabla\psi_i(y_t^{(i)}),\, y^{(i)} - y_t^{(i)}\right\rangle\right) \\
&= \frac{\tau}{n}\sum_{i=1}^n\left(\psi_i(\bar{y}_{t+1}^{(i)}) - \frac{n}{S}\psi_i(y_{t+1}^{(i)}) + \frac{n-S}{S}\psi_i(y_t^{(i)})\right) + \underbrace{\frac{\tau}{n}\sum_{i=1}^n\left\langle -\frac{n}{S}\nabla\psi_i(y_{t+1}^{(i)}) + \nabla\psi_i(\bar{y}_{t+1}^{(i)}) + \frac{n-S}{S}\nabla\psi_i(y_t^{(i)}),\, y^{(i)}\right\rangle}_{\natural} \\
&\quad + \frac{\tau}{S}\sum_{i=1}^n\left\langle\nabla\psi_i(y_{t+1}^{(i)}),\, y_{t+1}^{(i)}\right\rangle - \frac{\tau}{n}\sum_{i=1}^n\left\langle\nabla\psi_i(\bar{y}_{t+1}^{(i)}),\, \bar{y}_{t+1}^{(i)}\right\rangle + \frac{(S-n)\tau}{nS}\sum_{i=1}^n\left\langle\nabla\psi_i(y_t^{(i)}),\, y_t^{(i)}\right\rangle.
\end{aligned}$$

Note that both $\bar{y}_{t+1}^{(i)}$ and $y_t^{(i)}$ are independent of $\mathcal{S}_t$ such that

$$\begin{aligned}
&\mathbf{E}[\psi_i(y_{t+1}^{(i)}) \mid \mathcal{G}_t] = \frac{S}{n}\psi_i(\bar{y}_{t+1}^{(i)}) + \frac{n-S}{n}\psi_i(y_t^{(i)}),\\
&\mathbf{E}\left[\left\langle \nabla\psi_i(y_{t+1}^{(i)}), y_{t+1}^{(i)}\right\rangle \mid \mathcal{G}_t\right] = \frac{S}{n}\left\langle \nabla\psi_i(\bar{y}_{t+1}^{(i)}), \bar{y}_{t+1}^{(i)}\right\rangle + \frac{n-S}{n}\left\langle \nabla\psi_i(y_t^{(i)}), y_t^{(i)}\right\rangle,\\
&\mathbf{E}\left[\nabla\psi_i(y_{t+1}^{(i)}) \mid \mathcal{G}_t\right] = \frac{S}{n}\nabla\psi_i(\bar{y}_{t+1}^{(i)}) + \frac{n-S}{n}\nabla\psi_i(y_t^{(i)}).
\end{aligned}$$

Apply Lemma 12 to $\natural$ with $\Delta_t^{(i)} \coloneqq -\frac{n}{S}\nabla\psi_i(y_{t+1}^{(i)}) + \nabla\psi_i(\bar{y}_{t+1}^{(i)}) + \frac{n-S}{S}\nabla\psi_i(y_t^{(i)})$, $\hat{y}_{t+1}^{(i)} = \arg\min_v \left\langle -\Delta_t^{(i)}, v\right\rangle + \alpha U_{\psi_i}(v, \hat{y}_t^{(i)})$ (α to be determined) such that

$$\mathbf{E}\left[\left\langle \Delta_t^{(i)}, y^{(i)}\right\rangle\right] \le \mathbf{E}\left[\alpha U_{\psi_i}(y^{(i)}, \hat{y}_t^{(i)}) - \alpha U_{\psi_i}(y^{(i)}, \hat{y}_{t+1}^{(i)})\right] + \frac{1}{2\mu_\psi\alpha}\mathbf{E}\left[\left\|\Delta_t^{(i)}\right\|_*^2\right].$$

Sum both sides from 1 to n and divide n on both sides

$$\mathbf{E}[\natural] \le \mathbf{E}\left[\frac{\alpha\tau}{n}(U_\psi(y, \hat{y}_t) - U_\psi(y, \hat{y}_{t+1}))\right] + \frac{\tau}{2n\mu_\psi\alpha}\mathbf{E}\left[\sum_{i=1}^n \left\|\Delta_t^{(i)}\right\|_*^2\right].$$

Note that $\mathbf{E}[(\nabla\psi_i(y_{t+1}^{(i)}) - \nabla\psi_i(y_t^{(i)})) \mid \mathcal{G}_t] = \frac{S}{n}(\nabla\psi_i(\bar{y}_{t+1}^{(i)}) - \nabla\psi_i(y_t^{(i)}))$ such that

$$\begin{aligned}
\mathbf{E}\left[\left\|\Delta_t^{(i)}\right\|_*^2\right] &= \mathbf{E}\left\|(\nabla\psi_i(\bar{y}_{t+1}^{(i)}) - \nabla\psi_i(y_t^{(i)})) - \frac{n}{S}(\nabla\psi_i(y_{t+1}^{(i)}) - \nabla\psi_i(y_t^{(i)})\right\|_*^2\\
&\le \frac{n^2}{S^2}\mathbf{E}\left\|\nabla\psi_i(y_{t+1}^{(i)}) - \nabla\psi_i(y_t^{(i)})\right\|_*^2.
\end{aligned}$$

Thus, we have

$$\mathbf{E}[\natural] \le \mathbf{E}\left[\frac{\alpha\tau}{n}(U_\psi(y, \hat{y}_t) - U_\psi(y, \hat{y}_{t+1}))\right] + \underbrace{\frac{\tau n}{2\mu_\psi\alpha S^2}\mathbf{E}\left[\sum_{i=1}^n \left\|\nabla\psi_i(y_{t+1}^{(i)}) - \nabla\psi_i(y_t^{(i)})\right\|_*^2\right]}_{\spadesuit}. \tag{D.2}$$

We need to handle the $\spadesuit$ term.

$$\frac{\tau n}{2\mu_\psi\alpha S^2}\mathbf{E}\left[\sum_{i=1}^n \left\|\nabla\psi_i(y_{t+1}^{(i)}) - \nabla\psi_i(y_t^{(i)})\right\|_*^2\right] \le \frac{\tau n L_\psi^2}{2\mu_\psi\alpha S^2}\mathbf{E}\left[\sum_{i=1}^n \left\|y_{t+1}^{(i)} - y_t^{(i)}\right\|^2\right].$$

Choose $\alpha = \frac{n\lambda_1}{S}$ for some $\lambda_1 > 0$. According to (D.1) and (D.2) and $\mathbf{E}[\|y_{t+1} - y_t\|^2 \mid \mathcal{G}_t] = \frac{S}{n}\|\bar{y}_{t+1} - y_t\|^2 \le \frac{2S}{n\mu_\psi}U_\psi(\bar{y}_{t+1}, y_t)$, we can finish the proof. □

D.2 A Supporting Lemma

Lemma 19. *Suppose that Assumptions 4, 5, 6 hold. For any $\lambda_2, \lambda_3, \lambda_4, \lambda_5 > 0$ and any $y \in \mathcal{Y}$, Algorithm 1 with $\theta = 1$ satisfies that*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\langle g_i(x_{t+1}) - \tilde{g}_t^{(i)}, y^{(i)} - \bar{y}_{t+1}^{(i)}\right\rangle \qquad (D.3)\\
&= \mathbf{E}[\Gamma_{t+1} - \Gamma_t] + \frac{2\lambda_2}{n}\mathbf{E}[U_\psi(y, \hat{\mathring{y}}_t) - U_\psi(y, \hat{\mathring{y}}_{t+1})] + \frac{\lambda_5}{n}\mathbf{E}[U_\psi(y, \breve{y}_t) - U_\psi(y, \breve{y}_{t+1})]\\
&\quad + \frac{(\lambda_3+\lambda_4)\mathbf{E}[U_\psi(\bar{y}_{t+1}, y_t)]}{\mu_\psi n} + \frac{C_g^2\mathbf{E}\|x_{t+1} - x_t\|^2}{2\lambda_3} + \frac{C_g^2\mathbf{E}\|x_t - x_{t-1}\|^2}{2\lambda_4} + \frac{9\sigma_0^2}{\tau\mu_\psi B} + \frac{\sigma_0^2}{\lambda_2\mu_\psi B} + \frac{\sigma_0^2}{2\lambda_5\mu_\psi B}.
\end{aligned}$$

where $\Gamma_t \coloneqq \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_{t-1}), y^{(i)} - y_t^{(i)}\right\rangle$, $\{\hat{\mathring{y}}_t\}_{t\ge0}$, $\{\breve{y}_t\}_{t\ge0}$ are virtual sequences and $\hat{\mathring{y}}_t, \breve{y}_t \in \mathcal{Y}$.

Proof. The $\frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - \tilde{g}_t^{(i)},\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle$ term can be decomposed as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - \tilde{g}_t^{(i)},\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle \qquad \text{(D.4)}\\
&= \underbrace{\frac{1+\theta}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle}_{\text{I}} + \underbrace{\frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle - \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle}_{\text{II}}\\
&\quad + \underbrace{\frac{\theta}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}) - g_i(x_t),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle}_{\text{III}} + \underbrace{\frac{\theta}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1};\mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle}_{\text{IV}}.
\end{aligned}$$

Define $\dot{y}_{t+1}^{(i)} \coloneqq \arg\max_{v\in\mathcal{Y}_i}\{\langle v,\, (1+\theta)g_i(x_t) - \theta g_i(x_{t-1})\rangle - f_i^*(v) - \tau U_{\psi_i}(v, y_t^{(i)})\}$, $\forall i \in [n]$. We decompose the I term in (D.4) as

$$\begin{aligned}
\text{I} &= \frac{1+\theta}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle\\
&= \frac{1+\theta}{n}\sum_{i=1}^n (g_i(x_t) - g_i(x_t;\mathcal{B}_t))(\dot{y}_{t+1}^{(i)} - \bar{y}_{t+1}^{(i)}) + \frac{1+\theta}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t),\, y^{(i)} \right\rangle\\
&\quad - \frac{1+\theta}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t),\, \dot{y}_{t+1}^{(i)} \right\rangle.
\end{aligned}$$

Since $f_i^* + \tau U_{\psi_i}(y^{(i)}, y_t^{(i)})$ is $\tau\mu_\psi$-strongly convex, Lemma 13 implies that

$$\begin{aligned}
&\frac{1}{n}\mathbf{E}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t),\, \dot{y}_{t+1}^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle \le \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\| g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}) \right\|_* \left\| \dot{y}_{t+1} - \bar{y}_{t+1} \right\|\\
&\le \frac{1}{n\tau\mu_\psi}\sum_{i=1}^n \mathbf{E}\left[\left\| g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}) \right\|_* \left((1+\theta)\left\| g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}) \right\|_* + \theta \left\| g_i(x_{t-1}) - g_i(x_{t-1};\mathcal{B}_t^{(i)}) \right\|_* \right) \right]\\
&\le \frac{1}{n\tau\mu_\psi}\sum_{i=1}^n \mathbf{E}\left[(1+1.5\theta)\left\| g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}) \right\|_*^2 + 0.5\theta \left\| g_i(x_{t-1}) - g_i(x_{t-1};\mathcal{B}_t^{(i)}) \right\|_*^2 \right] \le \frac{(1+2\theta)\sigma_0^2}{\tau B\mu_\psi}.
\end{aligned}$$

Apply Lemma 12 to the term $\frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t),\, y^{(i)} \right\rangle$. For any $\lambda_2 > 0$ and some auxiliary sequence $\{\hat{y}_t\}_{t\ge 0}$, $\hat{y}_{t+1}^{(i)} = \arg\min_{v\in\mathcal{Y}_i}\{\langle g_i(x_t;\mathcal{B}_t) - g_i(x_t),\, v\rangle + \lambda_2 U_{\psi_i}(v, \hat{y}_t^{(i)})\}$, we have

$$\frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t),\, y^{(i)} \right\rangle \le \frac{\lambda_2}{n}\mathbf{E}[U_\psi(y, \hat{y}_t) - U_\psi(y, \hat{y}_{t+1})] + \frac{1}{2\lambda_2\mu_\psi n}\mathbf{E}\left\| \ell(x_t) - \ell(x_t;\mathcal{B}_t) \right\|_*^2.$$

Lastly, $\mathbf{E}[\left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}),\, \dot{y}_{t+1}^{(i)} \right\rangle \mid \mathcal{F}_{t-1}^2] = 0$. Choose $\theta = 1$. Then, the I term in (D.4) can be bounded as

$$\mathbf{E}[\text{I}] \le \frac{2\lambda_2}{n}\mathbf{E}[U_\psi(y, \hat{y}_t) - U_\psi(y, \hat{y}_{t+1})] + \frac{\sigma_0^2}{\lambda_2\mu_\psi B} + \frac{6\sigma_0^2}{\tau\mu_\psi B}. \qquad \text{(D.5)}$$

Define $\Gamma_t \coloneqq \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_{t-1}),\, y^{(i)} - y_t^{(i)} \right\rangle$. For any $\lambda_3, \lambda_4 > 0$, II + III can be rewritten as

$$\begin{aligned}
\text{II} + \text{III} &= \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle - \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}) - g_i(x_t),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle \\
&= \Gamma_{t+1} - \Gamma_t + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - g_i(x_t),\, y_{t+1}^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}) - g_i(x_t),\, y_t^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle \\
&\le \Gamma_{t+1} - \Gamma_t + \frac{C_g^2 \left\|x_{t+1} - x_t\right\|_2^2}{2\lambda_3} + \frac{\lambda_3 \left\|y_{t+1} - \bar{y}_{t+1}\right\|^2}{2n} + \frac{C_g^2 \left\|x_t - x_{t-1}\right\|_2^2}{2\lambda_4} + \frac{\lambda_4 \left\|y_t - \bar{y}_{t+1}\right\|^2}{2n}
\end{aligned}$$

Note that $y_{t+1}^{(i)} = \bar{y}_{t+1}^{(i)}$ if $i \in \mathcal{S}_t$ and $y_{t+1}^{(i)} = y_t^{(i)}$ otherwise. Then, $\left\|y_{t+1} - \bar{y}_{t+1}\right\|^2 \le \left\|y_t - \bar{y}_{t+1}\right\|^2$ such that

$$\text{II} + \text{III} \le \Gamma_{t+1} - \Gamma_t + \frac{C_g^2 \left\|x_{t+1} - x_t\right\|_2^2}{2\lambda_3} + \frac{C_g^2 \left\|x_t - x_{t-1}\right\|_2^2}{2\lambda_4} + \frac{(\lambda_3 + \lambda_4) U_\psi(\bar{y}_{t+1}, y_t)}{\mu_\psi n}. \tag{D.6}$$

We decompose the IV term in (D.4) as

$$\begin{aligned}
\text{IV} &= \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle \\
&= \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, \dot{y}_{t+1}^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle + \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, y^{(i)} \right\rangle \\
&\quad - \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t-1}; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, \dot{y}_{t+1}^{(i)} \right\rangle.
\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\langle g_i(x_{t-1}; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, \dot{y}_{t+1}^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle\right] \le \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|g_i(x_{t-1}; \mathcal{B}_t^{(i)}) - g_i(x_{t-1})\right\|_* \left\|\dot{y}_{t+1}^{(i)} - \bar{y}_{t+1}^{(i)}\right\|\right].$$

Since $f_i^*(y^{(i)}) + \tau U_{\psi_i}(y^{(i)}, y_t^{(i)})$ is $\tau\mu_\psi$-strongly convex to $y^{(i)}$, Lemma 13 implies that

$$\left\|\dot{y}_{t+1}^{(i)} - \bar{y}_{t+1}^{(i)}\right\| \le \frac{(1+\theta)\left\|g_i(x_t) - g_i(x_t; \mathcal{B}_t^{(i)})\right\|_* + \theta\left\|g_i(x_{t-1}) - g_i(x_{t-1}; \mathcal{B}_t^{(i)})\right\|_*}{\tau\mu_\psi}.$$

Similar to (D.5), the following holds for any $\lambda_5 > 0$ and some auxiliary sequence $\{\breve{y}_t\}_{t\ge 0}$, where $\breve{y}_{t+1}^{(i)} = \arg\min_{v\in\mathcal{Y}_i}\{\langle g_i(x_{t-1}; \mathcal{B}_t) - g_i(x_{t-1}),\, v\rangle + \lambda_2 U_{\psi_i}(v, \breve{y}_t^{(i)})\}$.

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\langle g_i(x_{t-1}; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, y^{(i)} \right\rangle\right] \le \frac{\lambda_5}{n}\mathbf{E}[U_\psi(y, \breve{y}_t) - U_\psi(y, \breve{y}_{t+1})] + \frac{\sigma_0^2}{2\lambda_5\mu_\psi B}.$$

Consider that $\frac{1}{n}\sum_{i=1}^n \mathbf{E}[\left\langle g_i(x_{t-1}; \mathcal{B}_t^{(i)}) - g_i(x_{t-1}),\, \ddot{y}_{t+1}^{(i)} \right\rangle] = 0$.

$$\mathbf{E}[\text{IV}] \le \frac{\lambda_5}{n}\mathbf{E}[U_\psi(y, \hat{y}_t) - U_\psi(y, \hat{y}_{t+1})] + \frac{\sigma_0^2}{2\lambda_5\mu_\psi B} + \frac{3\sigma_0^2}{\tau\mu_\psi B}. \tag{D.7}$$

Combine (D.5), (D.6), (D.7).

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\langle g_i(x_{t+1}) - \tilde{g}_t^{(i)},\, y^{(i)} - \bar{y}_{t+1}^{(i)} \right\rangle \\
&\le \mathbf{E}[\Gamma_{t+1} - \Gamma_t] + \frac{2\lambda_2}{n}\mathbf{E}[U_\psi(y, \hat{\bar{y}}_t) - U_\psi(y, \hat{\bar{y}}_{t+1})] + \frac{\lambda_5}{n}\mathbf{E}[U_\psi(y, \breve{y}_t) - U_\psi(y, \breve{y}_{t+1})] \\
&\quad + \frac{(\lambda_3 + \lambda_4)\mathbf{E}[U_\psi(\bar{y}_{t+1}, y_t)]}{\mu_\psi n} + \frac{C_g^2 \mathbf{E}\left\|x_{t+1} - x_t\right\|^2}{2\lambda_3} + \frac{C_g^2 \mathbf{E}\left\|x_t - x_{t-1}\right\|^2}{2\lambda_4} + \frac{9\sigma_0^2}{\tau\mu_\psi B} + \frac{\sigma_0^2}{\lambda_2\mu_\psi B} + \frac{\sigma_0^2}{2\lambda_5\mu_\psi B}.
\end{aligned}$$

□

D.3 Proof of Theorem 6

Proof. If g_i is smooth, we combine (6.1), (6.3), (C.3), (D.3). Set $x = x_*$ and $x_0 = x_{-1}$.

$$
\begin{aligned}
&\mathbf{E}[L(x_{t+1}, y_{t+1}) - L(x_*, \bar{y}_{t+1})] \\
&\le \frac{\tau}{S}\mathbf{E}[U_\psi(y, y_t) - U_\psi(y, y_{t+1})] + \frac{\tau\lambda_1}{S}\mathbf{E}[U_\psi(y, \hat{y}_t) - U_\psi(y, \hat{y}_{t+1})] + \frac{\eta}{2}\mathbf{E}\left\|x_* - x_t\right\|_2^2 - \frac{\eta}{2}\mathbf{E}\left\|x_* - x_{t+1}\right\|_2^2 \\
&\quad + \mathbf{E}[\Gamma_{t+1} - \Gamma_t] + \frac{2\lambda_2}{n}\mathbf{E}[U_\psi(y, \hat{\hat{y}}_t) - U_\psi(y, \hat{\hat{y}}_{t+1})] + \frac{\lambda_5}{n}\mathbf{E}[U_\psi(y, \breve{y}_t) - U_\psi(y, \breve{y}_{t+1})] \\
&\quad - \left(\frac{\tau}{n} - \frac{\tau L_\psi^2}{n\lambda_1\mu_\psi^2 S} - \frac{\lambda_3 + \lambda_4}{\mu_\psi n}\right)\mathbf{E}[U_\psi(\bar{y}_{t+1}, y_t)] - \left(\frac{\eta}{2} - \frac{C_g^2}{2\lambda_3} - \frac{L_g C_f}{2}\right)\mathbf{E}\left\|x_{t+1} - x_t\right\|_2^2 + \frac{C_g^2}{2\lambda_4}\mathbf{E}\left\|x_t - x_{t-1}\right\|_2^2 \\
&\quad + \frac{9\sigma_0^2}{\tau\mu_\psi B} + \frac{\sigma_0^2}{\lambda_2\mu_\psi B} + \frac{\sigma_0^2}{2\lambda_5\mu_\psi B} + \frac{C_f^2\sigma_1^2}{\eta B} + \frac{\delta^2}{\eta S}. \qquad \text{(D.8)}
\end{aligned}
$$

Do telescoping sum from $t = 0$ to $T - 1$ for the equation above.

$$
\begin{aligned}
&\sum_{t=0}^{T-1}\mathbf{E}[L(x_{t+1}, y) - L(x_*, \bar{y}_{t+1})] \\
&\le \frac{\eta\mathbf{E}\left\|x_* - x_0\right\|_2^2}{2} + \frac{\tau}{S}\mathbf{E}[U_\psi(y, y_0)] + \frac{\tau\lambda_1}{S}\mathbf{E}[U_\psi(y, \hat{y}_0)] + \frac{2\lambda_2}{n}\mathbf{E}[U_\psi(y, \hat{\hat{y}}_0)] + \frac{\lambda_5}{n}\mathbf{E}[U_\psi(y, \breve{y}_0)] \\
&\quad - \left(\frac{\tau}{n} - \frac{\tau L_\psi^2}{n\lambda_1\mu_\psi^2 S} - \frac{\lambda_3 + \lambda_4}{\mu_\psi n}\right)\sum_{t=0}^{T-1}\mathbf{E}[U_\psi(\bar{y}_{t+1}, y_t)] - \left(\frac{\eta}{2} - \frac{L_g C_f}{2} - \frac{C_g^2}{2\lambda_3} - \frac{C_g^2}{2\lambda_4}\right)\sum_{t=0}^{T-1}\mathbf{E}\left\|x_{t+1} - x_t\right\|^2 \\
&\quad + \mathbf{E}[\Gamma_T] - \frac{\tau}{S}\mathbf{E}[U_\psi(y, y_T)] + \left(\frac{C_f^2\sigma_1^2}{B\eta} + \frac{\delta^2}{S\eta}\right)T + \frac{9\sigma_0^2 T}{\tau\mu_\psi B} + \frac{\sigma_0^2 T}{\lambda_2\mu_\psi B} + \frac{\sigma_0^2 T}{2\lambda_5\mu_\psi B}.
\end{aligned}
$$

Note that $\Gamma_0 = 0$, $\Gamma_T \le \frac{1}{n}\sum_{i=1}^n \left\|g_i(x_T) - g_i(x_{T-1})\right\|_* \left\|y^{(i)} - y_T^{(i)}\right\| \le \frac{C_g^2}{2\lambda_3}\left\|x_T - x_{T-1}\right\|_2^2 + \frac{\lambda_3}{2n\mu_\psi}U_\psi(y, y_T)$. Choose $\lambda_1 \asymp \frac{L_\psi^2}{S\mu_\psi^2}$, $\lambda_2 \asymp \frac{n\tau}{S}$, $\lambda_3 \asymp \frac{C_g\sqrt{S}}{\sqrt{n}}$, $\lambda_4 \asymp \frac{C_g\sqrt{S}}{\sqrt{n}}$, $\lambda_5 \asymp \frac{n\tau}{S}$, and let $1/\tau \le O\left(\frac{\sqrt{n}\mu_\psi}{C_g\sqrt{S}}\right)$ and $1/\eta \le O\left(\frac{\sqrt{S}}{C_g\sqrt{n}}\right)$. Since $L(x, y)$ is convex in x and linear in y, we have

$$
\mathbf{E}\max_y[L(\bar{x}_T, y) - L(x_*, \bar{\bar{y}}_T)] \le \mathbf{E}\max_y \frac{1}{T}\sum_{t=0}^{T-1}[L(x_{t+1}, y) - L(x_*, \bar{y}_{t+1})],
$$

where $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_{t+1}$, $\bar{\bar{y}}_T = \frac{1}{T}\sum_{t=0}^{T-1}\bar{y}_{t+1}$. Now work on the LHS.

$$
L(\bar{x}_T, y) - L(x_*, \bar{\bar{y}}_T) = \frac{1}{n}\sum_{i=1}^n\left(y^{(i)}g_i(\bar{x}_T) - f_i^*(y^{(i)})\right) + r(\bar{x}_T) - \frac{1}{n}\sum_{i=1}^n\left(\bar{\bar{y}}_T^{(i)}g_i(x_*) - f_i^*(\bar{\bar{y}}_T^{(i)})\right) - r(x_*)
$$

Choose $y^{(i)} = \tilde{y}_T^{(i)} \in \arg\max_v\{vg_i(\bar{x}_T) - f_i^*(v)\} \Leftrightarrow g_i(\bar{x}_T) \in \partial f_i^*(\tilde{y}_T^{(i)}) \Leftrightarrow \tilde{y}_T^{(i)} \in \partial f_i(g_i(\bar{x}_T))$ such that $\tilde{y}_T^{(i)}g_i(\bar{x}_T) - f_i^*(\tilde{y}_T^{(i)}) = f_i(g_i(\bar{x}_T))$. By Fenchel-Young, $-\bar{\bar{y}}_T^{(i)}g_i(x_*) + f_i^*(\bar{\bar{y}}_T^{(i)}) \ge -f_i(g_i(x_*))$. Thus, $\mathbf{E}[F(\bar{x}_T) - F(x_*)] \le \mathbf{E}\max_y[L(\bar{x}_T, y) - L(x_*, \bar{\bar{y}}_T)]$. Thus, we can make $\mathbf{E}[F(\bar{x}_T) - F(x_*)] \le \epsilon$ after $T = O\left(\frac{L_g C_f D_{\mathcal{X}}^2}{\epsilon} + \frac{\sqrt{n}C_g D_{\mathcal{X}}^2}{\sqrt{S}\epsilon} + \frac{C_g(1 + L_\psi^2/(S\mu_\psi^2))\sum_{i=1}^n D_{\psi_i,\mathcal{Y}_i}^2}{\mu_\psi\sqrt{nS}\epsilon} + \frac{D_{\mathcal{X}}^2\delta^2}{S\epsilon^2} + \frac{D_{\mathcal{X}}^2 C_f^2\sigma_1^2}{B\epsilon^2} + \frac{\sigma_0^2(1 + L_\psi^2/(S\mu_\psi^2))\sum_{i=1}^n D_{\psi_i,\mathcal{Y}_i}^2}{\mu_\psi BS\epsilon^2}\right)$ iterations by setting $\theta = 1$, $\tau = O\left(\frac{\sqrt{S}C_g}{\mu_\psi\sqrt{n}} \vee \frac{\sigma_0^2}{\mu_\psi B\epsilon}\right)$, $\eta = O\left(L_g C_f \vee \frac{\sqrt{n}C_g}{\sqrt{S}} \vee \frac{\delta^2}{S\epsilon} \vee \frac{C_f^2\sigma_1^2}{B\epsilon}\right)$.

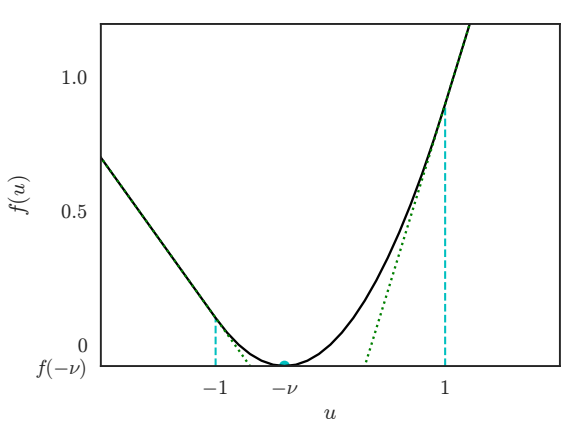
If g_i is non-smooth, we utilize (C.7) instead of (C.3). Correspondingly, the blue term $\frac{L_g C_f}{2}$ in (D.8) should be changed to $\frac{\eta}{4}$. Additionally, there is a $\frac{4C_f^2 C_g^2}{\eta}$ term on the right-hand side of (D.8). Following similar steps, we can get the iteration complexity to make $\mathbf{E}[F(\bar{x}_T) - F(x_*)] \leq \epsilon$ is $T = O\left(\frac{C_f^2 C_g^2 D_{\mathcal{X}}^2}{\epsilon^2} + \frac{\sqrt{n} C_g D_{\mathcal{X}}^2}{\sqrt{S}\epsilon} + \frac{C_g(1+L_\psi^2/(S\mu_\psi^2))\sum_{i=1}^n D_{\psi_i,\mathcal{Y}_i}^2}{\mu_\psi \sqrt{nS}\epsilon} + \frac{D_{\mathcal{X}}^2 \delta^2}{S\epsilon^2} + \frac{D_{\mathcal{X}}^2 C_f^2 \sigma_1^2}{B\epsilon^2} + \frac{\sigma_0^2(1+L_\psi^2/(S\mu_\psi^2))\sum_{i=1}^n D_{\psi_i,\mathcal{Y}_i}^2}{\mu_\psi BS\epsilon^2}\right)$ by setting $\theta = 1$, $\tau = O\left(\frac{\sqrt{S} C_g}{\mu_\psi \sqrt{n}} \vee \frac{\sigma_0^2}{\mu_\psi B\epsilon}\right)$, $\eta = O\left(\frac{\sqrt{n} C_g}{\sqrt{S}} \vee \frac{C_f^2 C_g^2}{\epsilon} \vee \frac{\delta^2}{S\epsilon} \vee \frac{C_f^2 \sigma_1^2}{B\epsilon}\right)$.

□

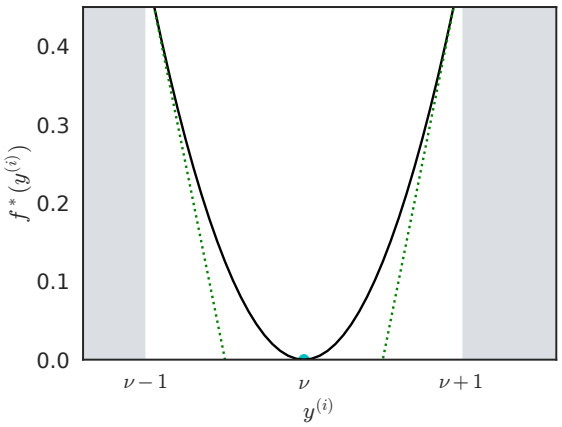
E Proof of the Lower Complexity Bounds in Theorem 10

Proof. We construct the hard problems for (i) smooth f_i; and (ii) non-smooth f_i separately.

(i) Smooth f_i and strongly convex r: First, we can consider the special instance that f_i is the identity mapping and $\delta = 0$ (e.g., $n = 1$), $\sigma_0 = 0$. Then, the cFCCO problem in (1.1) becomes the standard strongly convex minimization problem. Then, we can apply the information-theoretic lower bounds [71, 25] for the standard strongly convex minimization problem. Thus, any algorithm in the abstract scheme requires at least $\Omega\left(\frac{1}{\mu\epsilon}\right)$ iterations to find an $\bar{x}$ such that $\mathbf{E}\left[\frac{\mu}{2}\|\bar{x} - x_*\|_2^2\right] \leq \epsilon$.



(i) Visualization of f in (E.1)



(ii) Convex conjugate f^* in (E.2) of f. Note that $f^*(y^{(i)}) = +\infty$ in grey areas.

Next, we construct another "hard" instance to derive the second half of the lower bound in this case. Consider the following strongly convex FCCO problem

$$\min_{x\in\mathcal{X}} F(x) = \frac{1}{n}\sum_{i=1}^n f(g_i(x)) + r(x),$$
$$f(u) = \begin{cases} (\nu-1)u + \frac{1}{2}(\nu-1)^2 + \nu - 1 - \frac{\nu^2}{2}, & u \in (-\infty, -1) \\ \frac{1}{2}(u+\nu)^2 - \frac{\nu^2}{2}, & u \in [-1, 1] \\ (1+\nu)u + \frac{1}{2}(1+\nu)^2 - 1 - \nu - \frac{\nu^2}{2}, & u \in (1, \infty) \end{cases}, \quad r(x) = \frac{1}{4n}\|x\|_2^2 \tag{E.1}$$

where $\mathcal{X} = [-1, 1]^n$, the outer function $f : \mathbb{R} \to \mathbb{R}$ is smooth and Lipschitz continuous for $\nu < 1$. Besides, the inner function $g_i : \mathbb{R}^n \to \mathbb{R}$ is $g_i(x) = \mathbf{E}_\zeta[g_i(x;\zeta)]$ and $g_i(x;\zeta) = x^{(i)} + \zeta$, where ζ follows

$$\zeta = \begin{cases} -\nu & \text{w.p. } 1-p, \\ \nu(1-p)/p & \text{w.p. } p. \end{cases}, \quad \text{where } p \coloneqq \frac{\nu^2}{\sigma^2}.$$

As stated in Assumption 3, we do not require f to be monotonically non-decreasing when g_i is affine. We define that $F_i(x^{(i)}) \coloneqq f(g_i(x)) + \frac{1}{4}[x^{(i)}]^2$ such that $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x^{(i)})$. Thus, the problem $\min_x F(x)$ is equivalent to the problems $\min_{x^{(i)}} F_i(x^{(i)})$ on all coordinates $i \in [n]$. Since the problem is separable over the coordinates, we have $x_*^{(i)} = \arg\min_{x\in[-1,1]} F_i(x^{(i)})$ for $x_* = \arg\min_{x\in\mathcal{X}} F(x)$. Thus, we have $x_*^{(i)} = -\frac{2\nu}{3}$ and $F_i(x_*^{(i)}) = -\frac{\nu^2}{3}$. By the convex conjugate, for any $y^{(i)} \in \mathbb{R}$ we have

$$\begin{aligned} f^*(y^{(i)}) &= \max\left\{\sup_{u<-1}\left\{uy^{(i)} - \left((\nu-1)u + \frac{1}{2}(\nu-1)^2 + \nu - 1 - \frac{\nu^2}{2}\right)\right\}, \sup_{-1\le u\le 1}\left\{uy^{(i)} - \frac{1}{2}(u+\nu)^2 + \frac{\nu^2}{2}\right\},\right. \\ &\qquad \left.\sup_{u>1}\left\{uy^{(i)} - \left((1+\nu)u + \frac{1}{2}(1+\nu)^2 - 1 - \nu - \frac{\nu^2}{2}\right)\right\}\right\} \\ &= \begin{cases} +\infty, & y^{(i)} \in (-\infty, \nu-1)\cup(\nu+1,\infty) \\ \frac{1}{2}(y^{(i)}-\nu)^2, & y^{(i)} \in [\nu-1, \nu+1]. \end{cases} \end{aligned} \tag{E.2}$$

Note that the proximal mapping with $\psi_i(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ can be efficiently solved for this f_i^*. Since $\mathbb{P}_i = \mathbb{P}$ in the "hard" problem (E.1) and we only consider the inner mini-batch size $B = 1$, the abstract scheme (Algorithm 2) only needs to sample shared $\zeta_t, \tilde{\zeta}_t \sim \mathbb{P}$ for all coordinates $i \in \mathcal{S}_t$ in the t-th iteration. For an $i \in [n]$, suppose that $\mathfrak{g}_\tau^{(i)} = \varnothing$ or $\{-\nu\}$, $\mathfrak{Y}_\tau^{(i)} = \{0\}$, $\mathfrak{X}_\tau^{(i)} = \{0\}$ for all $\tau \le t$. Then,

- If $i \notin \mathcal{S}_t$, the abstract scheme (Algorithm 2) leads to

$$\mathfrak{g}_{t+1}^{(i)} = \varnothing \text{ or } \{-\nu\}, \quad \mathfrak{Y}_{t+1}^{(i)} = \{0\}, \quad \mathfrak{X}_{t+1}^{(i)} = \{0\}.$$

- If $i \in \mathcal{S}_t$ and $\zeta_t = -\nu$, the abstract scheme (Algorithm 2) proceeds as

$$\begin{aligned} \mathfrak{g}_{t+1}^{(i)} &= \mathfrak{g}_t^{(i)} + \operatorname{span}\left\{\hat{x}^{(i)} + \zeta_t \mid \hat{x}^{(i)} \in \mathfrak{X}_t^{(i)}\right\}, \\ \mathfrak{Y}_{t+1}^{(i)} &= \mathfrak{Y}_t^{(i)} + \operatorname{span}\left\{\operatorname*{arg\,max}_{y^{(i)}\in[\nu-1,\nu+1]}\left\{y^{(i)}(\hat{g}^{(i)}+\nu) - \frac{1}{2}\left(y^{(i)}\right)^2 - \tau\left(y^{(i)} - \hat{y}^{(i)}\right)\right)^2\right\} \mid \hat{g}^{(i)} \in \mathfrak{g}_{t+1}^{(i)}, \hat{y}^{(i)} \in \mathfrak{Y}_t^{(i)}\right\}, \\ \mathfrak{X}_{t+1}^{(i)} &= \mathfrak{X}_t^{(i)} + \operatorname{span}\left\{\operatorname*{arg\,min}_{x^{(i)}\in[-1,1]}\left\{\frac{1}{S}\hat{y}^{(i)}x^{(i)} + \frac{1}{n}[x^{(i)}]^2 + \frac{\eta}{2}\left(x^{(i)} - \hat{x}^{(i)}\right)^2\right\} \mid \hat{y}^{(i)} \in \mathfrak{Y}_{t+1}^{(i)}, \hat{x}^{(i)} \in \mathfrak{X}_t^{(i)}\right\}. \end{aligned}$$

Then, we can derive that $\mathfrak{g}_{t+1}^{(i)} = \varnothing$ or $\{-\nu\}$, $\mathfrak{Y}_{t+1}^{(i)} = \{0\}$, and $\mathfrak{X}_{t+1}^{(i)} = \{0\}$.

To sum up, given the event $\bigcap_{\tau=1}^t\{\mathfrak{g}_\tau^{(i)} = \varnothing \text{ or } \{-\nu\}, \mathfrak{Y}_\tau^{(i)} = \{0\}, \mathfrak{X}_\tau^{(i)} = \{0\}\}$, we can make sure that $\{\mathfrak{g}_{t+1}^{(i)} = \varnothing \text{ or } \{-\nu\} \wedge \mathfrak{Y}_{t+1}^{(i)} = \{0\} \wedge \mathfrak{X}_{t+1}^{(i)} = \{0\}\}$ for the abstract scheme in Algorithm 2 when one of the following mutually exclusive events happens:

- Event I: $i \notin \mathcal{S}_t$;
- Event II: $i \in \mathcal{S}_t$ and $\zeta_t = -\nu$.

Note that the random variable ζ_t is independent of $\mathcal{S}_t$. Thus, the probability of the event $E_{t+1}^{(i)} \coloneqq \{\mathfrak{g}_{t+1}^{(i)} = \varnothing \text{ or } \{-\nu\} \wedge \mathfrak{Y}_{t+1}^{(i)} = \{0\} \wedge \mathfrak{X}_{t+1}^{(i)} = \{0\}\}$ conditioned on $\bigcap_{\tau=1}^t E_\tau^{(i)}$ can be bounded as

$$\begin{aligned} \mathbf{P}\left[E_{t+1}^{(i)} \mid \bigcap_{\tau=1}^t E_\tau^{(i)}\right] &= \mathbf{P}\left[\left\{\mathfrak{g}_{t+1}^{(i)} = \varnothing \text{ or } \{-\nu\} \wedge \mathfrak{Y}_{t+1}^{(i)} = \{0\} \wedge \mathfrak{X}_{t+1}^{(i)} = \{0\}\right\} \mid \bigcap_{\tau=1}^t E_\tau^{(i)}\right] \\ &\ge \mathbf{P}\left[\{i \notin \mathcal{S}_t\}\right] + \mathbf{P}\left[\{\{i \in \mathcal{S}_t\} \wedge \{\zeta_t = -\nu\}\}\right] \\ &= \mathbf{P}\left[\{i \notin \mathcal{S}_t\}\right] + \mathbf{P}\left[\{i \in \mathcal{S}_t\}\right]\mathbf{P}\left[\{\zeta_t = -\nu\}\right] = \left(1 - \frac{S}{n}\right) + \frac{S}{n}(1-p) = 1 - \frac{Sp}{n}. \end{aligned}$$

Since $\mathcal{S}_t$ and ζ_t in different iterations t are mutually independent, we have

$$\mathbf{P}\left[E_T^{(i)}\right] \geq \mathbf{P}\left[\bigcap_{t=0}^{T-1} E_{t+1}^{(i)}\right] = \prod_{t=0}^{T-1} \mathbf{P}\left[E_{t+1}^{(i)} \mid \bigcap_{t=1}^{t} E_t^{(i)}\right] = \left(1 - \frac{Sp}{n}\right)^T > 3/4 - \frac{TSp}{n}.$$

Thus, letting $T < \frac{n}{4Sp}$ can make $\mathbf{P}\left[E_T^{(i)}\right] > \frac{1}{2}$. Choose $\nu = 3\sqrt{2\epsilon}$, and $\sigma = \sigma_0$ such that $p = \frac{\nu^2}{\sigma^2} = \frac{18\epsilon}{\sigma_0^2}$. For any $i \in [n]$ and any output $\bar{x}^{(i)} \in \mathfrak{X}_T^{(i)}$, we have

$$\begin{aligned}
\mathbf{E}\left[\left(\bar{x}^{(i)} - x_*^{(i)}\right)^2\right] &= \mathbf{E}\left[\mathbb{I}_{E_T^{(i)}}\left(\bar{x}^{(i)} - x_*^{(i)}\right)^2 + \mathbb{I}_{\overline{E_T^{(i)}}}\left(\bar{x}^{(i)} - x_*^{(i)}\right)^2\right] \\
&\geq \mathbf{E}\left[\mathbb{I}_{E_T^{(i)}}\left(\bar{x}^{(i)} - x_*^{(i)}\right)^2\right] \\
&= \mathbf{E}\left[\mathbb{I}_{E_T^{(i)}}\left(x_*^{(i)}\right)^2\right] = \mathbf{P}\left[E_T^{(i)}\right]\left(x_*^{(i)}\right)^2 > \frac{2\nu^2}{9} = 4\epsilon.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\mathbf{E}[F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})] &= \mathbf{E}\left[\mathbb{I}_{E_T^{(i)}}\left(F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})\right) + \mathbb{I}_{\overline{E_T^{(i)}}}\left(F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})\right)\right] \\
&\geq \mathbf{E}\left[\mathbb{I}_{E_T^{(i)}}\left(F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})\right)\right] \\
&= \mathbf{E}\left[\mathbb{I}_{E_T^{(i)}}\left(F_i(0) - F_i(x_*^{(i)})\right)\right] = \mathbf{P}[E_T^{(i)}]\left(F_i(0) - F_i(x_*^{(i)})\right) > \frac{\nu^2}{6} > \epsilon.
\end{aligned}$$

Since the derivations above hold for arbitrary $i \in [S]$ and the $r(x)$ in (E.1) is $\frac{1}{2n}$-strongly convex ($\mu = \frac{1}{2n}$), we can derive that

$$\begin{aligned}
\mathbf{E}\left[\frac{\mu}{2}\left\|\bar{x} - x_*\right\|_2^2\right] &= \mathbf{E}\left[\frac{1}{4n}\left\|\bar{x} - x_*\right\|_2^2\right] = \frac{1}{4n}\sum_{i=1}^n \mathbf{E}\left[\left(\bar{x}^{(i)} - x_*^{(i)}\right)^2\right] > \epsilon, \\
\mathbf{E}\left[F(\bar{x}) - F(x_*)\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})\right] > \epsilon.
\end{aligned}$$

Thus, to find an output $\bar{x}$ that satisfies $\mathbf{E}\left[\frac{\mu}{2}\left\|\bar{x} - x_*\right\|_2^2\right] \leq \epsilon$ or $\mathbf{E}\left[F(\bar{x}) - F(x_*)\right] \leq \epsilon$, the abstract scheme requires at least $T \geq \frac{n}{4Sp} = \frac{n\sigma_0^2}{72S\epsilon}$ iterations.

(ii) Non-smooth f_i: We borrow the construction $f(\cdot) = \beta\max\{\cdot, -\nu\}$ from Zhang and Lan [22]. We define that $F_i(x^{(i)}) \coloneqq f(g_i(x)) + \frac{\alpha}{2}[x^{(i)}]^2 = \beta\max\{x^{(i)}, -\nu\} + \frac{\alpha}{2}[x^{(i)}]^2$ such that $F(x) = \frac{1}{n}\sum_{i=1}^n F_i(x^{(i)})$. Thus, the problem $\min_x F(x)$ is equivalent to the problems $\min_{x^{(i)}} F_i(x^{(i)})$ on all coordinates $i \in [n]$. Let the domain $\mathcal{X}$ be $[-2\nu, 2\nu]^n$. Since the problem is separable over the coordinates, we have $x_*^{(i)} = \arg\min_{x\in[-2\nu,2\nu]} F_i(x^{(i)}) = \arg\min_{x\in[-2\nu,2\nu]}\left\{\beta\max\{x^{(i)}, -\nu\} + \frac{\alpha}{2}[x^{(i)}]^2\right\}$ for $x_* = \arg\min_{x\in\mathcal{X}} F(x)$. Considering $F_i(x^{(i)}) = \begin{cases}\beta x^{(i)} + \frac{\alpha}{2}[x^{(i)}]^2 & x^{(i)} \geq -\nu \\ -\beta\nu + \frac{\alpha}{2}[x^{(i)}]^2 & x^{(i)} < -\nu\end{cases}$, we have

$$x_*^{(i)} = \begin{cases}-\beta/\alpha & \text{if } \alpha > \beta/\nu \\ -\nu & \text{if } \alpha \in \frac{\beta}{\nu}[0,1]\end{cases}, \quad F_i(x_*^{(i)}) \leq \begin{cases}-\beta^2/(2\alpha) & \text{if } \alpha > \beta/\nu \\ -\beta\nu/2 & \text{if } \alpha \in \frac{\beta}{\nu}[0,1].\end{cases}$$

Since $F_i(0) = 0$, we can derive that $F_i(0) - F_i(x_*^{(i)}) \geq \frac{1}{2}\min\{\beta\nu, \beta^2/\alpha\}$. By the convex conjugate, we have

$$f(\hat{g}^{(i)}) = \max_{y^{(i)}\in[0,\beta]}\left\{y^{(i)}\hat{g}^{(i)} - \nu(\beta - y^{(i)})\right\}.$$

Consider an arbitrary $i \in [n]$. Suppose that $\mathfrak{g}_\tau^{(i)} = \varnothing$ or $\{-\nu\}$, $\mathfrak{X}_\tau^{(i)} = \{0\}$, $\mathfrak{Y}_\tau^{(i)} = \{0\}$ for all $\tau \le t$. Note that f is structured non-smooth such that we can select $\psi_i(\cdot) = \frac{1}{2}\|\cdot\|_2^2$.

- If $i \notin \mathcal{S}_t$, the abstract scheme (Algorithm 2) leads to

$$\mathfrak{g}_{t+1}^{(i)} = \varnothing \text{ or } \{-\nu\}, \quad \mathfrak{Y}_{t+1}^{(i)} = \{0\}, \quad \mathfrak{X}_{t+1}^{(i)} = \{0\}.$$

- If $i \in \mathcal{S}_t$, the abstract scheme (Algorithm 2) proceeds as

$$\mathfrak{g}_{t+1}^{(i)} = \mathfrak{g}_t^{(i)} + \operatorname{span}\left\{\hat{x}^{(i)} + \zeta_t \mid \hat{x}^{(i)} \in \mathfrak{X}_t^{(i)}\right\},$$

$$\mathfrak{Y}_{t+1}^{(i)} = \mathfrak{Y}_t^{(i)} + \operatorname{span}\left\{\operatorname*{arg\,max}_{y^{(i)}\in[0,\beta]}\left\{y^{(i)}\hat{g}^{(i)} - \nu(\beta - y^{(i)}) - \tau\left(y^{(i)} - \hat{y}^{(i)}\right)^2\right\} \mid \hat{g}^{(i)} \in \mathfrak{g}_{t+1}^{(i)}, \hat{y}^{(i)} \in \mathfrak{Y}_t^{(i)}\right\},$$

$$\mathfrak{X}_{t+1}^{(i)} = \mathfrak{X}_t^{(i)} + \operatorname{span}\left\{\operatorname*{arg\,min}_{x^{(i)}\in[-2\nu,2\nu]}\left\{\frac{1}{S}\hat{y}^{(i)}x^{(i)} + \frac{1}{n}[x^{(i)}]^2 + \frac{\eta}{2}\left(x^{(i)} - \hat{x}^{(i)}\right)^2\right\} \mid \hat{y}^{(i)} \in \mathfrak{Y}_{t+1}^{(i)}, \hat{x}^{(i)} \in \mathfrak{X}_t^{(i)}\right\}.$$

Due to the same reason as in the smooth f_i case, the probability of the event $E_T^{(i)} \coloneqq \{\mathfrak{g}_T^{(i)} = \varnothing \text{ or } \{-\nu\} \wedge \mathfrak{Y}_T^{(i)} = \{0\} \wedge \mathfrak{X}_T^{(i)} = \{0\}\}$ can be bounded as

$$\mathbf{P}\left[E_T^{(i)}\right] \ge \mathbf{P}\left[\bigcap_{t=0}^{T-1} E_{t+1}^{(i)}\right] = \prod_{t=0}^{T-1}\mathbf{P}\left[E_{t+1}^{(i)} \mid \bigcap_{t=1}^{t} E_t^{(i)}\right] = \left(1 - \frac{Sp}{n}\right)^T > 3/4 - \frac{TSp}{n}.$$

Thus, letting $T < \frac{n}{4Sp}$ can make $\mathbf{P}\left[E_T^{(i)}\right] > \frac{1}{2}$. Choose $\beta = C_f$, $\nu = \frac{4\epsilon}{C_f}$, and $\sigma = \sigma_0$ such that $p \coloneqq \frac{\nu^2}{\sigma^2} = \frac{16\epsilon^2}{C_f^2\sigma_0^2}$. For any $i \in [n]$ and any output $\bar{x}^{(i)} \in \mathfrak{X}_T^{(i)}$, we have

$$\begin{aligned}
\mathbf{E}[F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})] &= \mathbf{E}\left[\mathbb{I}_{E_T}\left(F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})\right) + \mathbb{I}_{\overline{E_T}}\left(F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})\right)\right] \\
&\ge \mathbf{E}\left[\mathbb{I}_{E_T}\left(F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})\right)\right] \\
&= \mathbf{E}\left[\mathbb{I}_{E_T}\left(F_i(0) - F_i(x_*^{(i)})\right)\right] \\
&= \mathbf{P}[E_T]\left(F_i(0) - F_i(x_*^{(i)})\right) > \min\{\beta\nu, \beta^2/\alpha\}/4 = \epsilon.
\end{aligned}$$

Since the derivations above hold for arbitrary $i \in [S]$, we can derive that

$$\mathbf{E}[F(\bar{x}) - F(x_*)] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[F_i(\bar{x}^{(i)}) - F_i(x_*^{(i)})] > \epsilon.$$

Thus, to find an output $\bar{x}$ that satisfies $\mathbf{E}[F(\bar{x}) - F(x_*)] \le \epsilon$, the abstract scheme requires at least $T \ge \frac{n}{4Sp} = \frac{nC_f^2\sigma_0^2}{64S\epsilon^2}$ iterations. □

F More Details of Experiments

All algorithms are implemented using the PyTorch framework. For the projection onto capped simplex in OOA, we borrow a Python implementation[9] of the efficient algorithm by Lim and Wright [60]. Experiments are conducted on a workstation with the 12th Gen Intel(R) Core(TM) i7-12700K CPU with 20 logical cores.

[9] https://github.com/mblondel/projection-losses/blob/master/polytopes.py

F.1 Group DRO

Here we provide the omitted details and results of our Group DRO experiments.

F.1.1 Data Preprocessing

Adult dataset: We construct 83 groups for the Adult dataset according to income (“>50K”, “≤50K”), race (“white”, “black”, “other”), sex (“female”, “male”), age (“<=30”, “30-45”, “>45”), relationship (“single”, “not_single”), and education (“higher”, “others”), where we discard those groups with less than 50 data points. Following [67], we transform both continuous and categorical features into binary features, resulting in a 122-dimensional feature vector for each data point.

CelebA dataset: We construct 160 groups for this dataset according to 4 binary attributes (“blond hair”, “male”, “mouth slightly open”, “smiling”) and 10 types of additive Gaussian noises (means -0.08:0.02:0.1 and variance 0.08) to the images. Each image of the CelebA dataset is resized to 224×224×3, normalized, and center-cropped. Then, we extract 512-dimensional feature vectors for those preprocessed images from the last convolutional layer of a ResNet18 pre-trained on ImageNet.

F.1.2 Parameter Tuning

We tune the step sizes of all algorithms in the range $\{2, 5, 10\} \times 10^{\{-3,-2,-1\}}$. Additionally, for primal-dual algorithms such as ALEXR and OOA, we adjust the step size for the dual variable within the same range. For SOX and SONX, we also tune the momentum parameter (τ in the SONX paper [11] and γ in the SOX paper [5]) in the range $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ following their papers. For ALEXR, we choose the extrapolation parameter $\theta \in \{0.1, 1.0\}$ and the generating function $\psi_i(\cdot) = \frac{1}{2}(\cdot)^2$. For all algorithms, we choose the weight decay parameter 0.05 on the Adult dataset and 0.1 on the CelebA dataset to improve the testing performance. We execute all algorithms for 5 runs with different random seeds and each run contains 2500 iterations for the Adult dataset and 15000 iterations for the CelebA dataset. For a fair comparison, each algorithm samples 64 data points in each iteration. For SGD, these data points are sampled from the entire training dataset, whereas for other algorithms, they are sampled from 8 sampled groups.

F.1.3 Additional Results

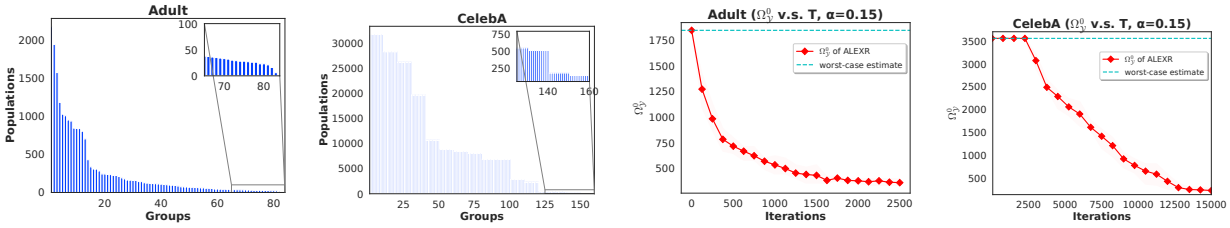


Figure 3: Group sizes and the estimated values of $\Omega_{\mathcal{Y}}^0$.

The first two columns of Figure 3 show the existence of rare groups in the datasets. The last two columns of Figure 3 demonstrate that the actual value of $\Omega_{\mathcal{Y}}^0$ is indeed much smaller than its worst-case estimate $\frac{n}{2\alpha^2}$, which verifies the claims in Section 6.2 and Section 7.1.

F.2 Partial AUC Maximization with Restricted TPR

F.2.1 Dataset Statistics and Preprocessing

To create imbalanced datasets, we randomly remove 99.5% positive data from the Covtype dataset and 99.9% positive data from the Higgs dataset. For the Covtype dataset, we randomly allocate 60% of the data for training, 20% for validation, and another 20% for testing. For the Higgs dataset, we randomly select 500,000 data points for validation, 500,000 data points for testing, and the rest as training data. The Cardiomegaly and Lung-mass datasets are naturally balanced and the train/val/test split is pre-defined. We vectorize each 28×28 image in Cardiomegaly/Lung-mass datasets into a 784-dim feature. We list the detailed statistics of those datasets in Table 6.

Table 6: Statistics of datasets used in the partial AUC maximization experiments. Here n_+ and n_- refer to the numbers of positive and negative data in the train/val/test splt.

Datasets	Train		Val		Test	
	n_+	n_-	n_+	n_-	n_+	n_-
Covtype	889	178,587	252	59,573	275	59,551
Higgs	4,676	4,172,030	582	499,418	571	499,429
Cardiomegaly	1,950	76,518	240	10,979	582	21,851
Lung-mass	3,988	74,480	625	10,594	1,133	21,300

F.2.2 Parameter Tuning

For the step sizes and momentum/extrapolation parameters, we tune them in the same way as in Appendix F.1.2. We execute all algorithms for 5 runs with different random seeds and each run contains 750 iterations for the Cardiomegaly/Lung-mass datasets and 1500 iterations for the Covtype/Higgs datasets. In each iteration, each algorithm randomly samples 16 positive data points and 16 negative data points.

G Convergence Rates of Baseline Algorithms

In Table 1 and 2, some of the baseline algorithms were originally proposed for stochastic compositional optimization (SCO) and convex-concave min-max optimization. In this section, we show how to derive their convergence rates on our FCCO problems.

G.1 SCO Algorithms

The FCCO problem in (1.1) can be reformulated as an SCO problem $F(x) = \hat{f}(g(x))$, $\hat{f} = \frac{1}{n}\sum_{i=1}^n \hat{f}_i$, $\hat{f}_i(u) = f_i(u^{(i)})$ for $u \in \mathbb{R}^n$, $g = [g_1, \ldots, g_n]^\top$.

SCGD [1]/ASC-PG [2] These two algorithms maintain a sequence $\{u_t\}_{t=1}^T$, $u_t \in \mathbb{R}^{nm}$ to estimate the inner function $g(x)$, which requires n zeroth-order oracles in each iteration. To update the variable x, the stochastic gradient is computed as $\nabla g(x_t;\zeta_t)\nabla \hat{f}_{i_t}(u_{t+1}) = \nabla g_{i_t}(x_t;\zeta_t^{(i_t)})\nabla f_{i_t}(u_{t+1}^{(i_t)})$, which requires one first-order oracles in each iteration. All their proofs still go through and the convergence rates of SCGD/ASC-PG on FCCO are the same as those of the SCO problem.

SSD [22] Both π_1 and π_2 in their paper are now n-dimensional. Then, steps 3 and 4 in their Algorithm 1 are done for each coordinate $i \in [n]$, which leads to $O(n)$ zeroth-order oracles in each iteration. To update the variable x, the stochastic gradient is computed as $\nabla g(\underline{y}_2^t; \zeta_t) \nabla \hat{f}_{i_t}(\underline{y}_1^t) = \nabla g_{i_t}(\underline{y}_2^t; \zeta_t^{(i_t)}) \nabla f_{i_t}(\underline{y}_1^t)$, which requires one first-order oracles in each iteration. We only need to handle the $\tilde{\sigma}_x^2$ term in (2.35) in their paper, which now becomes $\delta^2 + C_f \sigma_1^2$ under our assumptions.

G.2 Min-Max Algorithms

The primal-dual formulation in (1.2) of cFCCO can be viewed as a convex-concave min-max optimization $\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^n} \Phi(x, y) - \sum_{i=1}^n f_i^*(y^{(i)}) + r(x)$, where $\Phi(x, y) = \frac{1}{n} \sum_{i=1}^n y^{(i)} g_i(x)$.

SAPD [15] In this paper, they assume that the coupling term $\Phi(x, y)$ is $(L_{xx}, L_{xy}, L_{yx}, L_{yy})$-smooth.

$$\begin{aligned}
\left\| \nabla_x \Phi(x, y) - \nabla_x \Phi(x', y') \right\| &\le L_{xx} \left\| x - x' \right\| + L_{xy} \left\| y - y' \right\|, \\
\left\| \nabla_y \Phi(x, y) - \nabla_y \Phi(x', y') \right\| &\le L_{yx} \left\| x - x' \right\| + L_{yy} \left\| y - y' \right\|.
\end{aligned}$$

Considering the FCCO problem, we have $L_{xx} = C_f L_g$, $L_{xy} = L_{yx} = \frac{C_g}{n}$, $L_{yy} = 0$. Besides, the strong convexity moduli μ_x, μ_y in their paper are $\mu, \frac{1}{nL_f}$ in our paper. When applied to the FCCO problem, SAPD computes stochastic estimators $\tilde{\nabla}_y \Phi(x_t, y_t) = \frac{1}{n}[g_1(x_t; \tilde{\zeta}_t^{(1)}), \ldots, g_n(x_t; \tilde{\zeta}_t^{(n)})]^\top$ and $\tilde{\nabla}_y \Phi(x_t, y_t) = \frac{1}{n}[g_1(x_{t-1}; \tilde{\zeta}_{t-1}^{(1)}), \ldots, g_n(x_{t-1}; \tilde{\zeta}_{t-1}^{(n)})]^\top$ to update y while computing $\tilde{\nabla}_x \Phi(x_t, y_{t+1}) = y_{t+1}^{(i_t)} \nabla g_{i_t}(x_t; \tilde{\zeta}_t)$ to update x. Thus, δ_x^2 and δ_y^2 in their paper are $C_f^2 \sigma_1^2 + \delta^2$ and $\frac{\sigma_0^2}{n}$ under our assumptions.

H Convergence Analysis of ALEXR with $\theta = 0$

As an ablation study, we provide the convergence analysis of our ALEXR algorithm with $\theta = 0$.

H.1 Strongly Convex Case

Theorem 20. *Suppose that Assumptions 1, 2, 3, 4, 5, 6 hold. Moveover, r is μ-strongly convex with $\mu > 0$ while ρ in Proposition 1 satisfies that $\rho > 0$, i.e., f_i is L_f-smooth, $L_f := \frac{1}{\mu_\psi \rho}$. If g_i is L_g-smooth, ALEXR with $\theta = 0$, $\eta = \frac{\mu \upsilon}{1-\upsilon}$, $\tau = \frac{S}{2n(1-\upsilon)}$, and a specific $\upsilon < 1$ can make $\frac{\mu}{2} \mathbf{E} \left\| x_T - x_* \right\|_2^2 \le \epsilon$ after $T = \tilde{O}\left(\frac{n}{S} + \frac{L_g C_f}{\mu} + \frac{C_g \sqrt{n L_f}}{\sqrt{S\mu}} + \frac{C_g^2 L_f}{\mu} + \frac{n L_f \sigma_0^2}{BS\epsilon} + \frac{C_f^2 \sigma_1^2}{\mu B \epsilon} + \frac{\delta^2}{\mu S \epsilon} \right)$ iterations. If g_i is non-smooth, the iteration complexity is $T = \tilde{O}\left(\frac{n}{S} + \frac{C_g \sqrt{n L_f}}{\sqrt{S\mu}} + \frac{C_g^2 L_f}{\mu} + \frac{n L_f \sigma_0^2}{BS\epsilon} + \frac{C_f^2 \sigma_1^2}{\mu B \epsilon} + \frac{\delta^2}{\mu S \epsilon} + \frac{C_f^2 C_g^2}{\mu \epsilon} \right)$.*

Remark 21. Compared to the results of ALEXR with $\theta \in (0, 1)$ in Theorem 3, there is an extra term $\tilde{O}\left(\frac{C_g^2 L_f}{\mu} \right)$ term, which makes the iteration complexity of SOX has a worse dependence on the Lipschitz constants C_g and L_f when $C_g, L_f \ge 1$.

Proof. Plug $\theta = 0$ into (C.9).

$$\begin{aligned}
&\mathbf{E}[L(x_{t+1}, y_*) - L(x_*, \bar{y}_{t+1})] \qquad \text{(H.1)}\\
&\le \frac{\tau + \rho\left(1 - \frac{S}{n}\right)}{S}\mathbf{E}[U_\psi(y_*, y_t)] - \frac{\tau+\rho}{S}\mathbf{E}[U_\psi(y_*, y_{t+1})] + \frac{\eta}{2}\mathbf{E}\left\|x_* - x_t\right\|_2^2 - \frac{\eta+\mu}{2}\mathbf{E}\left\|x_* - x_{t+1}\right\|_2^2\\
&\quad - \left(\frac{\tau}{n} - \frac{\lambda_2}{\mu_\psi n}\right)\mathbf{E}\left[U_\psi(\bar{y}_{t+1}, y_t)\right] - \left(\frac{\eta}{2} - \frac{C_g^2}{2\lambda_2} - \frac{L_g C_f}{2}\right)\mathbf{E}\left\|x_{t+1} - x_t\right\|_2^2 + \mathbf{E}[\Gamma_{t+1}] + \frac{2\sigma_0^2}{B\mu_\psi(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu},
\end{aligned}$$

where $\Gamma_t \coloneqq \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_t) - g_i(x_{t-1}),\, y_*^{(i)} - y_t^{(i)}\right\rangle$. We bound the $\mathbf{E}[\Gamma_{t+1}]$ term by

$$\begin{aligned}
\mathbf{E}[\Gamma_{t+1}] &= \frac{1}{n}\sum_{i=1}^n \left\langle g_i(x_{t+1}) - g_i(x_t),\, y_*^{(i)} - y_{t+1}^{(i)}\right\rangle \le \frac{C_g}{n}\left\|x_{t+1} - x_t\right\|_2 \left\|y_* - y_{t+1}\right\|\\
&\le \frac{\rho}{2n^2}\mathbf{E}[U_\psi(y_*, y_{t+1})] + \frac{C_g^2}{2\mu_\psi\rho}\left\|x_{t+1} - x_t\right\|_2^2.
\end{aligned}$$

Choose $\lambda_2 = n\mu_\psi\rho$ and such that

$$\begin{aligned}
&\mathbf{E}[L(x_{t+1}, y_*) - L(x_*, \bar{y}_{t+1})]\\
&\le \frac{\tau + \rho\left(1 - \frac{S}{n}\right)}{S}\mathbf{E}[U_\psi(y_*, y_t)] - \frac{\tau+\rho\left(1-\frac{S}{n^2}\right)}{S}\mathbf{E}[U_\psi(y_*, y_{t+1})] + \frac{\eta}{2}\mathbf{E}\left\|x_* - x_t\right\|_2^2 - \frac{\eta+\mu}{2}\mathbf{E}\left\|x_* - x_{t+1}\right\|_2^2\\
&\quad - \left(\frac{\tau}{n} - \frac{\lambda_2}{\mu_\psi n}\right)\mathbf{E}\left[U_\psi(\bar{y}_{t+1}, y_t)\right] - \left(\frac{\eta}{2} - \frac{C_g^2}{2\lambda_2} - \frac{C_g^2}{2\mu_\psi\rho} - \frac{L_g C_f}{2}\right)\mathbf{E}\left\|x_{t+1} - x_t\right\|_2^2 + \frac{2\sigma_0^2}{B\mu_\psi(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}.
\end{aligned}$$

Define $\Upsilon_t^x \coloneqq \frac{1}{2}\mathbf{E}\left\|x_* - x_t\right\|_2^2$ and $\Upsilon_t^y = \frac{1}{S}\mathbf{E}[U_\psi(y_*, y_t)]$. Note that $L(x_{t+1}, y_*) - L(x_*, \bar{y}_{t+1}) \ge 0$. Multiply both sides of (C.9) by υ^{-t} for some $\upsilon > 0$ and do telescoping sum from $t = 0$ to $T-1$. Add $\eta\upsilon^{-T}\Upsilon_T^x$ to both sides.

$$\begin{aligned}
\eta\upsilon^{-T}\Upsilon_T^x &\le \sum_{t=0}^{T-1}\upsilon^{-t}\left(\left(\eta\Upsilon_t^x + \left(\tau + \rho\left(1 - \frac{S}{n}\right)\right)\Upsilon_t^y\right) - \left((\eta+\mu)\Upsilon_{t+1}^x + \left(\tau + \rho\left(1 - \frac{S}{n^2}\right)\right)\Upsilon_{t+1}^y\right)\right)\\
&\quad + \eta\upsilon^{-T}\Upsilon_T^x + \left(\frac{2\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B} + \frac{\delta^2}{S}}{\eta+\mu}\right)\sum_{t=0}^{T-1}\upsilon^{-t} - \sum_{t=0}^{T-1}\upsilon^{-t}\left(\frac{\tau}{n} - \frac{\lambda_2}{\mu_\psi n}\right)\mathbf{E}[U_\psi(\bar{y}_{t+1}, y_t)]\\
&\quad - \sum_{t=0}^{T-1}\upsilon^{-t}\left(\frac{\eta}{2} - \frac{C_g^2}{2\lambda_2} - \frac{C_g^2}{2\mu_\psi\rho} - \frac{L_g C_f}{2}\right)\mathbf{E}\left\|x_{t+1} - x_t\right\|_2^2.
\end{aligned}$$

Let $\eta \ge \frac{\mu\upsilon}{1-\upsilon}$ such that $\upsilon \le \frac{\eta}{\eta+\mu}$ and $\tau \ge \frac{\rho S}{n(1-\upsilon)}$ such that $\upsilon \le \frac{\tau+\rho\left(1-\frac{S}{n}\right)}{\tau+\rho\left(1-\frac{S}{n^2}\right)}$. Then,

$$\sum_{t=0}^{T-1}\theta^{-t}\left(\left(\eta\Upsilon_t^x + \left(\tau + \rho\left(1 - \frac{S}{n}\right)\right)\Upsilon_t^y\right) - \left((\eta+\mu)\Upsilon_{t+1}^x + (\tau+\rho)\Upsilon_{t+1}^y\right)\right) \le \eta\Upsilon_0^x + \left(\tau + \rho\left(1 - \frac{S}{n}\right)\right)\Upsilon_0^y.$$

We choose $\lambda_2 \asymp \frac{C_g\sqrt{S\rho\mu_\psi}}{\sqrt{n\mu}}$ and $1/\tau \le O\left(\frac{\sqrt{n\mu\mu_\psi}}{C_g\sqrt{S\rho}}\right)$, $1/\eta \le O\left(\frac{\mu_\psi\rho}{C_g^2} \vee \frac{\sqrt{S\rho\mu_\psi}}{C_g\sqrt{n\mu}} \wedge \frac{1}{L_gC_f}\right)$. Then,

$$\begin{aligned}
\mu\Upsilon_T^x &\le \mu\upsilon^T\Upsilon_0^x + \frac{\left(\tau+\rho\left(1-\frac{S}{n}\right)\right)(1-\upsilon)}{\upsilon}\upsilon^T\Upsilon_0^y + \left(\frac{2\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\eta+\mu}\right)\\
&\le \mu\upsilon^T\Upsilon_0^x + (\tau+\rho)(1-\upsilon)\upsilon^T\Upsilon_0^y + \left(\frac{2\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\eta+\mu}\right)\\
&= \mu\upsilon^T\Upsilon_0^x + \rho\left(\frac{S}{2n}+(1-\upsilon)\right)\upsilon^T\Upsilon_0^y + \left(\frac{2\sigma_0^2}{\mu_\psi B(\rho+\tau)} + \frac{\frac{C_f^2\sigma_1^2}{B}+\frac{\delta^2}{S}}{\eta+\mu}\right).
\end{aligned}$$

We select $\eta = \frac{\mu\upsilon}{1-\upsilon}$, $\tau = \frac{\rho S}{n(1-\upsilon)}$, and

$$\upsilon = O\left(1 - \frac{S}{n} \wedge \frac{\mu}{L_gC_f} \wedge \frac{\mu\rho\mu_\psi}{C_g^2} \wedge \sqrt{\frac{\mu\rho\mu_\psi S}{C_g^2 n}} \wedge \frac{\mu_\psi B\rho S\epsilon}{\sigma_0^2 n} \wedge \frac{B\mu\epsilon}{C_f^2\sigma_1^2} \wedge \frac{S\mu\epsilon}{\delta^2}\right).$$

Since $L_f \coloneqq \frac{1}{\mu_\psi\rho}$, the number of iterations needed by Algorithm 1 to make $\mu\Upsilon_T^x \le \epsilon$ is

$$T = \tilde{O}\left(\frac{n}{S} + \frac{L_gC_f}{\mu} + \frac{C_g\sqrt{nL_f}}{\sqrt{S\mu}} + \frac{C_g^2L_f}{\mu} + \frac{nL_f\sigma_0^2}{BS\epsilon} + \frac{C_f^2\sigma_1^2}{\mu B\epsilon} + \frac{\delta^2}{\mu S\epsilon}\right).$$

□

H.2 Convex Case

Theorem 22. *Under Assumptions 1, 2, 3, 4, 5, 6,* ALEXR *with $\theta = 0$ and an L_ψ-smooth ψ_i can make $\mathbf{E}[F(\bar{x}_T) - F(x_*)] \le \epsilon$, $\bar{x}_T = \frac{1}{T}\sum_{t=0}^{T-1} x_t$ after $T = O\left(\frac{C_f^2D_{\mathcal{X}}^2}{\epsilon^2} + \frac{\delta^2D_{\mathcal{X}}^2}{S\epsilon^2} + \frac{C_f^2\sigma_1^2}{B\epsilon^2} + \frac{\sigma_0^2(1+L_\psi^2/(S\mu_\psi^2))\sum_{i=1}^n D_{\psi_i,\mathcal{Y}_i}^2}{\mu_\psi BS\epsilon^2}\right)$ iterations by setting $\eta = O\left(\frac{C_f^2}{\epsilon} \vee \frac{\delta^2}{S\epsilon} \vee \frac{C_f^2\sigma_1^2}{B\epsilon}\right)$, $\tau = O\left(\frac{\sigma_0^2}{\mu_\psi B\epsilon}\right)$.*

Remark 23. Compared to the results of ALEXR with $\theta = 1$ in Theorem 6, the $O\left(\frac{1}{\epsilon^2}\right)$ term persists even in the case that g_i is smooth. Thus, ALEXR with $\theta = 0$ does not fully exhibit the parallel speed-up to batch sizes B, S, nor does it achieve the $O\left(\frac{1}{\epsilon}\right)$ rate when variances $\sigma_0^2, \sigma_1^2, \delta^2$ vanish.

Proof. For ALEXR with $\theta = 0$, we have $\tilde{g}_t^{(i)} = g_i(x_t; \mathcal{B}_t^{(i)})$. Then, for any $\lambda_4 > 0$ we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\langle g_i(x_{t+1}) - \tilde{g}_t,\, y^{(i)} - \bar{y}_{t+1}^{(i)}\right\rangle = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\langle g_i(x_{t+1}) - g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)} - \bar{y}_{t+1}^{(i)}\right\rangle\\
&= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\langle g_i(x_{t+1}) - g_i(x_t),\, y^{(i)} - \bar{y}_{t+1}^{(i)}\right\rangle\right] + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)} - \bar{y}_{t+1}^{(i)}\right\rangle\right]\\
&\le \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i(x_{t+1}) - g_i(x_t)\|_*\left\|y^{(i)} - \bar{y}_{t+1}^{(i)}\right\|\right] + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)} - \bar{y}_{t+1}^{(i)}\right\rangle\right]\\
&\le \frac{C_g^2\mathbf{E}\|x_{t+1}-x_t\|_2^2}{2\lambda_4} + 2\lambda_4\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\left\|y^{(i)}\right\|^2 + \left\|\bar{y}_{t+1}^{(i)}\right\|^2\right) + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)} - \bar{y}_{t+1}^{(i)}\right\rangle\right]\\
&\le \frac{C_g^2\mathbf{E}\left[\|x_{t+1}-x_t\|_2^2\right]}{2\lambda_4} + 4\lambda_4C_f^2 + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\langle g_i(x_t) - g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)} - \bar{y}_{t+1}^{(i)}\right\rangle\right]. \qquad \text{(H.2)}
\end{aligned}$$

The last term in (H.2) is bounded as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\left\langle g_i(x_t)-g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)}-\bar{y}_{t+1}^{(i)}\right\rangle\right]\\&=\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_t\left[\left\langle g_i(x_t)-g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)}-y_t^{(i)}\right\rangle\right]+\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\left\langle g_i(x_t)-g_i(x_t;\mathcal{B}_t^{(i)}),\, y_t^{(i)}-\bar{y}_{t+1}^{(i)}\right\rangle\right]\end{aligned}\tag{H.3}$$

To bound the first term in (H.3), we have $\mathbf{E}\left[\left\langle g_i(x_t)-g_i(x_t;\mathcal{B}_t^{(i)}),\, y_t^{(i)}\right\rangle \mid \mathcal{F}_{t-1}\right]=0$. Besides, Lemma 12 implies that for some $\lambda_2>0$ and sequence $\{\tilde{y}_t^{(i)}\}_t$

$$\mathbf{E}\left[\left\langle g_i(x_t)-g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)}\right\rangle\right]\le\mathbf{E}[\lambda_2 U_{\psi_i}(y^{(i)},\tilde{y}_t^{(i)})-\lambda_2 U_{\psi_i}(y^{(i)},\tilde{y}_{t+1}^{(i)})]+\frac{1}{2\lambda_2\mu_\psi}\mathbf{E}\left\|g_i(x_t)-g_i(x_t;\mathcal{B}_t^{(i)})\right\|_*^2$$

such that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\left\langle g_i(x_t)-g_i(x_t;\mathcal{B}_t^{(i)}),\, y^{(i)}\right\rangle\right]\le\frac{\lambda_2}{n}\mathbf{E}[U_\psi(y,\tilde{y}_t)-U_\psi(y,\tilde{y}_{t+1})]+\frac{\sigma_0^2}{2\lambda_2 B\mu_\psi}.\tag{H.4}$$

For any $\lambda_3>0$, the second term in (H.3) can be bounded as

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\left\langle g_i(x_t)-g_i(x_t;\mathcal{B}_t^{(i)}),\, y_t^{(i)}-\bar{y}_{t+1}^{(i)}\right\rangle\right]&\le\frac{\lambda_3}{2n}\sum_{i=1}^{n}\mathbf{E}\left[\left\|g_i(x_t)-g_i(x_t;\mathcal{B}_t^{(i)})\right\|_*^2\right]+\frac{\mathbf{E}\left[\|y_t-\bar{y}_{t+1}\|^2\right]}{2\lambda_3 n}\\&\le\frac{\lambda_3\sigma_0^2}{2B}+\frac{\mathbf{E}\left[U_\psi(\bar{y}_{t+1},y_t)\right]}{\lambda_3\mu_\psi n}.\end{aligned}\tag{H.5}$$

Put (H.2), (H.3), (H.4), (H.5) together

$$\begin{aligned}\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\langle g_i(x_{t+1})-\tilde{g}_t,\, y^{(i)}-\bar{y}_{t+1}^{(i)}\right\rangle\right]&\le\frac{C_g^2\mathbf{E}\left[\|x_{t+1}-x_t\|_2^2\right]}{2\lambda_4}+4\lambda_4 C_f^2+\frac{\lambda_2}{n}\mathbf{E}[U_\psi(y,\tilde{y}_t)-U_\psi(y,\tilde{y}_{t+1})]\\&\quad+\frac{\sigma_0^2}{2B\lambda_2\mu_\psi}+\frac{\lambda_3\sigma_0^2}{2B}+\frac{\mathbf{E}\left[U_\psi(\bar{y}_{t+1},y_t)\right]}{2\lambda_3\mu_\psi n}.\end{aligned}$$

Note that (6.1), (6.3), (C.3) still hold. Choose $\lambda_4=O(1/\eta)$, $\lambda_2=O(n\tau/(\mu_\psi S))$, $\lambda_3=O(1/(\tau\mu_\psi))$ and $\eta=O(1/\epsilon)$, $\tau=O(1/(B\mu_\psi\epsilon))$ and follow the steps in the proof of Theorem 6. □