Finite-Sum Coupled Compositional Stochastic Optimization: Theory and Applications

Bokun Wang [1] **Tianbao Yang** [1]

Abstract

This paper studies stochastic optimization for a sum of compositional functions, where the inner-level function of each summand is coupled with the corresponding summation index. We refer to this family of problems as finite-sum coupled compositional optimization (FCCO). It has broad applications in machine learning for optimizing non-convex or convex compositional measures/objectives such as average precision (AP), p-norm push, listwise ranking losses, neighborhood component analysis (NCA), deep survival analysis, deep latent variable models, etc., which deserves finer analysis. Yet, existing algorithms and analyses are restricted in one or other aspects. The contribution of this paper is to provide a comprehensive convergence analysis of a simple stochastic algorithm for both non-convex and convex objectives. Our key result is the **improved oracle complexity with the parallel speed-up** by using the moving-average based estimator with mini-batching. Our theoretical analysis also exhibits new insights for improving the practical implementation by sampling the batches of equal size for the outer and inner levels. Numerical experiments on AP maximization, NCA and p-norm push corroborate some aspects of the theory.

1. Introduction

A fundamental problem in machine learning (ML) that has been studied extensively is the empirical risk minimization (ERM), whose objective is a sum of individual losses on training examples, i.e.,

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i),$$

where $\mathbf{w}$ and Ω denotes the model parameter and its domain ($\Omega \subseteq \mathbb{R}^d$), $\mathcal{D}$ denotes the training set of n examples, and $\mathbf{z}_i$ denotes an individual data. However, ERM may hide the complexity of individual loss and gradient computation in many interesting measures/objectives. Instead, in this paper we study a new family of problems that aims to optimize the following compositional objective:

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i)), \qquad (1)$$

where $g : \Omega \mapsto \mathcal{R}$ [1], $f_i : \mathcal{R} \mapsto \mathbb{R}$, and $\mathcal{S}_i$ denotes another (finite or infinite [2]) set of examples that could be either dependent or independent of $\mathbf{z}_i$. We give an example for each case: 1) In the bipartite ranking problem, $\mathcal{D}$ represents the positive data while $\mathcal{S} = \mathcal{S}_i$ represents the negative data; 2) In the robust learning problem (e.g. the invariant logistic regression in Hu et al. 2020), $\mathcal{D}$ represents the training data set while $\mathcal{S}_i$ denotes the set of perturbed observations for data $\mathbf{z}_i \in \mathcal{D}$, where $\mathcal{S}_i$ depends on $\mathbf{z}_i$. We are particularly interested in the case that set $\mathcal{S}_i$ is infinite or contains a large number of items, and assume that an unbiased stochastic estimators of g and ∇g can be computed via sampling from $\mathcal{S}_i$. We refer to (1) as finite-sum coupled compositional optimization (FCCO) and its objective as finite-sum coupled compositional risk (FCCR), where for each data $\mathbf{z}_i$ the risk $f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$ is of a compositional form such that g couples each $\mathbf{z}_i$ and items in $\mathcal{S}_i$. It is notable that f_i could be stochastic or has a finite-sum structure that depends on a large set of items. For simplicity of presentation and discussion, we focus on the case that f_i is a simple deterministic function whose value and gradient can be easily computed, which covers many interesting objectives of interest. The algorithms and analysis can be extended to the case

[1]Department of Computer Science, The University of Iowa, IA, USA. Correspondence to: Bokun Wang <bokunw.wang@gmail.com>, Tianbao Yang <tianbao-yang@uiowa.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

[1] $\mathcal{R}$ refers to the d'-dimensional codomain of g ($d' \geq 1$).

[2] If $\mathcal{S}_i$ is a finite set, we define $g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i) = \frac{1}{|\mathcal{S}_i|}\sum_{\xi_{ij}\in\mathcal{S}_i} g(\mathbf{w};\mathbf{z}_i,\xi_{ij})$; If $\mathcal{S}_i$ is an infinite set, we define $g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i) = \mathbb{E}\left[g(\mathbf{w};\mathbf{z}_i,\xi_i) \mid \mathbf{z}_i\right]$, where $\xi_i \in \mathcal{S}_i$.

that $\{f_i, \nabla f_i\}$ are estimated by their unbiased stochastic versions using random samples (discussed in Appendix E).

Applications of FCCO. The Average Precision (AP) maximization problem studied in Qi et al. (2021) is an example of FCCO. Nevertheless, we notice that the application of FCCO is much broader beyond the AP maximization, including but are not limited to p-norm push optimization (Rudin, 2009), listwise ranking objectives (Cao et al., 2007; Xia et al., 2008) (e.g. ListNet, ListMLE, NDCG), neighborhood component analysis (NCA) (Goldberger et al., 2004), deep survival analysis (Katzman et al., 2018), deep latent variable models (Guu et al., 2020), etc. We postpone the details of some of these problems to Section 4 and 5. We would like to emphasize that efficient stochastic algorithms for these problems are lacking or under-developed when the involved set $\mathcal{S}_i$ is big and/or the predictive model is nonconvex.

In this paper, we propose the **S**tochastic **O**ptimization of the **X** objectives listed in Sections 4 and 5 (SOX) and establish its convergence guarantees for several classes of functions.

2. Related Work

In this section, we connect the FCCO problem to Conditional Stochastic Optimization (CSO) and Stochastic Compositional Optimization (SCO) in the literature and discuss the limitations of existing algorithms for FCCO. Then, we position SOX in previous studies and list our contributions.

2.1. Conditional Stochastic Optimization (CSO)

The most straightforward approach to solve the FCCO problem in (1) is to compute the gradient $\nabla F(\mathbf{w}) = \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \nabla g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i)\nabla f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$ and then use the gradient descent method. However, it can be seen that computing the gradient $\nabla F(\mathbf{w})$ is very expensive, if not infeasible, when $|\mathcal{D}|$ or $|\mathcal{S}_i|$ is large. Thus, a natural idea is to sample mini-batches $\mathcal{B}_1 \subset \mathcal{D}$ and $\mathcal{B}_{i,2} \subset \mathcal{S}_i$ and compute the stochastic gradient $\mathbf{v} = \frac{1}{|\mathcal{B}_1|}\sum_{\mathbf{z}_i\in\mathcal{B}_1} \nabla g(\mathbf{w};\mathbf{z}_i,\mathcal{B}_{i,2})\nabla f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{B}_{i,2}))$ and update the parameter as $\mathbf{w}\leftarrow\mathbf{w}-\eta\mathbf{v}$, where $g(\mathbf{w};\mathbf{z}_i,\mathcal{B}_{i,2}) := \frac{1}{|\mathcal{B}_{i,2}|}\sum_{\xi_{ij}\in\mathcal{B}_{i,2}} g(\mathbf{w};\mathbf{z}_i,\xi_{ij})$. The resulting algorithm is named as biased stochastic gradient descent (BSGD) in Hu et al. (2020) and the convergence guarantees of BSGD under different assumptions are established. Actually, Hu et al. (2020) study the more general problem $\mathbb{E}_\xi f_\xi(\mathbb{E}_{\zeta|\xi}[g_\zeta(\mathbf{w};\xi)])$, which is referred to as conditional stochastic optimization (CSO) and show that BSGD has the optimal oracle complexity for the general CSO problem. The FCCO problem can be mapped to the CSO problem by $\mathbf{z}_i=\xi$, $\zeta\in\mathcal{S}_i$ and the only difference between these two is the *finite-sum structure* of the outer-level function in FCCO. Unfortunately, BSGD requires unrealistically large batch sizes to ensure the convergence from the theoretical perspective (Please refer to columns 5 and 6 of Table 1). As a comparison, our approach explicitly exploits the finite support of the outer level and leads to improved oracle complexity with mini-batch sizes $|\mathcal{B}_1| = O(1)$, $|\mathcal{B}_{i,2}| = O(1)$.

2.2. Stochastic Compositional Optimization (SCO)

A closely related class of problems: stochastic compositional optimization (SCO) has been extensively studied in the literature. In particular, the SCO problem with the finite support in the outer level is in the form of $F(\mathbf{w}) = \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathcal{S}))$, where $\mathcal{S}$ might be finite or not. The difference between FCCO and SCO is that the inner function $g(\mathbf{w};\mathcal{S})$ in SCO does not depend on $\mathbf{z}_i$ of the outer summation. The SCGD algorithm (Wang et al., 2017) is a seminal work in this field, which tracks the unknown $g(\mathbf{w}_t;\mathcal{S})$ with an auxiliary variable u_t that is updated by the exponential moving average $u \leftarrow (1-\gamma)u + \gamma g(\mathbf{w};\mathcal{B}_2)$, $\gamma\in(0,1)$ based on an unbiased stochastic estimator $g(\mathbf{w};\mathcal{B}_2)$ of $g(\mathbf{w};\mathcal{S})$, which circumvents the unrealistically large batch size required by the sample average approximation approach. For example, we can sample a mini-batch $\mathcal{B}_2\subset\mathcal{S}$ and compute $g(\mathbf{w};\mathcal{B}_2) := \frac{1}{|\mathcal{B}_2|}\sum_{\xi_j\in\mathcal{B}_2} g(\mathbf{w};\xi_j)$. Then, the stochastic estimator of $\nabla F(\mathbf{w})$ can be computed as $\mathbf{v}_t = \frac{1}{|\mathcal{B}_2|}\sum_{\mathbf{z}_i\in\mathcal{B}_1}\nabla g(\mathbf{w};\mathcal{B}_2)\nabla f_i(u)$. More recently, the NASA algorithm (Ghadimi et al., 2020) modifies the SCGD algorithm by adding the exponential moving average (i.e., the momentum) to the gradient estimator, i.e., $\mathbf{v}\leftarrow(1-\beta)\mathbf{v} + \beta\frac{1}{|\mathcal{B}_1|}\sum_{\mathbf{z}_i\in\mathcal{B}_1}\nabla g(\mathbf{w};\mathcal{B}_2)\nabla f_i(u)$, $\beta\in(0,1)$, which improves upon the convergence rates of SCGD. When f_i is convex and monotone ($d'=1$) and g is convex, Zhang & Lan (2020) provide a more involved analysis for the two-batch SCGD[3] in its primal-dual equivalent form and derive the optimal rate for a special class of problems that inner function is convex while the outer function is convex and monotone.

SCO reformulation of FCCO. Given the union data set $\mathcal{S} = \mathcal{S}_1\cup\cdots\mathcal{S}_i\cdots\cup\mathcal{S}_n$, we can define $\mathbf{g}(\mathbf{w};\mathcal{S}) = [g(\mathbf{w};\mathbf{z}_1,\mathcal{S}_1)^\top,\dots,g(\mathbf{w};\mathbf{z}_n,\mathcal{S}_n)^\top]^\top\in\mathbb{R}^{nd'}$ and $\hat{f}_i(\cdot) := f_i(\mathbb{I}_i\cdot)$, $\mathbb{I}_i := [0_{d\times d},\dots,I_{d\times d},\dots,0_{d\times d}]\in\mathbb{R}^{d'\times nd'}$ (the i-th block in $\mathbb{I}_i$ is the identity matrix while the others are zeros), the FCCO problem $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$ can be reformulated as an SCO problem $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \hat{f}_i(\mathbf{g}(\mathbf{w};\mathcal{S}))$ such that the existing algorithms for the SCO problem can be directly applied to our FCCO problem. Unfortunately, applying SCGD and NASA on

[3]In the original SCGD algorithm (Wang et al., 2017), they use the same batch $\mathcal{B}_2$ to update u by $g(\mathbf{w};\mathcal{B}_2)$ and to compute the gradient estimator by $\nabla g(\mathbf{w};\mathcal{B}_2)$. In the work of Zhang & Lan (2020), they analyze the two-batch version SCGD which uses independent batches $\mathcal{B}_2$ and $\mathcal{B}_2'$ for $g(\mathbf{w};\mathcal{B}_2)$ and $\nabla g(\mathbf{w};\mathcal{B}_2')$. The two-batch version with independent $\mathcal{B}_2, \mathcal{B}_2'$ is definitely less efficient, but it considerably simplies the analysis.

Table 1. Summary of **iteration complexity** of different methods for different problems. "NC" means non-convexity of F, "C" means convexity of F,"SC" means strongly convex and "PL" means Polyak-Lojasiewicz condition. "N/A" means not applicable or not available. For the complexity of our method, we omit constants that are independent of n or the batch sizes. * denotes additional conditions that Ω is bounded, f is convex and monotone while g_i is convex (but the smoothness of g_i is not required). † assumes a stronger SC condition, which is imposed on a stochastic function instead of the original objective. We suppose $B_1 = |\mathcal{B}_1^t|$ and $B_2 = |\mathcal{B}_{i,2}^t|$ for simplicity.

Method	NC	C	SC (PL)	Outer Batch Size $\lvert\mathcal{B}_1\rvert$	Inner Batch Size $\lvert\mathcal{B}_{i,2}\rvert$	Parallel Speed-up
BSGD (Hu et al., 2020)	$O(\epsilon^{-4})$	$\mathcal{O}\left(\epsilon^{-2}\right)$	$\mathcal{O}\left(\mu^{-1}\epsilon^{-1}\right)^\dagger$	1	$O(\epsilon^{-2})$ (NC) $O(\epsilon^{-1})$ (C/SC)	N/A
SOAP (Qi et al., 2021)	$O(n\epsilon^{-5})$	-	-	1	1	N/A
MOAP (Wang et al., 2021)	$O\left(\frac{n\epsilon^{-4}}{B_1}\right)$	-	-	B_1	1	Partial
SOX/SOX-boost (this work)	$O\left(\frac{n\epsilon^{-4}}{B_1 B_2}\right)$	$O\left(\frac{n\epsilon^{-3}}{B_1 B_2}\right)$	$O\left(\frac{n\mu^{-2}\epsilon^{-1}}{B_1 B_2}\right)$	B_1	B_2	Yes
SOX ($\beta = 1$) (this work)	-	$O\left(\frac{n\epsilon^{-2}}{B_1}\right)^*$	-	B_1	B_2	Partial

the FCCO problem via the SCO reformulation need n oracles for the inner function $\mathbf{g}(\mathbf{w}; \mathcal{S})$ (one oracle for each $g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i)$) and update all n components of $\mathbf{u} = [u_1, \ldots, u_n]^\top$ at any time step t even if we only sample one data point $\mathbf{z}_i \in \mathcal{D}$ in the outer level, which could be expensive or even infeasible.

Apart from adopting the rather naïve SCO reformulation, we can also alter the algorithm according to the special structure of FCCO. SCGD and NASA algorithms could be better tailored for the FCCO problem if it selectively samples $\mathcal{B}_{i,2}$ and selectively updates those coordinates u_i for those sampled $\mathbf{z}_i \in \mathcal{B}_1$ at each time step, instead of sampling $\mathcal{B}_{i,2}$ for all $\mathbf{z}_i \in \mathcal{D}$ and update all n coordinates of $\mathbf{u} = [u_1^\top, \ldots, u_n^\top]^\top$. Formally, the update rule of $\mathbf{u} = [u_1^\top, \ldots, u_n^\top]^\top$ can be expressed as

$$u_i \leftarrow \begin{cases} (1-\gamma)u_i + \gamma g(\mathbf{w}; \mathbf{z}_i, \mathcal{B}_{i,2}), & i \in \mathcal{B}_1 \\ u_i, & i \notin \mathcal{B}_1 \end{cases}. \tag{2}$$

The update rule above has been exploited by some recent works (e.g. SOAP in Qi et al., 2021) to solve the average precision (AP) maximization problem, which is a special case of FCCO. However, the convergence guarantees of these algorithms are only established for smooth nonconvex problem and do not enjoy the parallel speed-up by mini-batching. In this work, we build convergence theory for a broader spectrum of problems and show the parallel speed-up effect. Moreover, we ressolve several issues of existing approaches from the algorithmic and theoretical perspectives (See Table 1 and Section 3.1 for details).

2.3. Our Contributions

Our contributions can be summarized as follows.

• On the convex and nonconvex problems, our SOX algorithm can guarantee the convergence but does not suffer from some limitations in previous methods such as the unrealistically large batch of BSGD (Hu et al., 2020), the two independent batches for oracles of the inner level in SCGD (Zhang & Lan, 2020), and the possibly inefficient/unstable update rule in MOAP (Wang et al., 2021).

• On the smooth nonconvex problem, SOX has an improved rate compared to SOAP and enjoys a better dependence on $|\mathcal{B}_{i,2}|$ compared to MOAP.

• Beyond the smooth nonconvex problem, we also establish the convergence guarantees of SOX for problems that F is convex/strongly convex/PL, which are better than BSGD in terms of oracle complexity.

• Moreover, we carefully analyze how mini-batching in the inner and outer levels improve the worst-case convergence guarantees of SOX in terms of iteration complexity, i.e., the parallel speed-up effect. The theoretical insights are numerically verified in our experiments.

3. Algorithm and Convergence Analysis

Notations. For machine learning applications, we let $\mathcal{D} = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ denote a set of training examples for general purpose, let $\mathbf{w} \in \Omega$ denote the model parameter (e.g., the weights of a deep neural network). Denote by $h_{\mathbf{w}}(\mathbf{z})$ a prediction score of the model on the data $\mathbf{z}$. A function f is Lipchitz continuous on the domain Ω if there exists $C > 0$ such that $\|f(\mathbf{w}) - f(\mathbf{w}')\| \leq C\|\mathbf{w} - \mathbf{w}'\|$ for any $\mathbf{w}, \mathbf{w}' \in \Omega$, and is smooth if its gradient is Lipchitz continuous. A function F is convex if it satisfies $F(\mathbf{w}) \geq F(\mathbf{w}') + \nabla F(\mathbf{w}')^\top(\mathbf{w} - \mathbf{w}')$ for all $\mathbf{w}, \mathbf{w}' \in \Omega$, is μ-strongly convex if there exists $\mu > 0$ such that $F(\mathbf{w}) \geq F(\mathbf{w}') + \nabla F(\mathbf{w}')^\top(\mathbf{w} - \mathbf{w}') + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}'\|^2$

Algorithm 1 SOX($\mathbf{w}^0$,$\mathbf{u}^0$,$\mathbf{v}^0$,η, β, γ, T)

1: **for** $t = 1, \ldots, T$ **do**
2: Draw a batch of samples $\mathcal{B}_1^t \subset \mathcal{D}$
3: **if** $\mathbf{z}_i \in \mathcal{B}_1^t$ **then**
4: Update the estimator of function value $g_i(\mathbf{w}^t)$

$$u_i^t = (1-\gamma)u_i^{t-1} + \gamma g(\mathbf{w}^t; \mathbf{z}_i, \mathcal{B}_{i,2}^t)$$

5: **end if**
6: Update the estimator of gradient $\nabla F(\mathbf{w}^t)$ by

$$\mathbf{v}^t = (1-\beta)\mathbf{v}^{t-1} + \beta \frac{1}{B_1} \sum_{\mathbf{z}_i \in \mathcal{B}_1^t} \nabla g(\mathbf{w}^t; \mathbf{z}_i, \mathcal{B}_{i,2}^t) \nabla f_i(u_i^{t-1})$$

7: Update the model parameter $\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t \mathbf{v}^t$
8: **end for**

for all $\mathbf{w}, \mathbf{w}' \in \Omega$. A smooth function F is said to satisfy μ-PL condition if there exists $\mu > 0$ such that $\|\nabla F(\mathbf{w})\|^2 \geq \mu(F(\mathbf{w}) - \min_{\mathbf{w}} F(\mathbf{w}))$, $\mathbf{w} \in \Omega$.

We make the following assumptions throughout the paper[4].

Assumption 1. We assume that (i) $f_i(\cdot)$ is differentiable, L_f-smooth and C_f-Lipchitz continuous; (ii) $g(\cdot; \mathbf{z}_i, \mathcal{S}_i)$ is differentiable, L_g-smooth and C_g-Lipchitz continuous for any $\mathbf{z}_i \in \mathcal{D}$; (iii) F is lower bouned by F^*.

Remark: If the assumption above is satisfied, it is easy to verify that $F(\mathbf{w})$ is L_F-smooth, where $L_F := C_f L_g + C_g^2 L_f$ (see Lemma 4.2 in Zhang & Xiao 2021). The assumption that f_i is smooth and Lipchitz continuous seems to be strong. However, the image of g_i is bounded on domain Ω in many applications (otherwise there might be a numerical issue), hence f_i is smooth and Lipchitz continuous in a bounded domain is enough for our results.

3.1. A Better Stochastic Algorithm for FCCO

We follow the idea of tailoring SCGD/NASA to solve the FCCO problem by *selective sampling* and *selective update* as described in Section 2.2. Next, we thoroughly discuss the relation of our SOX algorithm to the existing algorithms SOAP and MOAP for the FCCO problem.

SOAP algorithm (Qi et al., 2021) combines (2) with the gradient step $\mathbf{v} \leftarrow \frac{1}{|\mathcal{B}_1|}\sum_{\mathbf{z}_i \in \mathcal{B}_1} \nabla g(\mathbf{w}; \mathbf{z}_i, \mathcal{B}_{i,2}) \nabla f_i(u_i)$, $\mathbf{w} \leftarrow \mathbf{w} - \eta \mathbf{v}$. Wang et al. (2021) attempted to do the same adaptation for NASA (Ghadimi et al., 2020) by an algorithm called MOAP, which applies the uniform random sparsification (Wangni et al., 2018) or the uniform randomized block coordinate sampling (Nesterov, 2012) to the whole $\mathbf{g}(\mathbf{w})$ and derives the improved rate compared to SOAP. To be specific, the update rule of $\mathbf{u} = [u_1, \ldots, u_n]^\top$ in MOAP is

$$u_i \leftarrow \begin{cases} (1-\gamma)u_i + \gamma \frac{n}{|\mathcal{B}_1|} g(\mathbf{w}; \mathbf{z}_i, \mathcal{B}_{i,2}), & i \in \mathcal{B}_1 \\ (1-\gamma)u_i, & i \notin \mathcal{B}_1. \end{cases} \quad (3)$$

It is worth mentioning that the convergence guarantees for SOAP and MOAP are only established for the smooth nonconvex problems. Besides, SOAP and MOAP are only analyzed when $|\mathcal{B}_1^t| = 1$. The update rule (3) of MOAP also has several extra drawbacks: a) It requires extra costs to update all u_i at each iteration, while (2) only needs to update u_i for the sampled $\mathbf{z}_i \in \mathcal{B}_1$; b) For the large-scale problems (i.e., n is large), multiplying $g(\mathbf{w}; \mathbf{z}_i, \mathcal{B}_{i,2})$ by $\frac{n}{|\mathcal{B}_1|}$ might lead to numerical issue; c) Due to the property of random sparisifcation/block coordinate sampling (see Proposition 3.5 in Khirirat et al. 2018), it does not enjoy any benefit of mini-batch $\mathcal{B}_{i,2}$ in terms of iteration complexity.

Main idea of SOX: We make subtle modifications on SOAP — 1) directly adding the gradient momentum; 2) using u_i^{t-1} instead of u_i^t in step 6 of Algorithm 1, which are crucial for us to improve the convergence rate. In particular, taking expectation of the estimation error $\|u_i^{t-1} - g(\mathbf{w}^t; \mathbf{z}_i, \mathcal{S}_i)\|^2$ over the randomness in $\mathbf{z}_i \in \mathcal{B}_1^t$ (due to independence between $\mathbf{u}^{t-1}$ and the randomness in $\mathbf{z}_i \in \mathcal{B}_1^t$) leads to bounding the average error $\frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{D}} \|u_i^{t-1} - g(\mathbf{w}^t; \mathbf{z}_i, \mathcal{S}_i)\|^2$ over all $\mathbf{z}_i \in \mathcal{D}$, which can be decomposed into $\frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{D}} \|u_i^t - u_i^{t-1}\|^2$ and $\frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{D}} \|u_i^t - g(\mathbf{w}^t; \mathbf{z}_i, \mathcal{S}_i)\|^2$, where the latter term is bounded as in Lemma 1 and the first term is cancelled with the highlighted negative term in Lemma 1.

3.2. Improved Rate for the Nonconvex Problems

In this subsection, we present the convergence analysis for the smooth nonconvex problems. We will highlight the key differences from the previous analysis. We use the following assumption, which is also used in previous works (Qi et al., 2021; Wang et al., 2021).

Assumption 2. We assume that $\mathbb{E}[\|g(\mathbf{w}; \mathbf{z}_i, \xi_i) - g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i)\|^2 \mid \mathbf{z}_i] \leq \sigma^2$ and $\mathbb{E}[\|\nabla g(\mathbf{w}; \mathbf{z}_i, \xi_i) - \nabla g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i)\|^2 \mid \mathbf{z}_i] \leq \zeta^2$ for any $\mathbf{w}$, $\mathbf{z}_i \in \mathcal{D}$, and $\xi_i \in \mathcal{S}_i$.

We aim to find the approximate stationary points.

Definition 1. $\mathbf{w}$ is an ϵ-stationary point if $\|\nabla F(\mathbf{w})\| \leq \epsilon$.

The recursion for the variance of inner function value estimation $\Xi_t := \frac{1}{n}\|\mathbf{u}^t - \mathbf{g}(\mathbf{w}^t; \mathcal{S})\|^2$ is crucial for our analysis.

Lemma 1. *If $\gamma \leq 1/5$, the function value variance $\Xi_t := \frac{1}{n}\|\mathbf{u}^t - \mathbf{g}(\mathbf{w}^t; \mathcal{S})\|^2$ can be bounded as*

$$\mathbb{E}[\Xi_{t+1}] \leq \left(1 - \frac{\gamma B_1}{4n}\right)\mathbb{E}[\Xi_t] + \frac{5n\eta^2 C_g^2 \mathbb{E}[\|\mathbf{v}^t\|^2]}{\gamma B_1} + \frac{2\gamma^2\sigma^2 B_1}{nB_2} - \frac{1}{4n}\mathbb{E}\left[\sum_{\mathbf{z}_i \in \mathcal{B}_1^t} \|u_i^{t+1} - u_i^t\|^2\right].$$

[4]The result in Theorem 3 does not need g_i to be smooth.

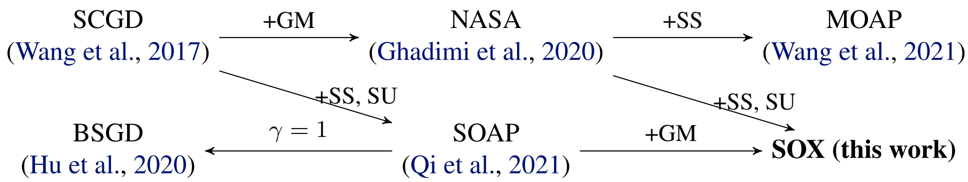


Figure 1. The algorithmic relationship among our SOX and the existing algorithms. "GM" refers to the gradient momentum; "SS" refers to the property that the algorithm only needs to selectively sample $\mathcal{B}_{i,2}$ for those $\mathbf{z}_i \in \mathcal{B}_1$ at each time step; "SU" refers to the property that the algorithm only needs to selectively update those coordinates u_i for $\mathbf{z}_i \in \mathcal{B}_1$.

Instead of sampling a singleton $\mathcal{B}_1^t = \{i_t\}$ at each iteration and bounding $\sum_t [\|u_{i_t}^t - g(\mathbf{w}^t; \mathbf{z}_{i_t}, \mathcal{S}_{i_t})\|^2]$ for $\mathbf{z}_{i_t}$ in Qi et al. (2021) [5], Lemma 1 bounds $\sum_t \|\mathbf{u}^t - \mathbf{g}(\mathbf{w}^t; \mathcal{S})\|^2$ that includes all coordinates of $\mathbf{u}^t$ at each iteration. To build the recursion, we consider a strongly convex minimization problem $\min_{\mathbf{u}} \frac{1}{2}\|\mathbf{u} - \mathbf{g}(\mathbf{w}^t; \mathcal{S})\|^2$ that is equivalent to

$$\min_{\mathbf{u}=[u_1,\ldots,u_n]^\top} \frac{1}{2} \sum_{\mathbf{z}_i \in \mathcal{D}} \|u_i - g(\mathbf{w}^t; \mathbf{z}_i, \mathcal{S}_i)\|^2. \quad (4)$$

Then, the step 4 in SOX can be viewed as stochastic block coordinate descent algorithm applied to (4), i.e.,

$$u_i^t = \begin{cases} u_i^{t-1} - \gamma\left(u_i^{t-1} - g(\mathbf{w}^t; \mathbf{z}_i, \mathcal{B}_{i,2}^t)\right), & \mathbf{z}_i \in \mathcal{B}_1^t \\ u_i^t, & \mathbf{z}_i \notin \mathcal{B}_1^t, \end{cases} \quad (5)$$

where $u_i^{t-1} - g(\mathbf{w}^t; \mathbf{z}_i, \mathcal{B}_{i,2}^t)$ is the stochastic gradient of the i-th coordinate in the objective (4). This enables to us to use the proof technique of stochastic block coordinate descent methods to build the recursion of Ξ_t and derive the improved rate compared to previous works listed in Table 1.

By combining the lemma above with Lemma 2 and Lemma 3 in the supplement, we prove the convergence to find an ϵ-stationary point, as stated in the following theorem.

Theorem 1. *Under Assumptions 1 and 2, SOX (Algorithm 1) with $\beta = O(\min\{B_1, B_2\}\epsilon^2)$, $\gamma = O(B_2\epsilon^2)$, $\eta = \min\left\{\frac{\beta}{4L_F}, \frac{\gamma B_1}{30 L_f n C_1 C_g}\right\}$ can find an ϵ-stationary point in $T = O\left(\max\left\{\frac{n}{B_1 B_2 \epsilon^4}, \frac{1}{\min\{B_1, B_2\}\epsilon^4}\right\}\right)$ iterations.*

Remark: The above theory suggests that given a budget on the mini-batch size $B_1 + B_2 = B$, the best value of B_1 is $B_1 = B/2$. We will verify this result in experiments.

3.3. Improved Rate for (Strongly) Convex Problems

In this subsection, we prove improved rates for SOX for convex and strongly convex objectives compared to previous work BSGD (Hu et al., 2020). One might directly analyze SOX with different decreasing step sizes for convex and strongly convex objectives separately as in Hu et al. (2020). However, to our knowledge, this strategy does not yield an optimal rate for strongly convex functions. To address this challenge, we provide a unified algorithmic framework for both convex and strongly convex functions and derive the improved rates. The idea is to use the stagewise framework given in Algorithm 2 to boost the convergence. Our strategy is to prove an improved rate for an objective that satisfies a μ-PL condition $\|\nabla F(\mathbf{w})\|^2 \geq \mu(F(\mathbf{w}) - F(\mathbf{w}^*))$, where $\mathbf{w}^*$ is a global minimum. Then, we use this result to derive the improved rates for (strongly) convex objectives.

Theorem 2. *Assume F satisfying the PL condition, by setting $\epsilon_k = O(1/2^{k-1})$, $\beta_k = O(\eta_k), \gamma_k = O(\frac{n\eta_k}{B_1}), T_k = O(\frac{1}{\mu\eta_k}), \eta_k = O(\min(\mu\min(B_1, B_2)\epsilon_k, \frac{\mu B_1 B_2 \epsilon_k}{n}))$, and $K = \log(1/\epsilon)$, SOX-boost ensures that $\mathbb{E}[F(\mathbf{w}^K) - F(\mathbf{w}^*)] \leq \epsilon$, which implies a total iteration complexity of $T = O(\max(\frac{n}{\mu^2 B_1 B_2 \epsilon}, \frac{1}{\mu^2 \min(B_1, B_2)\epsilon}))$.*

Specific values of the parameters in Theorem 2 can be found in Theorem 4 in the appendix. The result above directly implies the improved complexity for μ-strongly convex function, as it automatically satisfies the PL condition. For a convex function, we use a common trick to make it strongly convex by contructing $\hat{F}(\mathbf{w}) = F(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w}\|^2$, then we use SOX-boost to optimize $\hat{F}(\mathbf{w})$ with a small μ. Its convergence is summarized by the following corollary.

Corollary 1. *Assume F is convex, by setting $\mu = O(\epsilon)$, $\eta_k, \gamma_k, \beta_k, T_k$ according to Theorem 2, then after $K = \log(1/\epsilon)$-stages SOX-boost for optimizing $\hat{F}$ ensures that $\mathbb{E}[\hat{F}(\mathbf{w}^K) - \min_{\mathbf{w}} \hat{F}(\mathbf{w})] \leq \epsilon$, which implies a total iteration complexity of $T = O(\max(\frac{n}{\mu^2 B_1 B_2 \epsilon}, \frac{1}{\mu^2 \min(B_1, B_2)\epsilon}))$ for ensuring $\mathbb{E}[F(\mathbf{w}^K) - F(\mathbf{w}^*)] \leq \epsilon$.*

Remark: The above result implies an complexity of $T = O(\max(\frac{n}{B_1 B_2 \epsilon^3}, \frac{1}{\min(B_1, B_2)\epsilon^3}))$ for a convex function.

3.4. Optimal Rate for A Class of Convex Problems

In this section, we consider a class of FCCO problems on a closed, non-empty, and convex domain Ω and $d' = 1$. For simplicity, we denote $g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i)$ by $g_i(\mathbf{w})$ and its stochastic estimator $g(\mathbf{w}; \mathbf{z}_i, \xi_i)$ by $g_i(\mathbf{w}; \xi_i)$ in this section.

[5] Please refer to the comments above (27) in Wang et al. (2021) for the issue of bounding $\sum_t [\|u_{i_t}^t - g(\mathbf{w}^t; \mathbf{z}_{i_t}, \mathcal{S}_{i_t})\|^2]$.

Algorithm 2 SOX-boost($\mathbf{w}_1$, $\mathbf{u}_1$, $\mathbf{v}_1$, K)

1: **for** epochs $k = 1, \ldots, K$ **do**
2: Update $\mathbf{w}$, $\mathbf{u}$, $\mathbf{v}$ by SOX($\mathbf{w}^k$, $\mathbf{u}^k$, $\mathbf{v}^k$, η_k, β_k, γ_k, T_k)
3: Update $\eta_k, \beta_k, \gamma_k, T_k$ according to Theorem 2
4: **end for**

We additionally make the assumption below.

Assumption 3. Assume that $d' = 1$, f_i is monotonically increasing and convex, while g_i is convex. The domain Ω is bounded such that $\max_{\mathbf{w}\in\Omega} \|\mathbf{w} - \mathbf{w}_*\| \leq C_\Omega$ and $\max_{\mathbf{w}\in\Omega} \|g(\mathbf{w}; \mathbf{z}_i, \xi_i)\| \leq D_g$ for any i and r.v. ξ_i.

This FCCO can be reformulated as a saddle point problem.

$$\min_{\mathbf{w}\in\Omega} \max_{\pi_1\in\Pi_1} \max_{\pi_2\in\Pi_2} \mathcal{L}(\mathbf{w}, \pi_1, \pi_2),$$

where $\mathcal{L}(\mathbf{w}, \pi_1, \pi_2) = \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \mathcal{L}_i(\mathbf{w}, \pi_{i,1}, \pi_{i,2})$, $\pi_1 = [\pi_{1,1}, \ldots, \pi_{n,1}]^\top$, $\pi_2 = [\pi_{1,2}, \ldots, \pi_{n,2}]^\top$, $\mathcal{L}_{i,1}(\mathbf{w}, \pi_{i,1}, \pi_{i,2}) = \pi_{i,1}\mathcal{L}_{i,2}(\mathbf{w}, \pi_{i,2}) - f^*(\pi_{i,1})$ and $\mathcal{L}_{i,2}(\mathbf{w}, \pi_{i,2}) = \langle \pi_{i,2}, \mathbf{w}\rangle - g_i^*(\pi_{i,2})$. Here $f^*(\cdot)$ and $g_i^*(\cdot)$ are the convex conjugates of f and g_i, respectively. We analyze the SOX algorithm with $\beta = 1$, $\gamma = \frac{1}{1+\tau}$ ($\tau > 0$) and the projection onto Ω, which is equivalent to the following primal-dual update formula:

$$\pi_{i,2}^{t+1} = \arg\max_{\pi_{i,2}} \left\langle \pi_{i,2}, \mathbf{w}^t \right\rangle - g_i^*(\pi_{i,2}), \quad \mathbf{z}_i \in \mathcal{B}_1^t, \tag{6}$$

$$\pi_{i,1}^{t+1} = \begin{cases} \arg\max_{\pi_{i,1}} \pi_{i,1}\mathcal{L}_{i,2}(\mathbf{w}^t, \pi_{i,2}^t(\mathcal{B}_{i,2}^t)) \\ \quad -f^*(\pi_{i,1}) - \tau D_{f^*}(\pi_{i,1}^t, \pi_{i,1}), & \mathbf{z}_i \in \mathcal{B}_1^t \\ \pi_{i,1}^t, & \mathbf{z}_i \notin \mathcal{B}_1^t, \end{cases}$$

$$\mathbf{w}^{t+1} = \arg\min_{\mathbf{w}\in\Omega} \tfrac{1}{B_1}\textstyle\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \pi_{i,1}^t \pi_{i,2}^{t+1}(\mathcal{B}_{i,2}^t)\mathbf{w} + \frac{\eta}{2}\left\|\mathbf{w} - \mathbf{w}^t\right\|^2,$$

where $\pi_{i,2}(\mathcal{B}_{i,2})$ is a stochastic estimation of $\pi_{i,2}$ based on the mini-batch $\mathcal{B}_{i,2}$ and $D_f(x, y) := f(y) - f(x) - \langle \nabla f(x), y - x\rangle$ is the Bregman divergence. We define that $\mathbf{D}_{f^*}(\pi_1, \pi_1') := \sum_{\mathbf{z}_i\in\mathcal{D}} D_{f^*}(\pi_{i,1}, \pi_{i,1}')$ for any $\pi_1, \pi_1' \in \Pi_1$. Note that (6) is equivalent to $\pi_{i,2}^{t+1} = \nabla g_i(\mathbf{w}^t)$. Besides, for $\mathbf{z}_i \in \mathcal{B}_1^t$ and $\pi_{i,1}^t = \nabla f(u_i^t)$ we have

$$\begin{aligned}\pi_{i,1}^{t+1} &= \arg\min_{\pi_{i,1}} - \pi_{i,1}\mathcal{L}_{i,2}(\mathbf{w}^t, \pi_{i,2}^t(\mathcal{B}_{i,2}^t)) + f^*(\pi_{i,1}) \\ &\qquad + \tau D_{f^*}(\pi_{i,1}^t, \pi_{i,1}) \\ &= \arg\min_{\pi_{i,1}} - \pi_{i,1}\frac{g_i(\mathbf{w}^t; \mathcal{B}_{i,2}^t) + \tau f(u_i^t)}{1+\tau} + f^*(\pi_{i,1}).\end{aligned}$$

The last equation above is due to $g_i(\mathbf{w}^t; \mathcal{B}_{i,2}^t) = \mathcal{L}_{i,2}(\mathbf{w}^t, \pi_{i,2}^t(\mathcal{B}_{i,2}^t))$ and $f(u_i^t) = \pi_{i,1}^t u_i^t - f^*(\pi_{i,1}^t)$. Then, we can conclude that $\pi_{i,1}^{t+1} = \nabla f(u_i^{t+1})$ if we define $u_i^{t+1} = (1-\gamma)f(u_i^t) + \gamma g_i(\mathbf{w}^t; \mathcal{B}_{i,2}^t)$ and $\gamma = \frac{1}{1+\tau}$.

On this class of convex problems, we can establish an improved rate for SOX in the order of $O(1/\epsilon^2)$, which is optimal in terms of ϵ (but might not be optimal in terms of n). The analysis is inspired by Zhang & Lan (2020), which provide the optimal complexity for the traditional SCO problems. We extend their analysis to handle the selective sampling/update to accomodate the FCCO problem. Moreover, our analysis also gets rid of one drawback of Zhang & Lan (2020) that needs two independent batches to estimate $g_i(\mathbf{w})$ and $\nabla g_i(\mathbf{w})$, which is achieved by cancelling the highlighted terms in Lemma 10 and Lemma 11.

Theorem 3. *Assume f_i is monotone, convex, smooth and Lipschitz-continuous while g_i is convex and Lipschitz-continuous. If $\eta = O(\min(\min\{B_1, B_2\}\epsilon, \frac{B_1\epsilon}{n}))$, $\gamma = O(B_2\epsilon)$, $\beta = 1$, and $\bar{\mathbf{w}}_T = \sum_t \mathbf{w}_t/T$, SOX ensures that $\mathbb{E}[F(\bar{\mathbf{w}}_T) - F(\mathbf{w}^*)] \leq \epsilon$ after $O(\max(\frac{n}{B_1\epsilon^2}, \frac{n}{B_1B_2\epsilon^2}, \frac{1}{\min(B_1,B_2)\epsilon^2}))$ iterations.*

4. Experiments

In this section, we provide some experimental results to verify some aspects of our theory and compare SOX with other baselines for three applications: deep average precision (AP) maximization, p-norm push optimization with concentration at the top, and neighborhood component analysis (NCA).

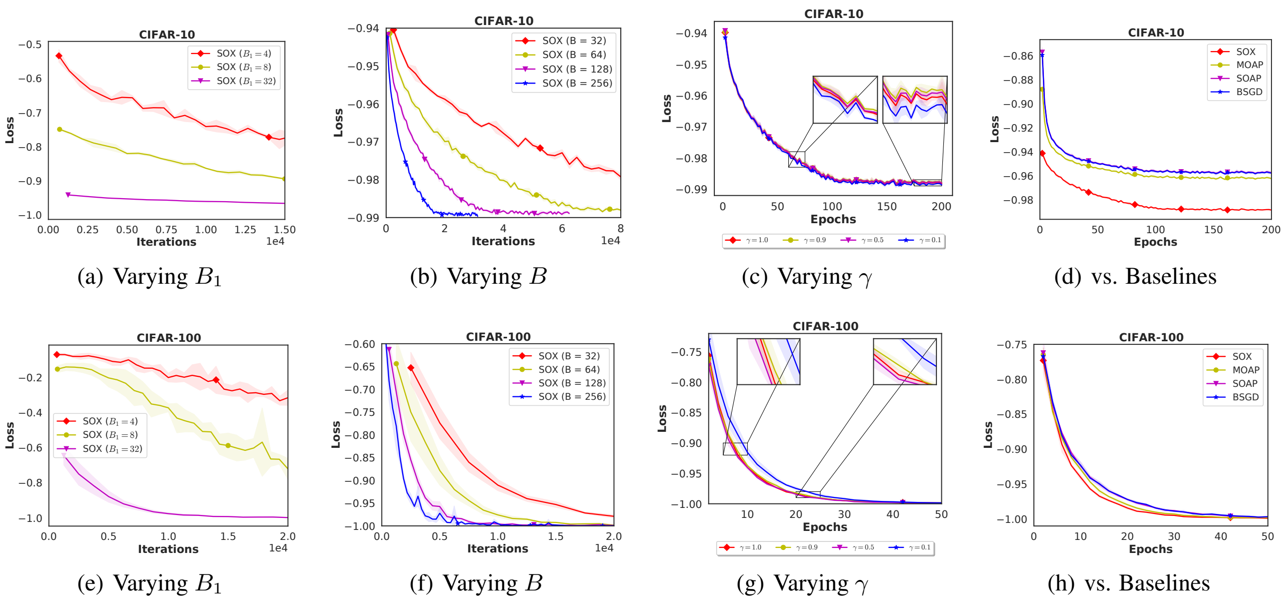
4.1. Deep AP Maximization

AP maximization in the form of FCCO has been considered in Qi et al. (2021); Wang et al. (2021). For a binary classification problem, let $\mathcal{S}_+, \mathcal{S}_-$ denote the set of positive and negative examples, respectively, $\mathcal{S} = \mathcal{S}_+ \cup \mathcal{S}_-$ denote the set of all examples. A smooth surrogate objective for maximizing AP can be formulated as:

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|}\sum_{\mathbf{x}_i\in\mathcal{S}_+} \frac{\sum_{\mathbf{x}\in\mathcal{S}_+} \ell(h_\mathbf{w}(\mathbf{x}) - h_\mathbf{w}(\mathbf{x}_i))}{\sum_{\mathbf{x}\in\mathcal{S}} \ell(h_\mathbf{w}(\mathbf{x}) - h_\mathbf{w}(\mathbf{x}_i))}, \tag{7}$$

where $\ell(\cdot)$ is a surrogate function that penalizes large input. It is a special case of FCCR by defining $g_i(\mathbf{w}) = [\sum_{\mathbf{x}\in\mathcal{S}_+} \ell(h_\mathbf{w}(\mathbf{x}) - h_\mathbf{w}(\mathbf{x}_i)), \sum_{\mathbf{x}\in\mathcal{S}} \ell(h_\mathbf{w}(\mathbf{x}) - h_\mathbf{w}(\mathbf{x}_i))]$ and $f(g_i(\mathbf{w})) = -\frac{[g_i(\mathbf{w})]_1}{[g_i(\mathbf{w})]_2}$.

Setting. We conduct experiments on two image datasets, namely CIFAR-10, CIFAR-100. We use the dataloader provided in the released code of Qi et al. (2021), which constructs the imbalanced versions of binarized CIFAR-10 and CIFAR-100. We consider two tasks: training ResNet18 on the CIFAR-10 data set and training ResNet34 on the CIFAR-100 data set. We follow the same procedure as in Qi et al. (2021) that first pre-trains the network by optimizing a cross-entropy loss and then fine-tunes all layers with the randomly initialized classification layer. We also use the same squared hinge loss as in Qi et al. (2021). We aim to answer the following four questions related to our theory: **Q1**: Given a batch size B, what is the best value for B_1, B_2, i.e., the sizes of $\mathcal{B}_1^t$ and $\mathcal{B}_2^t$? **Q2**: Is there parallel speed-up



(a) Varying B_1 (b) Varying B (c) Varying γ (d) vs. Baselines

(e) Varying B_1 (f) Varying B (g) Varying γ (h) vs. Baselines

Figure 2. Results of AP maximization on the CIFAR-10 and CIFAR-100 data sets.

by increasing the total batch size $B = B_1 + B_2$? **Q3**: What is the best value of γ? **Q4**: Does SOX converge faster than SOAP (SGD-style) and MOAP? In all experiments, we tune the initial learning rate in a range $10^{-4:1:-1}$ to achieve the best validation error, and decrease the learning rate at 50% and 75% of total epochs. The experiments are performed on a node of a cluster with single GeForce RTX 2080 Ti GPU. We tune the value of γ and fix $\beta = 0.1$ (same as the default value 0.9 of gradient momentum).

• To answer Q1, we fix the total batch size B as 64 and vary B_1 in the range $\{4, 8, 16, 32\}$. The curves of training losses are shown in Figure 2(a) and (e) on the two datasets. We can see that when $B_1 = 32 = B/2$ SOX has the fastest convergence in terms of number of iterations. This is consistent with our convergence theory.

• To answer Q2, we fix $B_1 = B_2$ and vary B in the range $\{32, 64, 128, 256\}$. The curves of training losses are shown in Figure 2(b) and (f) on the two datasets. We can see that the iteration complexity of SOX decreases as B increases, which is also consistent with our convergence theory.

• To answer Q3, we fix $B_1 = B_2 = B/2 = 32$ and run SOX with different values of γ. We can see that $\gamma = 1$ does not give the best result, which means the naïve mini-batch estimation of $g_i(\mathbf{w})$ is worse than the moving average estimator with a proper value of γ. Moreover, we also observe that the best value of γ depends on the task: $\gamma = 0.1, 0.5$ give the fastest convergence on training ResNet18 with CIFAR-10 and ResNet34 with CIFAR-100, respectively.

• The Figure 2 (d) and (h) answer Q4, which indicates that SOX converges faster than MOAP, which is faster than SOAP (SGD-style) and BSGD. The curves of average precision on the training data can be found in Figure 4 of the Appendix. We also report the test AP of SOX with baselines on CIFAR-10 and CIFAR-100 datasets in Table 2. Note that the CIFAR-10 and CIFAR-100 test datasets are *balanced* while our training datasets are imbalanced. Thus, there might be a distribution shift between the training and test datasets. To prevent overfitting, algorithms are early stopped when the validation loss reaches the minimum. The results indicate that SOX converges to a better solution using an overall fewer number of epochs.

Table 2. Test AP comparison among SOX and the baselines on the AP maximization task.

Dataset: CIFAR-10				
Metrics	MOAP	BSGD	SOAP	SOX
Test AP (↑)	0.763 ± 0.001	0.762 ± 0.001	0.762 ± 0.001	$\mathbf{0.765 \pm 0.001}$
#Epoch (↓)	13.0 ± 4.3	15.7 ± 1.9	15.7 ± 1.9	$\mathbf{7.0 \pm 3.3}$
Dataset: CIFAR-100				
Metrics	MOAP	BSGD	SOAP	SOX
Test AP (↑)	0.584 ± 0.010	0.582 ± 0.005	0.575 ± 0.017	$\mathbf{0.597 \pm 0.012}$
#Epoch (↓)	17.0 ± 1.6	$\mathbf{3.7 \pm 0.9}$	11.7 ± 6.6	5.0 ± 2.8

4.2. p-norm Push with Concentration at the Top

In the bipartite ranking problem, the p-norm push objective (Rudin, 2009) can be defined as

$$F(\mathbf{w}) = \frac{1}{|\mathcal{S}_-|} \sum_{\mathbf{z}_i \in \mathcal{S}_-} \left(\frac{1}{|\mathcal{S}_+|} \sum_{\mathbf{z}_j \in \mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{z}_j) - h_{\mathbf{w}}(\mathbf{z}_i)) \right)^p ,$$

where $p > 1$ and $\ell(\cdot)$ is similar as above. We can cast this function into FCCR by defining $\mathcal{D} = \mathcal{S}_+$, $\mathcal{S}_i = \mathcal{S}_-$, $g_i(\mathbf{w}) = \frac{1}{|\mathcal{S}_+|}\sum_{\mathbf{z}_j\in\mathcal{S}_+}\ell(h_\mathbf{w}(\mathbf{z}_j) - h_\mathbf{w}(\mathbf{z}_i))$ that couples each positive example $\mathbf{z}_i$ with all negative samples, $f(g) = g^p$. Note that f is monotonically increasing and convex while g_i is convex given that ℓ is convex. Rudin (2009) only provide a boosting-style p-norm push algorithm (BS-PnP), which is not scalable because it processes all $|\mathcal{S}_+|$ positive and $|\mathcal{S}_-|$ negative instances at each iteration. We compare SOX with the BS-PnP, and the baselines BSGD (Hu et al., 2020). Besides, SOAP (Qi et al., 2021) and MOAP (Wang et al., 2021), which were originally designed for the AP maximization, can also be applied to the p-norm push problem [6]. Following Rudin (2009), we choose $\ell(\cdot)$ to be the exponential function. We conduct our experiment on two LibSVM datasets: covtype and ijcnn1. For both datasets, we randomly choose 90% of the data for training and the rest of data is for testing. For this experiment, we learn a linear ranking function $h_\mathbf{w}(\mathbf{x}) = \langle\mathbf{w},\mathbf{x}\rangle$ and $p = 4$. For each algorithm, we run it with 5 different random seeds and report the average test loss with standard deviation. Besides, we also report the running time. For the stochastic algorithms (BSGD, SOAP, MOAP, SOX), we choose $B = 64$ and $B_1 = B_2$. The algorithms are implemented with Python and run on a server with 12-core Intel(R) Xeon(R) CPU E5-2697 v2 @ 2.70GHz.

Table 3. Comparison among SOX and the baselines BS-PnP, BSGD for optimizing p-norm Push for learning a linear model.

covtype			
Algorithms	BS-PnP	BSGD	SOX
Test Loss (↓)	0.778	0.625 ± 0.018	**0.516 ± 0.003**
Time (s) (↓)	6043.90	**4.20 ± 0.08**	4.62 ± 0.10
ijcnn1			
Algorithms	BS-PnP	BSGD	SOX
Test Loss (↓)	0.268	0.202 ± 0.001	**0.128 ± 0.002**
Time (s) (↓)	648.06	**4.02 ± 0.04**	4.15 ± 0.06

As shown in Table 3, the BS-PnP algorithm is indeed not scalable and takes much longer time than the stochastic algorithms. Moreover, our SOX is consistently better than BSGD in terms of test loss.

4.3. Neighborhood Component Analysis

Neighborhood Component Analysis (NCA) was proposed in Goldberger et al. (2004) for learning a Mahalanobis distance measure. Given a set of data points $\mathcal{D} = \{\mathbf{x}_1,\dots,\mathbf{x}_n\}$, where each data $\mathbf{x}_i$ has a class label y_i. The objective of NCA is defined as

$$F(A) = -\sum_{\mathbf{x}_i\in\mathcal{D}} \frac{\sum_{\mathbf{x}\in\mathcal{C}_i}\exp(-\|A\mathbf{x}_i - A\mathbf{x}\|^2)}{\sum_{\mathbf{x}\in\mathcal{S}_i}\exp(-\|A\mathbf{x}_i - A\mathbf{x}\|^2)}, \quad (8)$$

where $\mathcal{C}_i = \{\mathbf{x}_j \in \mathcal{D} : y_j = y_i\}$ and $\mathcal{S}_i = \mathcal{D}\setminus\{\mathbf{x}_i\}$. We can map the above objective as an FCCR by defining $g_i(A) = [\sum_{\mathbf{x}\in\mathcal{C}_i}\exp(-\|A\mathbf{x}_i - A\mathbf{x}\|^2), \sum_{\mathbf{x}\in\mathcal{S}_i}\exp(-\|A\mathbf{x}_i - A\mathbf{x}\|^2)]$ and $f(g_i(A)) = -\frac{[g_i(A)]_1}{[g_i(A)]_2}$. The problem (8) can be solved by the gradient descent method. However, the exact gradient computation could be expensive or even infeasible when $|\mathcal{D}|$ is large. A widely used stochastic algorithm is to sample a mini-batch $\mathcal{B}\subseteq\mathcal{D}$ and replace $\mathcal{C}_i$ and $\mathcal{S}_i$ by $\mathcal{C}_i\cap\mathcal{B}$ and $\mathcal{S}_i\cap\mathcal{B}$, respectively, which is equivalent to the BSGD algorithm (Hu et al., 2020). Besides, SOAP (Qi et al., 2021), MOAP (Wang et al., 2021) and our SOX algorithm are also applicable to (8).

The experiment is performed on three datasets: sensorless, usps, and mnist from the LibSVM (Chang & Lin, 2011). For each dataset, we randomly choose 90% of the data for training and the rest as test data. Each algorithm is executed for 5 runs with different random seeds. We report the average test loss with standard deviation. For all algorithms, we choose batch size to be 64. As shown in Figure 3, our SOX method outperforms previous methods on those datasets.

5. More Applications of SOX

In this section, we present more applications of the proposed algorithm in machine learning, and highlight the potential of the proposed algorithm in addressing their computational challenges. Providing experimental results of these applications is beyond the scope of this paper.

Listwise Ranking Objectives/Measures. In learning to rank (LTR), we are given a set of queries $\mathcal{Q} = \{\mathbf{q}_1,\dots,\mathbf{q}_n\}$. For each query, a set of items with relevance scores are provided $\mathcal{S}_q = \{(\mathbf{x}_1^q, y_1^q),\dots,(\mathbf{x}_{n_q}^q, y_{n_q}^q)\}$, where $\mathbf{x}_i^q$ denotes the input data, and $y_i^q\in\mathbb{R}^+$ denotes its a relevance score with $y_i^q = 0$ meaning irrelevant. For LTR, there are many listwise objectives and measures that can be formulated as FCCR, e.g., ListNet (Cao et al., 2007), ListMLE (Xia et al., 2008) , NDCG (Wang et al., 2013). Due to the limited space, we only consider that of ListNet. The objective function of ListNet can be defined by a cross-entropy loss between two probabilities of list of scores:

$$F(\mathbf{w}) = -\sum_q\sum_{\mathbf{x}_i^q\in\mathcal{S}_q} P(y_i^q)\log\frac{\exp(h_\mathbf{w}(\mathbf{x}_i^q;\mathbf{q}))}{\sum_{\mathbf{x}\in\mathcal{S}_q}\exp(h_\mathbf{w}(\mathbf{x};\mathbf{q}))},$$

where $h_\mathbf{w}(\mathbf{x}_i^q;\mathbf{q})$ denotes the prediction score of the item $\mathbf{x}_i^q$ with respect to the query $\mathbf{q}$, $P(y_i^q)$ denotes a probability for a relevance score y_i^q (e.g., $P(y_i^q)\propto y_i^q$). We can map the above function into FCCR, where

[6] Due to limited space, the comparison with SOAP and MOAP can be found in Table 4 of the Appendix.

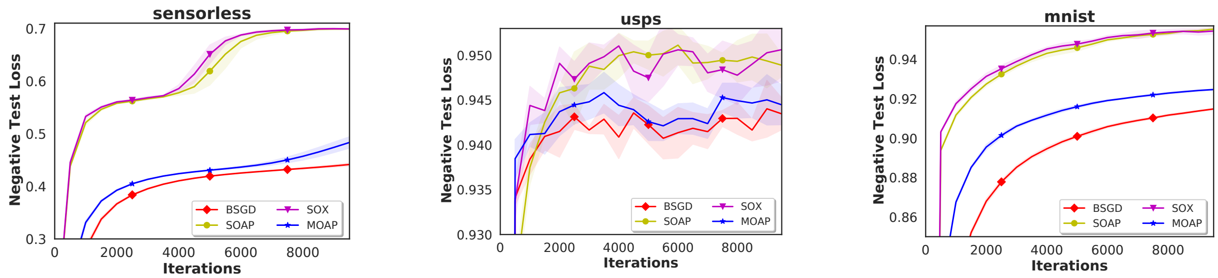


Figure 3. Results of neighborhood component analysis on three datasets.

$g(\mathbf{w}; \mathbf{x}_i^q, \mathcal{S}_q) = \frac{1}{|\mathcal{S}_q|}\sum_{\mathbf{x}\in\mathcal{S}_q}\exp(h_{\mathbf{w}}(\mathbf{x};\mathbf{q}) - h_{\mathbf{w}}(\mathbf{x}_i^q;\mathbf{q}))$ and $f(g) = \log(g)$, $\mathcal{D} = \{(\mathbf{q}, \mathbf{x}_i^q) : P(y_i^q) > 0\}$. The original paper of ListNet uses a gradient method for optimizing the above objective, which has a complexity of $O(|\mathcal{Q}||\mathcal{S}_q|)$ and is inefficient when $\mathcal{S}_q$ contains a large number of items.

Deep Survival Analysis (DSA). The survival analysis in medicine is to explore and understand the relationships between patients' covariates (e.g., clinical and genetic features) and the effectiveness of various treatment options. Using the Cox model for modeling the hazard function, the negative log-likelihood can be written as (Katzman et al., 2018):

$$F(\mathbf{w}) = \frac{1}{n}\sum_{i:E_i=1}\log\left(\sum_{j\in\mathcal{S}(T_i)}\exp(h_{\mathbf{w}}(\mathbf{x}_j) - h_{\mathbf{w}}(\mathbf{x}_i)\right),$$

where $\mathbf{x}_i$ denote the input feature of a patient, $h_{\mathbf{w}}(\mathbf{x}_i)$ denotes the risk value predicted by the network, $E_i = 1$ denotes an observable event of interest (e.g., death), T_i denotes the time interval between the time in which the baseline data was collected and the time of the event occurring, and $\mathcal{S}(t) = \{i : T_i \geq t\}$ denotes the set of patients still at risk of failure at time t. This is similar to the objective of ListMLE. The proposed algorithm is appropriate when both $\{i : E_i = 1\}$ and $\mathcal{S}(T_i)$ are large.

Deep Latent Variable Models (DLVM). Latent variable models refer to a family of generative models that use latent variables to model the observed data, where we consider the supervised learning setting. In particular, given a set of observed data $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, we model the probability of $\Pr(y|\mathbf{x})$ by introducing a discrete latent variable $\mathbf{z}$, i.e., $\Pr(y|\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}}\Pr(y|\mathbf{x},\mathbf{z})\Pr(\mathbf{z}|\mathbf{x})$, where $\mathcal{Z}$ denotes the support set of the latent variable $\mathbf{z}$ and both $\Pr(y|\mathbf{x},\mathbf{z})$ and $\Pr(\mathbf{z}|\mathbf{x})$ could be parameterized by a deep neural network. Then by minimizing negative log-likelihood of observed data, we have the objective function $F(\mathbf{w}) = -\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}}\log\sum_{\mathbf{z}\in\mathcal{Z}}\Pr(y_i|\mathbf{z},\mathbf{x}_i)\Pr(\mathbf{z}|\mathbf{x}_i)$. When $\mathcal{Z}$ is a large set, evaluating the inner sum is expensive. While the above problem is traditionally solved by EM-type algorithms, a stochastic algorithm based on backpropogation is used more often in modern deep learning. We consider an application in NLP for retrieve-and-predict language model pre-training (Guu et al., 2020). In particular, $\mathbf{x}_i$ denotes an masked input sentence, y denotes masked tokens, $\mathbf{z}$ denotes a document from a large corpus $\mathcal{Z}$ (e.g., wikipedia). In Guu et al. (2020), $\Pr(\mathbf{z}|\mathbf{x}_i) = \frac{\exp(E(\mathbf{x})^\top E(\mathbf{z}))}{\sum_{\mathbf{z}'\in\mathcal{Z}}\exp(E(\mathbf{x})^\top E(\mathbf{z}'))}$, where $E(\cdot)$ is a document embedding network, and $\Pr(y|\mathbf{x},\mathbf{z})$ is computed by a masked language model that a joint embedding $\mathbf{x}, \mathbf{z}$ is used to make the prediction. Hence, we can write $F(\mathbf{w})$ as

$$F(\mathbf{w}) = -\sum_{i=1}^n\log\sum_{\mathbf{z}\in\mathcal{Z}}\Pr(y_i|\mathbf{z},\mathbf{x}_i)\exp(E(\mathbf{x}_i)^\top E(\mathbf{z})))$$
$$+\sum_{i=1}^n\log(\sum_{\mathbf{z}'\in\mathcal{Z}}\exp(E(\mathbf{x}_i)^\top E(\mathbf{z}'))).$$

Note that both terms in the above is a special case of FCCR. The proposed algorithm gives an efficient way to solve this problem when $\mathcal{Z}$ is very large. Guu et al. (2020) address the challenge by approximating the inner summation by summing over the top k documents with highest probability under $\Pr(\mathbf{z}|\mathbf{x})$, which is retrieved by using maximum inner product search with a running time and storage space that scale sub-linearly with the number of documents. In contrast, SOX has a complexity independent of the number of documents per-iteration, which depends on the batch size.

Softmax Functions. One might notice that in the considered problems ListNet, ListMLE, NCA, DSA, DLVM, a common function that causes the difficulty in optimization is the softmax function in the form $\frac{\exp(h(\mathbf{x}_i))}{\sum_{\mathbf{x}\in\mathcal{X}}\exp(h(\mathbf{x}))}$ for a target item $\mathbf{x}_i$ out of a large number items in $\mathcal{X}$. This also occurs in NLP pre-training methods that predicts masked tokens out of billions/trillions of tokens (Borgeaud et al., 2022). Taking the logarithmic of the softmax function gives the coupled compositional form $\log\sum_{\mathbf{x}'\in\mathcal{X}}\exp(h(\mathbf{x})-h(\mathbf{x}_i))$, and summing over all items gives the considered FCCR.

Acknowledgements

We thank anonymous reviewers and Yao Yao (UIowa) for spotting several mistakes in the proof and Gang Li (UIowa) for discussing the experiments on p-norm push optimization. This work is partially supported by NSF Grant 2110545, NSF Career Award 1844403.

References

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J. W., Elsen, E., and Sifre, L. Improving language models by retrieving from trillions of tokens, 2022. (Cited on page 9)

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pp. 129–136, 2007. (Cited on pages 2 and 8)

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *TIST*, 2(3):27, 2011. (Cited on page 8)

Dang, C. D. and Lan, G. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881, 2015. (Cited on page 15)

Ghadimi, S., Ruszczy'nski, A., and Wang, M. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM J. Optim.*, 30:960–979, 2020. (Cited on pages 2, 4, and 5)

Goldberger, J., Hinton, G. E., Roweis, S., and Salakhutdinov, R. R. Neighbourhood components analysis. *Advances in neural information processing systems*, 17, 2004. (Cited on pages 2 and 8)

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL http://proceedings.mlr.press/v119/guu20a.html. (Cited on pages 2 and 9)

Hu, Y., Zhang, S., Chen, X., and He, N. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. *Advances in Neural Information Processing Systems*, 33, 2020. (Cited on pages 1, 2, 3, 5, and 8)

Juditsky, A., Nemirovski, A., Tauvel, C., et al. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011. (Cited on page 21)

Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology volume*, 18, 06 2018. (Cited on pages 2 and 9)

Khirirat, S., Feyzmahdavian, H. R., and Johansson, M. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018. (Cited on page 4)

Li, Z., Bao, H., Zhang, X., and Richtárik, P. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pp. 6286–6295. PMLR, 2021. (Cited on page 13)

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. (Cited on page 4)

Qi, Q., Luo, Y., Xu, Z., Ji, S., and Yang, T. Stochastic optimization of areas under precision-recall curves with provable convergence. *Advances in Neural Information Processing Systems*, 34, 2021. (Cited on pages 2, 3, 4, 5, 6, and 8)

Rudin, C. 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. (Cited on pages 2, 7, and 8)

Wang, G., Yang, M., Zhang, L., and Yang, T. Momentum accelerates the convergence of stochastic auprc maximization. *arXiv preprint arXiv:2107.01173*, 2021. (Cited on pages 3, 4, 5, 6, and 8)

Wang, M., Fang, E. X., and Liu, H. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017. (Cited on pages 2 and 5)

Wang, Y., Wang, L., Li, Y., He, D., Chen, W., and Liu, T.-Y. A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, pp. 6. Citeseer, 2013. (Cited on page 8)

Wangni, J., Wang, J., Liu, J., and Zhang, T. Gradient sparsification for communication-efficient distributed optimization. *Advances in Neural Information Processing Systems*, 31, 2018. (Cited on page 4)

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pp. 1192–1199, 2008. (Cited on pages 2 and 8)

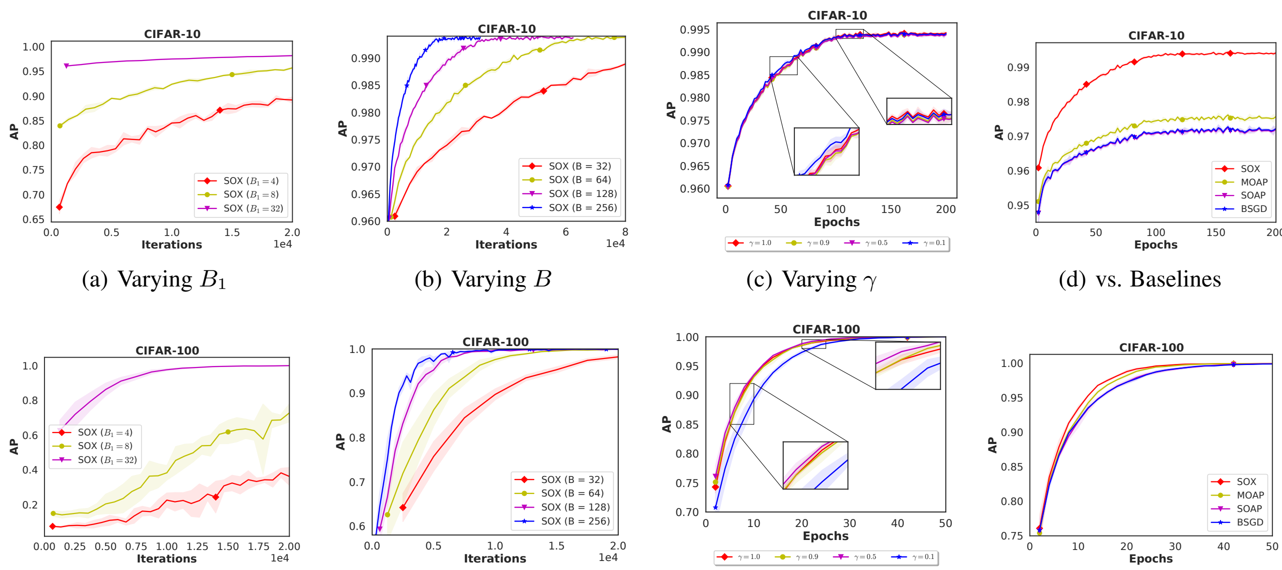
Zhang, J. and Xiao, L. Multilevel composite stochastic optimization via nested variance reduction. *SIAM Journal on Optimization*, 31(2):1131–1157, 2021. (Cited on page 4)

Zhang, Z. and Lan, G. Optimal algorithms for convex nested stochastic composite optimization. *arXiv preprint arXiv:2011.10076*, 2020. (Cited on pages 2, 3, 6, 21, 22, and 25)

A. Omitted Experimental Results

A.1. AP Maximization

We provide the curves of training loss in Figure 2. Here we also present the curves of training average precision.



(a) Varying B_1 (b) Varying B (c) Varying γ (d) vs. Baselines

(e) Varying B_1 (f) Varying B (g) Varying γ (h) vs. Baselines

Figure 4. Training average precision curves of AP maximization on the CIFAR-10 and CIFAR-100 data sets.

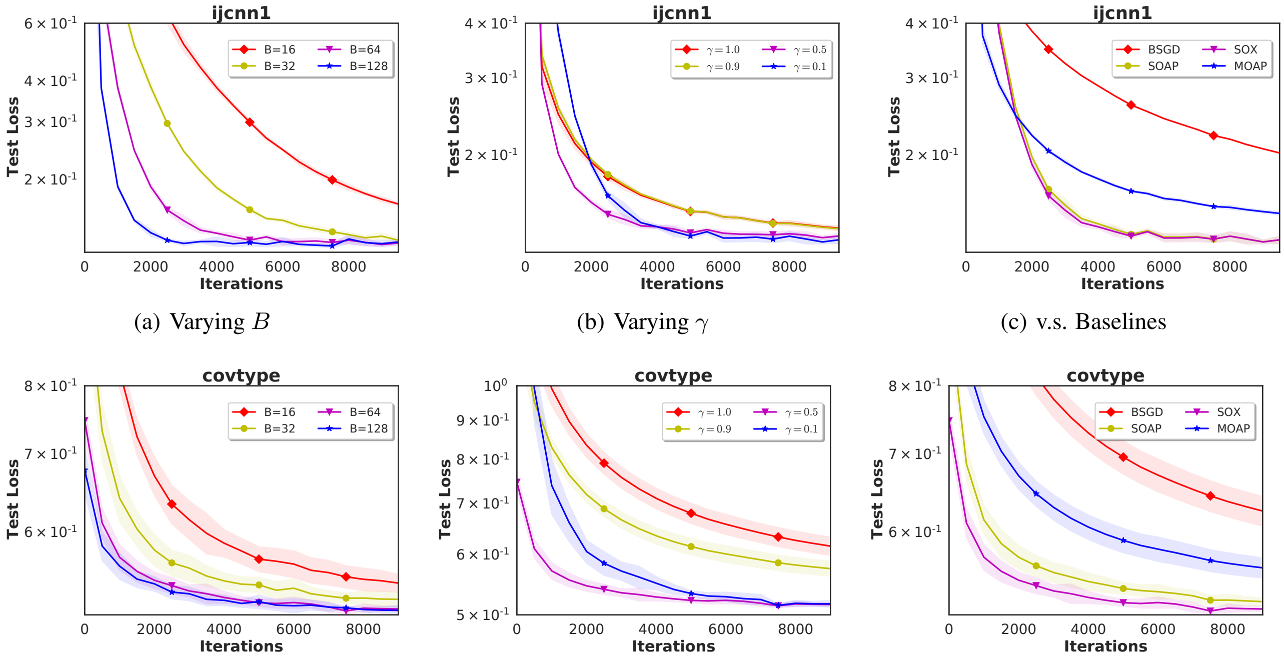
A.2. Minimizing p-norm Push

Table 4 and Figure 5 (c)&(f) show that SOX consistently outperforms BS-PnP/BSGD/MOAP in terms of p-norm push loss on the test data. SOX has better performance than SOAP on the covtype data while match its performance on ijcnn1.

Table 4. Comparison among SOX and the baselines BS-PnP, BSGD for optimizing p-norm Push for learning a linear model.

covtype					
Algorithms	BS-PnP	BSGD	SOAP	MOAP	SOX
Test Loss (↓)	0.778	0.625 ± 0.018	0.523 ± 0.004	0.559 ± 0.011	$\mathbf{0.516 \pm 0.003}$
Time (s) (↓)	6043.90	$\mathbf{4.20 \pm 0.08}$	4.32 ± 0.15	4.89 ± 0.06	4.62 ± 0.10
ijcnn1					
Algorithms	BS-PnP	BSGD	SOAP	MOAP	SOX
Test Loss (↓)	0.268	0.202 ± 0.001	$\mathbf{0.128 \pm 0.002}$	0.147 ± 0.001	$\mathbf{0.128 \pm 0.002}$
Time (s) (↓)	648.06	$\mathbf{4.02 \pm 0.04}$	4.04 ± 0.11	4.42 ± 0.05	4.15 ± 0.06

Besides, we also empirically verify other aspects of the theory. In Figure 5 (a)&(d), we show that the iteration complexity of SOX decreases when B increases; In Figure 5 (b)&(e), we show the effect of moving-average based estimation: $\gamma = 1$ is not the best choice.



(a) Varying B (b) Varying γ (c) v.s. Baselines

(d) Varying B (e) Varying γ (f) v.s. Baselines

Figure 5. Test loss curves of the p-norm push optimization task.

B. Proof of Theorem 1

Lemma 2 (Lemma 2 in Li et al. 2021)**.** *Consider a sequence $\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \mathbf{v}^t$ and the L_F-smooth function F and the step size $\eta L_F \leq 1/2$.*

$$F(\mathbf{w}^{t+1}) \leq F(\mathbf{w}^t) + \frac{\eta}{2}\Delta^t - \frac{\eta}{2}\left\|\nabla F(\mathbf{w}^t)\right\|^2 - \frac{\eta}{4}\left\|\mathbf{v}^t\right\|^2, \tag{9}$$

where $\Delta^t := \|\mathbf{v}^t - \nabla F(\mathbf{w}^t)\|^2$.

We build a recursion for the gradient variance $\Delta^t := \|\mathbf{v}^t - \nabla F(\mathbf{w}^t)\|^2$ by proving the following lemma.

Lemma 3. *If $\beta \leq \frac{2}{7}$, the gradient variance $\Delta^t := \|\mathbf{v}^t - \nabla F(\mathbf{w}^t)\|^2$ can be bounded as*

$$\mathbb{E}\left[\Delta^{t+1}\right] \leq (1-\beta)\mathbb{E}\left[\Delta^t\right] + \frac{2L_F^2\eta^2\mathbb{E}\left[\|\mathbf{v}^t\|^2\right]}{\beta} + \frac{3L_f^2C_1^2}{n}\mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\|u_i^{t+1}-u_i^t\|^2\right] + \frac{2\beta^2C_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}} + 5\beta L_f^2C_1^2\mathbb{E}\left[\Xi_{t+1}\right], \tag{10}$$

where $\Xi_t = \|\mathbf{u}^t - \mathbf{g}(\mathbf{w}^t;\mathcal{S})\|^2$, $\mathbf{u}^t = [u_1^t, \ldots, u_n^t]^\top$, $\mathbf{g}(\mathbf{w}^t;\mathcal{S}) = [g(\mathbf{w}^t;\mathbf{z}_1,\mathcal{S}_1), \ldots, g(\mathbf{w}^t;\mathbf{z}_n,\mathcal{S}_n)]^\top$.

Proof. We define that $\Delta^t := \left\|\mathbf{v}^t - \nabla F(\mathbf{w}^t)\right\|^2$ and $G(\mathbf{w}^{t+1}) = \frac{1}{B_1}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\nabla g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{B}_{i,2}^t)\nabla f(u_i^t)$. Based on the rule of update $\mathbf{v}^{t+1} = (1-\beta)\mathbf{v}^t + \beta G(\mathbf{w}^{t+1})$, we have

$$\begin{aligned}\Delta^{t+1} &= \left\|\mathbf{v}^{t+1} - \nabla F(\mathbf{w}^{t+1})\right\|^2 = \left\|(1-\beta)\mathbf{v}^t + \beta\frac{1}{B_1}\sum_{\mathbf{z}_i\in\mathcal{B}_1^{t+1}}\nabla g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{B}_{i,2}^{t+1})\nabla f(u_i^t) - \nabla F(\mathbf{w}^{t+1})\right\|^2\\ &= \left\|① + ② + ③ + ④\right\|^2,\end{aligned}$$

where ①, ②, ③, ④ are defined as

$$
\begin{aligned}
&① = (1-\beta)(\mathbf{v}^t - \nabla F(\mathbf{w}^t)), \quad ② = (1-\beta)(\nabla F(\mathbf{w}^t) - \nabla F(\mathbf{w}^{t+1})),\\
&③ = \beta \frac{1}{B_1} \sum_{\mathbf{z}_i \in \mathcal{B}_1^{t+1}} \left(\nabla g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{i,2}^{t+1}) \nabla f(u_i^t) - \nabla g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{i,2}^{t+1})\right) \nabla f(g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)),\\
&④ = \beta \left(\frac{1}{B_1} \sum_{\mathbf{z}_i \in \mathcal{B}_1^{t+1}} \nabla g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{i,2}^{t+1}) \nabla f(g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)) - \nabla F(\mathbf{w}^{t+1}) \right).
\end{aligned}
$$

Note that $\mathbb{E}_t[\langle ①, ④\rangle] = \mathbb{E}_t[\langle ②, ④\rangle] = 0$. Then, the Young's inequality for products implies that

$$
\begin{aligned}
&\mathbb{E}_t\left[\|① + ② + ③ + ④\|^2\right]\\
&= \|①\|^2 + \|②\|^2 + \mathbb{E}_t\left[\|③\|^2\right] + \mathbb{E}_t\left[\|④\|^2\right] + 2\langle ①, ②\rangle + 2\mathbb{E}_t[\langle ①, ③\rangle] + 2\mathbb{E}_t[\langle ②, ③\rangle] + 2\mathbb{E}_t[\langle ③, ④\rangle]\\
&\leq (1+\beta)\|①\|^2 + 2\left(1+\frac{1}{\beta}\right)\|②\|^2 + \frac{2+3\beta}{\beta}\mathbb{E}_t\left[\|③\|^2\right] + 2\mathbb{E}_t\left[\|④\|^2\right].
\end{aligned}
$$

Besides, we have

$$
\begin{aligned}
(1+\beta)\|①\|^2 &= (1+\beta)(1-\beta)^2\left\|\mathbf{v}^t - \nabla F(\mathbf{w}^t)\right\|^2 \leq (1-\beta)\left\|\mathbf{v}^t - \nabla F(\mathbf{w}^t)\right\|^2,\\
2\left(1+\frac{1}{\beta}\right)\|②\|^2 &= 2\left(1+\frac{1}{\beta}\right)(1-\beta)^2\left\|\nabla F(\mathbf{w}^t) - \nabla F(\mathbf{w}^{t+1})\right\|^2 \leq \frac{2L_F^2\eta^2}{\beta}\left\|\mathbf{v}^t\right\|^2,\\
\frac{2+3\beta}{\beta}\|③\|^2 &= \frac{2+3\beta}{\beta}\frac{\beta^2}{B_1}\sum_{\mathbf{z}_i \in \mathcal{B}_1 t+1}\left\|\nabla g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{i,2}^{t+1})\right\|^2\left\|\nabla f(u_i^t) - \nabla f(g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i))\right\|^2\\
&\leq \frac{(2+3\beta)\beta L_f^2}{B_1}\sum_{\mathbf{z}_i \in \mathcal{B}_1^{t+1}}\left\|\nabla g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{i,2}^{t+1})\right\|^2\left\|u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2.
\end{aligned}
$$

Consider that $\mathbf{w}^{t+1}$ and u_i^t do not depend on either $\mathcal{B}_1^{t+1}$ or $\mathcal{B}_{i,2}^{t+1}$.

$$
\begin{aligned}
&(2+3\beta)\beta L_f^2\mathbb{E}_t\left[\frac{1}{B_1}\sum_{\mathbf{z}_i \in \mathcal{B}_1^{t+1}}\left\|\nabla g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{i,2}^{t+1})\right\|^2\left\|u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2\right]\\
&= (2+3\beta)\beta L_f^2\mathbb{E}_t\left[\frac{1}{B_1}\sum_{\mathbf{z}_i \in \mathcal{B}_1^{t+1}}\mathbb{E}_t\left[\left\|\nabla g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{i,2}^{t+1})\right\|^2 \mid \mathbf{z}_i \in \mathcal{B}_1^{t+1}\right]\left\|u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2\right]\\
&\leq (2+3\beta)\beta L_f^2 C_1^2\mathbb{E}_t\left[\frac{1}{B_1}\sum_{\mathbf{z}_i \in \mathcal{B}_1^{t+1}}\left\|u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2\right]\\
&\leq \tfrac{(2+3\beta)\beta(1+\delta)L_f^2C_1^2}{n}\textstyle\sum_{\mathbf{z}_i \in \mathcal{D}}\mathbb{E}_t\left[\left\|u_i^{t+1} - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2\right] + \tfrac{(2+3\beta)\beta(1+1/\delta)L_f^2C_1^2}{n}\mathbb{E}_t\left[\textstyle\sum_{\mathbf{z}_i \in \mathcal{D}}\left\|u_i^{t+1} - u_i^t\right\|^2\right],
\end{aligned}
$$

where $C_1^2 := C_g^2 + \zeta^2/B$ and $\delta > 0$ is a constant to be determined later. Note that we have $u_i^{t+1} = u_i^t$ for all $i \notin \mathcal{B}_1^t$.

$$
\begin{aligned}
&(2+3\beta)\beta L_f^2\mathbb{E}\left[\frac{1}{B_1}\sum_{\mathbf{z}_i \in \mathcal{B}_1^t}\left\|\nabla g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{i,2}^t)\right\|^2\left\|u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2\right]\\
&\leq (2+3\beta)\beta\left((1+\delta)L_f^2C_1^2\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{D}}\left\|u_i^{t+1} - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2\right] + (1+1/\delta)L_f^2C_1^2\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{B}_1^t}\left\|u_i^{t+1} - u_i^t\right\|^2\right]\right).
\end{aligned}
$$

If $\beta \leq \frac{2}{7}$ and $\delta = \frac{3\beta}{2}$, we have $(2+3\beta)\beta(1+\delta) \leq 5\beta$ and $(2+3\beta)\beta(1+1/\delta) \leq 3$.

$$\mathbb{E}\left[\frac{2+3\beta}{\beta}\left\|③\right\|^2\right] \leq 5\beta L_f^2 C_1^2 \mathbb{E}\left[\Xi_{t+1}\right] + \frac{3L_f^2 C_1^2}{n}\mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1}-u_i^t\right\|^2\right].$$

Next, we upper bound the term $\mathbb{E}_t\left[\left\|④\right\|^2\right]$.

$$\begin{aligned}
\mathbb{E}_t\left[\left\|④\right\|^2\right] &= \beta^2\mathbb{E}_t\left[\left\|\tfrac{1}{B_1}\textstyle\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\nabla g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{B}_{i,2}^t)\nabla f(g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i)) - \tfrac{1}{n}\textstyle\sum_{\mathbf{z}_i\in\mathcal{D}}\nabla g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i)\nabla f(g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i))\right\|^2\right]\\
&= \beta^2\mathbb{E}_t\left[\left\|\frac{1}{B_1}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\nabla g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{B}_{i,2}^t)\nabla f(g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i)) - \frac{1}{B_1}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\nabla g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i)\nabla f(g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i))\right\|^2\right]\\
&\quad + \beta^2\mathbb{E}_t\left[\left\|\frac{1}{B_1}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\nabla g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i)\nabla f(g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i)) - \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\nabla g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i)\nabla f(g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{S}_i))\right\|^2\right]\\
&\leq \frac{\beta^2 C_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}}.
\end{aligned}$$

□

B.1. Proof of Lemma 1

Based on Algorithm 1, the update rule of u_i is

$$u_i^{t+1} = \begin{cases}(1-\gamma)u_i^t + \gamma g(\mathbf{w}_{t+1};\mathbf{z}_i,\mathcal{B}_{i,2}^t) & \mathbf{z}_i\in\mathcal{B}_1^t\\ u_i^t & \mathbf{z}_i\notin\mathcal{B}_1^t.\end{cases}$$

We can re-write it into the equivalent expression below.

$$u_i^{t+1} = \begin{cases}u_i^t - \gamma\left(u_i^t - g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{B}_{i,2}^t)\right) & \mathbf{z}_i\in\mathcal{B}_1^t\\ u_i^t & \mathbf{z}_i\notin\mathcal{B}_1^t.\end{cases}$$

Let us define $\phi_t(\mathbf{u}) = \frac{1}{2}\left\|\mathbf{u}-\mathbf{g}(\mathbf{w}^t)\right\|^2 = \frac{1}{2}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\|u_i - g(\mathbf{w}^t;\mathbf{z}_i,\mathcal{S}_i)\right\|^2$, which is a 1-strongly convex function. Then, the update rule (5) can be viewed as one step of the stochastic block coordinate descent algorithm (Algorithm 2 in Dang & Lan 2015) for minimizing $\phi_{t+1}(\mathbf{u})$, where the Bregman divergence is associated with the quadratic function. We follow the analysis of Dang & Lan (2015).

$$\begin{aligned}
\phi_{t+1}(\mathbf{u}^{t+1}) &= \frac{1}{2}\left\|\mathbf{u}^{t+1}-\mathbf{g}(\mathbf{w}^{t+1})\right\|^2 = \frac{1}{2}\left\|\mathbf{u}^t-\mathbf{g}(\mathbf{w}^{t+1})\right\|^2 + \left\langle \mathbf{u}^t - \mathbf{g}(\mathbf{w}^{t+1}), \mathbf{u}^{t+1}-\mathbf{u}^t\right\rangle + \frac{1}{2}\left\|\mathbf{u}^{t+1}-\mathbf{u}^t\right\|^2\\
&= \frac{1}{2}\left\|\mathbf{u}^t-\mathbf{g}(\mathbf{w}^{t+1})\right\|^2 + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\langle u_i^t - g(\mathbf{w}^{t+1};\mathbf{z}_i,\mathcal{B}_{i,2}^t), u_i^{t+1}-u_i^t\right\rangle + \frac{1}{2}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1}-u_i^t\right\|^2\\
&\quad + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\langle g(\mathbf{w}_{t+1};\mathbf{z}_i,\mathcal{B}_{i,2}^t) - g(\mathbf{w}_{t+1};\mathbf{z}_i,\mathcal{S}_i), u_i^{t+1}-u_i^t\right\rangle.
\end{aligned}$$

Note that $u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t) = (u_i^t - u_i^{t+1})/\gamma$ and $2\langle b-a, a-c\rangle \leq \|b-c\|^2 - \|a-b\|^2 - \|a-c\|^2$.

$$
\begin{aligned}
&\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\langle u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t), u_i^{t+1} - u_i^t \right\rangle \\
&= \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\langle u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t), g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i) - u_i^t \right\rangle + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\langle u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t), u_i^{t+1} - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i) \right\rangle \\
&= \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\langle u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_i^t), g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i) - u_i^t \right\rangle + \frac{1}{\gamma} \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\langle u_i^t - u_i^{t+1}, u_i^{t+1} - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i) \right\rangle \\
&\leq \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\langle u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t), g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i) - u_i^t \right\rangle \\
&\qquad + \frac{1}{2\gamma} \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left(\left\|u_i^t - g(\mathbf{w}_{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2 - \left\|u_i^{t+1} - u_i^t\right\|^2 - \left\|u_i^{t+1} - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2 \right).
\end{aligned}
$$

If $\gamma < \frac{1}{5}$, we have

$$
\begin{aligned}
&-\frac{1}{2}\left(\frac{1}{\gamma} - 1 - \frac{\gamma+1}{4\gamma}\right) \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|u_i^{t+1} - u_i^t\right\|^2 + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\langle g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t) - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i), u_i^{t+1} - u_i^t \right\rangle \\
&\leq -\frac{1}{4\gamma} \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|u_i^{t+1} - u_i^t\right\|^2 + \gamma \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t) - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2 + \frac{1}{4\gamma} \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|u_i^{t+1} - u_i^t\right\|^2 \\
&= \gamma \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t) - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2.
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
\tfrac{1}{2}\left\|\mathbf{u}^{t+1} - \mathbf{g}(\mathbf{w}^{t+1})\right\|^2 &\leq \tfrac{1}{2}\left\|\mathbf{u}^t - \mathbf{g}(\mathbf{w}^{t+1})\right\|^2 + \tfrac{1}{2\gamma}\textstyle\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2 - \tfrac{1}{2\gamma}\textstyle\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|u_i^{t+1} - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2 \\
&\quad - \frac{(\gamma+1)}{8\gamma} \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|u_i^{t+1} - u_i^t\right\|^2 + \gamma \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t) - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2 \\
&\quad + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\langle u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t), g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i) - u_i^t \right\rangle.
\end{aligned}
$$

Note that $\frac{1}{2\gamma}\sum_{i\notin\mathcal{B}_1^t} \left\|u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2 = \frac{1}{2\gamma}\sum_{i\notin\mathcal{B}_1^t} \left\|u_i^{t+1} - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2$ based on Algorithm 1, which implies that

$$
\frac{1}{2\gamma} \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left(\left\|u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2 - \left\|u_i^{t+1} - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2 \right) = \frac{1}{2\gamma_t}\left(\left\|\mathbf{u}^t - \mathbf{g}(\mathbf{w}^t)\right\|^2 - \left\|\mathbf{u}^{t+1} - \mathbf{g}(\mathbf{w}^t)\right\|^2 \right).
$$

Besides, we also have $\mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t) - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2\right] \leq \frac{B_1\sigma^2}{B_2}$ and

$$
\begin{aligned}
\mathbb{E}_t\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\langle \mathbf{u}_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{B}_{2,i}^t), g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i) - u_i^t \right\rangle \right] &= \frac{B_1}{n} \sum_{\mathbf{z}_i\in\mathcal{D}} \left\langle u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i), g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i) - u_i^t \right\rangle \\
&= -\frac{B_1}{n}\left\|u_i^t - g(\mathbf{w}^{t+1}; \mathbf{z}_i, \mathcal{S}_i)\right\|^2.
\end{aligned}
$$

Then, we can obtain

$$
\frac{\gamma+1}{2}\mathbb{E}\left[\left\|\mathbf{u}^{t+1} - \mathbf{g}(\mathbf{w}^{t+1})\right\|^2\right] \leq \frac{\gamma\left(1 - \frac{B_1}{n}\right) + 1}{2}\mathbb{E}\left[\left\|\mathbf{u}^t - \mathbf{g}(\mathbf{w}^{t+1})\right\|^2\right] + \frac{\gamma^2 B_1\sigma^2}{B_2} - \frac{(\gamma+1)}{8} \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \left\|u_i^{t+1} - u_i^t\right\|^2.
$$

Further divide $\frac{\gamma+1}{2}$ and take full expectation on both sides

$$\mathbb{E}\left[\left\|\mathbf{u}^{t+1}-\mathbf{g}(\mathbf{w}^{t+1})\right\|^2\right] \leq \frac{\gamma\left(1-\frac{B_1}{n}\right)+1}{\gamma+1}\mathbb{E}\left[\left\|\mathbf{u}^{t}-\mathbf{g}(\mathbf{w}^{t+1})\right\|^2\right] + \frac{2}{1+\gamma}\frac{\gamma^2\sigma^2 B_1}{B_2} - \frac{1}{4}\mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1}-u_i^t\right\|^2\right].$$

Note that $\frac{\gamma(1-\frac{B_1}{n})+1}{\gamma+1} = 1-\frac{\gamma B_1}{(\gamma+1)n} \leq 1-\frac{\gamma B_1}{2n}$ and $\frac{1}{1+\gamma}\leq 1$ for $\gamma\in(0,1]$. Besides, we have $\left\|\mathbf{u}^t-\mathbf{g}(\mathbf{w}^{t+1})\right\|^2 \leq (1+\frac{\gamma B_1}{4n})\left\|\mathbf{u}^t-\mathbf{g}(\mathbf{w}^t)\right\|^2 + (1+\frac{4n}{\gamma B_1})\left\|\mathbf{g}(\mathbf{w}^{t+1})-\mathbf{g}(\mathbf{w}^t)\right\|^2$ due to Young's inequality and $\left\|\mathbf{g}(\mathbf{w}^t)-\mathbf{g}(\mathbf{w}^{t-1})\right\|^2 \leq nC_g^2\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2 = n\eta^2 C_g^2\left\|\mathbf{v}^t\right\|^2$.

$$\begin{aligned}
&\mathbb{E}\left[\Xi_{t+1}\right] = \mathbb{E}\left[\frac{1}{n}\left\|\mathbf{u}^{t+1}-\mathbf{g}(\mathbf{w}^{t+1})\right\|^2\right]\\
&\leq \left(1-\frac{\gamma B_1}{2n}\right)\mathbb{E}\left[\frac{1}{n}\left\|\mathbf{u}^{t}-\mathbf{g}(\mathbf{w}^{t+1})\right\|^2\right] + \frac{2\gamma^2\sigma^2 B_1}{nB_2} - \frac{1}{4n}\mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1}-u_i^t\right\|^2\right]\\
&\leq \left(1-\tfrac{\gamma B_1}{2n}\right)\left(1+\tfrac{\gamma B_1}{4n}\right)\mathbb{E}\left[\tfrac{1}{n}\left\|\mathbf{u}^{t}-\mathbf{g}(\mathbf{w}^{t})\right\|^2\right] + \tfrac{5nC_g^2\mathbb{E}\left[\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2\right]}{\gamma B_1} + \tfrac{2\gamma^2\sigma^2B_1}{nB_2} - \tfrac{1}{4n}\mathbb{E}\left[\textstyle\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1}-u_i^t\right\|^2\right]\\
&\leq \left(1-\frac{\gamma B_1}{4n}\right)\mathbb{E}\left[\Xi_t\right] + \frac{5n\eta^2C_g^2\mathbb{E}\left[\left\|\mathbf{v}^t\right\|^2\right]}{\gamma B_1} + \frac{2\gamma^2\sigma^2 B_1}{nB_2} - \frac{1}{4n}\mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1}-u_i^t\right\|^2\right].
\end{aligned}$$

B.2. Proof of Theorem 1

Based on Lemma 2, Lemma 3, and Lemma 1, we have

$$\mathbb{E}\left[F(\mathbf{w}^{t+1})-F^*\right] \leq \mathbb{E}\left[F(\mathbf{w}^t)-F^*\right] + \frac{\eta}{2}\mathbb{E}\left[\Delta_t\right] - \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right] - \frac{\eta}{4}\mathbb{E}\left[\left\|\mathbf{v}^t\right\|^2\right], \tag{11}$$

$$\mathbb{E}\left[\Delta_{t+1}\right] \leq (1-\beta)\mathbb{E}\left[\Delta_t\right] + \tfrac{2L_F^2\eta^2}{\beta}\mathbb{E}\left[\left\|\mathbf{v}^t\right\|^2\right] + 5\beta L_f^2C_1^2\mathbb{E}\left[\Xi_{t+1}\right] + \tfrac{2\beta^2C_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}} + \tfrac{3L_f^2C_1^2}{n}\mathbb{E}\left[\textstyle\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1}-u_i^t\right\|^2\right], \tag{12}$$

$$\mathbb{E}\left[\Xi_{t+1}\right] \leq \left(1-\frac{\gamma B_1}{4n}\right)\mathbb{E}\left[\Xi_t\right] + \frac{5n\eta^2C_g^2\mathbb{E}\left[\left\|\mathbf{v}^t\right\|^2\right]}{\gamma B_1} + \frac{2\gamma^2\sigma^2B_1}{nB_2} - \frac{1}{4n}\mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1}-u_i^t\right\|^2\right]. \tag{13}$$

Summing (11), $\frac{\eta}{\beta}\times$(12), and $\frac{20L_f^2C_1^2n\eta}{\gamma B_1}\times$(13) leads to

$$\begin{aligned}
&\mathbb{E}\left[F(\mathbf{w}^{t+1})-F^* + \frac{\eta}{\beta}\Delta_{t+1} + \frac{20L_f^2C_1^2n\eta}{\gamma B_1}\left(1-\frac{\gamma B_1}{4n}\right)\Xi_{t+1}\right]\\
&\leq \mathbb{E}\left[F(\mathbf{w}^t)-F^* + \frac{\eta}{\beta}\left(1-\frac{\beta}{2}\right)\Delta_t + \frac{20L_f^2C_1^2n\eta}{\gamma B_1}\left(1-\frac{\gamma B_1}{4n}\right)\Xi_t\right] - L_f^2C_1^2\eta\left(\frac{5n}{\gamma B_1}-\frac{3}{\beta}\right)\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1}-u_i^t\right\|^2\right]\\
&\qquad - \frac{\eta}{2}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right] - \eta\left(\frac{1}{4}-\frac{2L_F^2\eta^2}{\beta^2}-\frac{100L_f^2n^2C_1^2\eta^2C_g^2}{\gamma^2B_1^2}\right)\mathbb{E}\left[\left\|\mathbf{v}^t\right\|^2\right] + \frac{2\beta\eta C_f^2(\zeta^2+C_g^2)}{\min\{B_2,B_2\}} + \frac{40\eta\gamma L_f^2C_1^2\sigma^2}{B_2}.
\end{aligned}$$

If $\gamma\leq\frac{5n}{3B_1}\beta$, we have $\frac{5n}{\gamma B_1}-\frac{3}{\beta}\geq 0$. Set $\beta = \min\{\frac{\min\{B_1,B_2\}\epsilon^2}{12C_f^2(\zeta^2+C_g^2)},\frac{2}{7}\}$, $\gamma=\min\left\{\frac{B_2\epsilon^2}{240L_f^2C_1^2\sigma^2},\frac{1}{5},\frac{5n}{3B_1}\beta\right\}$, and $\eta=\min\left\{\frac{\beta}{4L_F},\frac{\gamma B_1}{30L_fnC_1C_g}\right\}$. Define the Lyapunov function as $\Phi_t \coloneqq F(\mathbf{w}^t)-F^*+\frac{\eta}{\beta}\Delta_t+\frac{20L_f^2C_1^2}{B_1}\frac{\eta}{\gamma}\left(1-\frac{\gamma B_1}{4n}\right)\Xi_t$. If we initialize $\mathbf{u}^1$ and $\mathbf{v}^1$ as $\mathbf{u}_i^1 = g_i(\mathbf{w}^1;\mathcal{B}_{i,2}^1)$ for $\mathbf{z}_i\in\mathcal{B}_1^t$ and $\mathbf{v}^1=0$, we have $\mathbb{E}\left[\Delta_1\right]\leq C_f^2C_g^2$ and $\mathbb{E}\left[\Xi_1\right]\leq\frac{\sigma^2}{B_2}$. Then,

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right] \leq \frac{2\Lambda_\Phi^1}{\eta T} + \frac{4\beta C_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}} + \frac{80\gamma L_f^2C_1^2\sigma^2}{B_2}, \tag{14}$$

where we define $\Lambda_\Phi^1 \coloneqq \Delta_F + \frac{1}{4L_F}C_f^2 C_g^2 + \frac{2L_f C_1\sigma^2}{3C_g B_2} \geq \mathbb{E}[\Phi_1]$. After

$$T = \frac{6\Lambda_\Phi^1}{\epsilon^2}\max\left\{\frac{48C_f^2(\zeta^2+C_g^2)L_F}{\min\{B_1,B_2\}\epsilon^2}, 14L_F, \frac{7200nL_f^3C_1^3C_g\sigma^2}{B_1B_2\epsilon^2}, \frac{150L_f nC_1C_g}{B_1}, 63L_fC_1C_g, \frac{216L_fC_1C_gC_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}\epsilon^2}\right\}$$

iterations, we have $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right] \leq \epsilon^2$.

C. Proof of Theorem 2

Lemma 4. *Under assumptions 1, 2 and the μ-PL of F, the k-th epoch of applying Algorithm 2 leads to*

$$\mathbb{E}[\Gamma_{k+1}] \leq \frac{\mathbb{E}[\Gamma_k]}{\mu\eta_k T_k} + \frac{\mathbb{E}[\Delta_k]}{\mu\beta_k T_k} + \frac{20nL_f^2C_1^2\mathbb{E}[\Xi_k]}{B_1\mu\gamma_k T_k} + \frac{\beta_k C_2^2}{\mu\min\{B_1,B_2\}} + \frac{\gamma_k C_3^2}{\mu B_2}, \tag{15}$$

where $\Gamma_k \coloneqq F(\mathbf{w}^k) - F(\mathbf{w}^*)$, $C_1^2 \coloneqq C_g^2 + \zeta^2/B_2$, $C_2^2 \coloneqq 2C_f^2(\zeta^2 + C_g^2)$, $C_3^2 \coloneqq 40L_f^2C_1^2\sigma^2$.

Proof. Since F is μ-PL, we have $F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{1}{2\mu}\left\|\nabla F(\mathbf{w})\right\|^2$. We define $\Gamma_k \coloneqq F(\mathbf{w}^k) - F(\mathbf{w}^*)$, $\Delta_k \coloneqq \left\|\mathbf{v}^k - \nabla F(\mathbf{w}^k)\right\|^2$ and $\Xi_k \coloneqq \frac{1}{n}\left\|\mathbf{u}^k - \mathbf{g}(\mathbf{w}^k)\right\|^2$. Applying PL condition and Theorem 1 to one epoch of SOX-boost leads to $\mathbb{E}[\Gamma_{k+1}] \leq \frac{1}{2\mu}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{k+1})\right\|^2\right]$ and

$$\frac{1}{2\mu}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^{k+1})\right\|^2\right] \leq \frac{1}{2\mu T_k}\sum_{t=1}^{T_k}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right] \leq \frac{\mathbb{E}[\Gamma_k]}{\mu\eta_k T_k} + \frac{\mathbb{E}[\Delta_k]}{\mu\beta_k T_k} + \frac{20nL_f^2C_1^2\mathbb{E}[\Xi_k]}{B_1\mu\gamma_k T_k} + \frac{\beta_k C_2^2}{\mu\min\{B_1,B_2\}} + \frac{\gamma_k C_3^2}{\mu B_2},$$

where we define $C_2^2 \coloneqq 2C_f^2(\zeta^2 + C_g^2)$, $C_3^2 \coloneqq 40L_f^2C_1^2\sigma^2$. □

Lemma 5. *Under assumptions 1, 2, the k-th epoch of Algorithm 2 leads to*

$$\mathbb{E}[\Delta_{k+1} + C_5\Xi_{k+1}] \leq \frac{\mathbb{E}[6\Gamma_k + 10C_4\Delta_k + 7C_4C_5\Xi_k]}{\eta_k T_k} + \frac{10\eta_k C_2^2}{C_4\min\{B_1,B_2\}} + \frac{80nC_3^2\eta_k}{3B_1B_2C_4},$$

where $\beta_k \leq \min\left\{\frac{3B_1}{50n}, \frac{2}{7}\right\}$, $\gamma_k = \frac{10n}{3B_1}\beta_k$, $\eta_k = \beta_k C_4$, $C_4 \coloneqq \min\{1/4L_F, 1/9L_fC_1C_g\}$, $C_5 \coloneqq 12L_f^2C_1^2$.

Proof. Applying Lemma 3 to single iteration in any epoch of SOX-boost with $\beta_k \leq \frac{2}{7}$ leads to

$$\begin{aligned}\mathbb{E}[\Delta_{t+1}] \leq (1-\beta_k)\mathbb{E}[\Delta_t] + \frac{4L_F^2\eta_k^2}{\beta}\mathbb{E}[\Delta_t] + \frac{4L_F^2\eta_k^2}{\beta_k}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right] \\ + 5\beta_k L_f^2C_1^2\mathbb{E}[\Xi_{t+1}] + \frac{2\beta_k^2C_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}} + \frac{3L_f^2C_1^2}{n}\mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1} - u_i^t\right\|^2\right].\end{aligned}$$

Applying Lemma 1 to one iteration in any epoch of SOX-boost with $\gamma_k \leq \frac{1}{5}$ leads to

$$\mathbb{E}[\Xi_{t+1}] \leq \left(1 - \frac{\gamma_k B_1}{4n}\right)\mathbb{E}[\Xi_t] + \frac{10n\eta_k^2C_g^2\mathbb{E}[\Delta_t]}{\gamma_k B_1} + \frac{10n\eta_k^2C_g^2\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right]}{\gamma_k B_1} + \frac{2\gamma_k^2\sigma^2 D}{nB_2} - \frac{1}{4n}\mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1} - u_i^t\right\|^2\right].$$

The following holds by summing up $\mathbb{E}[\Delta_{t+1}]$ and $\frac{40L_f^2C_1^2n\beta_k}{\gamma_k B_1} \times \mathbb{E}[\Xi_{t+1}]$ and noticing $\left(1 - \frac{\gamma_k B_1}{4n}\right) \leq \left(1 - \frac{\gamma_k B_1}{8n}\right)^2$.

$$\begin{aligned}\mathbb{E}\left[\Delta_{t+1} + \frac{40L_f^2C_1^2n\beta_k}{\gamma_k B_1}\Xi_{t+1}\right] \leq \mathbb{E}\left[\left(1 - \beta_k + \frac{4L_F^2\eta_k^2}{\beta_k} + \frac{400n^2\eta_k^2\beta_k C_g^2L_f^2C_1^2}{\gamma_k^2B_1^2}\right)\Delta_t + \frac{40L_f^2C_1^2n\beta_k}{\gamma_k B_1}\left(1 - \frac{\gamma_k B_1}{8n}\right)\Xi_t\right] \\ + \left(\frac{4L_F^2\eta_k^2}{\beta_k} + \frac{400n^2\eta_k^2\beta_k C_g^2L_f^2C_1^2}{\gamma_k^2B_1^2}\right)\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right] + \frac{2\beta_k^2C_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}} \\ + \frac{80\gamma_k\beta L_f^2C_1^2\sigma^2}{B_2} - L_f^2C_1^2\left(\frac{10n\beta_k}{\gamma_k B_1} - 3\right)\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\left\|u_i^{t+1} - u_i^t\right\|^2\right]\end{aligned}$$

If we set $\gamma_k \leq \frac{10n}{3B_1}\beta_k$, $\eta_k \leq \frac{\beta_k}{4L_F}$ and $\eta_k \leq \frac{\gamma_k B_1}{30L_f n C_1 C_g}$.

$$1-\beta_k+\frac{4L_F^2\eta_k^2}{\beta_k}+\frac{400n^2\eta_k^2\beta_k C_g^2 L_f^2 C_1^2}{\gamma_k^2 B_1^2} \leq 1-\beta_k+\frac{\beta_k}{4}+\frac{4\beta_k}{9} \leq 1-\frac{\beta_k}{4}.$$

If $\beta_k \leq \min\left\{\frac{3B_1}{50n}, \frac{2}{7}\right\}$, we can set $\gamma_k = \frac{10n}{3B_1}\beta_k$ and $\eta_k = \beta_k C_4$, where $C_4 := \min\{1/4L_F, 1/9L_f C_1 C_g\}$. Then, $1-\frac{\gamma_k B_1}{8n} = 1-\frac{5\beta_k}{12} \leq 1-\frac{\beta_k}{4}$. Besides, we define $C_5 := \frac{40L_f^2 C_1^2 n\beta_k}{\gamma_k B_1} = 12L_f^2 C_1^2$.

$$\mathbb{E}\left[\Delta_{t+1}+C_5\Xi_{t+1}\right] \leq \left(1-\frac{\beta_k}{4}\right)\mathbb{E}\left[\Delta_t+C_5\Xi_t\right]+\frac{3\beta_k}{4}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right]+\frac{2\beta_k^2 C_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}}+\frac{2\gamma_k\beta_k C_3^2}{B_2}.$$

Telescoping the equation above from 1 to T_k iterations in epoch k leads to

$$\begin{aligned}\mathbb{E}\left[\Delta_{k+1}+C_5\Xi_{k+1}\right] &= \mathbb{E}\left[\frac{1}{T_k}\sum_{t=0}^{T_k}\Delta_{t+1}+C_5\frac{1}{T_k}\sum_{t=0}^{T_k}\Xi_{t+1}\right] \\ &\leq \frac{4C_4\mathbb{E}\left[\Delta_k+C_5\Xi_k\right]}{\eta_k T_k}+3\frac{1}{T_k}\sum_{t=1}^{T_k}\mathbb{E}\left[\left\|\nabla F(\mathbf{w}^t)\right\|^2\right]+\frac{4\beta_k C_2^2}{\min\{B_1,B_2\}}+\frac{20nC_3^2\eta_k}{3B_1B_2C_4}.\end{aligned}$$

Applying (14), we can further derive that

$$\mathbb{E}\left[\Delta_{k+1}+C_5\Xi_{k+1}\right] \leq \frac{\mathbb{E}\left[6\Gamma_k+10C_4\Delta_k+7C_4C_5\Xi_k\right]}{\eta_k T_k}+\frac{10\eta_k C_2^2}{C_4\min\{B_1,B_2\}}+\frac{80nC_3^2\eta_k}{3B_1B_2C_4}.$$

□

Lemma 6. *If we set* $\eta_k = \min\left\{\frac{\mu C_4\min\{B_1,B_2\}\varepsilon_k}{6C_2^2}, \frac{\min\{B_1,B_2\}\varepsilon_k}{60C_2^2}, \frac{\mu C_4 B_1 B_2\varepsilon_k}{20nC_3^2}, \frac{B_1B_2C_4\varepsilon_k}{200nC_3^2}, \frac{3B_1C_4}{50n}, \frac{2C_4}{7}\right\}$ *and* $T_k = \frac{\max\left\{\frac{12}{\mu}, 96C_4\right\}}{\eta_k}$, $\gamma_k = \frac{10n\eta_k}{3B_1C_4}$, $\beta_k = \frac{\eta_k}{C_4}$, *we can conclude that* $\mathbb{E}\left[\Gamma_k\right] \leq \varepsilon_k$ *and* $\mathbb{E}\left[\Delta_k+C_5\Xi_k\right] \leq \frac{\varepsilon_k}{C_4}$, *where* $\varepsilon_1 = \max\left\{\Delta_F, C_4(C_f^2C_g^2+C_5\sigma^2/B_2)\right\}$ *and* $\varepsilon_k = \varepsilon_1/2^{k-1}$ *for* $k \geq 1$.

Proof. We prove this lemma by induction. First, we define $\Delta_F := F(\mathbf{w}^1)-F(\mathbf{w}^*)$, $\Gamma_k = F(\mathbf{w}^k)-F(\mathbf{w}^*)$, and $\varepsilon_1 = \max\left\{\Delta_F, C_4(C_f^2C_g^2+C_5\sigma^2/B_2)\right\}$, where $C_4 := \min\{1/4L_F, 1/9L_fC_1C_g\}$, $C_5 := 12L_f^2C_1^2$. If we initialize $\mathbf{u}^1$ and $\mathbf{v}^1$ as $u_i^1 = g_i(\mathbf{w}^1;\mathcal{B}_{i,2}^1)$ for $i \in \mathcal{D}_t$ and $\mathbf{v}^1 = 0$, we have $\mathbb{E}\left[\Gamma_1\right] \leq \varepsilon_1$ and $\mathbb{E}\left[\Delta_1+C_5\Xi_1\right] \leq \frac{\varepsilon_1}{C_4}$.

Next, we consider the $k \geq 2$ case. Assume $\mathbb{E}\left[\Gamma_k\right] \leq \varepsilon_k$ and $\mathbb{E}\left[\Delta_k+C_5\Xi_k\right] \leq \frac{\varepsilon_k}{C_4}$. We define $\varepsilon_k = \varepsilon_1/2^{k-1}$ for $k \geq 2$. We choose $\beta_k \leq \min\left\{\frac{3D}{50n}, \frac{2}{7}\right\}$, $\gamma_k = \frac{10n}{3D}\beta_k$, $\eta_k = \beta_k C_4$. Based on Lemma 4, we have

$$\mathbb{E}\left[\Gamma_{k+1}\right] \leq \frac{\mathbb{E}\left[\Gamma_k+C_4(\Delta_k+C_5\Xi_k)\right]}{\mu\eta_k T_k}+\frac{\eta_k C_2^2}{\mu C_4\min\{B_1,B_2\}}+\frac{10nC_3^2\eta_k}{3\mu B_1B_2C_4}.$$

Besides, Lemma 5 implies that

$$\mathbb{E}\left[\Delta_{k+1}+C_5\Xi_{k+1}\right] \leq \frac{\mathbb{E}\left[6\Gamma_k+10C_4(\Delta_k+C_5\Xi_k)\right]}{\eta_k T_k}+\frac{10\eta_k C_2^2}{C_4\min\{B_1,B_2\}}+\frac{80n\eta_k C_3^2}{3B_1B_2C_4}.$$

The following choices of η_k and T_k makes $\mathbb{E}\left[\Gamma_{k+1}\right] \leq \varepsilon_{k+1} = \frac{\varepsilon_k}{2}$ and $\mathbb{E}\left[\Delta_{k+1}+C_5\Xi_{k+1}\right] \leq \frac{\varepsilon_{k+1}}{C_4} = \frac{\varepsilon_k}{2C_4}$. We define $C_6 := \min\{\lambda C_4, \frac{1}{10}\}$.

$$\begin{aligned}\eta_k &= \min\left\{\frac{\mu C_4\min\{B_1,B_2\}\varepsilon_k}{6C_2^2}, \frac{\min\{B_1,B_2\}\varepsilon_k}{60C_2^2}, \frac{\mu C_4B_1B_2\varepsilon_k}{20nC_3^2}, \frac{B_1B_2C_4\varepsilon_k}{200nC_3^2}, \frac{3B_1C_4}{50n}, \frac{2C_4}{7}\right\}, \\ T &= \frac{\max\left\{\frac{12}{\mu}, 96C_4\right\}}{\eta_k}.\end{aligned}$$

□

Theorem 4 (Detailed Version of Theorem 2)**.** *Under assumptions 1, 2 and the μ-PL of F, SOX-boost (Algorithm 2) can find an $\mathbf{w}$ satisfying that $\mathbb{E}\left[F(\mathbf{w})-F(\mathbf{w}^*)\right]\leq 2\epsilon$ after*

$$T=C_{1/\mu}\max\left\{\frac{6C_2^2}{\mu C_4\min\{B_1,B_2\}\epsilon},\frac{60C_2^2}{\min\{B_1,B_2\}\epsilon},\frac{20nC_3^2}{\mu C_4B_1B_2\epsilon},\frac{200nC_3^2}{B_1B_2C_4\epsilon},\frac{50n\log(\varepsilon_1/\epsilon)}{3B_1C_4},\frac{7\log(\varepsilon_1/\epsilon)}{2C_4}\right\}$$

iterations, where $C_{1/\mu}=\max\left\{\frac{12}{\mu},96C_4\right\}$.

Proof. According to Lemma 6, the total number of iterations to achieve target accuracy $\mathbb{E}\left[\Gamma_k\right]\leq\epsilon$ can be represented as:

$$\begin{aligned}T&=\sum_{k=1}^{\log(\varepsilon_1/\epsilon)}T_k\\&=C_{1/\mu}\max\left\{\frac{6C_2^2}{\mu C_4\min\{B_1,B_2\}\epsilon},\frac{60C_2^2}{\min\{B_1,B_2\}\epsilon},\frac{20nC_3^2}{\mu C_4B_1B_2\epsilon},\frac{200nC_3^2}{B_1B_2C_4\epsilon},\frac{50n\log(\varepsilon_1/\epsilon)}{3B_1C_4},\frac{7\log(\varepsilon_1/\epsilon)}{2C_4}\right\},\end{aligned}$$

where $C_{1/\mu}=\max\left\{\frac{12}{\mu},96C_4\right\}$. ☐

Proof of Corollary 1. Suppose that $\mathbf{w}^*$ is a minimum of F and $\hat{\mathbf{w}}^*$ is the minimum of the strongly convexified $\hat{F}$. If $\mathbb{E}\left[\hat{F}(\mathbf{w})-\hat{F}(\hat{\mathbf{w}}^*)\right]\leq\epsilon$, we have

$$\mathbb{E}\left[F(\mathbf{w})\right]\leq\mathbb{E}\left[\hat{F}(\mathbf{w})\right]\leq\hat{F}(\hat{\mathbf{w}}^*)+\epsilon\leq\hat{F}(\mathbf{w}^*)+\epsilon=F(\mathbf{w}^*)+\frac{\lambda}{2}\left\|\mathbf{w}^*\right\|^2+\epsilon.$$

Thus, if the minimum $\mathbf{w}^*$ of F is in a bounded domain $\|\mathbf{w}^*\|\leq C_*$ and we choose $\lambda=\frac{2\epsilon}{C_*^2}$, we also have $\mathbb{E}\left[F(\mathbf{w})-F(\mathbf{w}^*)\right]\leq 2\epsilon$. ☐

D. Proof of Theorem 3

For any iteration t, we can define $\mathbf{y}^t=(\mathbf{w}^t;\pi_1^{t+1};\pi_2^{t+1})$ and $\pi_1^{t+1}=[\pi_{1,1}^{t+1},\dots,\pi_{n,1}^{t+1}]^\top$, $\pi_2^{t+1}=[\pi_{1,2}^{t+1},\dots,\pi_{n,2}^{t+1}]^\top$, where $\pi_{i,1}^{t+1}=\nabla f(u_i^{t+1})$ and $\pi_{i,2}^{t+1}=\nabla g_i(\mathbf{w}_t)$. We define that $\mathbf{y}^*:=(\mathbf{w}^*;\bar{\pi}_1^*;\bar{\pi}_2^*)$ where $\bar{\pi}_1^*$, $\bar{\pi}_2^*$ are defined as $\bar{\pi}_{i,1}^*=\arg\max_{\pi_{i,1}}\left\langle\pi_{i,1},\mathcal{L}_{i,2}(\bar{\mathbf{w}}^T,\bar{\pi}_{i,2}^*)\right\rangle-f^*(\pi_{i,1})$ and $\bar{\pi}_{i,2}^*=\arg\max_{\pi_{i,2}}\left\langle\pi_{i,2},\bar{\mathbf{w}}^T\right\rangle-g_i^*(\pi_{i,2})$, $\bar{\mathbf{w}}:=\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{w}^t$. we can define the gap $Q(\mathbf{y}^t,\mathbf{y}^*)$ as

$$Q(\mathbf{y}^t,\mathbf{y}^*)=\mathcal{L}(\mathbf{w}^t,\bar{\pi}_1^*,\bar{\pi}_2^*)-\mathcal{L}(\mathbf{w}^*,\pi_1^{t+1},\pi_2^{t+1})=\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left(\underbrace{\mathcal{L}_{i,1}(\mathbf{w}^t,\bar{\pi}_{i,1}^*,\bar{\pi}_{i,2}^*)-\mathcal{L}_{i,1}(\mathbf{w}^*,\pi_{i,1}^{t+1},\pi_{i,2}^{t+1})}_{:=Q_i(\mathbf{y}^t,\mathbf{y}^*)}\right).$$

The following lemmas are needed.

Lemma 7. *For any $\pi_{i,1}$ such that $\pi_{i,1}=\nabla f(u_i)$ for some bounded u_i, then there exists C_{f_*} such that $|f^*(\pi_{i,1})|\leq C_{f_*}$.*

Proof. Due to the definition of convex conjugate, we have $f(u_i)+f^*(\pi_{i,1})=\pi_{i,1}u_i$. Due to that $\pi_{i,1}$ and u_i are bounded and $f(u_i)$ is bounded due to its Lipchitz continity. As a result, $f^*(\pi_{i,1})$ is bounded by some constant C_{f_*}. ☐

Lemma 8. *For any $\pi_1\in\Pi$ and the sequences $\{\tilde{\pi}_1^t\}$ and $\hat{\pi}_1^t$ defined as $\tilde{\pi}_1^t=\arg\min_{\pi_1}\left\langle\hat{g}(\mathbf{w}^t;\tilde{\mathcal{B}}_2^t)-g(\mathbf{w}^t),\pi_1\right\rangle+\tau'\mathbf{D}_{f^*}(\tilde{\pi}_1^t,\pi_1)$, $\hat{\pi}_1^{t+1}=\arg\min_{\pi_1}\langle\hat{g}(\mathbf{w}^t;\mathcal{B}_2^t)-g(\mathbf{w}^t),\pi_1\rangle+\tau'\mathbf{D}_{f^*}(\hat{\pi}_1^t,\pi_1)$, we have*

$$\left\langle\hat{g}(\mathbf{w}^t;\mathcal{B}_2^t)-g(\mathbf{w}^t),\pi_1-\tilde{\pi}_1^t\right\rangle\geq\tau'(\mathbf{D}_{f^*}(\hat{\pi}_1^{t+1},\pi_1)-\mathbf{D}_{f^*}(\hat{\pi}_1^t,\pi_1))-\frac{L_f}{2(\tau')^2}\left\|\hat{g}(\mathbf{w}^t;\mathcal{B}_2^t)-\hat{g}(\mathbf{w}^t;\tilde{\mathcal{B}}_2^t)\right\|^2.$$

Proof. The proof of this lemma is almost the same to that of Lemma 4 in Juditsky et al. (2011). Due to the three-point inequality (Lemma 1 in Zhang & Lan 2020), we have:

$$\mathbf{D}_{f^*}(\hat{\pi}_1^t, \hat{\pi}_1^{t+1}) \geq \mathbf{D}_{f^*}(\tilde{\pi}_1^t, \hat{\pi}_1^{t+1}) - \frac{1}{\tau'}\left\langle \hat{g}(\mathbf{w}^t; \tilde{\mathcal{B}}_2^t) - g(\mathbf{w}^t), \hat{\pi}_1^{t+1} - \tilde{\pi}_1^t \right\rangle + \mathbf{D}_{f^*}(\hat{\pi}_1^t, \tilde{\pi}_1^t), \tag{16}$$

$$\mathbf{D}_{f^*}(\hat{\pi}_1^{t+1}, \pi_1) \leq \mathbf{D}_{f^*}(\hat{\pi}_1^t, \pi_1) + \frac{1}{\tau'}\left\langle \hat{g}(\mathbf{w}^t; \mathcal{B}_2^t) - g(\mathbf{w}^t), \pi_1 - \hat{\pi}_1^{t+1} \right\rangle - \mathbf{D}_{f^*}(\hat{\pi}_1^t, \hat{\pi}_1^{t+1}), \tag{17}$$

where π_1 could be any $\pi_1 \in \Pi_1$. The last term on the R.H.S. of (17) can be upper bounded by (16).

$$\begin{aligned}\mathbf{D}_{f^*}(\hat{\pi}_1^{t+1}, \pi_1) \leq \mathbf{D}_{f^*}(\hat{\pi}_1^t, \pi_1) &+ \frac{1}{\tau'}\left\langle \hat{g}(\mathbf{w}^t; \mathcal{B}_2^t) - g(\mathbf{w}^t), \pi_1 - \tilde{\pi}_1^t \right\rangle \\ &+ \frac{1}{\tau'}\left\langle \hat{g}(\mathbf{w}^t; \mathcal{B}_2^t) - \hat{g}(\mathbf{w}^t; \tilde{\mathcal{B}}_2^t), \tilde{\pi}_1^t - \hat{\pi}_1^{t+1} \right\rangle - \mathbf{D}_{f^*}(\hat{\pi}_1^t, \tilde{\pi}_1^t) - \mathbf{D}_{f^*}(\tilde{\pi}_1^t, \hat{\pi}_1^{t+1})\end{aligned}$$

Considering the strong convexity, we have

$$\begin{aligned}\mathbf{D}_{f^*}(\hat{\pi}_1^{t+1}, \pi_1) \leq \mathbf{D}_{f^*}(\hat{\pi}_1^t, \pi_1) &+ \frac{1}{\tau'}\left\langle \hat{g}(\mathbf{w}^t; \mathcal{B}_2^t) - g(\mathbf{w}^t), \pi_1 - \tilde{\pi}_1^t \right\rangle \\ &+ \frac{1}{\tau'}\left\langle \hat{g}(\mathbf{w}^t; \mathcal{B}_2^t) - \hat{g}(\mathbf{w}^t; \tilde{\mathcal{B}}_2^t), \tilde{\pi}_1^t - \hat{\pi}_1^{t+1} \right\rangle - \frac{1}{2L_f}\left\|\hat{\pi}_1^t - \tilde{\pi}_1^t\right\|^2 - \frac{1}{2L_f}\left\|\tilde{\pi}_1^t - \hat{\pi}_1^{t+1}\right\|^2\end{aligned}$$

Based on the Young's inequality, we further have

$$\frac{1}{\tau'}\left\langle \hat{g}(\mathbf{w}^t; \mathcal{B}_2^t) - \hat{g}(\mathbf{w}^t; \tilde{\mathcal{B}}_2^t), \tilde{\pi}_1^t - \hat{\pi}_1^{t+1} \right\rangle \leq \frac{L_f}{2(\tau')^2}\left\|\hat{g}(\mathbf{w}^t; \mathcal{B}_2^t) - \hat{g}(\mathbf{w}^t; \tilde{\mathcal{B}}_2^t)\right\|^2 + \frac{1}{2L_f}\left\|\tilde{\pi}_1^t - \hat{\pi}_1^{t+1}\right\|^2.$$

Re-arranging the terms leads to

$$\left\langle \hat{g}(\mathbf{w}^t; \mathcal{B}_2^t) - g(\mathbf{w}^t), \pi_1 - \tilde{\pi}_1^t \right\rangle \geq \tau'(\mathbf{D}_{f^*}(\hat{\pi}_1^{t+1}, \pi_1) - \mathbf{D}_{f^*}(\hat{\pi}_1^t, \pi_1)) - \frac{L_f}{2(\tau')^2}\left\|\hat{g}(\mathbf{w}^t; \mathcal{B}_2^t) - \hat{g}(\mathbf{w}^t; \tilde{\mathcal{B}}_2^t)\right\|^2.$$

□

We can decompose $Q_i(\mathbf{y}^t, \mathbf{y}^*)$ into three terms

$$\begin{aligned}Q_{i,2}(\mathbf{y}^t, \mathbf{y}^*) &= \mathcal{L}_{i,1}(\mathbf{w}^t, \bar{\pi}_{i,1}^*, \bar{\pi}_{i,2}^*) - \mathcal{L}_{i,1}(\mathbf{w}^t, \bar{\pi}_{i,1}^*, \pi_{i,2}^{t+1}) = \bar{\pi}_{i,1}^*\left(\left\langle \bar{\pi}_{i,2}^*, \mathbf{w}^t \right\rangle - g_i^*(\bar{\pi}_{i,2}^*) - \left\langle \pi_{i,2}^{t+1}, \mathbf{w}^t \right\rangle + g_i^*(\pi_{i,2}^{t+1})\right), \\ Q_{i,1}(\mathbf{y}^t, \mathbf{y}^*) &= \mathcal{L}_{i,1}(\mathbf{w}^t, \bar{\pi}_{i,1}^*, \pi_{i,2}^{t+1}) - \mathcal{L}_{i,2}(\mathbf{w}^t, \pi_{i,1}^{t+1}, \pi_{i,2}^{t+1}) \\ &= \bar{\pi}_{i,1}^*\mathcal{L}_{i,2}(\mathbf{w}^t, \pi_{i,2}^{t+1}) - f^*(\bar{\pi}_{i,1}^*) - \pi_{i,1}^{t+1}\mathcal{L}_{i,2}(\mathbf{w}^t, \pi_{i,2}^{t+1}) + f^*(\pi_{i,1}^{t+1}), \\ Q_{i,0}(\mathbf{y}^t, \mathbf{y}^*) &= \mathcal{L}_{i,1}(\mathbf{w}^t, \pi_{i,1}^{t+1}, \pi_{i,2}^{t+1}) - \mathcal{L}_{i,1}(\mathbf{w}^*; \pi_{i,1}^{t+1}, \pi_{i,2}^{t+1}) = \pi_i^{t+1}\left\langle \pi_{i,2}^{t+1}, \mathbf{w}^t - \mathbf{w}^* \right\rangle.\end{aligned}$$

We upper bound these terms one by one by the following lemmas.

Lemma 9. *We have $Q_{i,2}(\mathbf{y}^t, \mathbf{y}^*) \leq 0$ for any $\mathbf{z}_i \in \mathcal{D}$ and any $t = 0, \ldots, T-1$.*

Proof. Since f is Lipschitz-continuous, convex and monotonically increasing, we have $0 \leq \bar{\pi}_{i,1}^* \leq C_f$. Besides, $\left\langle \bar{\pi}_{i,2}^*, \mathbf{w}_t \right\rangle - g_i^*(\bar{\pi}_{i,2}^*) \leq \left\langle \pi_{i,2}^{t+1}, \mathbf{w}_t \right\rangle - g_i^*(\pi_{i,2}^{t+1})$ due to $\pi_{i,2}^{t+1} = \arg\max_{\pi_{i,2}} \pi_{i,2}^\top \mathbf{w}_t - g_i^*(\pi_{i,2})$. We can conclude that $Q_{i,2}(\mathbf{y}_t, \mathbf{y}^*) \leq 0$. □

Lemma 10. *The term $\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} Q_{i,1}(\mathbf{y}^t, \mathbf{y}^*)\right]$ can be upper bounded as*

$$\begin{aligned}&\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} Q_{i,1}(\mathbf{y}^t, \mathbf{y}^*)\right] \\ &\leq \frac{2n(C_fC_gC_\Omega + C_{f^*})}{B_1} + \frac{n\eta CT}{B_1} + \frac{\tau n L_f C_g^2 C_\Omega^2}{B_1} + \frac{L_f T\sigma^2}{\tau B_2} + \frac{B_1 T L_f \sigma^2}{n\tau B_2} - \frac{\tau}{2B_1}\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbf{D}_{f^*}(\pi_1^t, \pi_1^{t+1})\right].\end{aligned}$$

Proof. We define that $\mathbf{f}^*(\pi_1) := \sum_{\mathbf{z}_i\in\mathcal{D}} f^*(\pi_{i,1})$ for any $\pi_1 \in \Pi_1$, $\phi(\mathbf{w}^t,\pi_1) := -\sum_{\mathbf{z}_i\in\mathcal{D}} \pi_{i,1}\mathcal{L}_2^i(\mathbf{w}^t;\pi_{i,2}^{t+1}) = -\sum_{\mathbf{z}_i\in\mathcal{D}} \pi_{i,1} g_i(\mathbf{w}^t)$, $h(\mathbf{w}^t,\pi_1) := \phi(\mathbf{w}^t,\pi_1) + \sum_{\mathbf{z}_i\in\mathcal{D}} f^*(\pi_{i,1})$. Due to the update of rule of $\pi_{i,1}^t$ ($\pi_{i,1}^{t+1} = \pi_{i,1}^t$ for $i\notin\mathcal{B}_1^t$) and the convexity, we have

$$\begin{aligned}h(\mathbf{w}^t,\pi_1^{t+1}) &= \phi(\mathbf{w}^t,\pi_1^t) + (\phi(\mathbf{w}^t,\pi_1^{t+1}) - \phi(\mathbf{w}^t,\pi_1^t)) + \mathbf{f}^*(\pi_1^{t+1})\\ &\leq \phi(\mathbf{w}^t,\pi_1^t) \underbrace{- \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t)(\pi_{i,1}^{t+1} - \pi_{i,1}^t) + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} f^*(\pi_{i,1}^{t+1})}_{:=¶} + \sum_{\mathbf{z}_i\notin\mathcal{B}_1^t} f^*(\pi_{i,1}^{t+1})\\ &\quad + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} (g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t) - g_i(\mathbf{w}^t))(\pi_{i,1}^{t+1} - \pi_{i,1}^t)\end{aligned}$$

Applying the three-point inequality (e.g. Lemma 1 of Zhang & Lan 2020) leads to

$$\begin{aligned}&-\pi_{i,1}^{t+1} g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t) + f^*(\pi_{i,1}^{t+1}) + \tau D_{f^*}(\pi_{i,1}^{t+1},\bar{\pi}_{i,1}^*) + \tau D_{f^*}(\pi_{i,1}^t,\pi_{i,1}^{t+1})\\ &\leq -\bar{\pi}_{i,1}^* g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t) + f^*(\bar{\pi}_{i,1}^*) + \tau D_{f^*}(\pi_{i,1}^t,\bar{\pi}_{i,1}^*), \quad \mathbf{z}_i\in\mathcal{B}_1^t.\end{aligned}$$

Add $\pi_{i,1}^t g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t)$ on both sides and re-arrange the terms. For $\mathbf{z}_i\in\mathcal{B}_1^t$, we have

$$\begin{aligned}&-g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t)(\pi_{i,1}^{t+1} - \pi_{i,1}^t) + f^*(\pi_{i,1}^{t+1})\\ &\leq -g_{i_t}(\mathbf{w}^t;\mathcal{B}_{i,2}^t)(\bar{\pi}_{i,1}^* - \pi_{i,1}^t) + f^*(\bar{\pi}_{i,1}^*) + \tau D_{f^*}(\pi_{i,1}^t,\bar{\pi}_{i,1}^*) - \tau D_{f^*}(\pi_{i,1}^{t+1},\bar{\pi}_{i,1}^*) - \tau D_{f^*}(\pi_{i,1}^t,\pi_{i,1}^{t+1}).\end{aligned}$$

The ¶ term can be upper bounded by summing the inequality above over all $\mathbf{z}_i\in\mathcal{B}_1^t$. Besides, note that $\pi_{i,1}^{t+1} = \pi_{i,1}^t$ for $\mathbf{z}_i\notin\mathcal{B}_1^t$ such that $\sum_{\mathbf{z}_i\notin\mathcal{B}_1^t} f^*(\pi_{i,1}^{t+1}) = \sum_{\mathbf{z}_i\notin\mathcal{B}_1^t} f^*(\pi_{i,1}^t)$.

$$\begin{aligned}h(\mathbf{w}^t,\pi_1^{t+1}) &\leq \phi(\mathbf{w}^t,\pi_1^t) - \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t)(\bar{\pi}_{i,1}^* - \pi_{i,1}^t) + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} f^*(\bar{\pi}_{i,1}^*) + \sum_{\mathbf{z}_i\notin\mathcal{B}_1^t} f^*(\pi_{i,1}^t) + \tau\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} D_{f^*}(\pi_{i,1}^t,\bar{\pi}_{i,1}^*)\\ &\quad - \tau\sum_{\mathbf{z}_{i_t}\in\mathcal{B}_1^t} D_{f^*}(\pi_{i,1}^{t+1},\bar{\pi}_{i,1}^*) - \tau\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} D_{f^*}(\pi_{i,1}^t,\pi_{i,1}^{t+1}) + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \underbrace{(g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t) - g_i(\mathbf{w}^t))(\pi_{i,1}^{t+1} - \pi_{i,1}^t)}_{:=\circledast}\end{aligned}$$

Based on the $\frac{1}{L_f}$-strong convexity of $f^*(\pi_{i,1})$, the $\circledast$ term for $\mathbf{z}_i\in\mathcal{B}_1^t$ can be bounded as

$$\begin{aligned}(g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t) - g_i(\mathbf{w}^t))(\pi_{i,1}^{t+1} - \pi_{i,1}^t) &\leq \frac{L_f}{\tau}\left\|g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t) - g_i(\mathbf{w}^t)\right\|^2 + \frac{\tau}{4L_f}\left\|\pi_{i,1}^{t+1} - \pi_{i,1}^t\right\|^2\\ &\leq \frac{L_f}{\tau}\left\|g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t) - g_i(\mathbf{w}^t)\right\|^2 + \frac{\tau}{2} D_{f^*}(\pi_{i,1}^t,\pi_{i,1}^{t+1}).\end{aligned}$$

Taking the upper bound of $\circledast$ into consideration leads to

$$\begin{aligned}h(\mathbf{w}^t,\pi_1^{t+1}) &\leq \phi(\mathbf{w}^t,\pi_1^t) \underbrace{- \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t)(\bar{\pi}_{i,1}^* - \pi_{i,1}^t)}_{:=\heartsuit} + \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} f^*(\bar{\pi}_{i,1}^*) + \sum_{\mathbf{z}_i\notin\mathcal{B}_1^t} f^*(\pi_{i,1}^t) - \frac{\tau}{2}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} D_{f^*}(\pi_{i,1}^t,\pi_{i,1}^{t+1})\\ &\quad \underbrace{+ \tau\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} D_{f^*}(\pi_{i,1}^t,\bar{\pi}_{i,1}^*) - \tau\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} D_{f^*}(\pi_{i,1}^{t+1},\bar{\pi}_{i,1}^*)}_{:=\blacklozenge} + \frac{L_f}{\tau}\sum_{i\in\mathcal{B}_1^t}\left\|g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t) - g_i(\mathbf{w}^t)\right\|^2.\end{aligned}$$

Define that $\hat{g}(\mathbf{w}^t;\mathcal{B}_2^t) = \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} \frac{n}{B_1} g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t)\mathbf{e}_i$ and $g(\mathbf{w}^t) = \sum_{\mathbf{z}_i\in\mathcal{D}} g_i(\mathbf{w}^t)\mathbf{e}_i$, where $\mathbf{e}_i\in\mathbb{R}^n$ is the indicator vector that only the i-th element is 1 while the others are 0. Note that $\mathbb{E}\left[\hat{g}(\mathbf{w}^t;\mathcal{B}_2^t)\right] = g(\mathbf{w}^t)$. Then, ❤ can be decomposed as

$$
\begin{aligned}
❤ &= -\frac{B_1}{n}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\frac{n}{B_1}g_i(\mathbf{w}^t;\mathcal{B}_{i,2}^t)(\bar{\pi}_{i,1}^* - \pi_{i,1}^t)\\
&= -\frac{B_1}{n}\left\langle \hat{g}(\mathbf{w}^t;\mathcal{B}_2^t), \bar{\pi}_1^* - \pi_1^t\right\rangle\\
&= -\frac{B_1}{n}\left\langle g(\mathbf{w}^t), \bar{\pi}_1^* - \pi_1^t\right\rangle - \frac{B_1}{n}\left\langle \hat{g}(\mathbf{w}^t;\mathcal{B}_2^t) - g(\mathbf{w}^t), \bar{\pi}_1^* - \pi_1^t\right\rangle\\
&= -\frac{B_1}{n}\left\langle g(\mathbf{w}^t), \bar{\pi}_1^* - \pi_1^t\right\rangle - \frac{B_1}{n}\underbrace{\left\langle \hat{g}(\mathbf{w}^t;\mathcal{B}_2^t) - g(\mathbf{w}^t), \bar{\pi}_1^* - \tilde{\pi}_1^t\right\rangle}_{:=♣} - \frac{B_1}{n}\underbrace{\left\langle \hat{g}(\mathbf{w}^t;\mathcal{B}_2^t) - g(\mathbf{w}^t), \tilde{\pi}_1^t - \pi_1^t\right\rangle}_{:=❦},
\end{aligned}
$$

where $\tilde{\pi}_1^t$ is defined as

$$
\begin{aligned}
\tilde{\pi}_1^t &= \arg\min_{\pi_1}\left\langle \hat{g}(\mathbf{w}^t;\tilde{\mathcal{B}}_2^t) - g(\mathbf{w}^t), \pi_1\right\rangle + \tau'\mathbf{D}_{f^*}(\hat{\pi}_1^t, \pi_1),\\
\hat{\pi}_1^{t+1} &= \arg\min_{\pi_1}\left\langle \hat{g}(\mathbf{w}^t;\mathcal{B}_2^t) - g(\mathbf{w}^t), \pi_1\right\rangle + \tau'\mathbf{D}_{f^*}(\hat{\pi}_1^t, \pi_1),
\end{aligned}
$$

where $\tau' > 0$, $\tilde{\mathcal{B}}_2^t$ is a "virtual batch" (never sampled in the algorithm) that is independent of but has the same size as $\mathcal{B}_2^t$. Based on Lemma 8, ♣ can be lower bounded as

$$
\left\langle \hat{g}(\mathbf{w}^t;\mathcal{B}_2^t) - g(\mathbf{w}^t), \bar{\pi}_1^* - \tilde{\pi}_1^t\right\rangle \geq -\tau'\mathbf{D}_{f^*}(\hat{\pi}_1^t, \bar{\pi}_1^*) + \tau'\mathbf{D}_{f^*}(\hat{\pi}_1^{t+1}, \bar{\pi}_1^*) - \frac{L_f}{2(\tau')^2}\left\|\hat{g}(\mathbf{w}^t;\mathcal{B}_2^t) - \hat{g}(\mathbf{w}^t;\tilde{\mathcal{B}}_2^t)\right\|^2.
$$

Thus, taking the expectation of the equation above w.r.t. the randomness in iteration t leads to

$$
\mathbb{E}_t\left[\frac{B_1}{n}\left\langle \hat{g}(\mathbf{w}^t;\mathcal{B}_2^t) - g(\mathbf{w}^t), \bar{\pi}_1^* - \tilde{\pi}_1^t\right\rangle\right] \geq \frac{\tau' B_1}{n}\left(-\mathbb{E}_t\left[\mathbf{D}_{f^*}(\hat{\pi}_1^t, \bar{\pi}_1^*)\right] + \mathbb{E}_t\left[\mathbf{D}_{f^*}(\hat{\pi}_1^{t+1}, \bar{\pi}_1^*)\right]\right) - \frac{B_1\sigma^2 L_f}{(\tau')^2 B_2}.
$$

Note that $\mathbb{E}_t\left[❦\right] = \mathbb{E}_t\left[\langle \hat{g}(\mathbf{w}^t;\mathcal{B}_2^t) - g(\mathbf{w}^t), \tilde{\pi}_1^t - \pi_1^t\rangle\right] = 0$ since both π_1^t and $\tilde{\pi}_1^t$ are independent of $\mathcal{B}_2^t$. Besides, we have $♦ = \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} D_{f^*}(\pi_{i,1}^t, \bar{\pi}_{i,1}^*) - \sum_{\mathbf{z}_i\in\mathcal{B}_1^t} D_{f^*}(\pi_{i,1}^{t+1}, \bar{\pi}_{i,1}^*) = \sum_{\mathbf{z}_i\in\mathcal{D}} D_{f^*}(\pi_{i,1}^t, \bar{\pi}_{i,1}^*) - \sum_{\mathbf{z}_i\in\mathcal{D}} D_{f^*}(\pi_{i,1}^{t+1}, \bar{\pi}_{i,1}^*) = \mathbf{D}_{f^*}(\pi_1^t, \bar{\pi}_1^*) - \mathbf{D}_{f^*}(\pi_1^{t+1}, \bar{\pi}_1^*)$ and $\sum_{\mathbf{z}_i\in\mathcal{B}_1^t} D_{f^*}(\pi_{i,1}^t, \pi_{i,1}^{t+1}) = \sum_{\mathbf{z}_i\in\mathcal{D}} D_{f^*}(\pi_{i,1}^t, \pi_{i,1}^{t+1}) = \mathbf{D}_{f^*}(\pi_1^t, \pi_1^{t+1})$ because $\pi_{i,1}^{t+1} = \pi_{i,1}^t$ for $i \notin \mathcal{B}_1^t$.

$$
\begin{aligned}
&\mathbb{E}_t\left[h(\mathbf{w}^t, \pi_1^{t+1})\right]\\
&\leq \phi(\mathbf{w}^t, \pi_1^t) + \left(1 - \frac{B_1}{n}\right)\mathbf{f}^*(\pi_1^t) + \frac{B_1(\mathbb{E}_t\left[\phi(\mathbf{w}^t, \bar{\pi}_1^*)\right] - \phi(\mathbf{w}^t, \pi_1^t))}{n} + \frac{B_1}{n}\mathbb{E}_t\left[\mathbf{f}^*(\bar{\pi}_1^*)\right] - \frac{\tau}{2}\mathbb{E}_t\left[\mathbf{D}_{f^*}(\pi_1^t, \pi_1^{t+1})\right]\\
&+ \frac{\tau' B_1}{n}(\mathbb{E}_t\left[\mathbf{D}_{f^*}(\hat{\pi}_1^t, \bar{\pi}_1^*)\right] - \mathbb{E}_t\left[\mathbf{D}_{f^*}(\hat{\pi}_1^{t+1}, \bar{\pi}_1^*)\right]) + \frac{nL_f\sigma^2}{(\tau')^2 B_2} + \tau\mathbb{E}_t\left[\mathbf{D}_{f^*}(\pi_1^t, \bar{\pi}_1^*)\right] - \tau\mathbb{E}_t\left[\mathbf{D}_{f^*}(\pi_1^{t+1}, \bar{\pi}_1^*)\right] + \frac{L_f B_1\sigma^2}{\tau B_2}\\
&= \left(1 - \frac{B_1}{n}\right)h(\mathbf{w}^t, \pi_1^t) + \frac{B_1}{n}\mathbb{E}_t\left[h(\mathbf{w}^t, \bar{\pi}_1^*)\right] + \tau(\mathbb{E}_t\left[\mathbf{D}_{f^*}(\pi_1^t, \bar{\pi}_1^*)\right] - \mathbb{E}_t\left[\mathbf{D}_{f^*}(\pi_1^{t+1}, \bar{\pi}_1^*)\right]) - \frac{\tau}{2}\mathbb{E}_t\left[\mathbf{D}_{f^*}(\pi_1^t, \pi_1^{t+1})\right]\\
&+ \frac{\tau' B_1}{n}(\mathbb{E}_t\left[\mathbf{D}_{f^*}(\hat{\pi}_1^t, \bar{\pi}_1^*)\right] - \mathbb{E}_t\left[\mathbf{D}_{f^*}(\hat{\pi}_1^{t+1}, \bar{\pi}_1^*)\right]) + \frac{nL_f\sigma^2}{(\tau')^2 B_2} + \tau\mathbb{E}_t\left[\mathbf{D}_{f^*}(\pi_1^t, \bar{\pi}_1^*)\right] - \tau\mathbb{E}_t\left[\mathbf{D}_{f^*}(\pi_1^{t+1}, \bar{\pi}_1^*)\right] + \frac{L_f B_1\sigma^2}{\tau B_2}.
\end{aligned}
$$

Subtract $\mathbb{E}_t\left[h(\mathbf{w}^t, \bar{\pi}_1^*)\right]$ from both sides and use the tower property of conditional expectation.

$$
\begin{aligned}
\mathbb{E}\left[h(\mathbf{w}^t, \pi_1^{t+1}) - h(\mathbf{w}^t, \bar{\pi}_1^*)\right] &\leq \left(1 - \frac{B_1}{n}\right)\mathbb{E}\left[h(\mathbf{w}^t, \pi_1^t) - h(\mathbf{w}^t, \bar{\pi}_1^*)\right] + \tau\mathbb{E}\left[\mathbf{D}_{f^*}(\pi_1^t, \bar{\pi}_1^*) - \mathbf{D}_{f^*}(\pi_1^{t+1}, \bar{\pi}_1^*)\right]\\
&\quad - \frac{\tau}{2}\mathbb{E}\left[\mathbf{D}_{f^*}(\pi_1^t, \pi_1^{t+1})\right] + \frac{L_f B_1\sigma^2}{\tau B_2} + \frac{L_f B_1\sigma^2}{(\tau')^2 B_2} + \frac{\tau' B_1}{n}\mathbb{E}\left[\mathbf{D}_{f^*}(\hat{\pi}_1^t, \bar{\pi}_1^*) - \mathbf{D}_{f^*}(\hat{\pi}_1^{t+1}, \bar{\pi}_1^*)\right].
\end{aligned}
$$

Let $\Delta^t = h(\mathbf{w}^t, \pi_1^{t+1}) - h(\mathbf{w}^t, \bar{\pi}_1^*)$. Thus,

$$
\begin{aligned}
&\mathbb{E}\left[h(\mathbf{w}^t, \pi_1^t) - h(\mathbf{w}^t, \bar{\pi}_1^*) - \Delta^{t-1}\right] = \mathbb{E}\left[h(\mathbf{w}^t, \pi_1^t) - h(\mathbf{w}^t, \bar{\pi}_1^*) - h(\mathbf{w}^{t-1}, \pi_1^t) + h(\mathbf{w}^{t-1}, \bar{\pi}_1^*)\right]\\
&= \mathbb{E}\left[\sum_{\mathbf{z}_i\in\mathcal{D}}(\bar{\pi}_{i,1}^* - \pi_{i,1}^t)(g_i(\mathbf{w}^t) - g_i(\mathbf{w}^{t-1}))\right] \leq nC_f C_g\eta\mathbb{E}\left[\left\|\mathbf{w}^t - \mathbf{w}^{t-1}\right\|\right] \leq n\eta C_f C_g\sqrt{C_f^2(C_g^2 + \zeta^2/B_2)}.
\end{aligned}
$$

We define $C = C_f C_g \sqrt{C_f^2(C_g^2 + \zeta^2/B_2)}$. Do the telescoping sum for $t = 1, \dots, T$.

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \frac{B_1}{n}\Delta^t\right]$$
$$\leq (\Delta^0 - \Delta^T) + n\eta CT + \tau \mathbf{D}_{f^*}(\pi_1^0, \bar{\pi}_1^*) - \frac{\tau}{2}\mathbb{E}\left[\sum_{t=0}^{T-1} \mathbf{D}_{f^*}(\pi_1^t, \pi_1^{t+1})\right] + \frac{L_f B_1 \sigma^2 T}{\tau B_2} + \frac{B_1 T L_f \sigma^2}{(\tau')^2 B_2} + \frac{\tau B_1}{n}\mathbf{D}_{f^*}(\hat{\pi}_1^0, \bar{\pi}_1^*).$$

Consider that $\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} Q_{i,1}(\mathbf{y}^t, \mathbf{y}^*) = \frac{1}{n}\left(h(\mathbf{w}^t, \pi_1^{t+1}) - h(\mathbf{w}^t, \bar{\pi}_1^*)\right) = \frac{\Delta^t}{n}$.

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} Q_{i,1}(\mathbf{y}^t, \mathbf{y}^*)\right] \leq \frac{\Delta^0 - \Delta^T}{B_1} + \frac{n\eta CT}{B_1} + \frac{\tau \mathbf{D}_{f^*}(\pi_1^0, \bar{\pi}_1^*)}{B_1} + \frac{L_f T\sigma^2}{\tau B_2} + \frac{T L_f \sigma^2}{(\tau')^2 B_2} + \frac{\tau' \mathbf{D}_{f^*}(\hat{\pi}_1^0, \bar{\pi}_1^*)}{n}$$
$$- \frac{\tau}{2B_1}\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbf{D}_{f^*}(\pi_1^t, \pi_1^{t+1})\right].$$

The numerator in the first term on the right hand side can be upper bounded as follows.

$$\begin{aligned}
\Delta^0 - \Delta^T &= h(\mathbf{w}^0, \pi_1^1) - h(\mathbf{w}^0, \bar{\pi}_1^*) - h(\mathbf{w}^T, \pi_1^{T+1}) + h(\mathbf{w}^T, \bar{\pi}_1^*) \\
&= -\sum_{\mathbf{z}_i\in\mathcal{D}} (\pi_{i,1}^1 - \bar{\pi}_{i,1}^*) g_i(\mathbf{w}^0) - \sum_{\mathbf{z}_i\in\mathcal{D}} (\bar{\pi}_{i,1}^* - \pi_{i,1}^{T+1}) g_i(\mathbf{w}^T) + \sum_{\mathbf{z}_i\in\mathcal{D}} f^*(\pi_{i,1}^1) - \sum_{\mathbf{z}_i\in\mathcal{D}} f^*(\pi_{i,1}^{T+1}) \\
&\leq 2nC_f C_g C_\Omega + 2nC_{f^*},
\end{aligned}$$

On the other hand, if we set $\hat{\pi}_{i,1}^0 = \nabla f(u_i^0)$ and $u_i^0 = g_i(\mathbf{w}^0)$, we have $D_{f^*}(\pi_{i,1}^0, \bar{\pi}_{i,1}^*) = D_{f^*}(\nabla f(u_i^0), \nabla f(g_i(\bar{\mathbf{w}}))) = D_f(u_i^0, g_i(\bar{\mathbf{w}})) \leq \frac{L_f}{2}\left\|u_i^0 - g_i(\bar{\mathbf{w}})\right\|^2 \leq \frac{L_f C_g^2 C_\Omega^2}{2}$ such that $\mathbf{D}_{f^*}(\pi_1^0, \bar{\pi}_1^*) = \sum_{\mathbf{z}_i\in\mathcal{D}} D_{f^*}(\pi_{i,1}^0, \bar{\pi}_{i,1}^*) \leq \frac{nL_f C_g^2 C_\Omega^2}{2}$. The proof concludes by setting $\tau' = \frac{n\tau}{B_1} > 1$. □

Lemma 11. *We have*

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} Q_{i,0}(\mathbf{y}^t, \mathbf{y}^*)\right] \leq \eta T C_f^2 C_g^2 + \frac{\eta T C_f^2(\zeta^2 + C_g^2)}{\min\{B_1, B_2\}} + \frac{C_\Omega^2}{2\eta} + \frac{\tau}{2B_1}\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbf{D}_{f^*}(\pi_1^t, \pi_1^{t+1})\right] + \frac{B_1 L_f C_g^2 C_\Omega^2 T}{n\tau}.$$

Proof. Based on the definition of $Q_{i,0}(\mathbf{y}^t, \mathbf{y}^*)$, we can derive that

$$\begin{aligned}
\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} Q_{i,0}(\mathbf{y}^t, \mathbf{y}^*)\right] &= \sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\langle \pi_{i,1}^{t+1}\pi_{i,2}^{t+1}, \mathbf{w}^t - \mathbf{w}^*\right\rangle\right] \\
&= \sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\langle \pi_{i,1}^{t}\pi_{i,2}^{t+1}, \mathbf{w}^{t+1} - \mathbf{w}^*\right\rangle\right] + \sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\langle \pi_{i,1}^{t+1}\pi_{i,2}^{t+1}, \mathbf{w}^t - \mathbf{w}^{t+1}\right\rangle\right] \\
&\quad + \sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\langle (\pi_{i,1}^{t+1} - \pi_{i,1}^t)\pi_{i,2}^{t+1}, \mathbf{w}^{t+1} - \mathbf{w}^*\right\rangle\right]. \qquad (18)
\end{aligned}$$

The second term on the right hand side of (18) can be upper bounded by

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\langle \pi_{i,1}^{t+1}\pi_{i,2}^{t+1}, \mathbf{w}^t - \mathbf{w}^{t+1}\right\rangle\right] \leq \mathbb{E}\left[\sum_{t=0}^{T-1}\left(\eta C_f^2 C_g^2 + \frac{1}{4\eta}\left\|\mathbf{w}^{t+1} - \mathbf{w}^t\right\|^2\right)\right].$$

The first term on the right hand side of (18) can be upper bounded by

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\langle\pi_{i,1}^t\pi_{i,2}^{t+1},\mathbf{w}^{t+1}-\mathbf{w}^*\right\rangle\right] &= \sum_{t=0}^{T-1}\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\pi_{i,1}^t\pi_{i,2}^{t+1}-\frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t}\pi_{i,1}^t\pi_{i,2}^{t+1}(\mathcal{B}_{i,2}^t),\mathbf{w}^{t+1}-\mathbf{w}^*\right\rangle\right]\\ &\quad+\sum_{t=0}^{T-1}\mathbb{E}\left[\left\langle\frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t}\pi_{i,1}^t\pi_{i,2}^{t+1}(\mathcal{B}_{i,2}^t),\mathbf{w}^{t+1}-\mathbf{w}^*\right\rangle\right]\\ &= \sum_{t=0}^{T-1}\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\pi_{i,1}^t\pi_{i,2}^{t+1}-\frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t}\pi_{i,1}^t\pi_{i,2}^{t+1}(\mathcal{B}_{i,2}^t),\mathbf{w}^{t+1}-\mathbf{w}^t\right\rangle\right]\\ &\quad+\sum_{t=0}^{T-1}\mathbb{E}\left[\left\langle\frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t}\pi_{i,1}^t\pi_{i,2}^{t+1}(\mathcal{B}_{i,2}^t),\mathbf{w}^{t+1}-\mathbf{w}^*\right\rangle\right].\end{aligned}$$

The last equality above uses $\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\pi_{i,1}^t\pi_{i,2}^{t+1}-\frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t}\pi_{i,1}^t\pi_{i,2}^{t+1}(\mathcal{B}_{i,2}^t),\mathbf{w}^t-\mathbf{w}^*\right\rangle\right]=0$. Moreover,

$$\mathbb{E}\left[\left\langle\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\pi_{i,1}^t\pi_{i,2}^{t+1}-\frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t}\pi_{i,1}^t\pi_{i,2}^{t+1}(\mathcal{B}_{i,2}^t),\mathbf{w}^{t+1}-\mathbf{w}^t\right\rangle\right]\leq\frac{\eta C_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}}+\frac{1}{4\eta}\left\|\mathbf{w}^{t+1}-\mathbf{w}^t\right\|^2. \tag{19}$$

According to the three-point inequality (Lemma 1 in Zhang & Lan 2020), we have

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\left\langle\frac{1}{B_1}\sum_{i\in\mathcal{B}_1^t}\pi_{i,1}^t\pi_{i,2}^{t+1}(\mathcal{B}_{i,2}^t),\mathbf{w}^{t+1}-\mathbf{w}^*\right\rangle\right]\leq\frac{1}{2\eta}\left\|\mathbf{w}^t-\mathbf{w}^*\right\|^2-\frac{1}{2\eta}\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2-\frac{1}{2\eta}\left\|\mathbf{w}^t-\mathbf{w}^{t+1}\right\|^2. \tag{20}$$

Besides, the third term on the R.H.S. of (18) can be upper bounded as follows based on the Young's inequality with a constant $\rho>0$.

$$\begin{aligned}&\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\langle(\pi_{i,1}^{t+1}-\pi_{i,1}^t)\pi_{i,2}^{t+1},\mathbf{w}^{t+1}-\mathbf{w}^*\right\rangle\right]\leq\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{C_g^2\rho}{2n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\|\pi_{i,1}^{t+1}-\pi_{i,1}^t\right\|^2+\frac{\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2}{2\rho}\right]\\ &\leq\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{L_fC_g^2\rho}{n}\mathbf{D}_{f^*}(\pi_1^t,\pi_1^{t+1})+\frac{\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2}{2\rho}\right],\end{aligned}$$

where the last inequality is due to L_f-smoothness of f. Choose $\rho=\frac{n\tau}{2B_1L_fC_g^2}$. Besides, we also have $\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2\leq C_\Omega^2$.

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}\left\langle(\pi_{i,1}^{t+1}-\pi_{i,1}^t)\pi_{i,2}^{t+1},\mathbf{w}^{t+1}-\mathbf{w}^*\right\rangle\right]\leq\frac{\tau}{2B_1}\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbf{D}_{f^*}(\pi_1^t,\pi_1^{t+1})\right]+\frac{B_1L_fC_g^2C_\Omega^2T}{n\tau}. \tag{21}$$

Plugging (19), (20), and (21) into (18) leads to

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}}Q_{i,0}(\mathbf{y}^t,\mathbf{y}^*)\right]\leq\eta TC_f^2C_g^2+\frac{\eta TC_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}}+\frac{C_\Omega^2}{2\eta}+\frac{\tau}{2B_1}\mathbb{E}\left[\sum_{t=0}^{T-1}\mathbf{D}_{f^*}(\pi_1^t,\pi_1^{t+1})\right]+\frac{B_1L_fC_g^2C_\Omega^2T}{n\tau}.$$

□

D.1. Proof of Theorem 3

Theorem 5 (Detailed Version of Theorem 3)**.** *Under assumptions 2, 3 and that f is convex, monotone, Lipschitz-continuous, and smooth while g_i is convex and Lipschitz-continuous, SOX with $\beta=1$ can find an $\mathbf{w}$ that $\mathbb{E}[F(\mathbf{w})-F(\mathbf{w}^*)]\leq\epsilon$ after*

$$T=\max\left\{\frac{18n(C_fC_gC_\Omega+C_{f^*})}{B_1\epsilon},\frac{81nL_f^2C_g^2C_\Omega^2\sigma^2}{B_1B_2\epsilon^2},\frac{81L_fC_g^2C_\Omega^2}{B_2\epsilon^2},\frac{81L_f^2C_f^2C_g^4C_\Omega^4}{\epsilon^2},\frac{81nCC_\Omega^2}{2B_1\epsilon^2},\frac{81C_f^2C_g^2C_\Omega^2}{2\epsilon^2},\frac{81C_f^2C_\Omega^2(\zeta^2+C_g^2)}{2\min\{B_1,B_2\}\epsilon^2}\right\}$$

iterations by setting $\eta=\min\left\{\frac{B_1\epsilon}{9nC},\frac{\epsilon}{9C_f^2C_g^2},\frac{\min\{B_1,B_2\}\epsilon}{9C_f^2(\zeta^2+C_g^2)}\right\}$ *and* $\tau=\max\left\{\frac{9L_f\sigma^2}{B_2\epsilon},\frac{9B_1}{nB_2\epsilon},\frac{9B_1L_fC_g^2C_\Omega^2}{n\epsilon}\right\}$.

Proof. Based on Lemma 9, Lemma 10, and Lemma 11, we have

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0} Q(\mathbf{y}^t, \mathbf{y}^*)\right]$$
$$\leq \frac{2n(C_f C_g C_\Omega + C_{f^*})}{B_1 T} + \frac{n\eta C}{B_1} + \frac{\tau n L_f C_g^2 C_\Omega^2}{B_1 T} + \frac{L_f \sigma^2}{\tau B_2} + \frac{B_1}{n\tau B_2} + \eta C_f^2 C_g^2 + \frac{\eta C_f^2(\zeta^2 + C_g^2)}{\min\{B_1, B_2\}} + \frac{C_\Omega^2}{2\eta T} + \frac{B_1 L_f C_g^2 C_\Omega^2}{n\tau},$$

where $F(\bar{\mathbf{w}}^T) = \mathcal{L}(\bar{\mathbf{w}}, \bar{\pi}_1^*, \bar{\pi}_2^*)$, $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{w}^t$ and $\mathcal{L}_{i,1}(\mathbf{w}^*, \pi_{i,1}^{t+1}, \pi_{i,2}^{t+1}) \leq f(g_i(\mathbf{w}^*))$ such that we have $\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \mathcal{L}_{i,1}(\mathbf{w}^*, \pi_{i,1}^{t+1}, \pi_{i,2}^{t+1}) \leq \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f(g_i(\mathbf{w}^*)) = F(\mathbf{w}^*)$.

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0} Q(\mathbf{y}^t, \mathbf{y}^*)\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=0}(\mathcal{L}(\mathbf{w}^t, \hat{\pi}_1, \hat{\pi}_2) - \mathcal{L}(\mathbf{w}^*, \pi_1^{t+1}, \pi_2^{t+1}))\right] \geq \mathbb{E}\left[F(\bar{\mathbf{w}}) - F(\mathbf{w}^*)\right]$$

Thus, to ensure that $\mathbb{E}\left[F(\bar{\mathbf{w}}) - F(\mathbf{w}^*)\right] \leq \epsilon$, we can make

$$\eta = \min\left\{\frac{B_1\epsilon}{9nC}, \frac{\epsilon}{9C_f^2C_g^2}, \frac{\min\{B_1,B_2\}\epsilon}{9C_f^2(\zeta^2 + C_g^2)}\right\}, \quad \tau = \max\left\{\frac{9L_f\sigma^2}{B_2\epsilon}, \frac{9B_1}{nB_2\epsilon}, \frac{9B_1 L_f C_g^2 C_\Omega^2}{n\epsilon}\right\},$$
$$T = \max\left\{\frac{18n(C_fC_gC_\Omega + C_{f^*})}{B_1\epsilon}, \frac{81nL_f^2C_g^2C_\Omega^2\sigma^2}{B_1B_2\epsilon^2}, \frac{81L_fC_g^2C_\Omega^2}{B_2\epsilon^2}, \frac{81L_f^2C_g^4C_\Omega^4}{\epsilon^2}, \frac{81nCC_\Omega^2}{2B_1\epsilon^2}, \frac{81C_f^2C_g^2C_\Omega^2}{2\epsilon^2}, \frac{81C_f^2C_\Omega^2(\zeta^2+C_g^2)}{2\min\{B_1,B_2\}\epsilon^2}\right\}.$$

□

E. Extensions for A More General Class of Problems

In this section, we briefly discuss the extension when f_i is also a stochastic function such that we can only get an unbiased estimate of its gradient, which has an application in MAML. To this end, we assume a stochastic oracle of f_i that given any $g(\cdot)$ returns $\nabla f_i(g(\cdot);\iota)$ such that $\mathbb{E}[\nabla f_i(g(\cdot);\iota)] = \nabla f_i(g)$, $\mathbb{E}[\|\nabla f_i(g(\cdot);\iota) - \nabla f_i(g(\cdot))\|^2] \leq \chi^2$. We can extend our results for the smooth nonconvex problems by the modifications as follows: First, we need to assume that $\nabla f_i(\cdot;\iota)$ is Lipschitz-continuous; Second, the $\frac{2\beta^2C_f^2(\zeta^2+C_g^2)}{\min\{B_1,B_2\}}$ term in Lemma 3 should be replaced by $\frac{2\beta^2(\chi^2C_g^2+(C_f^2+\chi^2/B_3))(\zeta^2+C_g^2)}{\min\{B_1,B_2,B_3\}}$, where B_3 is the batch size for sampling ι. Note that Lemma 1 remains the same and Theorem 1 does not change (up to a constant factor).