An Online Method for A Class of Distributionally Robust Optimization with Non-Convex Objectives

Qi Qi†∗, Zhishuai Guo†∗, Yi Xu‡, Rong Jin†, Tianbao Yang†
† Department of Computer Science, University of Iowa, Iowa City, IA 52242
‡ Machine Intelligence Technology, Alibaba Group
{qi-qi, zhishuai-guo, tianbao-yang}@uiowa.edu, {yixu, jinrong.jr}@alibaba-inc.com

Abstract

In this paper, we propose a practical online method for solving a class of distributionally robust optimization (DRO) with non-convex objectives, which has important applications in machine learning for improving the robustness of neural networks. In the literature, most methods for solving DRO are based on stochastic primal-dual methods. However, primal-dual methods for DRO suffer from several drawbacks: (1) manipulating a high-dimensional dual variable corresponding to the size of data is time expensive; (2) they are not friendly to online learning where data is coming sequentially. To address these issues, we consider a class of DRO with an KL divergence regularization on the dual variables, transform the min-max problem into a compositional minimization problem, and propose practical duality-free online stochastic methods without requiring a large mini-batch size. We establish the state-of-the-art complexities of the proposed methods with and without a Polyak-Łojasiewicz (PL) condition of the objective. Empirical studies on large-scale deep learning tasks (i) demonstrate that our method can speed up the training by more than 2 times than baseline methods and save days of training time on a large-scale dataset with ∼ 265K images, and (ii) verify the supreme performance of DRO over Empirical Risk Minimization (ERM) on imbalanced datasets. Of independent interest, the proposed method can be also used for solving a family of stochastic compositional problems with state-of-the-art complexities.

1 Introduction

Distributionally robust optimization (DRO) has received tremendous attention in machine learning due to its capability to handle noisy data, adversarial data and imbalanced classification data [42, 33, 4]. Given a set of observed data \( \{z_1, \ldots, z_n\} \), where \( z_i = (x_i, y_i) \), a DRO formulation can be written as:

\[
\min_{w \in \mathbb{R}^d} \max_{p \in \Delta_n} F_p(w) = \sum_{i=1}^{n} p_i \ell(w; z_i) - h(p, 1/n) + r(w),
\]

where \( w \) denotes the model parameter, \( \Delta_n = \{p \in \mathbb{R}^n : \sum_i p_i = 1, p_i \geq 0\} \) denotes a \( n \)-dimensional simplex, \( \ell(w; z) \) denotes a loss function on data \( z \), \( h(p, 1/n) \) is a divergence measure between \( p \) and uniform probabilities \( 1/n \), and \( r(w) \) is convex regularizer of \( w \). When \( \ell(w; z) \) is a convex function (e.g., for learning a linear model), many stochastic primal-dual methods can be employed for solving the above min-max problem [35][20][49][48][32]. When \( \ell(w; z) \) is a non-convex function (e.g., for learning a deep neural network), some recent studies also proposed stochastic primal-dual methods [41][48].

*The first two authors make equal contributions. Correspondence to tianbao-yang@uiowa.edu.
However, stochastic primal-dual methods for solving DRO problems with a non-convex $\ell(w; z)$ loss function (e.g., the predictive model is a deep neural network) suffer from several drawbacks. First, primal-dual methods need to maintain and update a high-dimensional dual variable $p \in \mathbb{R}^n$ for large-scale data, whose memory cost is as high as $O(n)$ per-iteration. Second, existing primal-dual methods usually need to sample data according to probabilities $p$ in order to update $w$, which brings additional costs than random sampling. Although random sampling can be used for computing the stochastic gradient in terms of $w$, the resulting stochastic gradient could have $n$-times larger variance than using non-uniform sampling according to $p$ (please refer to the supplement for a simple illustration). Third, due to the constraint on $p \in \Delta_n$, the min-max formulation (1) is not friendly to online learning in which the data is received sequentially and $n$ is rarely known in prior.

**Can we design an efficient online algorithm to address the DRO formulation (1) without dealing with $p \in \mathbb{R}^n$ for a non-convex objective that is applicable to deep learning?**

To address this question, we restrict our attention to a family of DRO problems, in which the KL divergence $h(p, 1/n) = \lambda \sum_i p_i \log(n p_i)$ is used for regularizing the dual variables $p$, where $\lambda > 0$ is a regularization parameter. We note that this consideration does not impose strong restriction to the modeling capability. It has been shown that for a family of divergence functions $h(p, 1/n)$, different DRO formulations are statistically equivalent to a certain degree [12]. The proposed method is based on an equivalent minimization formulation for $h(p, 1/n) = \lambda \sum_i p_i \log(n p_i)$. In particular, by maximizing over $p$ exactly, (1) is equivalent to

$$\min_{w \in \mathbb{R}^d} \left\{ F_{dro}(w) = \lambda \log \left(1/n \sum_{i=1}^n \exp(\ell(w; z_i)/\lambda)\right) + r(w) \right\}.$$  

(2)

In an online learning setting, we can consider a more general formulation:

$$\min_{w \in \mathbb{R}^d} \left\{ F_{dro}(w) = \lambda \log (E_{z} \exp (\ell(w; z)/\lambda)) + r(w) \right\}.$$  

(3)

The above problem is an instance of stochastic compositional problems of the following form:

$$\min_{w \in \mathbb{R}^d} F(w) := f(E_{z}[g_{z}(w)]) + r(w),$$  

(4)

by setting $f(s) = \lambda \log(s), s \geq 1$ and $g_{z}(w) = \exp(\ell(w; z)/\lambda)$. Stochastic algorithms have been developed for solving the above compositional problems. [44] proposed the first stochastic algorithms for solving (4), which are easy to implement. However, their sample complexities are sub-optimal for solving (4). Recently, a series of works have tried to improve the convergence rate by using advanced variance reduction techniques (e.g., SVRG [19], SPIDER [13], SARAH [36]). However, most of them require using a mega mini-batch size in the order of $O(n)$ or $O(1/\epsilon)$ at every iteration or many iterations for updating $w$, which hinders their applications on large-scale problems. In addition, these algorithms usually use a constant step size, which may harm the generalization performance.

This paper aims to develop more practical stochastic algorithms for solving (4) without suffering from the above issues in order to enable practitioners to explore the capability of DRO for deep learning with irregular data (e.g., imbalanced data, noisy data). To this end, we proposed an online stochastic method (COVER) and its restarted variant (RECOVER). We establish a state-of-the-art complexity of COVER for finding an $\epsilon$-stationary solution and a state-of-the-art complexity of RECOVER under a Polyak–Łojasiewicz (PL) condition of the problem. PL condition has been widely explored for developing practical optimization algorithms for deep learning [52]. Compared with other stochastic algorithms, the practical advantages of RECOVER are:

1. **RECOVER** is an online duality-free algorithm for addressing large-scale KL regularized DRO problem that is independent of the high dimensional dual variable $p \in \mathbb{R}^n$, which makes it suitable for deep learning applications.

2. **RECOVER** also enjoys the benefits of stagewise training similar to existing stochastic methods for deep learning [52], i.e., the step size is decreased geometrically in a stagewise manner.

$^2\epsilon$ is either the objective gap accuracy $F(w) - \min F(w) \leq \epsilon$ or the gradient norm square bound $\|\nabla F(w)\|^2 \leq \epsilon$
Table 1: Summary of properties of state-of-the-art algorithms for solving our DRO problem. The sample complexity is measured in terms of finding an $\epsilon$-stationary point w/o PL condition, i.e., $\|\nabla F(w)\|^2 \leq \epsilon$, for a class of two-level compositional problems in the order of $O(1/\epsilon^2)$ without a large mini-batch size, which is better than existing results [44, 45, 14, 5].

1. We establish a nearly optimal complexity for finding $\epsilon$-stationary point, i.e., $\|\nabla F(w)\|^2 \leq \epsilon$, for a class of two-level compositional problems in the order of $O(1/\epsilon^2)$ without a large mini-batch size, which is better than existing results [44, 45, 14, 5].

2. We establish an optimal complexity for finding $\epsilon$-optimal solution under an $\mu$−PL condition for a class of two-level compositional problems in the order of $O(1/(\mu\epsilon))$ without a large mini-batch size, which is better than existing results [53].

A theoretical comparison between our results and existing results is shown in Table 1. Empirical studies vividly demonstrate the effectiveness of RECOVER for deep learning on imbalanced data.

2 Related Work

DRO has been extensively studied in machine learning [31, 11, 40], statistics, and operations research [42]. In [33], the authors proved that minimizing the DRO formulation with a quadratic regularization in a constraint form is equivalent to minimizing the sum of the empirical loss and a variance regularization defined on itself. Variance regularization can enjoy better generalization error compared with the empirical loss minimization [33], and was also observed to be effective for imbalanced data [33, 50]. Recently, [12] also establishes this equivalence for a broader family of regularization function $h(p, 1/n)$ including the KL divergence.

Several recent studies have developed stochastic primal-dual methods for solving DRO with a non-convex loss function $\ell(w, z)$ assuming it is smooth or weakly convex [44, 20, 30, 48, 41]. They proposed the first primal-dual methods for solving weakly convex concave problems. For online problems, their algorithms for finding an $\epsilon$-stationary solution whose gradient norm square is less than $\epsilon$ have a sample complexity of $O(1/\epsilon^2)$ or $O(1/\epsilon^3)$ with or without leveraging the strong concavity of $h(p, 1/n)$ for finding an $\epsilon$-stationary point. Recently Liu et al. [28] proposed to leverage the PL condition of the objective function to improve the convergence for a non-convex min-max formulation of AUC maximization. Then, a PES-SGDA algorithm is proposed to solve a more general class of non-convex min-max problems by leveraging the PL condition [15]. Both [28] and [15] have used geometrically decreasing step sizes in a stagewise manner. However, their algorithms’ complexity is in the order of $O(1/(\mu\epsilon^2))$, which is worse than $O(1/(\mu\epsilon))$ achieved in this paper. Similarly, [50] also leveraged PL conditions to solve non-convex min-max problems and has a sample complexity of $O(1/(\mu\epsilon^2))$. Nevertheless, the step size of their algorithm is decreased polynomially in the order of $O(1/t)$, which usually yields poor performance for deep learning.

All the methods reviewed above require maintaining and updating both the primal variable $w$ and a high dimensional dual variable $p \in \mathbb{R}^n$. Recently, Levy et al. [23] considered different formulations of DRO, which includes our considered KL-regularized DRO formulation as a special case. Their assumed that the loss function is convex and proposed a stochastic method with a sample complexity $O(1/\epsilon^3)$ for solving the KL-regularized DRO formulation. In contrast, we provide a better sample
complexity in the order of $O(1/\epsilon)$ under a PL condition without convexity assumption. Additionally, their method requires a large batch size in the order of $O(1/\epsilon)$, while our method only requires a constant batch size which is more practical. We also notice that a recent work \cite{24} and its extended version \cite{25} have considered a formulation similar to (2) and proposed a stochastic algorithm. However, their algorithm has a slower convergence rate with an $O(1/\epsilon^2)$ complexity for finding an $\epsilon$-stationary point and an $O(1/(\mu^2\epsilon))$ complexity for finding an $\epsilon$-optimal solution under a PL condition. Our work is a concurrent work appearing online earlier than \cite{24}. To the best of our knowledge, this is the first work trying to solve the non-convex DRO problem with a duality-free stochastic method by formulating the min-max formulation into an equivalent stochastic compositional problem.

There are extensive studies for solving stochastic compositional problems. \cite{44} considered a more general family of stochastic compositional problems and proposed two algorithms. When the objective function is non-convex, their algorithm’s complexity is $O(1/\epsilon^{7/2})$ for finding an $\epsilon$-stationary solution. This complexity was improved in their later works \cite{14}, reducing to $O(1/\epsilon^2)$. When the objective is smooth, several papers proposed to use variance reduction techniques (e.g., SPIDER, SARAH) to improve the complexity for finding a stationary point \cite{53} \cite{18} \cite{55} \cite{27}. The best sample complexity achieved for online problems is $O(1/\epsilon^{3/2})$ \cite{53} \cite{55}. \cite{53} also considered the PL condition for developing a faster algorithm called restarted CIVR, whose sample complexity is $O(1/\mu \epsilon)$. However, these variance reduction-based methods require using a very large mini-batch size at many iterations, which has detrimental influence on training deep neural networks \cite{43}. To address this issue, \cite{9} proposed a new technique called STORM that integrates momentum and the recursive variance reduction technique for solving stochastic smooth non-convex optimization. Their algorithm does not require a large mini-batch size at every iteration and enjoys a sample complexity of $O(\log^{2/3}(1/\epsilon)/\epsilon^{3/2})$ for finding an $\epsilon$-stationary point. However, their algorithm uses a polynomially decreasing step size, which is not practical for deep learning, and is not directly applicable to stochastic composite problems.

3 Preliminaries

In this section, we provide some definitions and assumptions for next section. For more generality, we consider the stochastic compositional problem \cite{4}:

$$\min_{w \in \mathbb{R}^d} F(w) := f(E[z|g_z(w)]) + r(w)$$ (5)

where $g_z(w) : \mathbb{R}^d \rightarrow \mathbb{R}^p$. Define $g(w) = E_z[g_z(w)]$. Let $\| \cdot \|$ denote the Euclidean norm of a vector or the Frobenius norm of a matrix. We make the following standard assumptions regarding the problem (5).

Assumption 1. Let $C_f, L_f, C_g$ and $L_g$ be positive constants. Assume that

(a) $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is a $C_f$-Lipschitz function and its gradient $\nabla f$ is $L_f$-Lipschitz.
(b) $g_z : \mathbb{R}^d \rightarrow \mathbb{R}^p$ satisfies $\mathbb{E}\|g_z(w_1) - g_z(w_2)\|^2 \leq C_g^2\|w_1 - w_2\|^2$ for any $w_1, w_2$ and its Jacobian $\nabla g_z$ satisfies $\mathbb{E}\|\nabla g_z(w_1) - \nabla g_z(w_2)\|^2 \leq L_g^2\|w_1 - w_2\|^2$.
(c) $r : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$ is a convex and lower-semicontinuous function.
(d) $F_* = \inf_w F(w) \geq -\infty$ and $F(w_1) - F_* \leq \Delta_F$ for the initial solution $w_1$.

Remark: When $f(s) = s$ is a linear function, the assumption $\mathbb{E}\|g_z(w_1) - g_z(w_2)\|^2 \leq C_g^2\|w_1 - w_2\|^2$ is not needed. To upper bound continuity and smoothness of function $F$, we denote $L = 2\max\{L_g C_f L_f, C_f^2 L_f^2, C_f^2, C_g^2, C_g^2 L_f^2, C_f, C_g L_f, C_f, C_g L_f, C_g^2, C_g^2 L_f^2\}$ for simple derivation in the appendix.

Assumption 2. Let $\sigma_g$ and $\sigma_g^*$ be positive constants and $\sigma^2 = \sigma_g^2 + \sigma_g^2$. Assume that

$$\mathbb{E}_z\|g_z(w) - g(w)\|^2 \leq \sigma_g^2, \mathbb{E}_z\|\nabla g_z(w) - \nabla g(w)\|^2 \leq \sigma_g^2$$

Remark: We remark how the minimization formulation of DRO problem \cite{3} can satisfy Assumption 1 in particular Assumption 1(a) and (b). In order to satisfy Assumption 1(b), we can define a bounded loss function $\ell(w, z) \in [0, \ell_{\text{max}}]$ and then use a shifted loss $\ell(w; z) = \ell_{\text{max}}$ in (3). Then $g_z(w) = \exp((\ell(w; z) - \ell_{\text{max}})/\lambda)$ is Lipchitz continuous and smooth if $\ell(w; z)$ is Lipchitz and smooth. $f(s) = \lambda \log(s)$ is Lipchitz continuous and smooth since $s \geq \exp(-\ell_{\text{max}}/\lambda)$. 

4
For more generality, we allow for a non-smooth regularizer \( r(\cdot) \) in this section. To handle non-smoothness of \( r \), we can use the proximal operator of \( r \): \[ \text{prox}_{\eta}^r(\bar{w}) = \arg\min_{w} \frac{1}{2} \|w - \bar{w}\|^2 + \eta r(w). \]
When \( r = 0 \), the above operator reduces to the standard Euclidean projection. Correspondingly, we define the proximal gradient measure for the compositional problem \( \tilde{G}_\eta \):
\[ G_\eta(w) = \frac{1}{\eta} (w - \text{prox}_{\eta}^r(w - \eta \nabla g(w) \nabla f(w))). \]
When \( r = 0 \), the proximal gradient reduces to the standard gradient measure, i.e., \( G_\eta(w) = \nabla F(w) \).
To facilitate our discussion, we define sample complexity below.

**Definition 1.** The sample complexity is defined as the number of samples \( z \) in order to achieve \( \mathbb{E}[\|G_\eta(w)\|^2] \leq \epsilon \) for a certain \( \eta > 0 \) or \( \mathbb{E}[F(w) - F_*] \leq \epsilon. \)

## 4 Basic Algorithm: COVER

We present our Algorithm [1] which serves as the foundation for proving the convergence of the objective gap under a PL condition in next section. The convergence results in this section might be of independent interest to those who are interested in convergence analysis without a PL condition. The motivation is to develop a stochastic algorithm with fast convergence in terms of gradient norm. We refer to the algorithm as Compositional Optimal Variance Reduction (COVER). It will be clear shortly why it is called optimal variance reduction. Note that in order to compute a stochastic estimator of the gradient \( f(g(w)) \) given by \( \nabla g(w) \nabla f(g(w)) \), we maintain and update two estimators denoted by \( \{u_t\}_{t=1}^T \) and \( \{v_t\}_{t=1}^T \) sequence, respectively. The \( \{u_t\}_{t=1}^T \) sequence maintains an estimation of \( \{g(w_t)\}_{t=1}^T \) and the \( \{v_t\}_{t=1}^T \) sequence maintains an estimation of \( \{\nabla g(w_t)\}_{t=1}^T \). The strategy of maintaining and updating two individual sequences was first proposed in [44] and has been widely used for solving compositional problems [55] [53]. However, the key difference from previous algorithms lies in the method for updating the two sequences. COVER is inspired by the STORM technique [9]. To understand the update, let us consider update that applied to the DRO problem [3] by let \( f(\cdot) = \lambda \log(\cdot), g_\lambda(\cdot) = \exp(\ell(\cdot)/\lambda) \) and ignoring \( r \) for the moment. Plugging the gradient of \( f(\cdot) \) and \( g_\lambda(\cdot) \), we have
\[ w_{t+1} = w_t - \eta_t \frac{1}{u_t} \bar{v}_t, \]
\[ \bar{v}_{t+1} = \exp\left(-\frac{\ell(w_{t+1}; z_{t+1})}{\lambda}\right) \nabla \ell(w_{t+1}; z_{t+1}) + (1 - a_{t+1})(\bar{v}_t - \exp\left(-\frac{\ell(w_t; z_{t+1})}{\lambda}\right) \nabla \ell(w_t; z_{t+1})), \]
where \( u_t \) becomes a scalar, which is an online variance-reduced estimator of \( \mathbb{E}_u[\exp(\ell(w_t; z)/\lambda)] \), and \( \bar{v}_t \) is a scaled version of \( v_t \), which is an online variance-reduced estimator of \( \mathbb{E}_u[\exp(\ell(w_t; z)/\lambda)] \nabla \ell(w_t; z)]. \)

Finally, we notice that a similar method for updating the \( u_t \) sequence for estimating \( g(w_t) \) has been adopted in a recent work [6]. However, different from the present work they just use an unbiased stochastic gradient to estimate \( \nabla g(w_t) \), which yields a worse convergence rate.

### 4.1 Convergence of Proximal Gradient

In this section, we present the convergence result of COVER.

**Theorem 1.** Assume the Assumption [1] and [2] for any \( C > 0, k = \frac{C\sigma^2}{L}, c = 128L + \sigma^2/(7Lk^3), w = \max((16Lk^3), 2\sigma^2, (\frac{\kappa}{\rho})^3), \) and \( \eta_t = k/(w + \sigma^2 t)^{1/3} \). The output of COVER satisfies
\[ \mathbb{E}[\|G_\eta_t, (w_t)\|^2] \leq \tilde{O} \left( \frac{D_F}{T^{2/3}} + \frac{\sigma^2}{T^{2/3}} \right). \]
where \( t_* \) is sampled from \( \{1, \ldots, T\} \).

**Remark:** Theorem [1] implies that with a polynomially decreasing step size, COVER is able to find an \( \epsilon \)-stationary point, i.e., \( \mathbb{E}[\|G_\eta_t, (w_t)\|^2] \leq \epsilon \) for a regularized objective and \( \mathbb{E}[\|\nabla F(w)\|^2] \leq \epsilon \) for a non-regularized objective, with a near-optimal sample complexity \( \tilde{O}(1/\epsilon^{3/2}) \). Note that the complexity \( \tilde{O}(1/\epsilon^{3/2}) \) is optimal up to a logarithmic factor for making the (proximal) gradient’s norm smaller than \( \epsilon \) in expectation for solving non-convex smooth optimization problems [2].
Algorithm 1: COVER \((w_1, u_1, v_1, \{\eta_i\}, T, PL = \text{False})\)

1: Let \(a_t = c\eta_t^2\)
2: if not PL then
3:   Draw a samples \(z\) and construct the estimates: \(u_1 = g_x(w_1), \ v_1 = \nabla g_x(w_1)\)
4: end if
5: for \(t = 1, \ldots, T - 1\) do
6:   \(w_{t+1} = \text{prox}_{\eta_t}(w_t - \eta_t \nabla f(u_t))\)
7:   Draw a samples \(z_{t+1}\), and update
8:     \(u_{t+1} = g_x(w_{t+1}) + (1 - \alpha_{t+1})(u_t - g_x(w_t))\)
9:     \(v_{t+1} = \nabla g_x(w_{t+1}) + (1 - \alpha_{t+1})(v_t - \nabla g_x(w_t))\)
10: end for
11: Return: \((w_\tau, u_\tau, v_\tau)\) for randomly selected \(\tau \in \{1, \ldots, T\}\).

5 A Practical Variant (RECOVER) under a PL condition

The issue of COVER is that the polynomially decreasing step size is not practical for deep learning applications and obstacles its generalization performance [52]. A stagewise step size is widely and commonly used [17, 22, 52] for deep learning optimization. To this end, we develop a multi-stage REstarted version of COVER, called RECOVER, which uses a geometrically decreasing step size in a stagewise manner. In order to analyze RECOVER, we assume the following PL condition of the objective with a smooth regularization \(r\) term [52].

Assumption 3. \(F(w)\) satisfies the \(\mu\)-PL condition if there exists \(\mu > 0\) such that

\[
2\mu(F(w) - \min_{w \in \mathbb{R}^d} F(w)) \leq \|\nabla F(w)\|^2.
\]  

In the following, we simply consider the objective \(F(w) = f(E_x[g_x(w)])\), where \(r(\cdot)\) is absorbed into \(f(E_x[g_x(w)])\). As a result, \(G_n(w) = \nabla F(w)\).

Although the PL condition has been considered in various papers for developing stagewise algorithms and improving the convergence rate of non-convex optimization [52, 52, 28, 15]. In order to establish the improved rate, we have innovations in twofold (i) at the algorithmic level, we utilize the variance reduction techniques at the inner and outer level without using mega large mini-batch size at any iterations; (ii) at the analysis level, we innovatively prove that the estimation error of the two sequences, \(u\) and \(v\), are decreasing geometrically after a stage (Lemma 3). These innovations at two levels yield the optimal convergence rate in the order of \(O(1/\mu \epsilon)\).

5.1 Theoretical Verification of PL Assumption for KL-regularized DRO

Before presenting the proposed algorithm and its convergence, we discuss how the \(F_{\text{dro}}\) can satisfy Assumption 3. First, we note that a PL condition of the weighted loss implies that of the primal objective.

Lemma 1. Let \(F_p(w) = \sum_{i=1}^n p_i \ell(w; z_i)\). If for any \(p \in \Delta_n\), \(F_p(w)\) satisfies a \(\mu\)-PL condition, then \(F_{\text{dro}}(w) = \lambda \log(1/n) \sum_{i=1}^n \exp(\ell(w; z_i)/\lambda))\) satisfies the \(\mu\)-PL condition.

Remark: The assumption that the weighted loss satisfies a PL condition can be proven for a simple square loss \(\ell(w; z_i) = (w^T x_i - y_i)^2\), where \(z_i = (x_i, y_i)\) consists of a feature vector \(x_i\) and a label \(y_i\). In order to see this, we can write \(F_p(w) = \sum_{i=1}^n (w^T x_i \sqrt{p_i} - y_i \sqrt{p_i})^2 = \|Aw - b\|^2\), where \(A = (x_1 \sqrt{p_1}, \ldots, x_n \sqrt{p_n})^T, b = (y_1 \sqrt{p_1}, \ldots, y_n \sqrt{p_n})^T\). It has been shown in many previous studies that such \(F_p(w)\) satisfies a PL condition [47, 51, 54]. Hence, the above lemma indicates \(F_{\text{dro}}(w)\) satisfies a PL condition.

We can also justify that \(F_{\text{dro}}(w)\) satisfies a PL condition for deep learning with ReLU activation function in a neighborhood around a random initialized point following the result in [1].

Lemma 2. Assume that input \(\{(x_1, y_1), \ldots, (x_n, y_n)\}\) satisfies \(\|x_i\| = 1\ and \|x_i - x_j\| \geq \delta\), where \(x_n \in \mathbb{R}^{d_1}, y_i \in \mathbb{R}^{d_0}\ and \|y_i\| \leq O(1)\). Consider a deep neural network with \(h_{i,0} = \phi(Ax_i), h_{i,1} = \)
Algorithm 2: RECOVER($w_0, \epsilon_0, c$)

1: **Initialization:** Draw a sample $z_0$ and construct the estimates $\mathbf{u}_0 = g_{z_0}(w_0)$, $v_0 = \nabla g_{z_0}(w_0)$
2: for $k = 1, \ldots, K$ do
3: \( (w_k, u_k, v_k) = \text{COVER}(w_{k-1}, u_{k-1}, v_{k-1}, \eta_k, T_k, \text{True}) \)
4: \( \text{change } \eta_k, T_k \text{ according to Theorem } 2 \)
5: end for
6: Return: $w_K$

\( \phi(W_i h_{i,l-1}, l = 1, \ldots, \tilde{L}, \hat{y}_i = B h_{i,\tilde{L}} \) where \( A \in \mathbb{R}^{d_2 \times d_1}, W_l \in \mathbb{R}^{d_2 \times d_2}, B \in \mathbb{R}^{d_0 \times d_2}, \phi \) is the ReLU activation function, and \( \ell(W; z_i) = (\hat{y}_i - y_i)^2 \) is a square loss. Suppose that for any \( W \), \( p_i^* = \exp(\ell(W; z_i)/\lambda) / \sum_{i=1}^{\infty} \exp(\ell(W; z_i)/\lambda) \geq p_0 > 0 \), then with a high probability over randomness of \( W_0, A, B \) for every \( W \) with \( \|W - W_0\| \leq O(1/\text{poly}(n, \tilde{L}, p_0^{-1}, \delta^{-1})) \), there exists a small \( \mu > 0 \) such that \( \|\nabla F_{dro}(W)\|_{F}^2 + O(\epsilon) \geq \mu F_{dro}(W) - \min_W F_{dro}(W) \).

**Remark:** The \( O(\epsilon) \) term in the left side of the PL condition is caused by using the covering net argument for proving the high probability result. Nevertheless, it does not affect the final convergence rate.

### 5.2 Theoretical Analysis of RECOVER

Now, we are ready to present the proposed algorithm under the PL condition and its convergence result. The algorithm is described in Algorithm 2.

The first key feature of RECOVER is equipped with the practical geometrical decreases step size between stages. At each stage, we adopt a constant step size \( \eta_k \) rather than the polynomial decreases step size used by COVER as in Theorem 1. Another key feature of RECOVER is that it uses not only \( w_k \) for restarting but also \( u_k, v_k \), the corresponding online estimator of \( g(w_k) \) and \( \nabla g(w_k) \), for restarting the next stage. It is this feature that allows us to avoid the large batch size required in other variance reduction methods to achieve the optimal sample complexity. With this feature, we can show that the variance of \( u_k, v_k \) is decreased by a constant factor stagewise as shown in the following lemma.

**Lemma 3.** Define constants \( \epsilon_1 = \frac{c^2 \sigma^2}{64 \mu L} \) and \( \epsilon_k = \epsilon_1/2^{k-1} \), with \( \eta_k = \min\{ \sqrt{\frac{\epsilon_1 \mu L}{2c \sigma}}, \frac{1}{16L} \} \), \( T_k = O(\max\{ \frac{96c^2}{\mu^2 + \frac{16c^2}{L^2 \epsilon^2}}, \frac{2c^2}{\mu L^2 \epsilon^2}, \frac{L^2}{\sigma} \}) \), the variance of the stochastic estimator of Algorithm 2 at \( w_k \) satisfies:

\[
\mathbb{E}[\|u_k - g(w_k)\|^2 + \|v_k - \nabla g(w_k)\|^2] \leq \mu \epsilon_k. \tag{8}
\]

With the above lemma and the convergence bound for \( \mathbb{E}[\|\nabla F(w_k)\|^2] \) at the \( k \)-th stage, we can show that the objective gap \( \mathbb{E}[F(w_k) - F_\star] \) is decreased by a factor of 2 after each stage under the PL condition. Hence, we have the following convergence for RECOVER.

**Theorem 2.** Assume that assumptions 1-3 hold. Define constants \( \epsilon_1 = \frac{c^2 \sigma^2}{64 \mu L} \) and \( \epsilon_k = \epsilon_1/2^{k-1} \). By setting \( \eta_k = \min\{ \sqrt{\frac{\epsilon_1 \mu L}{2c \sigma}}, \frac{1}{16L} \} \), \( T_k = O(\max\{ \frac{96c^2}{\mu^2 + \frac{16c^2}{L^2 \epsilon^2}}, \frac{2c^2}{\mu L^2 \epsilon^2}, \frac{L^2}{\sigma} \}) \), \( c = 104L^2 \), then after \( K = O(\log(\epsilon_1/\epsilon)) \) stages, the output of RECOVER satisfies \( \mathbb{E}[F(w_K) - F_\star] \leq \epsilon \).

**Remark:** It is not difficult to derive the sample complexity of RECOVER is \( O(\max\{ \frac{1}{\mu} \sqrt{\frac{1}{2c \sigma}}, \frac{1}{\mu^{3/2}} \}) = O(\frac{1}{\mu \epsilon}) \) for \( \epsilon \leq \mu \). It is notable this complexity is optimal for the considered general stochastic compositional problem, which includes stochastic strongly convex optimization as a special case, whose lower bound is \( O(1/(\mu \epsilon)) \) [16].

In addition, it is notable that the proposed multi-stage algorithm is very different from many other multi-stage algorithms for non-convex optimization that are based on the proximal point framework [2, 43, 44, 45]. In particular, in these previous studies, a quadratic function \( \gamma/2 \|w - w_{k-1}\|^2 \) with an appropriate regularization parameter \( \gamma \) is added into the objective function at the \( k \)-th stage in order to convexify the objective function. In RECOVER, no such regularization is manually added. Nevertheless, we can still obtain strong convergence guarantee.
6 Experimental Results

We focus on the task of classification with imbalanced data in our experiments. Firstly, we compare RECOVER with five State-Of-The-Art (SOTA) baselines from two categories: (i) primal-dual algorithms for solving the primal-dual formulation of DRO \([1]\), and (ii) algorithms that are designed for the stochastic compositional formulation of DRO \([3]\). Secondly, we verify the advantages of DRO over Empirical Risk Minimization (ERM) for imbalanced data problems by comparing the test accuracy learned by optimizing DRO using RECOVER and optimizing ERM using SGD on the imbalanced datasets. Then we show the RECOVER is also an effective fine-tuning algorithm for large-scale imbalanced data training. The code for reproducing the results is released here \([39]\).

6.1 Comparison with SOTA DRO Baselines

We compare RECOVER with five baselines: Restarted CIVR \([53]\) (RCIVR), ASC-PG \([45]\), Stoc-AGDA \([50]\), PG-SMD2 \([41]\) and PES-SGDA \([15]\). RCIVR and ASC-PG are the state-of-the-art algorithms for solving stochastic compositional problems. RCIVR uses variance reduction techniques and leverages the PL condition, while ASC-PG does neither. Stoc-AGDA and PG-SMD2 are the primal-dual algorithms with and without leveraging the PL condition explicitly, respectively. PES-SGDA is a variant of PG-SMD2 and was proposed by leveraging the PL condition for achieving faster convergence. Please note that ASC-PG and Stoc-AGDA use polynomially decreasing step sizes, RECOVER, PG-SMD2 and PES-SGDA use stagewise decreasing step size, and RCIVR uses a constant step size. The parameters of each algorithm are appropriately tuned for the best performance. All the algorithms are implemented using Pytorch and run on GeForce GTX 1080 Ti GPU.

We conduct experiments on four datasets, namely STL10 \([8]\), CIFAR10, CIFAR100 \([21]\), and iNaturalist2019 \([46]\). The original STL10, and CIFAR10, CIFAR100 are balanced data, where STL10 has 10 classes and each class has 500 training images, CIFAR10 (resp. CIFAR100) has 10 (resp. 100) classes and each class has 5K (resp. 500) training images. For STL10, CIFAR10 and CIFAR100, we artificially construct imbalanced training data, where we only keep the last 100 images of each class for the first half classes. iNaturalist2019 itself is an imbalanced dataset that contains 265,213 images with 1010 classes. We train ResNet-20 on STL10, CIFAR10, CIFAR100, and Inception-V3 on iNaturalist2019.

For fair comparison, we use the same constant batch size \(b\) for all methods except for RCIVR in which the inner loop batch size \(b'\) and outer loop batch size \(B_0\) are hyperparameters that relate to convergence. We use the constant batch size \(b = 128\) on CIFAR10, CIFAR100, and \(b = 64\) on iNaturalist2019, and \(b = 32\) on STL. For RCIVR, both the fixed inner loop batch size \(b'\) and the initial outer loop batch size \(B_0\) are tuned in \([32, 64, 128]\). The outer loop mini-batch size \(B_k\) is also increased by a factor of 10 per-stage according to the theory.
We compare the generalization performance of DRO optimized by RECOVER with traditional ERM optimized by SGD for imbalance multi-classification tasks on STL10, CIFAR10, CIFAR100. The IMbalance RATIO (IMRATIO) is defined as the number of samples in the minority classes over

<table>
<thead>
<tr>
<th>IMRATIO</th>
<th>STL10</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SGD</td>
<td>RECOVER</td>
<td>SGD</td>
</tr>
<tr>
<td>0.02</td>
<td>37.97 (0.78)</td>
<td>38.08 (0.59)</td>
<td>65.36 (0.64)</td>
</tr>
<tr>
<td>0.05</td>
<td>41.12 (0.94)</td>
<td>42.68 (0.60)</td>
<td>74.74 (0.71)</td>
</tr>
<tr>
<td>0.1</td>
<td>46.03 (0.96)</td>
<td>48.94 (0.86)</td>
<td>79.32 (0.42)</td>
</tr>
<tr>
<td>0.2</td>
<td>51.75 (1.14)</td>
<td>56.06 (1.26)</td>
<td>84.84 (0.51)</td>
</tr>
</tbody>
</table>
the number of samples in the majority classes. We manually construct different training sets with different IMRATIO, i.e., we only keep the last IMRATIO portion of images in the first half of classes. Different from previous experiments, we tune $\lambda$ in a certain range $\{1, 5, 10, 20, 100\}$ by a cross-validation approach and report the best testing results. Other parameters of RECOVER is tuned according to the setting in previous experiments. We use ResNet-32 for CIFAR10, CIFAR100, and ResNet-20 for STL10. For SGD, the step size is set as $\eta_0$ in the first 60 epochs, and is decreased by a factor of 10 at 60, and 90 epochs following the practical strategy [17], where $\eta_0$ is tuned in $\{0.1, 0.5, 1\}$ and 1 epoch means one pass of training data.

We report averaged test accuracy over 5 runs with mean (variance) in Table 2. We can see that DRO with RECOVER achieves higher test accuracy with smaller variance over multiple runs on all datasets than ERM with SGD. In addition, we report the results over 5 runs of different $\lambda$ on CIFAR10 with different IMRATIO in Figure 3 (left). It is obvious to see that an appropriate regularization on the dual variable can improve the performance.

6.3 Effectiveness of RECOVER as a Fine-tuning Method

Fine-tuning high level layers from a pre-trained model is widely used for transfer learning and is also an effective method to update the models without increasing the computational cost too much when receiving new samples. For this purpose, we demonstrate that DRO is a better objective than the Cross Entropy (CE) loss and focal loss for fine-tuning on imbalanced datasets.

ImageNet-LT [29] and Places-LT [29] are two popular imbalanced data sets and are the Long-Tailed (LT) version of ImageNet-2012 [10] and Places-2 [54] by sampling a subset following the Pareto distribution [3] with the power value 6. ImageNet-LT has 115.8K images from 1000 categories, and Places-LT contains 62.5K training images from 365 classes. The head class is 4980 images and the tail class contains 5 images in both datasets.

To verify that DRO is a better objective and that RECOVER is an efficient optimization algorithm, we compare the test accuracy of the model trained with different objectives: DRO, CE loss and focal loss, where DRO is optimized by RECOVER and the other two losses are optimized by SGD. All methods start from the same pre-trained model. We apply the ImageNet pretrained ResNet512 as the pre-trained model for Places-LT. For ImageNet-LT, we train ResNet50 using CE loss for 90 epochs following the standard training strategy proposed in [17] as the pre-trained model. We then fine-tune the last block of the convolutions layer and the classifier layer for 30 epochs by using RECOVER for optimizing DRO and using SGD for optimizing ERM, respectively. The initial step size for RECOVER and SGD are both tuned in $\eta_0 \in \{0.1, 0.5, 1\}$. For DRO, $\lambda$ is tuned in $\{1, 5, 10\}$.

The test accuracy over 3 runs with mean (variance) is reported in Figure 2 (right). It is clear to see that DRO optimized by RECOVER outperforms ERM with the CE loss and focal loss optimized by SGD more than 1(%) on both datasets. This vividly verifies the effectiveness of RECOVER as a fine-tuning method on imbalanced data.

7 Conclusion

In this paper, we proposed a duality-free online method for solving a class of distributionally robust optimization problems. We used a KL divergence regularization on the dual variable and transformed the problem into a two-level stochastic compositional problem. By leveraging a practical PL condition, we developed a practical method RECOVER based on recursive variance-reduced estimators and established an optimal sample complexity. Experiments verify the effectiveness of the proposed algorithm in terms of both running time and prediction performance on large-scale imbalanced data. An open question remains is how to solve the DRO problem with a KL constraint on the dual variable by a practical stochastic algorithm without maintaining and updating the high dimensional dual variable. We plan to address this challenge in the future work.

Acknowledgments

The authors thank anonymous reviewers for constructive comments. This work was supported by NSF Career Award #1844403, NSF Award #2110545 and NSF Award #1933212.
References


[34] Ion Necula, Yuriy Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. LIDAM Reprints CORE 3000, Universite catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2019.


Appendix

Notations we refer the compositional stochastic gradient estimator $v_t^T \nabla f(u_t)$ of COVER (Algorithm 1) as $d_t$, i.e., $d_t = v_t^T \nabla f(u_t)$, where $u_t, w_t$ are the two estimator sequences maintained in COVER. The compositional stochastic variance introduced by $d_t$ as $\varepsilon_t = d_t - \nabla g(w_t)^T \nabla f(g(w_t))$, the stochastic variance introduced by $u_t$ denoted as $\varepsilon_{u_t} = u_t - g(w_t)$, the stochastic variance introduced by $v_t$ denoted as $\varepsilon_{v_t} = v_t - \nabla g(w_t)$. The stochastic proximal gradient measure of COVER is $\hat{G}_n(w_t) = \frac{1}{n}(w_{t+1} - w_t)$. And $L = \max\{L_0C_fL_f, C_fC_fL_f, C_f^2, L_0C_f, C_f^3, C_f^2L_f, C_f, C_f^2L_f, C_f^2, L_f^2\}.

8 Illustration of Variance Introduced by $p \in \mathbb{R}^n$

To see this, the variance of stochastic gradient in terms of $w$ with random sampling is given by $\text{Var}_r = \frac{1}{n} \sum_{i=1}^n \|np_i \nabla \ell(w; z_i) - \nabla_w L(w, p)\|^2 = \sum_{i=1}^n np_i^2 \|\nabla \ell(w; z_i)\|^2 - \|\nabla L(w, p)\|^2$, where $L(w, p) = \sum_{i=1}^n p_i \ell(w; z_i)$. In contrast, the variance of stochastic gradient in terms of $w$ with non-uniform sampling according to $p$ is given by $\text{Var}_n = \sum_{i=1}^n p_i \|\nabla \ell(w; z_i) - \nabla_w L(w, p)\|^2 = \sum_{i=1}^n p_i \|\nabla \ell(w; z_i)\|^2 - \|\nabla L(w, p)\|^2$. Let us consider an extreme case when $p_i = 1, p_j = 0, \forall j \neq i$, we have $\text{Var}_r = (n - 1) \|\nabla \ell(w; z_i)\|^2 \gg \text{Var}_n = 0$.

9 Proof of Section 4

Lemma 4. Suppose Assumption 1 and 2 hold, we have

$$\mathbb{E}[\|\varepsilon_t\|^2] \leq 2C_f^2 \mathbb{E}[\|\varepsilon_{v_t}\|^2] + 2C_f^2L_f^2 \mathbb{E}[\|\varepsilon_{u_t}\|^2]. \tag{9}$$

Remark: Plugging the definition of $L$ into it, we have $\mathbb{E}[\|\varepsilon_t\|^2] \leq L \mathbb{E}[\|\varepsilon_{v_t}\|^2] + L \mathbb{E}[\|\varepsilon_{u_t}\|^2]$.

Proof.

$$\begin{align*}
\mathbb{E}[\|d_t - \nabla g(w_t)^T \nabla f(g(w_t))\|^2] &= \mathbb{E}[\|v_t^T \nabla f(u_t) - \nabla g(w_t)^T \nabla f(g(w_t))\|^2] \\
&= \mathbb{E}[\|v_t^T \nabla f(u_t) - \nabla g(w_t)^T \nabla f(u_t) + \nabla g(w_t)^T \nabla f(u_t) - \nabla g(w_t)^T \nabla f(g(w_t))\|^2] \\
&\leq 2\mathbb{E}[\|v_t^T \nabla f(u_t) - \nabla g(w_t)^T \nabla f(u_t)\|^2] + 2\mathbb{E}[\|\nabla g(w_t)^T \nabla f(u_t) - \nabla g(w_t)^T \nabla f(g(w_t))\|^2] \\
&\leq 2C_f^2 \mathbb{E}[\|v_t - \nabla g(w_t)^T\|^2] + 2C_f^2L_f^2 \mathbb{E}[\|\nabla f(u_t) - \nabla f(g(w_t))\|^2] \\
&\leq 2C_f^2 \mathbb{E}[\|v_t - \nabla g(w_t)^T\|^2] + 2C_f^2L_f^2 \mathbb{E}[\|u_t - g(w_t)\|^2] \\
&= 2C_f^2 \mathbb{E}[\|\varepsilon_{v_t}\|^2] + 2C_f^2L_f^2 \mathbb{E}[\|\varepsilon_{u_t}\|^2], \tag{10}
\end{align*}$$

where the first inequality is due to $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, the second inequality is due to the $C_f$-Lipschitz continuous of $f$, i.e., $\|\nabla f(w_t)\|^2 \leq C_f^2$, and $C_g$-Lipschitz continuous of $g$, i.e., $\|\nabla g(w_t)\|^2 \leq C_g^2$. The third inequality is due to the $L_f$-smoothness of $f$ function.

Lemma 5. For the two gradient mappings $\|\hat{G}_n(w_t)\|^2, \|\tilde{G}_n(w_t)\|^2$, we have

$$\begin{align*}
\mathbb{E}[\|\hat{G}_n(w_t)\|^2] &\leq 2\mathbb{E}[\|\nabla g(w_t)^T \nabla f(g(w_t))\|^2] - \mathbb{E}[\|d_t\|^2], \\
\mathbb{E}[\|\tilde{G}_n(w_t)\|^2] &\leq 2\mathbb{E}[\|\nabla g(w_t)^T \nabla f(g(w_t))\|^2] - \mathbb{E}[\|d_t\|^2]. \tag{11}
\end{align*}$$

Remark: This lemma implies that

$$\begin{align*}
\mathbb{E}[\|w_{t+1} - w_t\|^2] &\leq \eta^2 \mathbb{E}[\|\hat{G}_n(w_t)\|^2] + 2\eta^2 \mathbb{E}[\|\nabla g(w_t)^T \nabla f(g(w_t))\|^2] - \mathbb{E}[\|d_t\|^2], \tag{12}
\end{align*}$$
Proof. Denote that $\hat{w}_{t+1} = \text{prox}^\eta_t(w_t - \eta \nabla g(w_t) \top \nabla f(g(w_t)))$. Then we have $\|w_t - \hat{w}_{t+1}\|^2 \leq 2\|w_t - w_{t+1}\|^2 + 2\|w_{t+1} - \hat{w}_{t+1}\|^2$. By the definition of $\|G_t(w_t)\|^2, \|G_t(w_t)\|^2$, we have

$$
E[\|G_t(w_t)\|^2] \leq 2E[\|G_t(w_t)\|^2] + \frac{2}{\eta^2} E[\|w_{t+1} - \hat{w}_{t+1}\|^2]
$$

$$
= 2E[\|G_t(w_t)\|^2] + \frac{2}{\eta^2} [\text{prox}^\eta_t(w_t - \eta d_t) - \text{prox}^\eta_t(w_t - \eta \nabla g(w_t) \top \nabla f(g(w_t)))\|^2]
$$

$$
\leq 2E[\|G_t(w_t)\|^2] + \frac{2}{\eta^2} E[\|w_t - \eta d_t - (w_t - \eta \nabla g(w_t) \top \nabla f(g(w_t)))\|^2]
$$

$$
= 2E[\|G_t(w_t)\|^2] + 2E[\|\nabla g(w_t) \top \nabla f(g(w_t)) - d_t\|^2],
$$

where the second inequality is due to the non-expansive property of proximal mapping. Similarly, by $\|w_t - w_{t+1}\|^2 \leq 2\|w_t - \hat{w}_{t+1}\|^2 + 2\|w_{t+1} - w_{t+1}\|^2$, following the same analysis as equation (13), we would have the second inequality in Lemma 5.

Lemma 6. Let sequence $\{x_t\}$ be generated by COVER and with $\eta_t \leq \frac{1}{2\eta}$ for all $t \geq 1$, the following inequality holds

$$
E[F(w_{t+1})] - E[F(w_t)] \leq -\frac{\eta_t}{8} E[\|G_t(w_t)\|^2] + \frac{3\eta_t L}{4} E[\|u_{v_{t}}\|^2] + \frac{3\eta_t L}{4} E[\|u_{v_{t}}\|^2].
$$

Proof. Denote $F(w_{t+1}) = f(g(w_{t+1})) + r(w_{t+1})$. First, show that $f(g(w))$ is smooth and $\nabla f(g(w))$ has Lipschitz constant with $L_{f(g)} = C_g^2 L_f + C_f L_g$. For any two variables $w, w' \in \mathbb{R}^d$

$$
\|\nabla g(w) \top \nabla f(g(w)) - \nabla g(w') \top \nabla f(g(w'))\|
$$

$$
= \|\nabla g(w) \top \nabla f(g(w)) - \nabla g(w) \top \nabla f(g(w)) - \nabla g(w) \top \nabla f(g(w')) + \nabla g(w') \top \nabla f(g(w')) - \nabla g(w') \top \nabla f(g(w'))\|
$$

$$
\leq \|\nabla g(w)\| \|\nabla f(g(w)) - \nabla f(g(w'))\| + \|\nabla g(w) \top \nabla f(g(w')) - \nabla g(w') \top \nabla f(g(w'))\|
$$

$$
\leq C_g L_f \|g(w) - g(w')\| + C_g \|\nabla f(g(w'))\| \|w - w'\|
$$

$$
\leq C_g^2 L_f \|w - w'\|^2 + C_g C_f \|w - w'\| \leq L \|w - w'\|.
$$

Then by above equation (15), we have

$$
f(g(w_{t+1})) + r(w_{t+1})
$$

$$
\leq f(g(w_t)) - \frac{1}{2\eta_t} \|w_{t+1} - w_t\|^2 + \frac{L}{2} \|w_{t+1} - w_t\|^2 + r(w_{t+1})
$$

$$
\leq f(g(w_t)) - \frac{1}{2\eta_t} \|w_{t+1} - w_t\|^2 + \frac{L}{2} \|w_{t+1} - w_t\|^2
$$

$$
= f(w_t) + \frac{\eta_t}{2} \|\nabla g(w_t) \top \nabla f(g(w_t)) - d_t\|^2 + \frac{L}{2} \|w_{t+1} - w_t\|^2
$$

where the proof of (a) will be shown shortly after we derive the claimed result of this lemma. By the setting $\eta_t \leq \frac{1}{2\eta}$, taking expectation on both sides and in combination with Lemma 5, we have

$$
E[F(w_{t+1}) - F(w_t)] \leq -\frac{\eta_t}{8} E[\|G_t(w_t)\|^2] + \frac{3\eta_t L}{4} \|\nabla g(w_t) \top \nabla f(g(w_t)) - d_t\|^2.
$$

Then applying the results of Lemma 5, we have the results.

Proof of (a): By the definition of $w_{t+1} = \text{prox}^\eta_t(w_t - \eta d_t) = \arg \min_w \{\frac{1}{2}\|w - (w_t - \eta d_t)\|^2 + \eta_t r(w)\} = \arg \min_w \{\frac{1}{\eta_t}\|w - (w_t - \eta d_t)\|^2 + r(w)\}$. Then by the $\frac{1}{\eta_t}$ strongly convexity of the
To prove the convergence of proximal gradient \(\|s^{t}(w_{t+1})\|^{2}\), we need to construct telescoping sum that depending on the Lemma 6 As a result, we need to bound the variance on the R.H.S of Lemma 6, i.e., \(E[\|\varepsilon_{u_{t}}\|^{2}]\), \(E[\|\varepsilon_{v_{t}}\|^{2}]\) with the following lemmas.

**Lemma 7.** With notations in COVER, we have

\[
\frac{E[\|s^{t+1}(w_{t+1})\|^{2}]}{\eta_{t}} \leq E\left[\frac{2\eta_{t}^{3}\sigma^{2} + (1 - \alpha_{t})^{2}(1 + 4\eta_{t}^{2}L^{2})\|\varepsilon_{u_{t}}\|^{2}}{\eta_{t}} + \frac{4\eta_{t}^{2}(1 - \alpha_{t})^{2}L^{2}\|\varepsilon_{v_{t}}\|^{2}}{\eta_{t}} + 4\eta_{t}(1 - \alpha_{t})^{2}L\|s^{t}(w_{t+1})\|^{2}\right].
\]

---

**Proof.**

\[
E[\|s^{t+1}(w_{t+1})\|^{2}] = \frac{E[\|s^{t+1}(w_{t+1})\|^{2}]}{\eta_{t}}
\]

\[
= E\left[\frac{|g_{x^{t+1}}(w_{t+1}) + (1 - \alpha_{t})(u_{t} - g_{x^{t+1}}(w_{t})) - g(w_{t+1})|^{2}}{\eta_{t}}\right]
\]

\[
= E\left[\frac{|a_{t}(g_{x^{t+1}}(w_{t+1}) - g(w_{t+1})) + (1 - a_{t})(u_{t} - g(w_{t}))}{\eta_{t}}\right]
\]

\[
= E\left[\frac{(1 - \alpha_{t})^{2}\|\varepsilon_{u_{t}}\|^{2}}{\eta_{t}} + |a_{t}(g_{x^{t+1}}(w_{t+1}) - g(w_{t+1})) + (1 - a_{t})(u_{t} - g(w_{t}))|^{2}}{\eta_{t}}\right]
\]

\[
\leq E\left[\frac{2\eta_{t}^{2}\|g_{x^{t+1}}(w_{t+1}) - g(w_{t+1})\|^{2}}{\eta_{t}} + \frac{(1 - \alpha_{t})^{2}\|\varepsilon_{u_{t}}\|^{2}}{\eta_{t}}\right]
\]

\[
+ 2(1 - \alpha_{t})^{2}\|g_{x^{t+1}}(w_{t+1}) - g_{x^{t+1}}(w_{t}) - (u_{t} - g(w_{t}))\|^{2}\right].
\]

\[
\leq E\left[\frac{2\eta_{t}^{2}\sigma^{2}}{\eta_{t}} + \frac{(1 - \alpha_{t})^{2}\|\varepsilon_{u_{t}}\|^{2}}{\eta_{t}} + \frac{2(1 - \alpha_{t})^{2}L\|w_{t+1} - w_{t}\|^{2}}{\eta_{t}}\right]
\]

\[
= E\left[\frac{2\sigma^{2}\eta_{t}^{3}}{\eta_{t}} + \frac{(1 - \alpha_{t})^{2}\|\varepsilon_{u_{t}}\|^{2}}{\eta_{t}} + \frac{2(1 - \alpha_{t})^{2}L\eta_{t}^{2}\|s^{t}(w_{t+1})\|^{2}}{\eta_{t}}\right]
\]

\[
\leq E\left[\frac{2\sigma^{2}\eta_{t}^{3}}{\eta_{t}} + \frac{(1 - \alpha_{t})^{2}\|\varepsilon_{u_{t}}\|^{2}}{\eta_{t}} + \frac{2(1 - \alpha_{t})^{2}L\eta_{t}^{2}\|s^{t}(w_{t+1})\|^{2}}{\eta_{t}}\right]
\]

\[
= E\left[\frac{2\sigma^{2}\eta_{t}^{3}}{\eta_{t}} + \frac{(1 - \alpha_{t})^{2}(1 + 4\eta_{t}^{2}L^{2})\|\varepsilon_{u_{t}}\|^{2}}{\eta_{t}} + \frac{4\eta_{t}^{2}(1 - \alpha_{t})^{2}L\|\varepsilon_{v_{t}}\|^{2}}{\eta_{t}} + \frac{4\eta_{t}^{2}(1 - \alpha_{t})^{2}L\|s^{t}(w_{t+1})\|^{2}}{\eta_{t}}\right].
\]
where the fourth equality is due to $E_t[g_{x_{t+1}}(w_{t+1})] - g(w_t)] = 0$ and $E_t[g_{x_{t+1}}(w_{t+1}) - g(w_t)] = 0$
with $E_t$ denoting an expectation conditioned on events until $t$-iteration; and the first inequality holds because $\|a + b\|^2 \leq 2a^2 + 2b^2$. Applying the same analysis, we are able to have the bound of $\mathbb{E}[\|v_{x_{t+1}}\|^2] = \mathbb{E}[\|v_{x_{t+1}} - g(w_{t+1})\|^2]$ in the lemma.

\[ \square \]

9.1 Proof of Theorem 1

Proof. After deriving Lemma 6 and 7, we are ready to prove Theorem 1. We construct Lyapunov function $\Gamma_t = F(w_t) + \frac{1}{c_0\eta_t-1}([\|v_t\|^2 + \|u_t\|^2]),$ where $c_0$ is a constant and can be derived in the following proof. According to equation (14)

\[ \mathbb{E}[\Gamma_{t+1} - \Gamma_t] \leq \frac{3\eta_t L}{4} \mathbb{E}[\|v_t\|^2 + \|u_t\|^2] + \frac{1}{c_0\eta_t} \mathbb{E}[\|v_{t+1}\|^2 + \|u_{t+1}\|^2] - \frac{1}{c_0\eta_t-1} \mathbb{E}[\|v_{t+1}\|^2 + \|u_{t+1}\|^2]. \]  

(20)

Then by telescoping sum from 1, \cdots, T, and rearranging terms we have

\[ \sum_{t=1}^{T} \frac{\eta_t}{8} \mathbb{E}[\|G_{\eta_t}(w_t)\|^2] \leq \mathbb{E}[\Gamma_1 - \Gamma_{T+1}] + \sum_{t=1}^{T} \frac{3\eta_t L}{4} \mathbb{E}[\|v_t\|^2 + \|u_t\|^2] \]

(a)

\[ + \sum_{t=1}^{T} \frac{1}{c_0\eta_t} \mathbb{E}[\|v_{t+1}\|^2 + \|u_{t+1}\|^2] - \frac{1}{c_0\eta_t-1} \mathbb{E}[\|v_{t+1}\|^2 + \|u_{t+1}\|^2]. \]

(b)

We want (b) \leq 0 such that it can be used to cancel the increasing cumulative variance of term (a).

Next we will upper bound (b) up to a negative level:

\[ \frac{1}{c_0\eta_t} \mathbb{E}[\|v_{t+1}\|^2 + \|u_{t+1}\|^2] - \frac{1}{c_0\eta_t-1} \mathbb{E}[\|v_t\|^2 + \|u_t\|^2] \]

\[ \leq \frac{1}{c_0} \mathbb{E}[4\eta_t^3 c^2 \sigma^2 + \frac{(1 - a_t)^2 (1 + 8\eta_t^2 L^2)}{\eta_t} - \frac{1}{\eta_t-1}][\|v_t\|^2 + \|u_t\|^2] \]

\[ + 8\eta_t (1 - a_t + 1) L \mathbb{E}[\|G_{\eta_t}(w_t)\|^2] \]

\[ \leq \frac{1}{c_0} \mathbb{E}[4\eta_t^3 c^2 \sigma^2 + \frac{(1 - a_t)(1 + 8\eta_t^2 L^2)}{\eta_t} - \frac{1}{\eta_t-1}][\|v_t\|^2 + \|u_t\|^2] \]

\[ + 8\eta_t L \mathbb{E}[\|G_{\eta_t}(w_t)\|^2]. \]

(B)

Next we upper bound (B):

\[ B_t \leq \eta_t^{-1} \eta_t^{-1} \eta_t^{-1} (8\eta_t^2 L^2 - a_t) \mathbb{E}[\|v_{t+1}\|^2 + \|v_{t+1}\|^2] \]

\[ = \eta_t^{-1} (1 - a_t + \eta_t (8L^2 - c)) \mathbb{E}[\|u_{t+1}\|^2 + \|v_{t+1}\|^2]. \]

(23)

For $\frac{1}{\eta_t} - \frac{1}{\eta_t-1}$, by applying $(x + y)^{1/3} - x^{1/3} \leq yx^{-2/3}/3$ and manipulating constant terms, we have

\[
\begin{align*}
\frac{1}{\eta_t} - \frac{1}{\eta_t-1} &= \frac{1}{k} (w + ta^2)^{1/3} - \frac{1}{k} (w + (t - 1)a^2)^{1/3} \\
&\leq \frac{\sigma^2}{3k(w + (t - 1)a^2)^{2/3}} \\
&\leq \frac{\sigma^2}{3k(w - \sigma^2 + ta^2)^{2/3}} \\
&\leq \frac{2^{2/3} \sigma^2}{3k^{3/2} - \eta_t} \leq \frac{2^{2/3}}{12k^3 \eta_t} \leq \frac{\sigma^2}{7Lk^3 \eta_t}.
\end{align*}
\]

(24)
where \( w \geq (16Lk)^3 \) to have \( \eta_t \leq \frac{1}{16L} \). Then by setting \( c = 104L^2 + \frac{2}{7Lk^4} \),
\[
\eta_t(8L^2 - c) \leq -96L^2 \eta_t - \sigma^2 \eta_t/(7Lk^3).
\]

Then we obtain
\[
B_t \leq -96L^2 \eta_t \| \varepsilon_{u_t} \|^2 + \| \varepsilon_{v_t} \|^2 \quad \text{(25)}
\]

Then plugging equation (25) into equation (22) and set \( c_0 = 128L \),
\[
\frac{1}{128\eta_t L} E[\| \varepsilon_{v_{t+1}} \|^2 + \| \varepsilon_{u_{t+1}} \|^2] - \frac{1}{128\eta_{t-1} L} E[\| \varepsilon_{v_t} \|^2 + \| \varepsilon_{u_t} \|^2] \\
\leq \frac{\eta_t^3 \sigma^2}{2} - \frac{3L\eta_t}{4} [\| \varepsilon_{u_t} \|^2 + \| \varepsilon_{v_t} \|^2] + \frac{\eta_t}{16} E[\| G_{\eta_t}(w_t) \|^2],
\]

Substituting equation (26) into equation (21), Dividing \( \eta_t^3 \) on both sides of equation (21) and substituting (26). We get
\[
\frac{\eta_t}{8} E[\| G_{\eta_t}(w_t) \|^2] \leq E[\Gamma_t - \Gamma_{t+1}] + \frac{3L\eta_t}{4} E[\| \varepsilon_{v_t} \|^2 + \| \varepsilon_{u_t} \|^2] \\
+ \frac{\eta_t^3 \sigma^2}{32L} - \frac{3L\eta_t}{4} [\| \varepsilon_{u_t} \|^2 + \| \varepsilon_{v_t} \|^2] + \frac{\eta_t}{16} E[\| G_{\eta_t}(w_t) \|^2] \\
\leq E[\Gamma_t - \Gamma_{t+1}] + \frac{\eta_t^3 \sigma^2}{32L} + \frac{\eta_t}{16} E[\| G_{\eta_t}(w_t) \|^2],
\]

In addition
\[
\frac{\eta_t}{16} E[\| G_{\eta_t}(w_t) \|^2] \leq E[\Gamma_t - \Gamma_{T+1}] + \sum_{t=1}^{T} \frac{\eta_t^3 \sigma^2}{32L}.
\]

where the first inequality is due to the assumption \( w \geq 2\sigma^2 \) and the second inequality applies \( \sum_{t=1}^{T} \frac{1}{\eta_{t+2}} \leq \ln(T + 2) \).

Then
\[
\sum_{t=1}^{T} \frac{\eta_t}{16} E[\| G_{\eta_t}(w_t) \|^2] \leq E[\Gamma_t - \Gamma_{T+1}] + \frac{c^2 \sigma^2 k^3}{32L} \ln(T + 2),
\]

\[
\sum_{t=1}^{T} \eta_t E[\| G_{\eta_t}(w_t) \|^2] \leq 16E[\Gamma_t - \Gamma_{T+1}] + \frac{c^2 \sigma^2 k^3}{2L} \ln(T + 2)
\]

Since \( \eta_t \) is decreasing, we get
\[
\frac{1}{T} \sum_{t=1}^{T} E[\| G_{\eta_t}(w_t) \|^2] \leq \frac{16(F(w_1) - F_*)}{\eta_t T} + \frac{16E[\| \varepsilon_{u_t} \|^2 + \| \varepsilon_{v_t} \|^2]}{c_0 \eta_0 \eta_T T} + \frac{c^2 k^3 \ln(T + 2)}{2L} \cdot \frac{T}{T \eta_T} \\
\leq O \left( \frac{16(F(w_1) - F_*)}{T^{2/3}} + \frac{32\sigma^2}{c_0 \eta_0 T^{2/3}} + \frac{c^2 k^3 \ln(T + 2)}{2L} \frac{T^{2/3}}{T^{2/3}} \right) \quad \text{(30)}
\]

where \( O \) suppresses constant scalars.
As a result, we want $b$ to be such that it can be used to cancel the increasing cumulative variance of the inner stage of RECOVER, in which a constant step size is used in each stage.

**Theorem 3.** At $k$-th stage, under the Assumption 7 and 2, let $c \geq 104L^2$ and the step size $\eta_k$, after $T_k$ iterations, the output of RECOVER satisfies,

$$
E[\|G_{\eta_k}(w_t)\|^2] \leq \frac{16c^2}{\eta_k T_k} + \frac{2c^2 + \eta_k^2}{2L} + \frac{E[\|u_{k-1} - g(w_{k-1})\|^2 + \|v_{k-1} - \nabla g(w_{k-1})\|^2]}{8\eta_k^2 LT_k}
$$

where $w_k$ is uniformly sampled from $\{w_t\}_{t=1}^{T_k}$ at $k$-th stage.

### 9.2 Proof of Theorem 3

**Proof of Theorem 3** We derive the theoretical analysis for the $k$-th stage based on Lemma 6 and 7. We construct Lyapunov function $\Gamma_t = F(w_t) + \frac{1}{c_0}(\|\varepsilon_{v_t}\|^2 + \|\varepsilon_{u_t}\|^2)$, where $c_0$ is a constant and can be derived in the following proof. According to equation (14),

$$
E[\Gamma_{t+1} - \Gamma_t] \leq E[-\frac{\eta_k}{8} E[\|G_{\eta_k}(w_t)\|^2]] + \frac{3\eta_k L}{4} E[\|\varepsilon_{v_t}\|^2 + \|\varepsilon_{u_t}\|^2] + \frac{1}{c_0\eta_k} E[\|\varepsilon_{v_{t+1}}\|^2 + \|\varepsilon_{u_{t+1}}\|^2] - \frac{1}{c_0\eta_k} E[\|\varepsilon_{v_t}\|^2 + \|\varepsilon_{u_t}\|^2].
$$

Then by telescoping sum and rearranging terms we have

$$
\sum_{t=1}^{T_k} \frac{\eta_k}{8} E[\|G_{\eta_k}(w_t)\|^2] \leq \sum_{t=1}^{T_k} \left( \frac{3\eta_k L}{4} E[\|\varepsilon_{v_t}\|^2 + \|\varepsilon_{u_t}\|^2] + \frac{1}{c_0\eta_k} E[\|\varepsilon_{v_{t+1}}\|^2 + \|\varepsilon_{u_{t+1}}\|^2] - \frac{1}{c_0\eta_k} E[\|\varepsilon_{v_t}\|^2 + \|\varepsilon_{u_t}\|^2] \right).
$$

As a result, we want $\frac{5}{8} \leq 0$ such that it can be used to cancel the increasing cumulative variance of term $\frac{3}{4}$.

Next we will upper bound $\frac{5}{8}$ up to a negative level by making use of Lemma 7 with $a_t$ to be fixed at $k$-th stage as $a_k = c_0\eta_k^2$.

Applying Lemma 7,

$$
\frac{1}{c_0\eta_k} E[\|\varepsilon_{v_{t+1}}\|^2 + \|\varepsilon_{u_{t+1}}\|^2] - \frac{1}{c_0\eta_k} E[\|\varepsilon_{v_{t}}\|^2 + \|\varepsilon_{u_{t}}\|^2] \leq \frac{1}{c_0} E \left[ 4\eta_k^3 c^2 + \frac{(1 - a)(1 + 8\eta_k^2 L^2)}{\eta_k} \left[ \|\varepsilon_{v_{t}}\|^2 + \|\varepsilon_{u_{t}}\|^2 \right] \right]
$$

$$
\leq \frac{1}{c_0} E \left[ 4\eta_k^3 c^2 + \frac{(1 - a)(1 + 8\eta_k^2 L^2)}{\eta_k} \left[ \|\varepsilon_{v_{t}}\|^2 + \|\varepsilon_{u_{t}}\|^2 \right] \right] + \frac{8\eta_k L\|G_{\eta_k}(w_t)\|^2}{c_t}.
$$

For $a_t$, by set $c = 104L^2$, we have

$$
B_t \leq \eta_k^{-1} - \frac{\eta_k^{-1} + \eta_k^{-1}(8\eta_k^2 L^2 - a)(\|\varepsilon_{u_t}\|^2 + \|\varepsilon_{v_t}\|^2)}{\eta_k^{-1} + \eta_k(8L^2 - c)(\|\varepsilon_{u_t}\|^2 + \|\varepsilon_{v_t}\|^2)} \leq -96L^2 \eta_k(\|\varepsilon_{u_t}\|^2 + \|\varepsilon_{v_t}\|^2).
$$

(35)
To satisfies $c\eta_k^2 \leq 1$, we should have $\eta_k \leq \frac{1}{4L}$. Then by setting $c_0 = 128L$, we have

$$
\sum_{t=1}^{T_k} \frac{1}{128\eta_k} E[\|\epsilon_{v_{t+1}}\|^2 + \|\epsilon_{u_{t+1}}\|^2] - \frac{1}{128\eta_k} E[\|\epsilon_{v_t}\|^2 + \|\epsilon_{u_t}\|^2] 
\leq \frac{c_0^2 \sigma^2}{32L} \sum_{t=1}^{T_k} 3L \eta_k E[\|\epsilon_{v_t}\|^2 + \|\epsilon_{u_t}\|^2] + \sum_{t=1}^{T_k} \frac{\eta_k}{16} E[\|G_{\eta_k}(w_t)\|^2].
$$

Plugging it into equation (33), we get

$$
E \left[ \frac{\eta_k}{8} \sum_{t=1}^{T_k} \|G_{\eta_k}(w_t)\|^2 \right] \leq E[\Gamma_1 - \Gamma_{T_k+1}] + E \left[ \frac{c_0^2 \sigma^2}{32L} \eta_k^3 T_k + \frac{\eta_k}{16} \sum_{t=1}^{T_k} \|G_{\eta_k}(w_t)\|^2 \right].
$$

Then we have

$$
E \left[ \frac{\eta_k}{16} \sum_{t=1}^{T_k} \|G_{\eta_k}(w_t)\|^2 \right] \leq \frac{c_0^2 \sigma^2}{32L} \eta_k^3 T_k + E[\Gamma_1 - \Gamma_{T_k+1}] 
\leq E[F(w_1) - F_*] + \frac{c_0^2 \sigma^2}{32L} \eta_k^3 T_k + \frac{\eta_k}{16} \sum_{t=1}^{T_k} \|G_{\eta_k}(w_t)\|^2 
\Rightarrow E[\|G_{\eta_k}(w_t)\|^2] \leq \frac{16E[F(w_1) - F_*]}{\eta_k T_k} + \frac{c_0^2 \sigma^2}{2L} \eta_k^2 + \frac{E[\|\epsilon_{v_t}\|^2 + \|\epsilon_{u_t}\|^2]}{8\eta_k^2 L T_k},
$$

where $w_k$ is uniformly sampled from $\{w_1, \ldots, w_T\}$.

\medskip

10 Proof of Section 5

10.1 Proof of Lemma 1

Proof. This proof follows Lemma A.3 of [50]. Note that

$$
F_{dro}(w) = \lambda \log \left( \frac{1}{n} \sum_{i=1}^{n} \exp \left( \frac{\ell(w; z_i)}{\lambda} \right) \right) = \max_{p \in \Delta_n} (F_p(w) - h(p, 1/n)).
$$

Denote $\psi(w, p) = F_p(w) - h(p, 1/n)$ and $p^*(w) = \arg \max_{p \in \Delta_n} \psi(w, p)$.

Thus, we have $F_{dro}(w) = \max_{p \in \Delta_n} \psi(w, p)$ and $p^*(w) = \arg \max_{p \in \Delta_n} \psi(w, p)$.

Since $F_p(w)$ satisfies a $\mu$-PL condition for any $p \in \Delta_n$, we have

$$
\|\nabla F_{dro}(w)\|^2 = \|\nabla F_{p^*(w)}(w)\|^2 
\geq 2\mu \left( F_{p^*(w)}(w) - \min_{w'} F_{p^*(w)}(w') \right)
= 2\mu \left( \psi(w, p^*(w)) - \min_{w'} \psi(w', p^*(w)) \right).
$$

For any $w'$,

$$
\psi(w', p^*(w)) \leq \max_{p'} \psi(w', p').
$$

Therefore,

$$
\min_{w'} \psi(w', p^*(w)) \leq \min_{w'} \max_{p'} \psi(w', p').
$$

Plug this into (40), we get

$$
\|\nabla F_{dro}(w)\|^2 \geq 2\mu \left( \psi(w, p^*(w)) - \min_{w'} \max_{p'} \psi(w', p') \right)
= 2\mu (F_{dro}(w) - \min_{w'} F_{dro}(w'))
$$

which means $F_{dro}$ satisfies the $\mu$-PL condition.
10.2 Proof of Lemma 2

Proof. Let us define scaled data $\mathbf{v}_i = \sqrt{p_i}\mathbf{x}_i, \ 1 \leq i \leq n$ with $p_i \geq p_0$. Then we have $\|\mathbf{v}_i - \mathbf{v}_j\| \geq \sqrt{p_i}p_j$ since $p_i \geq p_0$.

Taking $\{(\mathbf{v}_1, \sqrt{p_1}\gamma_1), ..., (\mathbf{v}_n, \sqrt{p_n}\gamma_n)\}$ as input to the defined network, then we accordingly denote the output of the first layer of the network as $\hat{h}_{i,0} = \phi(\sqrt{p_i}\mathbf{A}\mathbf{x}_i) = \sqrt{p_i}\phi(A\mathbf{x}_i) = \sqrt{p_i}h_{i,0}$, where the the second equality is due to the property of ReLU activation function. By induction, we see that the output of the $l$-th layer is $\hat{h}_{i,l} = \sqrt{p_i}h_{i,l}$. And then the output logit is $\hat{y}_i(\mathbf{v}_i) = \sqrt{p_i}\gamma_i$.

As a result, the weighted loss defined on the original data is the average of square loss on the scaled data,

$$F(W, p) = \frac{1}{n} \sum_{i=1}^{n} \ell(W; \mathbf{v}_i) = \frac{1}{n} \sum_{i=1}^{n} (\sqrt{p_i}\gamma_i - \sqrt{p_i}\gamma_i)^2 = \sum_{i=1}^{n} p_i(\hat{y}_i - y_i)^2$$

(44)

Then we plug in Theorem 3, Lemma 7.4 and Lemma 8.7 of [1] with $F(W)$ as the objective function and $\{(\mathbf{v}_1, \sqrt{p_1}\gamma_1), ..., (\mathbf{v}_n, \sqrt{p_n}\gamma_n)\}$ as input data. We obtain that for any fixed $\mathbf{p} \in \Delta, p_i \geq p_0$, with probability $1 - \exp(-\Omega(d_2/poly(n, L, \delta^{-1}))$, it holds for every $W$ with $\|W - W_0\|^2 \leq \frac{1}{poly(n, L, \delta^{-1})}$,

$$\|\nabla_W F(W, \mathbf{p})\|^2_{F} \geq \Omega \left( \frac{\sqrt{p_0}\delta d_2}{d_0n^2} (F(W, \mathbf{p}) - \min_{W'} F(W', \mathbf{p})) \right)$$

(45)

and

$$\|\hat{y}_i - y_i\|^2 \leq \text{poly}(d_2, d_0^{-1}, L), \|2(\hat{y}_i - y_i)\nabla_W \hat{y}_i\| \leq \text{poly}(d_2, d_0^{-1}, L).$$

(46)

To generalize this bound to all $\mathbf{p} \in \Delta, p_i \geq p_0$, we need to introduce $\epsilon$-net. A subset $N \subset \mathcal{K}$ is called an $\epsilon$-net of $\mathcal{K}$ if for every $W \in \mathcal{K}$ one can find $\hat{W} \in N$ so that $\|W - \hat{W}\| \leq \epsilon$. Let $N(\mathcal{K}, \epsilon)$ denote the $\epsilon$-net of a set $\mathcal{K}$ with minimal cardinality, which is referred to as the covering number. It can be seen that the set $\mathcal{P} = \{\mathbf{p} \in \Delta, p_i \geq p_0\}$ can be covered by a $n$-dimension unit ball $B$. Take $\epsilon' = O(\epsilon/poly(d_2, d_0^{-1}, L))$. According to a standard volume comparison argument [37], we have

$$\log |N(\mathcal{B}, \epsilon')| \leq n \log \frac{3}{\epsilon'}$$

(47)

Since we have $\mathcal{P} \subset \mathcal{B}$, it follows that

$$\log |N(\mathcal{P}, \epsilon')| \leq \log |N(\mathcal{B}, \epsilon')| \leq n \log \frac{6}{\epsilon'},$$

(48)

where the first inequality is due to that the covering numbers are (almost) increasing by inclusion [38]. Taking union bound over the $\epsilon'$-net $\mathcal{N}(\mathcal{P}, \epsilon')$, we obtain that with probability $1 - \exp(-\Omega(d_2/poly(n, L, \delta^{-1}))$, it holds for every $\mathbf{p} \in N(\mathcal{P}, \epsilon')$ and for every $W$ with $\|W - W_0\|^2 \leq \frac{1}{poly(n, L, \delta^{-1})}$,

$$\|\nabla_W F(W, \mathbf{p})\|^2_{F} \geq \Omega \left( \frac{\sqrt{p_0}\delta d_2}{d_0n^2} (F(W, \mathbf{p}) - \min_{W'} F(W', \mathbf{p})) \right)$$

(49)

and

$$\|\hat{y}_i - y_i\|^2 \leq \text{poly}(d_2, d_0^{-1}, L), \|2(\hat{y}_i - y_i)\nabla_W \hat{y}_i\| \leq \text{poly}(d_2, d_0^{-1}, L).$$

(50)

For $\mathbf{p}$ not in $N(\mathcal{P}, \epsilon')$, let $\mathbf{p}$ be a point in $N(\mathcal{P}, \epsilon')$ such that $\|\mathbf{p} - \mathbf{p}\| \leq \epsilon'$, we have

$$\|\nabla_W F(W, \mathbf{p})\|^2_{F} + O(\epsilon') \geq 2\|\nabla_W F(W, \mathbf{p})\|^2_{F} + 2\|\nabla_W F(W, \mathbf{p}) - \nabla_W F(W, \hat{\mathbf{p}})\|^2_{F}$$

$$\geq \|\nabla_W F(W, \hat{\mathbf{p}})\|^2_{F} \geq \Omega \left( \frac{\sqrt{p_0}\delta d_2}{d_0n^2} (F(W, \hat{\mathbf{p}}) - \min_{W'} F(W', \hat{\mathbf{p}})) \right)$$

(51)

$$\geq \Omega \left( \frac{\sqrt{p_0}\delta d_2}{d_0n^2} (F(W, \mathbf{p}) - \min_{W'} F(W', \mathbf{p})) \right) - O(\epsilon'),$$

where the first inequality uses the second part of (50) and $\epsilon' = O(\epsilon/poly(d_2, d_0^{-1}, L))$, and the last inequality uses the first part of (50).
We also have
\[ F_{dro}(W) = \max_{p \in \Delta, p_i \geq \phi_0} F(W, p) - h(p, 1/n) = F(W, p^*(W)) - h(p^*(W), 1/n), \]
\[ \nabla F_{dro}(W) = \nabla W F(W, p^*(W)), \]
where the second line uses standard property of min-max problem [26]. Thus (51) implies that, with probability\(1 - \exp(-\Omega(d_2/\text{poly}(n, L, \delta^{-1})))\), it holds for every\(W\) with \(\|W - W_0\|^2 \leq \frac{1}{\text{poly}(n, L, \delta^{-1})}\).
\[ \|\nabla F_{dro}(W)\|_F^2 + O(\epsilon) \]
\[ \geq \Omega\left( \frac{\sqrt{\theta_0} d_2}{d_0 n^2} \left( F(W, p^*(W)) - h(p^*(W), 1/n) - \min_{W'} (F(W', p^*(W)) - h(p^*(W), 1/n)) \right) \right) \]
\[ \geq \Omega\left( \frac{\sqrt{\theta_0} d_2}{d_0 n^2} (F_{dro}(W) - \min_{W'} F_{dro}(W')) \right), \]
where the second inequality holds due to the same reason as (41) and (42).

This means that \(F_{dro}(W)\) satisfies a \(\mu\)-PL condition with \(\mu \in O\left( \frac{\sqrt{\theta_0} d_2}{d_0 n^2} \right)\) with an extra addition term of \(O(\epsilon)\), which will be omitted later in the paper for simplicity.

\[ \square \]

**10.3 Reduced Variance (Proof of Lemma 3)**

*Proof.* This lemma implies that the variance also decreasing with the increasing of stages. By equation (56) and rearranging terms, the cumulative variance of \(k\)-th stage satisfies:
\[ E[\sum_{t=1}^{T_k} \frac{3L \eta_k}{4} (\|\varepsilon_{u_t}\|^2 + \|\varepsilon_{v_t}\|^2)] \leq \frac{1}{128 \eta_k L} E[\|\varepsilon_{v_1}\|^2 + \|\varepsilon_{u_1}\|^2] + \frac{\eta_k^2 \sigma^2 T_k}{32 L} + \sum_{t=1}^{T_k} \eta_k E[\|G_{\eta_k}(w_t)\|^2] \]
\[ \leq E[F(w_1) - F_\star] + \frac{c^2 \sigma^2 T_k}{4 L} + \frac{E[\|\varepsilon_{v_1}\|^2 + \|\varepsilon_{u_1}\|^2]}{64 \eta_k L}, \]

where the second inequality uses Theorem 3. Thus we have,
\[ E[\|\varepsilon_{u_1}\|^2 + \|\varepsilon_{v_1}\|^2] \leq \frac{2E[F(w_1) - F_\star]}{\eta_T T_1 L} + \frac{c^2 \sigma^2 \eta_T^2}{3 L^2} + \frac{E[\|\varepsilon_{v_1}\|^2 + \|\varepsilon_{u_1}\|^2]}{48 \eta_T^2 L^2 T_1}, \]

where \(T\) is randomly sampled from \(1, \ldots, T_k\).

Without loss of generality, let’s assume that \(\varepsilon_0 = \Delta_F \geq \frac{c^2 \sigma^2}{64 \mu L T}\), i.e., \(\frac{\sqrt{\theta_0} \sigma L}{2c \sigma} \geq \frac{1}{16 L}\). The case that \(\Delta_F \leq \frac{c^2 \sigma^2}{64 \mu L T}\) can be simply covered by our proof. Then, denote \(\epsilon_1 = \frac{c^2 \sigma^2}{64 \mu L T}\) and \(\epsilon_k = \frac{\epsilon_1}{2^{k-1}}\).

Let’s consider the first stage, we have initialization such that \(F(w_1) - F_\star = \Delta_F\) and \(E[\|\varepsilon_{v_1}\|^2 + \|\varepsilon_{u_1}\|^2] \leq \sigma^2\). Setting \(\eta_T = \frac{1}{\sqrt{T_1}}\) and \(T_1 = O(\text{max}(\frac{\Delta_F}{\frac{1}{L}}, 1))\). Note that in below the numerical subscripts denote the epoch index \((1, \ldots, K)\). We bound the the error of first stage’s output as follows,
\[ E[\|\varepsilon_{u_1}\|^2 + \|\varepsilon_{v_1}\|^2] \leq \frac{2E[F(w_1) - F_\star]}{\eta T_1 L} + \frac{c^2 \sigma^2 \eta_T^2}{3 L^2} + \frac{E[\|\varepsilon_{v_1}\|^2 + \|\varepsilon_{u_1}\|^2]}{48 \eta_T^2 L^2 T_1} \]
\[ = \frac{2E[F(w_1) - F_\star]}{\eta T_1 L} + \frac{c^2 \sigma^2 \eta_T^2}{3 L^2} + \frac{E[\|\varepsilon_{v_1}\|^2 + \|\varepsilon_{u_1}\|^2]}{48 \eta_T^2 L^2 T_1} \]
\[ \leq \frac{2\epsilon_0}{\eta T_1 L} + \frac{c^2 \sigma^2 \eta_T^2}{3 L^2} + \frac{\sigma^2}{24 \eta_T^2 L^2 T_1} \leq \frac{c^2 \sigma^2}{64 L^2} = \mu \epsilon_1. \]

Starting from the second stage, we will prove by induction. Suppose we are at \(k\)-th stage. Assuming that \(F(w_{k-1}) - F_\star \leq \epsilon_{k-1}\) and \(\|\varepsilon_{v_{k-1}}\|^2 + \|\varepsilon_{u_{k-1}}\|^2 \leq \mu \epsilon_{k-1}\) after the \((k-1)\)-th stage, we will
show that $\mathbb{E}[\|\varepsilon_{u_k}\|^2 + \|\varepsilon_{v_k}\|^2] \leq \mu \epsilon_k$ by induction. Note that the induction of $F(w_k) - F(w_0)$ will be addressed later in Theorem 2.

$$
\mathbb{E}[\|\varepsilon_{u_k}\|^2 + \|\varepsilon_{v_k}\|^2] \leq \frac{2\mathbb{E}[F(w_{k-1}) - F_*]}{\eta_k T_k L} + \frac{c^3 \sigma^2 \eta_k^2}{3L^2} + \frac{\mathbb{E}[\|\varepsilon_{v_{k-1}}\|^2 + \|\varepsilon_{u_{k-1}}\|^2]}{48\eta_k^2 L^2 T_k}
$$

$$
= \frac{2\epsilon_{k-1}}{\eta_k T_k L} + \frac{c^3 \sigma^2 \eta_k^2}{3L^2} + \frac{\mu \epsilon_{k-1} - \mu \epsilon_k}{2} = \mu \epsilon_k,
$$

where the last inequality follows from the setting that $\eta_k = \frac{\sqrt{\sigma \epsilon_L}}{2c\sigma} \leq \frac{1}{16L}$, and $T_k = \max\{\frac{96\sigma}{\mu^2 \sigma \sqrt{\epsilon L}}, \frac{16c^2 \sigma^2}{\mu L^4 \epsilon_k}\}$, where $c = 104L^2$.

\subsection*{10.4 Proof of Theorem 2}

\begin{proof}
Without loss of generality, let’s assume that $\epsilon_0 = \Delta_F \geq \frac{c^3 \sigma^2 \epsilon_L}{64\mu L^2}$, i.e., $\frac{\sqrt{\sigma \epsilon_L}}{2c\sigma} \geq \frac{1}{16L}$. The case that $\Delta_F < \frac{c^3 \sigma^2 \epsilon_L}{64\mu L^2}$ can be simply covered by our proof. Then, denote $\epsilon_1 = \frac{c^3 \sigma^2 \epsilon_L}{64\mu L^2}$ and $\epsilon_k = \epsilon_1/2^{k-1}$, $c = 104L^2$.

Note that in below the numerical subscripts denote the epoch index $(1, \ldots, K)$ (different from in proof of Lemma 3 which all are in one stage). Let’s consider the first stage, we have initialization such that $F(w_0) - F_* = \Delta_F$ and $\mathbb{E}[\|\varepsilon_{v_0}\|^2 + \|\varepsilon_{u_0}\|^2] \leq \sigma^2$. We bound the the error of first stage’s output as follows,

$$
\mathbb{E}[F(w_1) - F_*] \leq \frac{1}{2\mu} \mathbb{E}[\|G_0(w_1)\|^2]
$$

$$
\leq \frac{8\epsilon_{1}}{\mu \eta_1 T_1 L} + \frac{c^3 \sigma^2 \epsilon_1}{4L^2 \eta_1^2} + \frac{\mathbb{E}[\|\varepsilon_{v_1}\|^2 + \|\varepsilon_{u_1}\|^2]}{4L \eta_1 T_1}
$$

$$
\leq \frac{8\Delta_F}{\mu \eta_1 T_1 L} + \frac{c^3 \sigma^2 \epsilon_1}{4L^2 \eta_1^2} + \frac{\sigma^2}{8\mu \eta_1^2 T_1 L} \leq \frac{c^3 \sigma^2 \epsilon_1}{64\mu L^2} = \epsilon_1,
$$

where the first inequality uses PL condition, the second inequality use Theorem 3 and the fourth inequality uses the setting of $\eta_1 = \frac{1}{16L}$ and $T_1 = O(\max(\frac{\Delta_F}{\epsilon_1}, 1))$.

Starting from the second stage, we will prove by induction. Denote $\epsilon_1 = \frac{c^3 \sigma^2 \epsilon_L}{64\mu L^2}$ and $\epsilon_k \leq \epsilon_1/2^{k-1}$ for $k \geq 2$. Suppose at the beginning of $k$-stage ($k \geq 2$), we have $\mathbb{E}[F(w_{k-1}) - F_*] \leq \epsilon_{k-1}$ and $\mathbb{E}[\|\varepsilon_{v_{k-1}}\|^2 + \|\varepsilon_{u_{k-1}}\|^2] \leq \mu \epsilon_{k-1}$. When $k \geq 2$, we have $\eta_k = \frac{\sqrt{\sigma \epsilon_L}}{2c\sigma} \leq \frac{1}{16L}$. Then by Lemma 3 and Theorem 3, setting $T_k = \max\{\frac{96\sigma}{\mu^2 \sigma \sqrt{\epsilon L}}, \frac{16c^2 \sigma^2}{\mu L^4 \epsilon_k}\}$, RECOVER satisfies the following equations at the $k$-th stage,

$$
\mathbb{E}[F(w_k) - F_*] \leq \frac{1}{2\mu} \mathbb{E}[\|G_k(w_k)\|^2]
$$

$$
\leq \frac{8\epsilon_{k-1}}{\mu \eta_k T_k L} + \frac{c^3 \sigma^2 \epsilon_{k-1}}{4L^2 \eta_k^2} + \frac{\mathbb{E}[\|\varepsilon_{v_{k-1}}\|^2 + \|\varepsilon_{u_{k-1}}\|^2]}{4L \eta_k T_k}
$$

$$
\leq \frac{8\epsilon_{k-1}}{\mu \eta_k T_k L} + \frac{c^3 \sigma^2 \epsilon_{k-1}}{4L^2 \eta_k^2} + \frac{\mu \epsilon_{k-1} - \mu \epsilon_k}{2} = \mu \epsilon_k,
$$

where the forth inequality is implied by the induction hypothesis and the last inequality holds by the setting of $\eta_k$ and $T_k$.
Combing two cases, after $K \leq 1 + \log_2(\epsilon_1/\epsilon) \leq \log_2(\epsilon_0/\epsilon)$ stages, $\mathbb{E}[F(w_k) - F(w_*)] \leq \epsilon$.

By setting $c = 104L^2$, following the proof of Theorem 6, the sample complexity of RECOVER equals to the number of samples in the first stage, i.e., $T_1$ plus the number of samples in later stages, i.e. $\sum_{k=2}^{K} T_k$, which is

$$T_1 + \sum_{k=2}^{K} T_k = O \left( \frac{\Delta^F}{\sigma^2} + \sum_{k=2}^{K} T_k \right)$$

$$= O \left( \sum_{k=2}^{K} \left( \frac{c \sigma}{\mu^{3/2} \sqrt{\epsilon_k L}} + \frac{L^2 \sigma^2}{\mu \epsilon_k} \right) \right)$$

$$\leq O \left( \frac{c \sigma}{\mu^{3/2} \sqrt{\epsilon L}} + \frac{c^2 \sigma^2}{L^2 \mu \epsilon} \right) \mu \geq \epsilon \leq O \left( \frac{1}{\mu \epsilon} \right).$$

\[ \square \]

11 Derivation of the Compositional Formulation

Recall the problem:

$$\min_{w \in \mathbb{R}^d} \max_{p \in \Delta_n} F_p(w) = \sum_{i=1}^{n} p_i \ell(w; z_i) - h(p, 1/n) + r(w),$$

where $\Delta_n = \{ p \in \mathbb{R}^n : \sum_i p_i = 1, 0 \leq p_i \leq 1 \}$. In order to solve the inner maximization, we will fix $w$ and derive an optimal solution $p^*(w)$ that depends on $w$. To this end, we consider the following problem:

$$\min_{p \in \Delta_n} - \sum_{i=1}^{n} p_i \ell(w; z_i) + h(p, 1/n)$$

where $r(w)$ was neglected since it does not involve $p$. Note the expression of $h(p, 1/n) = \lambda \sum_i p_i \log(n p_i) = \lambda \sum_i p_i \log(p_i) + \lambda \log(n)$ due to $\sum_i p_i = 1$. There are three constraints to handle, i.e., $p_i \geq 0, \forall i$ and $p_i \leq 1, \forall i$ and $\sum_i p_i = 1$. Note that the constraint $p_i \geq 0$ is enforced by the term $p_i \log(p_i)$, otherwise the above objective will become infinity. As a result, the constraint $p_i < 1$ is automatically satisfied due to $\sum_i p_i = 1$ and $p_i \geq 0$. Hence, we only need to explicitly tackle the constraint $\sum_i p_i = 1$. To this end, we define the following Lagrangian function

$$L_w(p, \mu) = - \sum_{i=1}^{n} p_i \ell(w; z_i) + \lambda (\log(n) + \sum_i p_i \log(p_i)) + \mu (\sum_i p_i - 1)$$

where $\mu$ is the Lagrangian multiplier for the constraint $\sum_i p_i = 1$. The optimal solutions satisfy the KKT conditions:

$$- \ell(w; z_i) + \lambda (p^*_i(w) + 1) + \mu = 0,$$

$$\sum_i p^*_i(w) = 1$$

From the first equation, we can derive $p^*_i(w) \propto \exp(\ell(w; z_i)/\lambda)$. Due to the second equation, we can conclude that $p^*_i(w) = \frac{\exp(\ell(w; z_i)/\lambda)}{\sum_i \exp(\ell(w; z_i)/\lambda)}$. Plugging this optimal $p^*(w)$ into the original min-max objective, we have

$$\sum_{i=1}^{n} p^*_i(w) \ell(w; z_i) - \lambda (\log(n) + \sum_i p^*_i(w) \log(p^*_i(w))) + r(w) = \lambda \log \frac{1}{n} \sum_i \exp(\ell(w; z_i)/\lambda) + r(w),$$

which is the $F_{dro}(w)$ in the paper (the expression above Eq (2)).