Proximal Compositional Optimization for Distributionally Robust Learning

Prashant Khanduri [1] **Chengyin Li** [1] **Rafi Ibn Sultan** [1] **Yao Qiang** [1] **Joerg Kliewer** [2] **Dongxiao Zhu** [1]

Abstract

Recently, compositional optimization (CO) has gained popularity because of its applications in distributionally robust optimization (DRO) and many other machine learning problems. Often (non-smooth) regularization terms are added to an objective to impose some structure and/or improve the generalization performance of the learned model. However, when it comes to CO, there is a lack of efficient algorithms that can solve regularized CO problems. Moreover, current state-of-the-art methods to solve such problems rely on the computation of large batch gradients (depending on the solution accuracy) not feasible for most practical settings. To address these challenges, in this work, we consider a regularized version of the CO problem that often arises in DRO formulations and develop a proximal algorithm for solving the problem. We perform a Moreau envelope-based analysis and establish that without the need to compute large batch gradients `Prox-DRO` achieves $\mathcal{O}(\epsilon^{-2})$ sample complexity, that matches the vanilla SGD guarantees for solving non-CO problems. We corroborate our theoretical findings with empirical studies on large-scale DRO problems.

1. Introduction

Composite optimization (CO) problems deal with the minimization of the composition of functions. A standard CO problem takes the form

$$\min_{x\in\mathbb{R}^d} f(g(x)) \text{ with } g(x) := \mathbb{E}_{\zeta\sim\mathcal{D}_g}[g(x;\zeta)], \quad (1)$$

where $x \in \mathbb{R}^d$ is the optimization variable, the mappings $f : \mathbb{R}^{d_g} \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}^{d_g}$ are smooth functions, and $\zeta \sim \mathcal{D}_g$ represents a stochastic sample of $g(\cdot)$ from distribution $\mathcal{D}_g$. The problems of the form (1) find applications in a broad range of machine learning applications, including but not limited to distributionally robust optimization (DRO) (Qi et al., 2022), meta-learning (Finn et al., 2017), phase retrieval (Duchi & Ruan, 2019), portfolio optimization (Shapiro et al., 2021), and reinforcement learning (Wang et al., 2017).

Regularization terms are often added to the training objectives to impose some structure to the obtained solutions (Hoerl & Kennard, 1970; Tibshirani, 1996; Bennett & Mangasarian, 1992; Zou & Hastie, 2005). For example, a non-smooth ℓ_1-norm penalty is usually added to the optimization objective to enforce sparsity in the solutions (Beck & Teboulle, 2009). Proximal methods present a popular approach to tackle (non-smooth) regularization terms with the optimization objective for non-compositional problems (Ghadimi & Lan, 2012; Lewis & Wright, 2016). However, there is a lack of efficient algorithms for solving potentially non-smooth regularized CO problems. To fill this gap in the literature, in this work, we focus on a more challenging version of the CO problem in (1) that often arises in the DRO formulation. Specifically, the problems that jointly minimize the summation of a (non-smooth) regularization, a compositional, and a non-compositional objective.

DRO has recently garnered significant attention from the research community because of its capability of handling noisy labels (Chen et al., 2022), training fair machine learning models (Haddadpour et al., 2022), imbalanced (Qi et al., 2020a) and adversarial data (Chen & Paschalidis, 2018). A standard approach to solve DRO is to utilize primal-dual algorithms (Nemirovski et al., 2009) that are inherently slow because of a large number of stochastic constraints. The CO formulation enables the development of faster (dual-free) primal-only DRO algorithms (Haddadpour et al., 2022). A major drawback of current approaches to tackle CO problems is that they either rely on complicated, double-loop algorithms with very large gradient (and function) evaluations (Haddadpour et al., 2022) or are incapable of handling non-smooth regularization terms (Wang et al., 2017; Ghadimi et al., 2020; Chen et al., 2022). In this work, we address these challenges and develop, `Prox-DRO`, a proximal algorithm to solve typical versions of regularized CO problems that are often observed in DRO formulations. Major contributions of our work are listed below:

[1]Department of Computer Science, Wayne State University, MI, USA [2]Department of Electrical & Computer Engineering, NJIT, NJ, USA. Correspondence to: Prashant Khanduri <khanduri.prashant@wayne.edu>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

– We develop, `Prox-DRO`, a Proximal-SGD-type algorithm to solve potentially non-smooth CO problems. The proposed algorithm utilizes a hybrid momentum-based estimator to learn the compositional embedding, $g(\cdot)$, and combine it with the proximal stochastic gradient (SG) updates. This construction allows us to circumvent the need to compute large accuracy-dependent batch sizes for computing the gradients and the compositional function evaluations.

– The regularized CO problem may be potentially non-smooth, therefore, the standard notion of stationarity is not sufficient to characterize the quality of the obtained solutions. To address this, we adopt a Moreau envelope-based analysis. We show that to reach an ϵ-stationary point of the Moreau envelope of the non-smooth objective, `Prox-DRO` requires $\mathcal{O}(\epsilon^{-2})$ samples while computing the batch gradients that are independent of the solution accuracy. To the best of our knowledge, this is the first analysis to establish such a guarantee for solving general DRO problems.

– We conduct experiments on large-scale DRO problems to corroborate our theoretical findings. Our experiments establish the superior performance of `Prox-DRO` compared to state-of-the-art methods.

Notations: The expected value of a random variable (r.v) X is denoted by $\mathbb{E}[X]$. Conditioned on an event $\mathcal{F}$ the expectation of X is denoted by $\mathbb{E}[X|\mathcal{F}]$. We denote by $\mathbb{R}$ (resp. $\mathbb{R}^d$) the real line (resp. the d dimensional Euclidean space). The notation $\|\cdot\|$ defines a standard ℓ_2-norm. For a set B, $|B|$ denotes the cardinality of B. We use $\xi \sim \mathcal{D}_h$ and $\zeta \sim \mathcal{D}_g$ to denote the stochastic samples of $h(\cdot)$ and $g(\cdot)$ from distributions $\mathcal{D}_h$ and $\mathcal{D}_g$, respectively. A batch of samples of $h(\cdot)$ (resp. $g(\cdot)$) is denoted by b_h (resp. b_g). Joint samples of $h(\cdot)$ and $g(\cdot)$ is denoted by $\bar{\xi} = \{b_h, b_g\}$.

2. Problem

In general, a DRO problem aims to solve

$$\min_{x\in\mathbb{R}^d} \left\{ \Psi(x) := \max_{\xi\sim Q} \mathbb{E}_Q[\ell(x;\xi)] \right\} \tag{2}$$

where $\xi \sim Q$ represents a sample from distribution Q, $\ell(x;\xi)$ is the loss function and Q belongs to an uncertainty set $\mathcal{U}_m$ (Duchi & Ruan, 2019). It has been established in the past that particular reformulation of the DRO (see Section 2.1) problem in (2) can be equivalently stated as regularized CO that we tackle in this work (Haddadpour et al., 2022). Specifically, we consider the following CO problem

$$\min_{x\in\mathbb{R}^d} \left\{ \Psi(x) := r(x) + \underbrace{h(x) + f(g(x))}_{\Phi(x)} \right\} \tag{3}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ are $h(x) := \mathbb{E}_{\xi\sim\mathcal{D}_h}[h(x;\xi)]$ and $g(x) := \mathbb{E}_{\zeta\sim\mathcal{D}_g}[g(x;\zeta)]$, respectively. The mapping $r : \mathbb{R}^d \to \mathbb{R}$ is possibly a non-smooth closed and convex proximable function (see Definition 3.5) while $f(\cdot)$ is the same as in (1). Also, $\xi \sim \mathcal{D}_h$ (resp. $\zeta \sim \mathcal{D}_g$) represents a sample of $h(\cdot)$ (resp. $g(\cdot)$) from distribution $\mathcal{D}_h$ (resp. $\mathcal{D}_g$).

We note that in contrast to the standard CO problem in (1), the formulation in (3) represents a hybrid objective, $\Phi(\cdot)$, which is a combination of compositional and non-compositional objectives and is regularized by a potentially non-smooth regularizer $r(\cdot)$. Note that the joint function$\Psi(\cdot)$ in (3) may be non-smooth. Therefore, the standard notion of smoothness is not applicable to this problem. In this work, we develop a proximal algorithm for solving (3) and utilize a Moreau envelope-based notion of stationarity to evaluate the algorithm's performance (Davis & Drusvyatskiy, 2019). Next, we discuss various DRO formulations where problems of the form (3) often arise

2.1. Examples: Regularized CO to solve DRO Problems

In this section, we discuss different DRO formulations that can be efficiently solved using CO. A standard reformulation of (2) with a set of m training samples $\{\zeta_i\}_{i=1}^m$ is

$$\min_{x\in\mathcal{X}} \max_{\{\mathbf{p}\in P_m : D_*(\mathbf{p},\mathbf{1}/m)\le\rho\}} \sum_{i=1}^m p_i \ell(x;\zeta_i) - \lambda_0 D_*(\mathbf{p},\mathbf{1}/m) \tag{4}$$

where $x \in \mathbb{R}^d$ is the model parameter and $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set, $P_m := \{\mathbf{p} \in \mathbb{R}^m : \sum_{i=1}^m p_i = 1, p_i \ge 0\}$ denotes a m-dimensional simplex, $D_*(\mathbf{p},\mathbf{1}/m)$ denotes a divergence metric that measures the distance between $\mathbf{p}$ and uniform probability $\mathbf{1}/m \in \mathbb{R}^m$, $\ell(x,\zeta_i)$ denotes the loss function on sample ζ_i, ρ is a constraint parameter, and λ_0 is a hyperparameter. Next, we discuss two popular equivalent reformulations of (4) in the form of CO problems.

DRO with KL-Divergence. Problem (4) is referred to as a KL-regularized DRO when the distance metric $D_*(\mathbf{p},\mathbf{1}/m)$ utilized to measure the distance between $\mathbf{p}$ and $\mathbf{1}/m$ is the KL-Divergence, i.e. when we have $D_*(\mathbf{p},\mathbf{1}/m) = D_{\text{KL}}(\mathbf{p},\mathbf{1}/m)$ with $D_{\text{KL}}(\mathbf{p},\mathbf{1}/m) := \sum_{i=1}^m p_i \log(p_i m)$. For this case, an equivalent reformulation of problem (4) is (Qi et al., 2022)

$$\min_{x\in\mathcal{X}} \min_{\lambda\ge\lambda_0} \lambda \log\left(\frac{1}{m}\sum_{i=1}^m \exp\left(\frac{\ell(x;\zeta_i)}{\lambda}\right)\right) + (\lambda - \lambda_0)\rho,$$

a CO problem with joint parameter $[x^\top, \lambda]^\top \in \mathbb{R}^{d+1}$, $g(\cdot)$ defined as $g(x,\lambda) = \left[\lambda, 1/m\sum_{i=1}^m \exp(\ell(x;\zeta_i)/\lambda)\right] \in \mathbb{R}^2$, $f(g(x)) = g_1\log(g_2) + g_1(\rho)$ and $h(x) = 0$. Note here that function $r(x,\lambda)$ takes the form of indicator function

Table 1. Comparison of `Prox-DRO` with the existing works. Here, CO + Proj (resp. Prox) refers to the compositional optimization + projection (resp. proximal) updates. CO + Non-CO + Prox refers to compositional, non-compositional, and proximal objectives. SGD refers to stochastic gradient descent update for the model parameters. A-SGD (resp. M-SGD) refers to SGD update with acceleration (resp. momentum). M (resp. MVR) refers to the momentum (resp. momentum-based variance reduction) update for inner-function estimation. VR refers to variance reduction. (I) and (O) refers to the inner and outer loop, respectively.
* Theoretical guarantees for GCIVR exist only for the finite sample setting with m total samples.

ALGORITHM	SETTING	UPDATE	BATCH-SIZES	CONVERGENCE
SCGD (Wang et al., 2017)	CO + Proj	SGD + M	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-4})$
ASC-GD (Wang et al., 2016)	CO + Prox	A-SGD + M	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-2.25})$
NASA(Ghadimi et al., 2020)	CO + Proj	M-SGD + M	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-2})$
SCSC (Chen et al., 2021)	CO	SGD + MVR	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-2})$
GCIVR* (Haddadpour et al., 2022)	CO + Non-CO + Prox	VR	$\sqrt{m}$ (I), m (O)	$\mathcal{O}(\min\{\sqrt{m}\epsilon^{-1}, \epsilon^{-1.5}\})$
`Prox-DRO` (Ours)	CO + Non-CO + Prox	SGD + MVR	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-2})$

on the set $\{[x^\top, \lambda]^\top \in \mathbb{R}^{d+1} : x \in \mathcal{X}, \lambda \geq \lambda_0\}$ which is a non-smooth function.

DRO with χ^2-Divergence Similar to KL-regularized DRO, (4) is referred to as a χ^2-regularized DRO when $D_*(\mathbf{p}, \mathbf{1}/m)$ utilized to measure the distance between $\mathbf{p}$ and $\mathbf{1}/m$ is χ^2-Divergence, i.e., when we have $D_*(\mathbf{p}, \mathbf{1}/m) = D_{\chi^2}(\mathbf{p}, \mathbf{1}/m)$ with $D_{\chi^2}(\mathbf{p}, \mathbf{1}/m) := m/2 \sum_{i=1}^m (p_i - 1/m)^2$. For this case, an equivalent reformulation of problem (4) for $\rho = \infty$ is (Haddadpour et al., 2022)

$$\min_{x\in\mathcal{X}} \frac{1}{2\lambda_0 m} \sum_{i=1}^m \big(\ell(x;\zeta_i)\big)^2 + \frac{1}{2\lambda_0}\Big(\frac{1}{m}\sum_{i=1}^m \ell(x;\zeta_i)\Big)^2$$

a CO problem with $g(x) = 1/m \sum_{i=1}^m \ell(x;\zeta_i)$, $f(g) = g^2/2\lambda_0$, $h(x) = -\frac{1}{2\lambda_0 m}\sum_{i=1}^m \big(\ell(x;\zeta_i)\big)^2$ and $r(x)$ being the indicator function of set $\mathcal{X}$.

We note that DRO with Wasserstein distance can also be reformulated in the form (3). Please see Section 2 of (Haddadpour et al., 2022) for more details. It is also worth mentioning that in general one may include additional regularization terms (e.g., ℓ_1-penalty) with the objective to impose additional structure to the obtained solutions (Beck & Teboulle, 2009). Moreover, we would like to point out that the regularized CO is not limited to DRO formulations and can be utilized to solve problems in multiple domains (see Section 1).

2.2. Related Work

The first non-asymptotic analysis of stochastic CO problems was performed in (Wang et al., 2017) where the authors proposed Stochastic Compositional Gradient Descent (SCGD) a two-timescale algorithm for solving problem (1). The convergence of SCGD was improved in (Wang et al., 2016) where the authors considered a regularized CO problem and proposed Accelerated Stochastic Compositional Proximal Gradient (ASC-PG) an accelerated variant of SCGD. Both SCGD and ASC-PG achieved convergence rates strictly worse than SGD for solving non-compositional problems. Recently, (Ghadimi et al., 2020) and (Chen et al., 2021) developed single time-scale algorithms, Nested Averaged Stochastic Approximation (NASA) and Stochastically Corrected Stochastic Compositional gradient method (SCSC), respectively. Both NASA and SCSC matched the guarantees of SGD for solving non-CO problems without requiring the need to compute large batch gradients (or function evaluations). Variance-reduced algorithms for solving the CO problems have also been considered in the literature, however, a major drawback of such approaches is a double-loop structure and the reliance of batch size on the desired solution accuracy (Lian et al., 2017; Zhang & Xiao, 2019; Hu et al., 2019). Recently, (Haddadpour et al., 2022) developed Generalized Composite Incremental Variance Reduction (GCIVR) a variance-reduced double loop algorithm for solving problems of the form (3) in the finite sum setting. However, a major drawback of the GCIVR is its reliance on large batch gradients and the double-loop structure which is often less preferred compared to single-loop algorithms. Please see Table 1 for a summary of the discussion. Also, refer to Appendix A for a detailed literature review of DRO.

In summary, there is a lack of efficient CO algorithms to solve problems of the form (3). Key issues with the current approaches are

- Inability to handle potentially non-smooth (but proximable) regularization terms (Chen et al., 2021; Ghadimi et al., 2020; Wang et al., 2017).
- Worse performance of proximal CO algorithms compared to vanilla SGD implementations to solve non-compositional problems (Wang et al., 2016) or reliance on the computation of large accuracy-dependent batch sizes (Haddadpour et al., 2022).

In our work, we address these challenges and develop `Prox-DRO`, a proximal framework to solve hybrid optimization problems of the form (3).

3. Preliminaries

In this section, we introduce the assumptions, definitions, and preliminary lemmas utilized in the analysis of `Prox-DRO`.

Definition 3.1 (Lipschitzness)**.** For all $x_1, x_2 \in \mathbb{R}^d$, a function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is referred to as:

1. **Lipschitz smooth** if it is differentiable and $\|\nabla\Phi(x_1) - \nabla\Phi(x_2)\| \leq L_\Phi\|x_1 - x_2\|$ for some $L_\Phi > 0$.
2. **Lipschitz** if $\|\Phi(x_1) - \Phi(x_2)\| \leq B_\Phi\|x_1 - x_2\|$ for some $B_\Phi > 0$.
3. **Mean-Squared Lipschitz** if $\mathbb{E}_{\bar{\xi}}\|\Phi(x_1;\bar{\xi}) - \Phi(x_2;\bar{\xi})\|^2 \leq B_\Phi^2\|x_1 - x_2\|^2$ for some $B_\Phi > 0$ and where $\bar{\xi} \sim \mathcal{D}_\Phi$ represents a stochastic sample of Φ.

Next, we make the following assumptions.

Assumption 3.2 (Lipschitzness)**.** The following holds

1. The functions $f(\cdot)$, $h(\cdot)$, $g(\cdot)$ for all $k \in [K]$ are differentiable and Lipschitz-smooth with constants $L_f, L_h, L_g > 0$, respectively.
2. The function $f(\cdot)$ is Lipschitz with constant $B_f > 0$.
3. The functions $h(\cdot)$ and $g(\cdot)$ are mean-squared Lipschitz with constants $B_h > 0$ and $B_g > 0$, respectively.

Next, we introduce the unbiased and variance assumptions on the gradients and function evaluations.

Assumption 3.3 (Unbiased Gradient and Bounded Variance)**.** The stochastic gradients and function evaluations of the local functions at each client are unbiased and have bounded variance. Specifically, we have

$$\mathbb{E}_\xi[\nabla h(x;\xi)] = \nabla h(x),\ \mathbb{E}_\zeta[\nabla g(x;\zeta)] = \nabla g(x),$$
$$\mathbb{E}_\zeta[g(x;\zeta_k)] = g(x),$$
$$\mathbb{E}_\zeta[\nabla g(x;\zeta)\nabla f(y)] = \nabla g(x)\nabla f(y)$$

and

$$\mathbb{E}_\xi\|\nabla h(x;\xi) - \nabla h(x)\|^2 \leq \sigma_h^2$$
$$\mathbb{E}_\zeta\|\nabla g(x;\zeta) - \nabla g(x)\|^2 \leq \sigma_g^2$$
$$\mathbb{E}_\zeta\|g(x;\zeta) - g(x)\|^2 \leq \sigma_g^2,$$

for some $\sigma_h, \sigma_g > 0$.

A few comments regarding the assumptions are in order. We note that the above assumptions are commonplace in the context of CO problems. Specifically, Assumption 3.2 is required to establish Lipschitz smoothness of the composite objective $\Phi(\cdot)$ (please see Lemma 3.4) and is standard in the analyses of CO problems (Wang et al., 2017; Chen et al., 2021). Assumption 3.3 captures the effect of stochasticity in the data gradient and function evaluations. We note that these assumptions are standard and have been utilized in the past for solving many non-CO problems as well (Ghadimi & Lan, 2013; 2012). Next, we state a preliminary lemma and the performance metrics for analyzing the CO problems of the form (3).

Lemma 3.4 (Lipschitzness of Φ)**.** *Under Assumption 3.2 the composite function, $\Phi(\cdot)$, defined in (3) is Lipschitz smooth with constant: $L_\Phi := L_h + B_f L_g + B_g^2 L_f > 0$.*

Lemma 3.4 establishes Lipschitz smoothness (Definition 3.1) of the compositional function $\Phi(\cdot)$. Moreover, the function $\Psi(\cdot)$ is a non-convex function in general, and therefore, we cannot expect to globally solve (3). We instead rely on finding approximate stationary points of $\Psi(\cdot)$. However, note that problem (3) is a regularized CO problem with mapping $r(\cdot)$ being potentially non-smooth. This implies that the standard notion of stationarity might not be sufficient to characterize the solutions of the regularized CO problem in (3). For this purpose, we utilize a Moreau envelope-based definition of stationarity discussed next (Davis & Drusvyatskiy, 2019).

Definition 3.5 (Moreau Envelope)**.** We define the Moreau envelope of $\Psi(\cdot)$ for any $\lambda > 0$ as

$$\Psi_\lambda(x) := \arg\min_{z\in\mathbb{R}^d}\left\{\Psi(z) + \tfrac{1}{2\lambda}\|z - x\|^2\right\} \text{ for } x \in \mathbb{R}^d,$$

while the proximal operator is defined as

$$\text{prox}_\Psi^\lambda(x) := \arg\min_{z\in\mathbb{R}^d}\left\{\Psi(z) + \tfrac{1}{2\lambda}\|z - x\|^2\right\} \text{ for } x \in \mathbb{R}^d.$$

An important property of the Moreau envelope is that as long as $\lambda < 1/L_\Phi$ the mapping $\Psi_\lambda(x)$ is continuously differentiable. This smoothing of Ψ allows the development of an alternate notion of stationarity defined next.

Definition 3.6 (ϵ-stationary point - Reguralized CO Problem)**.** A point x generated by a stochastic algorithm is an ϵ-stationary point of possibly non-smooth function $\Psi(\cdot)$ (see (3)) if $\mathbb{E}\|\nabla\Psi_\lambda(x)\|^2 \leq \epsilon$ for some $\lambda < 1/L_\Phi$.

Note that the condition $\mathbb{E}\|\nabla\Psi_\lambda(x)\|^2 \leq \epsilon$ ensures the subgradient of $\Psi(\cdot)$ in a neighboring point of x will also be small (Davis & Drusvyatskiy, 2019). Please see Appendix B for further details.

4. Proximal CO: `Prox-DRO`

In this section, we develop `Prox-DRO`, an algorithm to solve the regularized CO problem (3). For `Prox-DRO`, we assume that $r(\cdot)$ is convex and proximable mapping. We utilize the proximal operator of the mapping $r(\cdot)$, $\text{prox}_r^\lambda(x)$ (see Definition 3.5), for updating the model parameters.

The detailed steps of `Prox-DRO` are listed in Algorithm 1. Specifically, for `Prox-DRO` we compute the stochastic

Algorithm 1 Algorithm: `Prox-DRO`

1: **Input**: Parameters: $\{\beta^t\}_{t=0}^{T-1}$, $\{\eta^t\}_{t=0}^{T-1}$,
2: **Initialize**: x^0, y^0
3: **for** $t = 0$ to $T-1$ **do**
4: Sample $\bar{\xi}^t = \{b_g^t, b_h^t\}$ uniformly randomly from $\mathcal{D}_g$ and $\mathcal{D}_h$, respectively
5: `Update`: $\begin{cases} y^t \text{ using (6)} \\ \text{Compute } \nabla\Phi(x^t;\bar{\xi}^t) \text{ using (5)} \\ x^{t+1} = \text{prox}_r^{\eta^t}(x^t - \eta^t \nabla\Phi(x^t;\bar{\xi}^t)) \end{cases}$
6: **end for**
7: **Return:** $x^{a(T)}$ where $a(T) \sim \mathcal{U}\{1, ..., T\}$.

gradient of $\Phi(x^t)$ denoted as $\nabla\Phi(x^t;\bar{\xi}^t)$ in each iteration $t \in \{0, 1, \ldots, T-1\}$ in Step 5 using the chain rule of differentiation as

$$\nabla\Phi(x^t;\bar{\xi}^t) = \frac{1}{|b_h^t|}\sum_{i\in b_h^t} \nabla h(x^t;\xi_i^t) + \frac{1}{|b_g^t|}\sum_{j\in b_g^t} \nabla g(x^t;\zeta_j^t)\nabla f(y^t), \tag{5}$$

where $\bar{\xi}^t = \{b_h^t, b_g^t\}$ represents the stochasticity of the gradient estimate with $b_h^t = \{\xi_i^t\}_{i=1}^{|b_h^t|}$ (resp. $b_g^t = \{\zeta_i^t\}_{i=1}^{|b_g^t|}$) as the batch of stochastic samples of $h(\cdot)$ (resp. $g(\cdot)$) utilized to compute the stochastic gradient estimate. Here, the variable y^t is an estimate of the function $g(x^t)$ for each $t \in \{0, 1, \ldots, T-1\}$ and is updated using the momentum-based estimator proposed in (Chen et al., 2021) as

$$y^t = (1-\beta^t)\Big(y^{t-1} - \frac{1}{|b_t^g|}\sum_{i\in b_t^g} g(x^{t-1};\zeta_i^t)\Big) + \frac{1}{|b_g^t|}\sum_{i\in b_g^t} g(x_t;\zeta_i^t), \tag{6}$$

where $\beta_t \in (0,1)$ is a momentum parameter. Finally, the model parameter x_t for all $t \in [T]$ is updated in Step 5 first by taking an SGD update and then evaluating the proximal operator of $r(\cdot)$ on the updated step.

Next, we characterize the convergence of `Prox-DRO`.

5. Main Result: Convergence of `Prox-DRO`

In this section, we present the convergence guarantees for `Prox-DRO`. In the following, we characterize the behavior of the gradient of the Moreau envelope of $\Psi(\cdot)$.

Theorem 5.1 (Convergence of `Prox-DRO`)**.** *For Algorithm 1, choosing the step-size η^t such that*

$$\eta^t \le \max\left\{\frac{1}{\bar{L}}, \frac{2\bar{L}_{\Phi,\gamma}}{L_\Phi^2}, \frac{1}{4}\Big(1+\frac{1}{\gamma}\Big)\right\},$$

where $\bar{L}_{\Phi,\gamma} := \max\{L_\Phi, 1+\gamma\}$ with constant $\gamma > 0$ and $\bar{L} \in (8\bar{L}_{\Phi,\gamma}, 16\bar{L}_{\Phi,\gamma}]$. Moreover, choosing the momentum parameter β^t as

$$\beta^t = 2\Big(1+\frac{1}{\gamma}\Big)B_g^2 L_f^2 \bar{L}\cdot\eta^t.$$

Then for the selection of batch-sizes $|b_h^t| = |b_g^t| = |b|$ for all $t \in \{0, 1, \ldots, T-1\}$ under Assumptions 3.2 and 3.3, the iterates generated by Algorithm 1 satisfy

$$\sum_{t=0}^{T-1}\eta^t \mathbb{E}\|\nabla\Psi_{1/\bar{L}}(x^t)\|^2 \le \left(\frac{2\bar{L}}{\bar{L}-8\bar{L}_{\Phi,\gamma}}\right)\Bigg[\big(\Psi_{1/\bar{L}}(x^0) - \Psi_{1/\bar{L}}^*\big) + \|y^0 - g(x^0)\|^2 + \frac{C_\Psi}{|b|}\sum_{t=0}^{T-1}(\eta^t)^2\Bigg],$$

where $\nabla\Psi_{1/\bar{L}}(x^t)$ is the Moreau envelope of $\Psi(\cdot)$ defined in Definition 3.5, $\Psi_{1/\bar{L}}^* = \min_x \Psi_{1/\bar{L}}(x)$ and C_Ψ is a constant defined in Appendix B.

The detailed proof of Theorem 5.1 is given in Appendix B. Compared to a standard non-compositional problem (Davis & Drusvyatskiy, 2019), where one can directly establish descent in the Moreau envelope of $\Psi(\cdot)$, our analysis utilizes a potential function-based analysis to prove Theorem 5.1. Specifically, the designed potential function depends on the Moreau envelope of $\Psi(\cdot)$ as well as on the bias in the computed stochastic gradients for the CO problem. Our analysis generalizes the results of (Davis & Drusvyatskiy, 2019) to CO problems via utilizing a momentum update (see (6)) that helps mitigate the gradient bias asymptotically. Next, we characterize the sample complexity of `Prox-DRO`.

Corollary 5.2 (Sample Complexity of `Prox-DRO`)**.** *Under the same setting as Theorem 5.1, choosing $\eta^t = \eta = \kappa\sqrt{\frac{|b|}{T}}$ for all $t \in \{0, 1, \ldots, T-1\}$ for some $\kappa > 0$ such that*

$$\kappa \le \sqrt{\frac{T}{|b|}}\cdot\max\left\{\frac{1}{\bar{L}}, \frac{2\bar{L}_{\Phi,\gamma}}{L_\Phi^2}, \frac{1}{4}\Big(1+\frac{1}{\gamma}\Big)\right\}.$$

Then for the choice $\bar{L} = 16\bar{L}_{\Phi,\gamma}$, the iterate $x^{a(T)}$ chosen according to Algorithm 1 satisfies

$$\mathbb{E}\|\nabla\Psi_{1/\bar{L}}(x^{a(T)})\|^2 \le 4\cdot\left[\frac{\big(\Psi_{1/\bar{L}}(x^0) - \Psi_{1/\bar{L}}^*\big) + \|y^0 - g(x^0)\|^2 + C_\Psi}{\kappa\sqrt{|b|T}}\right],$$

Moreover, this implies that the sample complexity of `Prox-DRO` is $\mathcal{O}(\epsilon^{-2})$.

A key consequence of Corollary 5.2 is that `Prox-DRO` achieves a sample complexity of $\mathcal{O}(\epsilon^{-2})$ without requiring large batch sizes of stochastic gradient (or function) evaluations. This is in key contrast to (Haddadpour et al., 2022) where the batch sizes depend on the total sample size. In addition, our analysis improves the guarantees in (Wang et al., 2016) and establishes that the regularized CO problems can be solved with the same sample complexity as the standard non-compositional SGD (Ghadimi & Lan, 2013).

6. Experiments

In this section, we evaluate the performance of `Prox-DRO` with popular baselines to solve DRO problems. The models are trained on an NVIDIA GeForce RTX 3090 GPU with 24 GB memory. All experiments are conducted using Python 3.9.16 and PyTorch 1.8. To evaluate the performance of `Prox-DRO`, we focus on two tasks: classification with an imbalanced dataset and learning with fairness constraints. The goal of the experiments is twofold, first, we establish superior performance of `Prox-DRO` in terms of training/testing accuracy. Second, we establish the fast convergence of `Prox-DRO` compared to competing baselines.

Classification with Imbalanced Dataset. For the image classification task, we utilize CIFAR10-ST and CIFAR100-ST datasets (Qi et al., 2020b) (imbalanced versions of CIFAR10 and CIFAR100 (Krizhevsky, 2009)), and evaluate the performance via training and testing accuracy achieved by different algorithms with ResNet20. The baselines adopted for comparison are a popular DRO method, FastDRO (Levy et al., 2020), a primal-dual SGD approach to solve constrained problems with many constraints, PDSGD (Xu, 2020), and a popular baseline minibatch SGD, MBSGD, customized for CO (Ghadimi & Lan, 2013). For each algorithm, we used a batch size of 128, and the learning rates are tuned from the set $\{0.001, 0.01, 0.05, 0.1\}$, the learning rate was dropped to $1/10^{\text{th}}$ after 90 iterations. As can be observed in Figure 1, `Prox-DRO` outperforms other baselines for both CIFAR10-ST and CIFAR100-ST in terms of guaranteeing higher training and testing accuracies while achieving faster convergence compared to other baselines.

Classification with Fairness Constraints. For the second task, we use the Adult dataset (Dua & Graff, 2017) for enforcing equality of opportunity (on protected classes) on tabular data classification (Hardt et al., 2016). We use GCIVR (Haddadpour et al., 2022) as the baseline model to compare with `Prox-DRO`, since like `Prox-DRO` it is the only algorithm that can deal with regularization, as well as compositional and non-compositional objectives at the same time. We also implement a simple parallel SGD algorithm as a baseline that ignores the fairness constraints, referred to as unconstrained in the experiments. We utilize a logistic regression model for the task. We adopt the parameter settings as suggested in (Haddadpour et al., 2022) and for `Prox-DRO` we keep the same setting as in the earlier task. For this setting, the performance is evaluated by training/testing accuracy, and the constraint violations, which are measured by the gap between the true positive rate of the overall data and the protected groups (Haddadpour et al., 2022). In Figure 2, we compare the training/testing accuracies and the max group violation during the training and testing phase of `Prox-DRO` against the baselines on the Adult dataset. We note that `Prox-DRO` clearly outperforms GCIVR and the baseline SGD in terms of training and testing performance. Moreover, we note that the `Prox-DRO` matches the constraint violation performance of GCIVR as the iteration count increases.

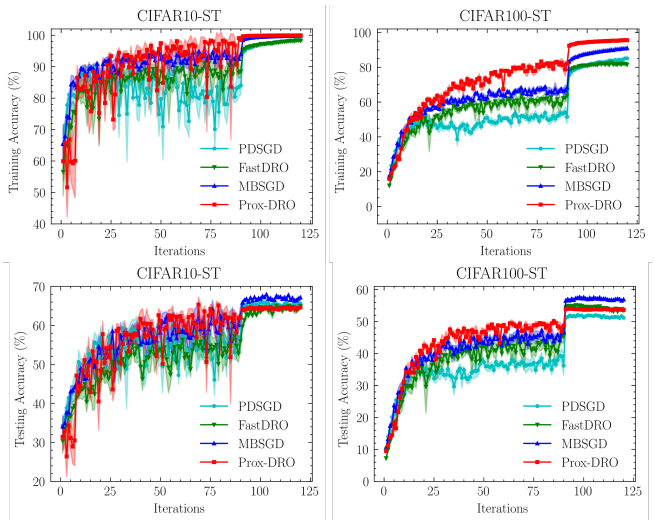

Figure 1. Training and testing accuracies of different algorithms on CIFAR10-ST and CIFAR100-ST datasets .

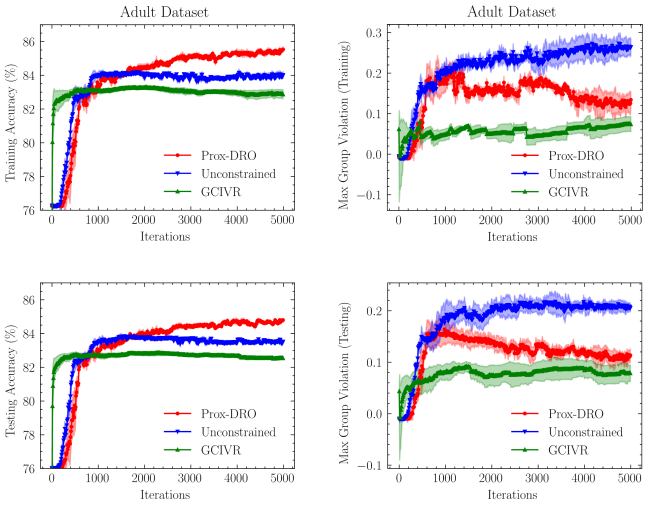

Figure 2. Training and testing accuracies and maximum group violation of different algorithms on the Adult dataset.

7. Conclusion

In this work, we proposed `Prox-DRO`, a proximal algorithm to solve CO problems of the form (3) that often arise in DRO formulations. Utilizing a Moreau envelope-based analysis, we established $\mathcal{O}(\epsilon^{-2})$ convergence of `Prox-DRO`, which matches the convergence of vanilla SGD for minimizing non-compositional objectives. Importantly, `Prox-DRO` achieves this performance without the computation of large accuracy-dependent batch gradients or function evaluations.

References

Alacaoglu, A., Cevher, V., and Wright, S. J. On the complexity of a practical primal-dual coordinate method. *arXiv preprint arXiv:2201.07684*, 2022.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Bennett, K. P. and Mangasarian, O. L. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1):23–34, 1992.

Bertsimas, D., Gupta, V., and Kallus, N. Data-driven robust optimization. *Mathematical Programming*, 167:235–292, 2018.

Chen, M., Zhao, Y., He, B., Han, Z., Wu, B., and Yao, J. Learning with noisy labels over imbalanced subpopulations. *arXiv preprint arXiv:2211.08722*, 2022.

Chen, R. and Paschalidis, I. C. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19 (13), 2018.

Chen, T., Sun, Y., and Yin, W. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Duchi, J. C. and Ruan, F. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

Duchi, J. C., Glynn, P. W., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

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

Ghadimi, S., Ruszczynski, A., and Wang, M. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, 2020.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Karbasi, A. Learning distributionally robust models at scale via composite optimization. *arXiv preprint arXiv:2203.09607*, 2022.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Hu, W., Li, C. J., Lian, X., Liu, J., and Yuan, H. Efficient smooth non-convex stochastic compositional optimization via stochastic recursive gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

Lewis, A. S. and Wright, S. J. A proximal method for composite minimization. *Mathematical Programming*, 158:501–546, 2016.

Lian, X., Wang, M., and Liu, J. Finite-sum composition optimization via variance reduced gradient descent. In *Artificial Intelligence and Statistics*, pp. 1159–1167. PMLR, 2017.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. *Advances in Neural Information Processing Systems*, 30, 2017.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Qi, Q., Xu, Y., Jin, R., Yin, W., and Yang, T. Attentional biased stochastic gradient for imbalanced classification. *arXiv preprint arXiv:2012.06951*, 2020a.

Qi, Q., Yan, Y., Wu, Z., Wang, X., and Yang, T. A simple and effective framework for pairwise deep metric learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pp. 375–391. Springer, 2020b.

Qi, Q., Lyu, J., Bai, E. W., Yang, T., et al. Stochastic constrained DRO with a complexity independent of sample size. *arXiv preprint arXiv:2210.05740*, 2022.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on stochastic programming: Modeling and theory*. SIAM, 2021.

Song, C., Wright, S. J., and Diakonikolas, J. Variance reduction via primal-dual accelerated dual averaging for nonsmooth convex finite-sums. In *International Conference on Machine Learning*, pp. 9824–9834. PMLR, 2021.

Staib, M. and Jegelka, S. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Tran Dinh, Q., Liu, D., and Nguyen, L. Hybrid variance-reduced SGD algorithms for minimax problems with nonconvex-linear function. *Advances in Neural Information Processing Systems*, 33:11096–11107, 2020.

Wang, M., Liu, J., and Fang, E. Accelerating stochastic composition optimization. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Xu, Y. Primal-dual stochastic gradient method for convex programs with many functional constraints. *SIAM Journal on Optimization*, 30(2):1664–1692, 2020.

Yan, Y., Xu, Y., Lin, Q., Zhang, L., and Yang, T. Stochastic primal-dual algorithms with faster convergence than $O(1/\sqrt{T})$ for problems without bilinear structure. *arXiv preprint arXiv:1904.10112*, 2019.

Zhang, J. and Xiao, L. A stochastic composite gradient method with incremental variance reduction. *Advances in Neural Information Processing Systems*, 32, 2019.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix

A. Related Work

DRO. DRO has been extensively studied in optimization, machine learning, and statistics literature (Ben-Tal et al., 2013; Bertsimas et al., 2018; Duchi et al., 2021; Namkoong & Duchi, 2017; Staib & Jegelka, 2019) Broadly, DRO problem formulation can be divided into two classes, one is a constrained formulation and the other is the regularized formulation (see (4)) (Levy et al., 2020; Duchi et al., 2021). A popular approach to solve the constrained DRO formulation is via primal-dual formulation where algorithms developed for min-max problems can directly be applied to solve constrained DRO (Yan et al., 2019; Namkoong & Duchi, 2017; Song et al., 2021; Alacaoglu et al., 2022; Tran Dinh et al., 2020). Many algorithms under different settings, e.g., convex, non-convex losses, and stochastic settings have been considered in the past to address such problems. However, primal-dual algorithms suffer from computational bottlenecks, since they require maintaining and updating the set of dual variables equal to the size of the dataset which can become particularly challenging, especially for large-scale machine learning tasks. Recently, (Levy et al., 2020) (Qi et al., 2022) (Haddadpour et al., 2022) have developed algorithms that are applicable to large-scale stochastic settings. Works (Levy et al., 2020) and (Qi et al., 2022) consider specific formulations of the DRO problem while (Haddadpour et al., 2022) considers a general formulation, however, as pointed out earlier the algorithms developed in (Haddadpour et al., 2022) are double loop and require accuracy-dependent batch sizes to guarantee convergence (see Table 1). In contrast, in this work, we develop algorithms that solve general instants of CO problems that often arise in DRO formulation. Importantly, the developed algorithm is amenable to large-scale implementation with algorithmic guarantees independent of accuracy-dependent batch sizes.

B. Proof of Theorem 5.1

Theorem B.1. *For Algorithm 1, choosing the step-size η^t such that*

$$\eta^t \leq \max\left\{\frac{1}{\bar{L}}, \frac{2\bar{L}_{\Phi,\gamma}}{L_\Phi^2}, \frac{1}{4}\left(1+\frac{1}{\gamma}\right)\right\}$$

where $\bar{L}_{\Phi,\gamma} := \max\{L_\Phi, 1+\gamma\}$ with constant $\gamma > 0$ and $\bar{L} \in (8\bar{L}_{\Phi,\gamma}, 16\bar{L}_{\Phi,\gamma}]$. Moreover, choosing the momentum parameter β^t as

$$\beta^t = C(L_f, B_g, \gamma)\cdot\eta^t \;\; \textit{with} \;\; C(L_f, B_g, \gamma) := 2\left(1+\frac{1}{\gamma}\right)B_g^2 L_f^2 \bar{L}.$$

Then for the selection of batch-sizes $|b_h^t| = |b_g^t| = |b|$ for all $t \in \{0, 1, \ldots, T-1\}$, under Assumptions 3.2 and 3.3, the iterates generated by Algorithm 1 satisfy

$$\sum_{t=0}^{T-1}\eta^t \mathbb{E}\|\nabla\Psi_{1/\bar{L}}(x^t)\|^2 \leq \left(\frac{2\bar{L}}{\bar{L}-8\bar{L}_{\Phi,\gamma}}\right)\left[\left(\Psi_{1/\bar{L}}(x^0) - \Psi^*_{1/\bar{L}}\right) + \|y^0 - g(x^0)\|^2 + \frac{C_\Psi}{|b|}\sum_{t=0}^{T-1}(\eta^t)^2\right],$$

where $\nabla\Psi_{1/\bar{L}}(x^t)$ is the Moreau envelope of $\Psi(\cdot)$ defined in Definition 3.5, $\Psi^*_{1/\bar{L}} = \min_x \Psi_{1/\bar{L}}(x)$ and C_Ψ is a constant defined as

$$C_\Psi := 2\sigma_h^2\bar{L} + 4B_f^2\sigma_g^2\bar{L} + 2C^2(L_f, B_g, \gamma)\sigma_g^2 + 2B_g^2 B_\Phi^2.$$

Proof. First, defining the Moreau envelope of $\Psi(\cdot)$ as

$$\Psi_\lambda(x) := \min_z\left\{\Psi(z) + \frac{1}{2\lambda}\|z-x\|^2\right\}.$$

Using the above definition, we have for $\widehat{x} := \text{prox}_\Psi^\lambda(x) := \arg\min_z\left\{\Psi(z) + \frac{1}{2\lambda}\|z-x\|^2\right\}$

$$\begin{cases} \|\widehat{x} - x\| = \lambda\|\nabla\Psi_\lambda(x)\| \\ \Psi(\widehat{x}) \leq \Psi(x) \\ \text{dist}(0, \partial\Psi(\widehat{x})) \leq \|\nabla\Psi_\lambda(x)\| \end{cases} \tag{7}$$

Recall that $\Psi(x) = r(x) + \Phi(x)$ is L_Φ-weakly convex. Let us define $\bar{L} > L_\Phi$ with $\widehat{x}^t := \text{prox}_\Psi^{1/\bar{L}}$, this implies that we have

$$\begin{aligned}
0 &\in \partial\Psi(\widehat{x}^t) + \bar{L}(\widehat{x}^t - x) \\
\bar{L}(\widehat{x}^t - x) &\in \partial r(\widehat{x}^t) + \nabla\Phi(\widehat{x}^t) \\
\eta^t\bar{L}(x^t - \widehat{x}^t) &\in \eta^t\partial r(\widehat{x}^t) + \eta^t\nabla\Phi(\widehat{x}^t) \\
\eta\bar{L}x^t - \eta^t\nabla\Phi(\widehat{x}^t) + (1-\eta^t\bar{L})\widehat{x}^t &\in \widehat{x}^t + \eta^t\partial r(\widehat{x}^t) \\
\widehat{x}^t &= \text{prox}_r^{\eta^t}\big(\eta^t\bar{L}x^t - \eta^t\nabla\Phi(\widehat{x}^t) + (1-\eta^t\bar{L})\widehat{x}^t\big).
\end{aligned}$$

Using the above fact, we have

$$\begin{aligned}
\mathbb{E}\|x^{t+1} - \widehat{x}^t\|^2 &= \mathbb{E}\|\text{prox}_r^{\eta^t}(x^t - \eta^t\nabla\Phi(x^t;\bar{\xi}^t)) - \text{prox}_r^{\eta^t}\big(\eta^t\bar{L}x^t - \eta^t\nabla\Phi(\widehat{x}^t) + (1-\eta^t\bar{L})\widehat{x}^t\big)\|^2 \\
&\le \mathbb{E}\|(x^t - \eta^t\nabla\Phi(x^t;\bar{\xi}^t)) - \big(\eta^t\bar{L}x^t - \eta^t\nabla\Phi(\widehat{x}^t) + (1-\eta^t\bar{L})\widehat{x}^t\big)\|^2 \\
&= \mathbb{E}\|(1-\eta^t\bar{L})(x^t - \widehat{x}^t) + \eta^t\big(\nabla\Phi(\widehat{x}^t) - \nabla\Phi(x^t;\bar{\xi}^t)\big)\|^2 \\
&= (1-\eta^t\bar{L})^2\mathbb{E}\|x^t - \widehat{x}^t\|^2 + \underbrace{(\eta^t)^2\mathbb{E}\|\nabla\Phi(\widehat{x}^t) - \nabla\Phi(x^t;\bar{\xi}^t)\|^2}_{\text{Term I}} \\
&\qquad \underbrace{-2\eta^t(1-\eta^t\bar{L})\mathbb{E}\big\langle x^t - \widehat{x}^t, \nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(\widehat{x}^t)\big\rangle}_{\text{Term II}}
\end{aligned}$$

Now, considering each term separately, we have

$$\begin{aligned}
\text{Term II} &= -2\eta^t(1-\eta^t\bar{L})\mathbb{E}\big\langle x^t - \widehat{x}^t, \nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(\widehat{x}^t)\big\rangle \\
&= -2\eta^t(1-\eta^t\bar{L})\mathbb{E}\big[\big\langle x^t - \widehat{x}^t, \nabla\Phi(x^t) - \nabla\Phi(\widehat{x}^t)\big\rangle + \big\langle x^t - \widehat{x}^t, \nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(x^t)\big\rangle\big] \\
&\le 2\eta^t L_\Phi(1-\eta^t\bar{L})\mathbb{E}\|x^t - \widehat{x}^t\|^2 - 2\eta^t(1-\eta^t\bar{L})\mathbb{E}\big\langle x^t - \widehat{x}^t, \nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(x^t)\big\rangle.
\end{aligned}$$

Next,

$$\begin{aligned}
\text{Term I} &= (\eta^t)^2\mathbb{E}\|\nabla\Phi(\widehat{x}^t) - \nabla\Phi(x^t;\bar{\xi}^t)\|^2 \\
&\le 2(\eta^t)^2\mathbb{E}\|\nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(x^t)\|^2 + 2(\eta^t)^2\mathbb{E}\|\nabla\Phi(x^t) - \nabla\Phi(\widehat{x}^t)\|^2 \\
&\le 2(\eta^t)^2\mathbb{E}\|\nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(x^t)\|^2 + 2(\eta^t)^2L_\Phi^2\mathbb{E}\|x^t - \widehat{x}^t\|^2.
\end{aligned}$$

Combining the terms, we get

$$\begin{aligned}
\mathbb{E}\|x^{t+1} - \widehat{x}^t\|^2 \le \Big[(1-\eta^t\bar{L})^2 + 2\eta^t L_\Phi(1-\eta^t\bar{L}) + 2(\eta^t)^2L_\Phi^2\Big]\mathbb{E}\|x^t - \widehat{x}^t\|^2 \\
+ 2(\eta^t)^2\underbrace{\mathbb{E}\|\nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(x^t)\|^2}_{\text{Term III}} - 2\eta^t(1-\eta^t\bar{L})\underbrace{\mathbb{E}\big\langle x^t - \widehat{x}^t, \nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(x^t)\big\rangle}_{\text{Term IV}}
\end{aligned}$$

Again considering Terms III and IV separately, we have

$$\begin{aligned}
\text{Term III} &= \mathbb{E}\|\nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(x^t)\|^2 \\
&= \mathbb{E}\left\|\left[\frac{1}{|b_h^t|}\sum_{i\in b_h^t}\nabla h(x^t;\xi_i^t) + \frac{1}{|b_g^t|}\sum_{i\in b_g^t}\nabla g(x^t;\zeta_i^t)\nabla f(y^t)\right] - \Big[\nabla h(x^t) + \nabla g(x^t)\nabla f(g(x^t))\Big]\right\|^2 \\
&\le \frac{2\sigma_h^2}{|b_h^t|} + 2\mathbb{E}\left\|\frac{1}{|b_g^t|}\sum_{i\in b_g^t}\nabla g(x^t;\zeta_i^t)\nabla f(y^t) - \nabla g(x^t)\nabla f(g(x^t))\right\|^2 \\
&\le \frac{2\sigma_h^2}{|b_h^t|} + \frac{4B_f^2\sigma_g^2}{|b_g^t|} + 4B_g^2L_f^2\|y^t - g(x^t)\|^2.
\end{aligned}$$

Next, we have

$$\begin{aligned}\text{Term IV} &= \mathbb{E}\big\langle x^t - \widehat{x}^t, \nabla\Phi(x^t;\bar{\xi}^t) - \nabla\Phi(x^t)\big\rangle \\ &= \mathbb{E}\Big\langle x^t - \widehat{x}^t, \frac{1}{|b_g^t|}\sum_{i\in b_g^t}\nabla g(x^t;\zeta_i^t)\nabla f(y^t) - \nabla g(x^t)\nabla f(g(x^t))\Big\rangle \\ &\leq (1+\gamma)\mathbb{E}\|x^t - \widehat{x}^t\|^2 + \left(1+\frac{1}{\gamma}\right)B_g^2 L_f^2\|y^t - g(x^t)\|^2.\end{aligned}$$

Combining all the terms we finally get

$$\begin{aligned}\mathbb{E}\|x^{t+1} - \widehat{x}^t\|^2 \leq \Big[(1-\eta^t\bar{L})^2 + 2\eta^t L_\Phi(1-\eta^t\bar{L}) + 2(\eta^t)^2L_\Phi^2 + 2\eta^t(1-\eta^t\bar{L})(1+\gamma)\Big]\mathbb{E}\|x^t - \widehat{x}^t\|^2 \\ + \Big[2\eta^t(1-\eta^t\bar{L})\left(1+\frac{1}{\gamma}\right)B_g^2L_f^2 + 8(\eta^t)^2B_g^2L_f^2\Big]\mathbb{E}\|y^t - g(x^t)\|^2 \\ + \frac{4(\eta^t)^2\sigma_h^2}{|b_h^t|} + \frac{8B_f^2(\eta^t)^2\sigma_g^2}{|b_g^t|}.\end{aligned}$$

Define $\bar{L}_{\Phi,\gamma} \coloneqq \max\{L_\Phi, 1+\gamma\}$, choosing η^t such that $\eta^t \leq \max\{1/\bar{L}, 2\bar{L}_{\Phi,\gamma}/L_\Phi^2, (1/4)(1+1/\gamma)\}$, and assuming $|b_h^t| = |b_h|$ and $|b_g^t| = |b_g|$ for all $t \in \{0,1,\ldots,T-1\}$

$$\begin{aligned}\mathbb{E}\|x^{t+1} - \widehat{x}^t\|^2 \leq \Big[(1-\eta^t(\bar{L} - 8\bar{L}_{\Phi,\gamma})\Big]\mathbb{E}\|x^t - \widehat{x}^t\|^2 + 4\eta^t\left(1+\frac{1}{\gamma}\right)B_g^2L_f^2\underbrace{\mathbb{E}\|y^t - g(x^t)\|^2}_{\text{Term V}} \\ + \frac{4(\eta^t)^2\sigma_h^2}{|b_h|} + \frac{8B_f^2(\eta^t)^2\sigma_g^2}{|b_g|}.\end{aligned}$$

Next, considering Term V for $t+1$, we get

$$\begin{aligned}\text{Term V} &= \mathbb{E}\|y^t - g(x^t)\|^2 \\ &= \mathbb{E}\Big\|(1-\beta^t)\Big(y^{t-1} + \frac{1}{|b_g|}\sum_{i\in b_g^t}g(x^t;\zeta_i^t) - \frac{1}{|b_g|}\sum_{i\in b_g^t}g(x^{t-1};\zeta_i^t)\Big) + \frac{\beta^t}{|b_g|}\sum_{i\in b_g^t}g(x^t;\zeta_i^t) - g(x^t)\Big\|^2 \\ &= (1-\beta^t)^2\,\mathbb{E}\|y^{t-1} - g(x^{t-1})\|^2 \\ &+ \Big\|(1-\beta^t)\Big[(g(x^{t-1}) - g(x^t)) - \frac{1}{|b_g|}\sum_{i\in b_g^t}\big(g(x^{t-1};\zeta_i^t) - g(x^t;\zeta_i^t)\big)\Big] + \beta_t\left(\frac{1}{|b_g|}\sum_{i\in b_g^t}g(x^t;\zeta_i^t) - g(x^t)\right)\Big\|^2 \\ &\leq (1-\beta^t)^2\mathbb{E}\|y^{t-1} - g(x^{t-1})\|^2 + \frac{2(\beta^t)^2\sigma_g^2}{|b_g|} \\ &\qquad + 2(1-\beta^t)^2\,\mathbb{E}\Big\|\big(g(x^{t-1}) - g(x^t)\big) - \frac{1}{|b_g|}\sum_{i\in b_g^t}\big(g(x^{t-1};\zeta_i^t) - g(x^t;\zeta_i^t)\big)\Big\|^2 \\ &\leq (1-\beta^t)^2\mathbb{E}\|y^{t-1} - g(x^{t-1})\|^2 + \frac{2(\beta^t)^2\sigma_g^2}{|b_g|} \\ &\qquad + \frac{2(1-\beta^t)^2}{|b_g|^2}\sum_{i\in b_g^t}\mathbb{E}\big\|\big(g(x^{t-1}) - g(x^t)\big) - \big(g(x^{t-1};\zeta_i^t) - g(x^t;\zeta_i^t)\big)\big\|^2 \\ &\leq (1-\beta^t)^2\mathbb{E}\|y^{t-1} - g(x^{t-1})\|^2 + \frac{2(\beta^t)^2\sigma_g^2}{|b_g|} + \frac{2(1-\beta^t)^2}{|b_g|^2}\sum_{i\in b_g^t}\mathbb{E}\big\|g(x^{t-1};\zeta_i^t) - g(x^t;\zeta_i^t)\big\|^2 \\ &\leq (1-\beta^t)^2\mathbb{E}\|y^{t-1} - g(x^{t-1})\|^2 + \frac{2(\beta^t)^2\sigma_g^2}{|b_g|} + \frac{2(1-\beta^t)^2B_g^2}{|b_g|}\mathbb{E}\big\|x^t - x^{t-1}\big\|^2 \\ &\leq (1-\beta^t)^2\mathbb{E}\|y^{t-1} - g(x^{t-1})\|^2 + \frac{2(\beta^t)^2\sigma_g^2}{|b_g|} + \frac{2(1-\beta^t)^2B_g^2(\eta^{t-1})^2}{|b_g|}\mathbb{E}\big\|\nabla\Phi(x^{t-1};\bar{\xi}^{t-1})\big\|^2\end{aligned}$$

$$\leq (1-\beta^t)^2\mathbb{E}\|y^{t-1}-g(x^{t-1})\|^2+\frac{2(\beta^t)^2\sigma_g^2}{|b_g|}+\frac{2B_g^2B_\Phi^2(\eta^{t-1})^2}{|b_g|}.$$

where B_Φ is defined as: $B_\Phi^2 \coloneqq 2B_h^2+2B_g^2B_f^2$. Finally, we have

$$\mathbb{E}[\Psi_{1/\bar{L}}(x^{t+1})] \leq \mathbb{E}\Big[\Psi(\widehat{x}^t)+\frac{\bar{L}}{2}\|\widehat{x}^t-x^{t+1}\|^2\Big]$$

Using the bound on $\mathbb{E}\|\widehat{x}^t-x^{t+1}\|^2$, we get

$$\begin{aligned}\mathbb{E}[\Psi_{1/\bar{L}}(x^{t+1})] &\leq \mathbb{E}\Big[\Psi(\widehat{x}^t)+\frac{\bar{L}}{2}\|\widehat{x}^t-x^t\|^2\Big]-\frac{\bar{L}}{2}\Big[\eta^t(\bar{L}-8\bar{L}_{\Phi,\gamma})\mathbb{E}\|\widehat{x}^t-x^t\|^2\\ &\quad+4\eta^t\Big(1+\frac{1}{\gamma}\Big)B_g^2L_f^2\mathbb{E}\|y^t-g(x^t)\|^2+\frac{4(\eta^t)^2\sigma_h^2}{|b_h|}+\frac{8B_f^2(\eta^t)^2\sigma_g^2}{|b_g|}\Big]\\ &= \Psi_{1/\bar{L}}(x^t)+\frac{\bar{L}}{2}\Big[-\eta^t(\bar{L}-8\bar{L}_{\Phi,\gamma})\mathbb{E}\|\bar{x}^t-x^t\|^2+4\eta^t\Big(1+\frac{1}{\gamma}\Big)B_g^2L_f^2\mathbb{E}\|y^t-g(x^t)\|^2\\ &\quad+\frac{4(\eta^t)^2\sigma_h^2}{|b_h|}+\frac{8B_f^2(\eta^t)^2\sigma_g^2}{|b_g|}\Big].\end{aligned}$$

Defining the potential function as: $P^{t+1} \coloneqq \mathbb{E}[\Psi_{1/\bar{L}}(x^{t+1})+\|y^{t+1}-g(x^{t+1})\|^2]$, we have

$$\begin{aligned}P^{t+1}-P^t &\leq -\frac{\eta^t\bar{L}}{2}(\bar{L}-8\bar{L}_{\Phi,\gamma})\mathbb{E}\|\widehat{x}^t-x^t\|^2+2\eta^t\bar{L}\Big(1+\frac{1}{\gamma}\Big)B_g^2L_f^2\mathbb{E}\|y^t-g(x^t)\|^2\\ &\quad+\frac{2(\eta^t)^2\sigma_h^2\bar{L}}{|b_h|}+\frac{4B_f^2(\eta^t)^2\sigma_g^2\bar{L}}{|b_g|}-\beta^t\mathbb{E}\|y^t-g(x^t)\|^2+\frac{2(\beta^t)^2\sigma_g^2}{|b_g|}+\frac{2B_g^2B_\Phi^2(\eta^t)^2}{|b_g|}\end{aligned}$$

Choosing

$$\beta^t=\underbrace{2\Big(1+\frac{1}{\gamma}\Big)B_g^2L_f^2\bar{L}}_{C(L_f,B_g,\gamma)}\cdot\eta^t$$

we get

$$P^{t+1}-P^t \leq -\frac{\eta^t\bar{L}}{2}(\bar{L}-8\bar{L}_{\Phi,\gamma})\mathbb{E}\|\widehat{x}^t-x^t\|^2+\frac{2(\eta^t)^2\sigma_h^2\bar{L}}{|b_h|}+\frac{4B_f^2(\eta^t)^2\sigma_g^2\bar{L}}{|b_g|}+\frac{2C^2(L_f,B_g,\gamma)(\eta^t)^2\sigma_g^2}{|b_g|}+\frac{2B_g^2B_\Phi^2(\eta^t)^2}{|b_g|},$$

Telescoping the sum over $t=\{0,\ldots,T-1\}$, we get

$$P^T-P^0 \leq \sum_{t=0}^{T-1}-\frac{\eta^t\bar{L}}{2}(\bar{L}-8\bar{L}_{\Phi,\gamma})\mathbb{E}\|\widehat{x}^t-x^t\|^2+\Big[\frac{2\sigma_h^2\bar{L}}{|b_h|}+\frac{4B_f^2\sigma_g^2\bar{L}}{|b_g|}+\frac{2C^2(L_f,B_g,\gamma)\sigma_g^2}{|b_g|}+\frac{2B_g^2B_\Phi^2}{|b_g|}\Big]\sum_{t=0}^{T-1}(\eta^t)^2$$

Choosing $|b_h|=|b_g|=|b|$, we get

$$P^T-P^0 \leq \sum_{t=0}^{T-1}-\frac{\eta^t\bar{L}}{2}(\bar{L}-8\bar{L}_{\Phi,\gamma})\mathbb{E}\|\widehat{x}^t-x^t\|^2+\underbrace{\Big[2\sigma_h^2\bar{L}+4B_f^2\sigma_g^2\bar{L}+2C^2(L_f,B_g,\gamma)\sigma_g^2+2B_g^2B_\Phi^2\Big]}_{C_\Psi}\frac{1}{|b|}\sum_{t=0}^{T-1}(\eta^t)^2$$

Rearranging the terms, we get

$$\frac{\bar{L}(\bar{L}-8\bar{L}_{\Phi,\gamma})}{2}\sum_{t=0}^{T-1}\eta^t\mathbb{E}\|\widehat{x}^t-x^t\|^2 \leq P^0-P^T+\frac{C_\Psi}{|b|}\sum_{t=0}^{T-1}(\eta^t)^2$$

Using (7) and multiplying both sides by $\frac{2\bar{L}}{\bar{L}-8\bar{L}_{\Phi,\gamma}}$, we get

$$\sum_{t=0}^{T-1}\eta^t\mathbb{E}\|\nabla\Psi_{1/\bar{L}}(x^t)\|^2 \leq \left(\frac{2\bar{L}}{\bar{L}-8\bar{L}_{\Phi,\gamma}}\right)\left(P^0 - P^T + \frac{C_\Psi}{|b|}\sum_{t=0}^{T-1}(\eta^t)^2\right)$$

Finally, using the definition of the potential function, we get

$$\sum_{t=0}^{T-1}\eta^t\mathbb{E}\|\nabla\Psi_{1/\bar{L}}(x^t)\|^2 \leq \left(\frac{2\bar{L}}{\bar{L}-8\bar{L}_{\Phi,\gamma}}\right)\left(\left(\Psi_{1/\bar{L}}(x^0) - \Psi^*_{1/\bar{L}}\right) + \|y^0 - g(x^0)\|^2 + \frac{C_\Psi}{|b|}\sum_{t=0}^{T-1}(\eta^t)^2\right)$$

where $\Psi^*_{1/\bar{L}} = \min_x \Psi_{1/\bar{L}}(x)$. □