Journal of Machine Learning Research xx (xxxx) 1-49 Submitted xxx; Revised xxx; Published xxx

Nested Stochastic Algorithm for Generalized Sinkhorn distance-Regularized Distributionally Robust Optimization

Yufeng Yang YUFENG.YANG@TAMU.EDU
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843, USA

Yi Zhou YI.ZHOU@TAMU.EDU
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843, USA

Zhaosong Lu ZHAOSONG@UMN.EDU
Department of Industrial and Systems Engineering
University of Minnesota
Minneapolis, MN 55455, USA

Editor: My editor

Abstract

Distributionally robust optimization (DRO) is a powerful technique to train robust models against data distribution shift. This paper aims to solve regularized nonconvex DRO problems, where the uncertainty set is modeled by a so-called generalized Sinkhorn distance and the loss function is nonconvex and possibly unbounded. Such a distance allows to model uncertainty of distributions with different probability supports and divergence functions. For this class of regularized DRO problems, we derive a novel dual formulation taking the form of nested stochastic optimization, where the dual variable depends on the data sample. To solve the dual problem, we provide theoretical evidence to design a nested stochastic gradient descent (SGD) algorithm, which leverages stochastic approximation to estimate the nested stochastic gradients. We study the convergence rate of nested SGD and establish polynomial iteration and sample complexities that are independent of the data size and parameter dimension, indicating its potential for solving large-scale DRO problems. We conduct numerical experiments to demonstrate the efficiency and robustness of the proposed algorithm.

Keywords: Distributionally Robust Optimization; Lagrange Duality; Nested Stochastic Optimization; First-order Algorithm; Adversarial Robustness.

1 Introduction

In classic machine learning, the primary goal is to achieve good predictive performance in the test set after training the model on a designated training set. The training problem is typically formulated as an empirical risk minimization (ERM) problem (Vapnik and Chervonenkis, 2015). However, empirical risk minimization assumes that the training set and the test set follow the same underlying data distribution, which is often unrealistic and may result in poor test performance when data distribution shift exists.

Data distribution shift is prevalent in real-world scenarios. It can be caused by many factors such as sampling bias, presence of anomalies, data merging and change of measurements, etc. To tackle this challenge, distributionally robust optimization (DRO) (Herbert, 1957) was proposed, which formulates the objective function as a min-max problem. DRO aims to learn a robust model by minimizing the expected risk over the worst-case data distribution within a predefined ambiguity set. This formulation offers a principled framework to learn the optimal resilient solution in the face of distribution uncertainty.

One key factor in DRO is the selection of an appropriate divergence measure for modeling the ambiguity set. Specifically, the divergence measure should not only be computationally tractable but also yield a solution that avoids excessive conservatism. In the existing literature, various divergence-based ambiguity sets have been studied. In Pflug and Wozabal (2007); Wozabal (2012); Shafieezadeh Abadeh et al. (2015); Esfahani and Kuhn (2018); Gao and Kleywegt (2023); Luo and Mehrotra (2019); Blanchet et al. (2023), the authors focused on reformulating the expressions of objective functions under the worst-case distributions into a tractable form and exploring possible algorithms to tackle DRO problems under Wasserstein-type ambiguity sets. For more information, we refer the readers to Kuhn et al. (2019) for a comprehensive survey on Wasserstein DRO. In Hu and Hong (2013); Bayraksan and Love (2015); Levy et al. (2020); Duchi and Namkoong (2020); Liu et al. (2023), the authors analyzed alternative expressions of objective functions under the worst-case distribution, developed algorithms to solve DRO problems under Kullback–Leibler (KL) and f-divergence based ambiguity sets. However, the aforementioned divergence measures have certain limitations. For example, it is known that DRO with Wasserstein distance requires high computational complexity (Pele and Werman, 2009; Ba et al., 2010). Also, both KL and f-divergence are not symmetric when assessing distributions. Furthermore, these two divergence measures require that the distributions share the same probability support, a strong condition that may fail to capture extreme distributions at certain points. We refer readers to check Examples 1 and 2 for some concrete applications.

The Sinkhorn distance, first introduced in Cuturi (2013), was designed to address the aforementioned limitations. Sinkhorn distance is symmetric and allows distributions from the same sample space to have different probability supports. Furthermore, Sinkhorn distance is a convex function with respect to distributions, ensuring computation tractability and efficiency for large-scale problems. In Wang et al. (2023); Blanchet et al. (2023); Azizian et al. (2023), constrained Sinkhorn DRO was initially investigated. Specifically, Wang et al. (2023) derived the equivalent formulation of constrained Sinkhorn DRO problem and solved it via stochastic mirror descent algorithm, marking the first work to solve constrained Sinkhorn DRO using first-order optimization methods. However, the convergence analysis conducted in their work assumed that the loss function is convex and bounded, which may not hold in practical modern machine learning applications. Furthermore, the log-exponential compositional structure induced by the conjugate dual of the KL-divergence makes the objective function difficult to optimize and hinders convergence.

In this work, we develop and study dual formulations of *generalized* Sinkhorn-distance regularized DRO problems (see formulation (1)). The problem takes the form of a nested stochastic optimization with a contextual variable. Unlike traditional stochastic optimization problems, our formulation makes dual decision variable dependent on data sample following distribution $\mathbb{P}$. Hu et al. (2023) studied contextual bilevel optimization, which can be applied to solve such problem. However, their convergence analysis is not directly applicable to proposed Sinkhorn DRO dual formulation (3). Specifically, their convergence guarantee relies on the strong convexity of the inner objective, a condition that does not hold in our setting. Moreover, from algorithmic perspective, while Multi-level Monte Carlo (MLMC) significantly improves the per-iteration sample complexity for computing a variance-reduced stochastic gradient estimator, it still requires storing a non-constant number of samples that scale with the target error, and second-order information remains unavoidable under bilevel optimization framework, leading to increased computational cost and unnecessary overall sample complexity for our problem. To tackle aforementioned challenges, based on tractable expression of gradient proved in Jin et al. (2021) (see Lemma 1) and structural relationship between outer problem's gradient and inner problem's gradient, we establish controllable approximation error of outer objective's gradient given the inner objective's gradient satisfying mild conditions (see Theorem 2). This observation enables us to further utilize sample-average approximation (SAA) to estimate stochastic gradients, solve dual problem efficiently via Nested SGD algorithm (see Algorithm 2) and establish convergence analysis without requiring large batch size of samples and additional assumptions including convex loss and strong convexity for inner-objective assumed in Wang et al. (2023); Hu et al. (2023). Finally, we train several models including logistic regression, LeNet (LeCun et al., 1998) over CIFAR-10 (Krizhevsky, 2009) and MNIST (Deng, 2012) dataset, and conduct experiments to evaluate test performance of proposed Sinkhorn DRO dual formulation (3) with other baseline methods over perturbed test dataset. Our results demonstrate that the Sinkhorn DRO formulation (3) and our proposed nested algorithm successfully improve models' robustness against distribution shift.

1.1 Summary of Contributions

- To preserve the advantages of Sinkhorn distance while broadening the class of divergence measures used to model the ambiguity set, we consider a generalized Sinkhorn distance based on the family of f-divergences. Building on this formulation, we transform the primal problem using inverse CDF sampling and derive the dual form of the regularized Sinkhorn DRO problem via Lagrangian duality, under which strong duality holds. The resulting dual problem takes a novel form of contextual nested stochastic optimization problem, where the variable of the inner stochastic sub-problem depends explicitly on the data samples.

- To solve the contextual nested stochastic optimization problem, we begin by reviewing the explicit gradient formula of the proposed Sinkhorn DRO objective. By exploiting the structural relationship between the objective and its sub-problem, we derive conditions under which the gradient approximation error can be made arbitrarily small. This insight motivates the development of a practical Nested-SGD algorithm (Algorithm 2), which simplifies the approach taken by the research community to solve contextual nested problems within the general bilevel optimization framework. To establish convergence guarantee, we further analyze the smoothness properties and derive second-moment upper bounds. Our convergence analysis reveals a standard sample complexity of $\mathcal{O}(\varepsilon^{-4})$ for each loop of the algorithm, resulting in a total sample complexity of $\mathcal{O}(\varepsilon^{-8})$ for non-convex loss. However, considering the one-dimensional nature of the sub-problem of the proposed Sinkhorn DRO formulation and by appropriately adjusting the batch size, the overall problem can be efficiently solved with $\mathcal{O}(\varepsilon^{-4})$ iteration complexity.

- To justify the efficiency of our proposed formulation and algorithm, we evaluate their practical performance on real-world datasets. The results, under distribution shifts simulated by various adversarial attack methods, demonstrate that our approach not only improves the robustness of diverse machine learning models but also scales effectively to large models and datasets.

2 Related Work

DRO. The DRO framework shares strong connections with contrastive learning (Wu et al., 2023), multiple instance learning (Sapkota et al., 2021), AUC maximization (Zhu et al., 2022), anomaly detection (Chen and Paschalidis, 2018a), and self-supervised learning (Qiu et al., 2024; Wei et al., 2025). The key challenge in modeling lies in transforming the problem into a tractable formulation given the chosen ambiguity set. The first stream focuses on using information divergence to construct ambiguity set. Commonly employed divergence measures include the Wasserstein metric (Shafieezadeh Abadeh et al., 2015; Esfahani and Kuhn, 2018; Gao and Kleywegt, 2023; Luo and Mehrotra, 2019; Liu et al., 2022; Zhu et al., 2024); KL divergence (Hu and Hong, 2013; Shapiro et al., 2023; Kocuk, 2020); f-divergence (Levy et al., 2020; Lam, 2019; Duchi and Namkoong, 2020; Jin et al., 2021; Namkoong and Duchi, 2016; Liu et al., 2023); Sinkhorn distance (Wang et al., 2023; Azizian et al., 2023) and its variants (Wang et al., 2024). Recently, Blanchet et al. (2023) further proposed a framework to unify aforementioned constrained DRO problem based on optimal transport theory with conditional moment constraints. Another stream for constructing ambiguity set is using special statistics, such as geometry shape constraints (Chen et al., 2021) and statistical moments (Herbert, 1957; Delage and Ye, 2010; Cheramin et al., 2022) etc.

Sinkhorn Distance. Sinkhorn distance has successful applications in areas like generative models (Patrini et al., 2020; Genevay et al., 2018), matrix factorization (Qian et al., 2016), image segmentation (Rabin and Papadakis, 2015) etc. In Cuturi (2013), Sinkhorn matrix scaling algorithms was proposed to compute optimal transport map under Sinkhorn distance objective. Later, in Altschuler et al. (2018); Aude et al. (2016), greedy and stochastic variants of Sinkhorn scaling algorithms were proposed to clarify relationship between algorithm convergence and input dimensions. Some works also study Sinkhorn distance computation over data samples with special structures. In Tenetov et al. (2018); Altschuler et al. (2019), they propose variant algorithms specially applying to data samples over compact Riemannian manifolds and Euclidean balls respectively.

Algorithms for Solving DRO. For Wasserstein metric ambiguity set, several works reformulate primal problem into tractable forms such as convex programming (Shafieezadeh Abadeh et al., 2015), semidefinite programming (Luo and Mehrotra, 2019) and mixed integer programming (Sapkota et al., 2021). Subsequent works (Shafieezadeh Abadeh et al., 2015; Sapkota et al., 2021) transform problems into the form which is directly solvable using software toolbox. In Luo and Mehrotra (2019), they use cutting-surface method to solve semi-definite programming for general nonlinear objective and branch-and-bound algorithms for bilinear objective. Another common technique to transform DRO is using duality theory (Gao and Kleywegt, 2023; Levy et al., 2020). Through this way, computation of shifted distribution can be avoided. In Qi et al. (2023), projected SGD and acceleration is used to solve the dual form of KL divergence constrained DRO. In Jin et al. (2021), normalized SGD with momentum is used to optimize the dual of f-divergence regularized DRO. Later Zhang et al. (2025) revisits f-divergence regularized DRO and propose double SGD with clipping to solve it. In Zhang et al. (2024), stochastic Frank-Wolfe method is used to solve approximation for the dual of general Cressie-Read family divergence constrained DRO. Wang et al. (2023) used stochastic mirror descent to solve the dual form of Sinkhorn distance constrained DRO and Wang et al. (2024) used a projected sub-gradient method to solve the dual form of unbalanced

Wasserstein distance constrained DRO. Hu et al. (2023) empirically applied contextual bilevel optimization methods to solve the primal formulation of Wasserstein DRO with side information.

3 Notations

Throughout this work, we denote $\xi \sim \mathbb{Q}$ as data drawn from the underlying distribution $\mathbb{Q}$, $\zeta \sim \mathbb{P}$ as data drawn from the nominal distribution $\mathbb{P}$, and their corresponding reference measures are denoted by μ (for $\mathbb{P}$) and ν (for $\mathbb{Q}$). We denote $\mathcal{N}(\cdot,\cdot)$ as normal distribution following certain mean and variance. For the primal problem (1), we denote $\ell(x;\xi)$ as the loss function associated with sample ξ and parameter $x \in \mathbf{R}^d$ (i.e., the weights of a linear model or neural network), λ as the regularization coefficient. For generalized Sinkhorn distance used to model the ambiguity set (see Definition 1), we denote $c(\cdot,\cdot)$ as the cost metric for measuring *proximity* between samples ζ and ξ, β as the regularization coefficient, D_f as the f-divergence. In primal-dual problem reformulation, we introduce η as Lagrange multiplier and $f^*(\cdot)$ as the conjugate dual function induced by the chosen information divergence. For the proposed dual formulation (3), when $\eta_x^*(\zeta) = \arg\min \mathcal{L}_\zeta(x,\eta)$, we denote $\Psi(x) = \mathcal{L}_\zeta(x,\eta_x^*(\zeta))$ for simplicity, and use $\nabla\Psi(x)$ to denote the gradient with respect to x. For functions including $\mathcal{L}_\zeta, \mathcal{L}_{\zeta,\xi} : \mathbf{R}^d \times \mathbf{R} \to \mathbf{R}$, which takes two arguments $x \in \mathbf{R}^d$, $\eta \in \mathbf{R}$, we use ∇_1, ∇_2 to denote the gradient with respect to x and η. For Algorithm 1, which optimizes the inner objective (5), we denote α_d as the learning rate, $v(\cdot)$ as the stochastic gradient estimator with respect to η, $\tilde{d}$ as the output index, $\tilde{B}$ as the batch size, and $\tilde{\varepsilon}$ as the scaled target error. For Algorithm 2, which optimizes the dual formulation (3), we denote γ_t as the learning rate, $\hat{g}^B(\cdot), g^B(\cdot)$ as inexact, exact stochastic gradient estimator of dual formulation (3) with respect to x, $\tilde{t}$ as the algorithm's output index, B as the batch size, and ε as the target error passed to Algorithm 2. Through this work, we denote $\|\cdot\|$, $\|\cdot\|_1$, and $\|\cdot\|_\infty$ as ℓ_2, ℓ_1, and ℓ_∞ norms over Euclidean space, respectively.

4 Regularized Nonconvex DRO with Generalized Sinkhorn Distance

In this section, we first introduce a class of *regularized* nonconvex distributionally-robust optimization (DRO) problems, where the data distribution uncertainty is quantified by generalized Sinkhorn distance. We then study its strong dual formulation in Theorem 1 and compare it with the strong dual formulation of *constrained* DRO quantified by Sinkhorn distance obtained in Wang et al. (2023).

4.1 Problem Formulation

In distributionally-robust optimization (DRO), the goal is to learn a model which achieves good and robust performance when the underlying data distribution is uncertain. Specifically, consider a machine learning problem with the nonconvex loss function denoted by $\ell(x;\xi)$, where $x \in \mathbf{R}^d$ denotes the collection of model parameters and ξ corresponds to a data sample that follows an underlying data distribution $\mathbb{Q}$. Then, with a regularization parameter $\lambda > 0$, we study the following regularized DRO problem, which is a popular formulation in robust machine learning (Chen and Paschalidis, 2018b; Gao et al., 2022; Sagawa et al., 2019).

$$\min_{x\in\mathbf{R}^d} \sup_{\mathbb{Q}} \Big\{ \mathbb{E}_{\xi\sim\mathbb{Q}}\big[\ell(x;\xi)\big] - \lambda W_\beta(\mathbb{P},\mathbb{Q}) \Big\}, \tag{1}$$

where $W_\beta(\mathbb{P}, \mathbb{Q})$ denotes a certain function (with parameter $\beta > 0$) that measures the distance between a nominal data distribution $\mathbb{P}$ and the underlying data distribution $\mathbb{Q}$. In particular, the operation $\min_x \sup_{\mathbb{Q}}$ aims to optimize the model x under the worst-case data distribution $\mathbb{Q}$ to enhance model robustness against the distribution shift from the nominal distribution $\mathbb{P}$. In the existing literature, many studies have considered KL-divergence (Hu and Hong, 2013; Shapiro et al., 2023), f-divergence (Levy et al., 2020; Duchi and Namkoong, 2020; Jin et al., 2021; Namkoong and Duchi, 2016; Liu et al., 2023) and Wasserstein distance (Shafieezadeh Abadeh et al., 2015; Esfahani and Kuhn, 2018; Gao and Kleywegt, 2023; Luo and Mehrotra, 2019; Blanchet et al., 2023) to quantify the above distribution shift. However, both the KL-divergence and the f-divergence require $\mathbb{Q}$ and $\mathbb{P}$ to have the same probability support, which is restrictive and undesirable for many machine learning applications. The following two examples illustrate the limitations of f-divergence.

Example 1 *In robust Markov Decision Process (MDP) (Wang et al., 2022), denote the underlying environment's transition kernel as $\mathbb{Q}$. Then, robust reinforcement learning aims to optimize the following robust state value function over the policy π.*

$$V^\pi(s) := \inf_{\mathbb{Q}: D_f(\mathbb{Q}|\mathbb{P}) \le \rho} \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi, s_0 = s\Big],$$

where γ is a discount factor, s corresponds to the state and r_t represents the reward obtained after the t-th state transition. Here, the robust value function $V^\pi(s)$ considers the worst-case environment transition kernel $\mathbb{Q}$ over the uncertainty set defined by the f-divergence, i.e., $\{\mathbb{Q} : D_f(\mathbb{Q}|\mathbb{P}) \le \rho\}$. In this setting, if the nominal transition kernel $\mathbb{P}$ cannot visit a certain crucial state s, then neither can the transition kernels $\mathbb{Q}$ from the uncertainty set visit that state s. This indicates that f-divergence does not handle "unknown" uncertainties (e.g., states that never visited by $\mathbb{P}$).

Example 2 *Consider the following f-divergence-regularized DRO problem.*

$$\min_{x \in \mathbf{R}^d} \sup_{\mathbb{Q}} \Big\{\mathbb{E}_{\xi \sim \mathbb{Q}}\big[\ell(x; \xi)\big] - \lambda D_f(\mathbb{Q}|\mathbb{P})\Big\},$$

where the distribution shift on data is characterized by the f-divergence. Suppose we want to train a face detection model. If the nominal data distribution $\mathbb{P}$ only covers face images collected from the majority group and excludes the minority groups, then the f-divergence DRO cannot yield a robust model over all the groups.

On the other hand, the classic Wasserstein distance does not require the distributions $\mathbb{P}, \mathbb{Q}$ to have the same probability support. However, it is known that Wasserstein distance suffers from computational intractability for high-dimension data (Pele and Werman, 2009; Ba et al., 2010), and hence is not suitable for large-scale problems in machine learning.

To tackle these challenges, inspired by the Wasserstein distance, Sinkhorn distance (Cuturi, 2013; Wang et al., 2023), we consider the following *generalized* Sinkhorn distance to quantify the data distribution shift. To elaborate, we consider a sample space Ω associated with σ-algebra $\mathcal{F}$. Furthermore, for distributions $\mathbb{Q}$, $\mathbb{P}$ over a measurable subset of $\mathcal{F}$, we assume they are absolutely continuous with regard to some reference measures ν and μ, i.e., $\mathbb{Q} \ll \nu$, $\mathbb{P} \ll \mu$.

Definition 1 (Generalized Sinkhorn Distance) *Consider probability distributions $\mathbb{Q}$ and $\mathbb{P}$ over $(\Omega, \mathcal{F})$ and let ν and μ be reference measures satisfying $\mathbb{Q} \ll \nu$,$\mathbb{P} \ll \mu$. Denote $\Gamma(\mathbb{P}, \mathbb{Q})$ as the* *set of joint distributions that have marginal distributions $\mathbb{P}, \mathbb{Q}$. For a fixed regularization parameter $\beta > 0$ and a cost metric $c : \Omega \times \Omega \to \mathbf{R}$, the generalized Sinkhorn distance is defined as*

$$W_\beta(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \Big\{ \mathbb{E}_{(\zeta,\xi)\sim\gamma}\big[c(\zeta,\xi)\big] + \beta D_f(\gamma|\mathbb{P}\otimes\nu) \Big\},$$

where D_f corresponds to the f-divergence, that is, $D_f(\gamma|\mathbb{P}\otimes\nu) = \int f\big(\frac{\mathrm{d}\gamma(\zeta,\xi)}{\mathrm{d}\mathbb{P}(\zeta)\mathrm{d}\nu(\xi)}\big)\mathrm{d}\nu(\xi)\mathrm{d}\mathbb{P}(\zeta)$, where the function $f : [0,+\infty) \to [-\infty,+\infty]$ is convex and satisfies $f(1) = 0$ and $f(0) = \lim_{t\to 0+} f(t)$, and $\frac{\mathrm{d}\gamma(\zeta,\xi)}{\mathrm{d}\mathbb{P}(\zeta)\mathrm{d}\nu(\xi)}$ represents the density ratio of γ with respect to $\mathbb{P}\otimes\nu$.

Remark 1 *The absolute continuity condition $\mathbb{Q} \ll \nu$, $\mathbb{P} \ll \mu$ is crucial to guarantee that the generalized Sinkhorn distance is well-defined. Typical choices of the reference measure ν include the Lebesgue measure or the Gaussian measure. In addition, when $\mathbb{Q} \ll \nu$, $D_f(\gamma|\mathbb{P}\otimes\nu)$ and $D_f(\gamma|\mathbb{P}\otimes\mathbb{Q})$ are equivalent up to a constant, which does not affect the optimal solution in the regularized setting. Thus, we consider the former term for simplicity.*

The considered generalized Sinkhorn distance is regularized by the f-divergence , which generalizes the KL-divergence regularization adopted in the definition of the standard Sinkhorn distance (Wang et al., 2023). Such generalization still allows the distributions $\mathbb{P}$ and $\mathbb{Q}$ to have different probability support. By adding f-divergence regularization, it preserves the joint convexity with respect to the probability distributions and thus guarantees the uniqueness of the optimal solution, which helps reduce the computation complexity. Moreover, the generalized Sinkhorn distance provides more flexibility to model data distribution uncertainty compared to other divergence-based measures (Levy et al., 2020; Jin et al., 2021).

4.2 Dual Formulation

With generalized Sinkhorn distance, the regularized DRO problem (1) can be rewritten as

$$\min_{x\in\mathbf{R}^d} \sup_{\mathbb{Q}} \Big\{ \mathbb{E}_{\xi\sim\mathbb{Q}}\big[\ell(x;\xi)\big] - \inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \big\{ \mathbb{E}_{(\zeta,\xi)\sim\gamma}\big[\lambda c(\xi,\zeta)\big] + \lambda\beta D_f(\gamma|\mathbb{P}\otimes\nu) \big\} \Big\}. \tag{2}$$

The primal Sinkhorn distance regularized DRO problem (2) is hard to solve, since it is challenging to obtain an analytical form of the worst-case distribution $\mathbb{Q}$. However, the generalized Sinkhorn distance involves special structures that can transform the primal regularized DRO problem (1) into a simpler dual form. The following theorem deduces an equivalent dual formulation (See Appendix E for proof details).

Theorem 1 (Dual formulation) *The DRO problem (2) has the following equivalent dual formulation*

$$\min_{x\in\mathbf{R}^\mathbf{d}} \mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x)], \text{ where } \Psi_\zeta(x) = \min_{\eta\in\mathbf{R}} \mathbb{E}_{\xi\sim\nu}\underbrace{\Big[\lambda\beta f^*\Big(\frac{\ell(x;\xi) - \lambda c(\zeta,\xi) - \eta}{\lambda\beta}\Big) + \eta\Big]}_{\mathcal{L}_{\xi,\zeta}(x,\eta)}, \tag{3}$$

and f^* denotes the conjugate function of f.

Remark 2 (Technical Novelty) *Proving the equivalence between the primal problem* (2) *and the dual problem* (3) *is crucial. To elaborate, we first decompose the joint distribution as* $\gamma(\zeta, \xi) = \gamma_\zeta(\xi)\mathbb{P}(\zeta)$*, where* γ_ζ *corresponds to the conditional distribution over* ξ*. Then, by the principle of interchangeability (Theorem 7.92, Chapter 7.3.2 in Shapiro et al. (2021)), we are able to swap the order between* $\mathbb{E}_{\zeta\sim\mathbb{P}}$ *and* $\sup_{\gamma_\zeta}$ *without changing the optimal value, which yields*

$$\min_{x\in\mathbf{R}^d} \mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\Psi_\zeta(x) = \sup_{\gamma(\xi|\zeta)} \big\{\mathbb{E}_{\xi\sim\gamma(\cdot|\zeta)}\big[\ell(x;\xi) - \lambda c(\zeta,\xi)\big] - \lambda\beta D_f\big(\gamma(\xi|\zeta)|\nu(\xi)\big)\big\}\Big].$$

Then, by utilizing techniques of data processing inequality, we show that $\Psi_\zeta(x)$ *is equivalent as follows auxiliary function*

$$\widetilde{\Psi}_\zeta(x) = \sup_{\mu_{\gamma|\zeta}}\Big\{\mathbb{E}_{\mu_{\gamma|\zeta}}\big[\ell(x;\xi) - \lambda c(\zeta,\xi)\big] - \lambda\beta D_f(\mu_{\gamma|\zeta}|\mu_\nu)\Big\}.$$

where $\sup_{\mu_{\gamma|\zeta}}$ *corresponds to the supremum over all possible distributions* $\mu_{\gamma|\zeta}$ *induced by* $\gamma(\xi|\zeta)$*. Last, by utilizing inverse c.d.f sampling on* $\widetilde{\Psi}_\zeta(x)$ *introduced in Levy et al. (2020); Duchi and Namkoong (2020), we are able to derive equivalent formulation* $\Psi_\zeta(x) = \min_{\eta\in\mathbf{R}} \int_0^1 \sup_{r\in\mathbf{R}_+} \big[rF^{-1}(u) - \eta(r-1) - \lambda\beta f(r)\big]\mathrm{d}u$*, apply Lagrange duality and definition of convex conjugate dual transforming* $\Psi_\zeta(x)$ *into desired formulation.*

We notice that in Wang et al. (2023); Azizian et al. (2023), the authors also present strong dual formulation for the constrained Sinkhorn DRO problem with D_f being the KL divergence and constraint radius ρ. Specifically, they showed the following equivalent formulation of the problem

$$\min_{x\in\mathbf{R}^d,\lambda>0} \Big\{\lambda\rho + \lambda\beta\mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\log\big(\mathbb{E}_{\xi\sim\nu}\big[\exp(\frac{\ell(x;\xi) - \lambda c(\xi;\zeta)}{\lambda\beta})\big]\big)\Big]\Big\}. \tag{4}$$

Such dual formulation takes a compositional structure of the form $\mathbb{E}_\zeta \log(\mathbb{E}_\xi(\exp(t)))$, where controlling the variance and bias of its stochastic gradient estimator is challenging. To elaborate, Wang et al. (2023) demonstrates that sample-average estimation of gradient leads to a suboptimal convergence rate under mild assumptions. Such limitation motivates them studying variance reduction technique-multilevel Monte Carlo method (MLMC), to improve the sample complexity thereafter.

As a comparison, for proposed Sinkhorn DRO dual formulation (3) derived using Lagrange duality, it involves nested minimization structure, where the inner minimization problem is with respect to the dual variable and sample ζ. Such nested structure is challenging as most nested problems are solved by using bilevel optimization frameworks, where computing hyper-gradient estimation is known to be time-consuming due to the requirement of second-order information (Franceschi et al., 2018). Moreover, the dependency between the inner minimizer η and ζ motivates us to reduce the number of sampled ζ to $\mathcal{O}(1)$-level per iteration. Thanks to Lemma 1 proved by Jin et al. (2021), which provides a tractable expression for evaluating gradients without requiring second-order information. As we later demonstrate in Theorem 2 and 4, one can efficiently control the gradient approximation error and establish convergence of nested SGD for solving proposed dual formulation (3) without querying a large batch of ζ and ξ per iteration.

5 Solving the Dual Problem via Nested SGD

To minimize optimization problem with contextual nested structure, Hu et al. (2023) studied bilevel SGD framework and attained near optimal sample complexity $\tilde{\mathcal{O}}(\varepsilon^{-4})$ when lower level problem is strongly convex and MLMC is used for hyper-gradient estimation. However, their analysis is not directly applicable to proposed Sinkhorn DRO dual formulation (3), as their framework assumes dependency between ξ and ζ, but does not account for the dependency between the dual variable η and the sample ζ. Additionally, most conjugate dual functions f^* for information divergences are not globally strongly convex, which further limits the applicability of their analysis. As we mentioned before, Wang et al. (2023); Hu et al. (2023) showed that MLMC requires storing multiple samples of ζ up to $\mathcal{O}(\log(\varepsilon^{-1}))$ to construct a stochastic estimator with $\mathcal{O}(\varepsilon)$-level approximation error, and second-order derivative remains unavoidable for evaluating hyper-gradient in contextual bilevel objective.

Motivated by these bottlenecks, in this section, we present a tractable expression for computing the gradient of proposed Sinkhorn DRO dual formulation (3) and we then find, by exploring the structural relationship between the inner and outer objective gradients, that the requirements of strong convexity assumption and sampling multiple ζ can be eliminated to control the approximation error between $\nabla\mathbb{E}_\zeta[\Psi_\zeta(x)]$ and $\nabla_1\mathbb{E}_\zeta[\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))]$ given $\eta_x^{\tilde{d}}(\zeta)$ is an output of stochastic oracle satisfying mild conditions. Recall that the dual formulation (3) consists of two stochastic optimization problems. Throughout, we denote the objective function of the inner problem in (3) as

$$\mathcal{L}_\zeta(x,\eta) = \mathbb{E}_{\xi\sim\nu}\big[\lambda\beta f^*\big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta}{\lambda\beta}\big)\big]+\eta. \tag{5}$$

For simplicity, we denote $\eta_x^*(\zeta) = \arg\min_\eta L_\zeta(x,\eta)$ to highlight its dependence on the fixed parameter x and data sample ζ. We also denote $\mathbb{E}_\zeta\big[\Psi_\zeta(x)\big]$ as the objective function of the outer problem.

To analyze the problem structure, we adopt the following standard assumptions on the loss functions and convex conjugate dual of chosen f-divergence.

Assumption 1 *The functions in Sinkhorn DRO dual formulation* (3) *satisfy:*

- *For every ξ, $\ell(\cdot;\xi)$ is G-Lipschitz continuous. i.e., $\big\|\ell(x;\xi)-\ell(y;\xi)\big\| \le G\big\|x-y\big\|$.*
- *For every ξ, $\ell(\cdot;\xi)$ is continuously differentiable and L-smooth. i.e., $\big\|\nabla\ell(x;\xi)-\nabla\ell(y;\xi)\big\| \le L\big\|x-y\big\|$.*
- *The conjugate function f^* is continuously differentiable and M-smooth.*
- *The objective function $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(\cdot)\big]$ is bounded below.*

Remark 3 *Note that the loss function $\ell(x;\xi)$ is generally nonconvex. Regarding the smooth assumption on f^* in the third item, some typical examples of f-divergence include the χ^2-divergence, smoothed CVaR divergence (Jin et al., 2021), where their corresponding conjugate functions are given by $f^*(y) = -1+\frac{1}{4}(y+2)_+^2$ and $f^*(t) = \frac{1}{\alpha}\log(1-\alpha+\alpha\exp(t))$. To analyze the dual formulation* (3) *with KL-divergence, we need to adopt assumption $\sup_x \exp((\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta)/\lambda\beta) < \infty$ holds almost-surely for every ξ given ζ. Under these assumptions, one can ensure that the convex conjugate dual $f^*(t) = \exp(t)-1$ is locally M-smooth within a bounded domain. We will later demonstrate in our ablation study (Appendix D) that the f^* induced by the KL-divergence yields similar convergence behavior when optimizing dual formulation* (3)*, compared to choices of*

f^* that satisfy the M-smoothness property, suggesting that this assumption does not limit practical applicability.

Since the dual problem takes a nested form, we need an efficient way to compute the gradient of the objective function. The following lemma, proved by Jin et al. (2021), provides an analytical formula for computing the exact gradient. (See Lemma 2.6 in Jin et al. (2021) for proof details)

Lemma 1 (Computation of $\nabla\Psi_\zeta(x)$ (Jin et al., 2021)) *Let Assumption 1 hold and consider fixed x and given ζ. Then, the function $\Psi_\zeta(x)$ is differentiable and satisfies $\nabla\Psi_\zeta(x) = \nabla_1\mathcal{L}_\zeta(x, \eta_x^*(\zeta))$, where $\eta_x^*(\zeta) \in \arg\min_\eta \mathcal{L}_\zeta(x, \eta)$.*

This lemma shows that, given the exact minimizer $\eta_x^*(\zeta)$ of the inner problem, one can directly evaluate the gradient of dual formulation (3). Notably, it eliminates the need to acquire second-order derivatives for computing $\nabla_1\eta_x^*(\zeta)$, which are typically required under bilevel optimization framework (Franceschi et al., 2018; Ghadimi and Wang, 2018; Hu et al., 2023). Motivated by Lemma 1, we aim to develop conditions and algorithms to estimate $\eta_x^*(\zeta)$ with arbitrarily small error, thereby enabling the construction of an inexact gradient estimator for proposed Sinkhorn DRO dual formulation (3) with an ε-level approximation error. The following theorem shows that, for any fixed x and ζ, the approximation error in estimating $\nabla\Psi_\zeta(x)$ can be made arbitrarily small by querying an inner solution $\eta_x^{\tilde{d}}(\zeta)$ that is accurate on average. (See Appendix G for proof details)

Theorem 2 (Gradient approximation error bound) *Consider a stochastic algorithm minimizing (5). If the stochastic oracle outputs an $\eta_x^{\tilde{d}}(\zeta)$ converging to $\nabla_2\mathcal{L}_\zeta(x_t, \eta_x^*(\zeta))$ with scaled small target error $\tilde{\varepsilon} = \varepsilon/G$, i.e.,*

$$\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}\left|\nabla_2\mathcal{L}_\zeta(x, \eta_x^{\tilde{d}}(\zeta))\right|^2 \le \tilde{\varepsilon}^2, \tag{6}$$

then the gradient $\nabla_1\mathcal{L}_\zeta(x, \eta_x^{\tilde{d}}(\zeta))$ approximates full gradient $\nabla\Psi(x)$ with error up to ε, i.e.,

$$\left\|\nabla\Psi_\zeta(x) - \mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}[\nabla_1\mathcal{L}_\zeta(x, \eta_x^{\tilde{d}}(\zeta))]\right\|^2 \le \varepsilon^2, \forall\zeta \sim \mathbb{P}. \tag{7}$$

Remark 4 (Technical Novelty) *The novelty of our proof lies in the usage of the monotonicity property of $(f^*)'$, which enables us to perform an equivalence transformation by moving the expectation $\mathbb{E}_{\xi\sim\nu}$ into the norm $\|\cdot\|$. Thanks to special structure of formulation (3), such operation swap doesn't change its value since each inner problem $\min_\eta \mathcal{L}_\zeta(x, \eta)$ depends on a fixed ζ. In the proof, we also utilize the convexity of the conjugate function $f^*(\frac{\ell(x;\xi)-c(\zeta;\xi)-\eta}{\lambda\beta})$ in η, albeit the loss function $\ell(x;\xi)$ is generally nonconvex.*

Note that Theorem 2 requires an algorithm capable of generating an $\eta_x^{\tilde{d}}(\zeta)$ that satisfies condition (6), which can be achieved by the vanilla SGD algorithm (Ghadimi and Lan, 2013) under mild assumptions. And the condition (6) cannot be further reduced to other forms, as the second-order condition (6) will be reused when estimating the second moment of the inexact gradient (See Lemma 5).

Additionally, condition (7) characterizes the relationship between the gradients $\nabla\Psi_\zeta(x)$ and $\nabla_1\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}[\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))]$ when the inner problem (5) is solved approximately by an inexact minimizer $\eta_x^{\tilde{d}}(\zeta)$. This observation motivates the design of a nested-type stochastic algorithm (detailed in the next subsection), in which the inner algorithm computes an inexact minimizer $\eta_x^{\tilde{d}}(\zeta)$ for $\mathcal{L}_\zeta(x,\eta)$, and the outer stochastic algorithm subsequently optimizes $\min_x \mathbb{E}_{\zeta\sim\mathbb{P}}[\mathcal{L}_\zeta(x,\eta)]$. This approach is justified by taking the expectation over $\zeta\sim\mathbb{P}$ and applying Jensen's inequality to (7), which further implies

$$\begin{aligned}&\big\|\nabla_1\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta)}\big[\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))\big]-\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]\big\|^2\\&\le\mathbb{E}_{\zeta\sim\mathbb{P}}\big\|\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))-\nabla\Psi_\zeta(x)\big\|^2\le\varepsilon^2.\end{aligned}\tag{8}$$

Through condition (8), we conclude that one can establish bias and variance guarantee with respect to the true gradient $\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x)]\|$ by sub-sampling $\mathcal{O}(1)$-batches of ζ.

6 Algorithm design and convergence analysis

In this section, we propose a nested procedure composed of Algorithm 1 and Algorithm 2 to sequentially optimize x and estimate $\eta_x^*(\zeta)$ via SGD-type algorithms. To facilitate the convergence analysis of these algorithms, we impose the following bounded variance assumptions on the loss function $\ell(x;\xi)$ and the cost metric $c(\zeta;\xi)$. Through the work, to account for randomness arising from high-dimensional data, we use the variance definition $Var_\varpi(\rho(\varpi))=\mathbb{E}\big[\rho(\varpi)-\mathbb{E}\big[\rho(\varpi)\big]\big]^2$ in our assumption, where $\rho(\cdot):\mathbf{R}^d\to\mathbf{R}$ denotes a function mapping a random variable $\varpi\in\mathbf{R}^d$ to a scalar.

Assumption 2 *There exists $\sigma,\delta>0$ such that:*

- *For every x, the variance of $\ell(x;\cdot)$ over ξ is bounded by σ^2, i.e., $Var_\xi\big(\ell(x;\xi)\big)\le\sigma^2$.*
- *For every ζ, the variance of $c(\zeta,\cdot)$ over ξ is bounded by δ^2, i.e., $Var_\xi\big(c(\zeta,\xi)\big)\le\delta^2$.*
- *For every ξ, the variance of $c(\cdot,\xi)$ over ζ is bounded by δ^2, i.e., $Var_\zeta\big(c(\zeta,\xi)\big)\le\delta^2$.*

6.1 Inexact Estimation of Inner Minimizer via SGD

Recall that the inner problem of the dual problem (3) takes the following form.

$$\min_{\eta\in\mathbf{R}}\mathcal{L}_\zeta(x,\eta)=\mathbb{E}_{\xi\sim\nu}\big[\lambda\beta f^*\big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta}{\lambda\beta}\big)+\eta\big],\tag{9}$$

which is a stochastic optimization problem. Inspired by condition (6) stated in Theorem 2, we utilize a SGD-type algorithm (Algorithm 1) for computing an inexact estimation of $\eta_x^*(\zeta)$. Specifically, for fixed x and sample ζ, we compute the mini-batch stochastic gradient estimator of $\nabla_2\mathcal{L}_\zeta(x,\eta)$ at each iteration as follows

$$v_{x,\zeta}^{\tilde{B}}(\eta)=1-\frac{1}{\tilde{B}}\sum_{i=1}^{\tilde{B}}(f^*)'\big(\frac{\ell(x;\xi_i)-c(\zeta,\xi_i)-\eta}{\lambda\beta}\big).\tag{10}$$

For simplicity, we simplify notation $v_{x,\zeta}^{\tilde{B}}(\eta)$ as $v^{\tilde{B}}(\eta)$ when the dependence on x and ζ is clear from the context.

Algorithm 1 SGD for estimating $\eta_x^*(\zeta)$

Input: Initialization η^0; sample ζ; number of iteration D; batch size $\tilde{B}$.
while $d < D$ **do**
  Draw samples $\{\xi\}_{\tilde{B}} \sim \nu$ with batch size $\tilde{B}$.
  Compute $v(\eta_{x_t}^d(\zeta))$ via (10).
  Update $\eta_{x_t}^{d+1}(\zeta) = \eta_{x_t}^d(\zeta) - \alpha_d v^{\tilde{B}}(\eta_{x_t}^d(\zeta))$.
end while
Output: $\eta_x^{\tilde{d}}(\zeta)$, where $\tilde{d}$ uniformly sampled from $\{0, 1, 2, ...D-1\}$.

To analyze the convergence of the inner Algorithm 1 based on Assumptions 1 and 2, we first obtain the following expected smoothness property (See Appendix H for proof details)

Lemma 2 (K'-smoothness of inner objective (5)) *Let Assumption 1 hold and denote* $\mathcal{L}_{\zeta,\xi}(x,\eta) = \lambda\beta\mathbb{E}_{\xi\sim\nu}\big[f^*(\frac{\ell(x;\xi)-c(\zeta,\xi)-\eta}{\lambda\beta})\big] + \eta$. *Then, for any η and η', we have*

$$\mathbb{E}_{\xi\sim\nu}\big\|\nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta) - \nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta')\big\|^2 \leq (K')^2\big\|\eta-\eta'\big\|^2,$$

where $K' = M(\lambda\beta)^{-1}$.

We then obtain the upper bound estimate for second moment of $\nabla_2\mathcal{L}_\zeta(x,\eta)$ as follows (See Appendix I for proof details).

Lemma 3 (Second moment bound for $\nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta)$) *Let Assumption 2 hold. Then, $v^{\tilde{B}}(\eta)$ is the unbiased estimator for $\nabla_2\mathcal{L}_\zeta(x,\eta)$, and the second moment of $\nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta)$ satisfies*

$$\mathbb{E}_{\xi\sim\nu}\left\|\nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta)\right\|^2 \leq R_2 + \left\|\nabla_2\mathcal{L}_\zeta(x,\eta)\right\|^2, \tag{11}$$

where $R_2 = 2M^2(\lambda\beta)^{-2}(\sigma^2+\lambda^2\delta^2)$.

Based on K'-smooth and affine bounded second moment shown above, we establish the following convergence result of Algorithm 1 (See Appendix J for proof details).

Corollary 3 (Convergence of Algorithm 1) *Let Assumptions 1 and 2 hold, denote* $\widehat{\Delta} = \sup_\zeta\{L_\zeta(x_t,\eta^0) - \Psi_\zeta(x_t)\}$. *Apply Algorithm 1 to solve the inner problem (5) with learning rate* $\alpha_d = \min\{\frac{1}{K'}, \frac{\tilde{\varepsilon}^2}{R_2K'}\}$ *and batch size $\tilde{B} = 1$, then Algorithm 1 outputs an $\eta_x^{\tilde{d}}(\zeta)$ satisfying*

$$\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}\big|\nabla_2\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))\big|^2 \leq \tilde{\varepsilon}^2. \tag{12}$$

In particular, it takes $D = \mathcal{O}(\hat{\Delta}K'R_2\tilde{\varepsilon}^{-4}) = \mathcal{O}(\hat{\Delta}K'R_2G^4\varepsilon^{-4})$ *number of iterations to obtain an $\tilde{\varepsilon}$-stationary point, and the stochastic gradient oracle complexity is* $\mathcal{O}(\widehat{\Delta}K'R_2G^4\varepsilon^{-4})$.

Remark 5 *Notice that solving the inner optimization problem (5) to obtain an $\eta_x^{\tilde{d}}(\zeta)$ only requires solving a one-dimensional optimization problem, where evaluating its stochastic gradient does not require expensive tensor computation or backpropagation. Consequently, employing a large batch size can substantially reduce the iteration complexity without significantly increasing computational overhead. For example, by selecting batch size as* $\tilde{B} = \Theta(G^{-2}\tilde{\varepsilon}^{-2}) = \Theta(\varepsilon^{-2})$ *and setting the learning rate* $\alpha_d = \min\{\frac{1}{K'}, \frac{1}{2K'R_2G^2}\}$, *Algorithm 1 generates an $\eta_x^{\tilde{d}}(\zeta)$ satisfying condition (6) after* $D \geq 8R_2\hat{\Delta}K'G^4\varepsilon^{-2} = \mathcal{O}(R_2\hat{\Delta}K'G^4\varepsilon^{-2})$ *iterations.*

6.2 Nested SGD and its convergence analysis

The previous section has shown that SGD-type algorithms can successfully generate an $\eta_x^{\tilde{d}}(\zeta)$ satisfying the conditions required by Theorem 2. In this section, we introduce the nested SGD algorithm (Algorithm 2) and establish its convergence guarantees. Specifically, at each iteration, Algorithm 2 samples a single ζ along with a mini-batch $\{\xi\}_B$, and obtain an inexact estimate of $\eta_x^*(\zeta)$ by calling Algorithm 1. It is worth noting that the sampled ξ from Algorithm 1 can be reused to save computation overhead. Algorithm 2 then evaluates a mini-batch gradient estimator of $\nabla\mathbb{E}_\zeta[\Psi(x_t)]$ as follows

$$\hat{g}(x_t;\zeta;\xi_B) = \frac{1}{B}\sum_{i=1}^{B}(f^*)'\big(\frac{\ell(x;\xi_i) - c(\zeta,\xi_i) - \eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\big)\nabla\ell(x_t;\xi_i). \tag{13}$$

Remark 6 *In the stochastic gradient estimator* (13)*, note that we only apply mini-batch sampling over ξ. This is because the inexact minimizer $\eta_x^{\tilde{d}}(\zeta)$ explicitly depends on ζ and thus can vary over different ζ. We notice that Wang et al. (2023) proposes RT-MLMC estimator, which is specifically designed to reduce the upper bound of the second moment using fewer samples. In our setting, their RT-MLMC gradient estimator can be rewritten as*

$$\begin{aligned} g^{\textit{RT-MLMC}}(x_t) &= \frac{1}{n_q^\circ}\sum_{i=1}^{n_q^\circ}\frac{1}{\mathbb{P}(\iota=\iota_i)}A^{\iota_i}(x_t;\zeta^{\iota_i}),\ \textit{where} \\ A^{\iota_i}(x_t;\zeta^{\iota_i}) &= \hat{g}(x_t;\zeta^{\iota_i};\xi_{1:2^l}) - \frac{1}{2}\hat{g}(x_t;\zeta^{\iota_i};\xi_{1:2^{l-1}}) - \frac{1}{2}\hat{g}(x_t;\zeta^{\iota_i};\xi_{2^{l-1}+1:2^l}), \end{aligned} \tag{14}$$

where n_q° denotes the randomly sampled levels $\iota_1,\dots,\iota_{n_q^\circ}$, where each level ι_i is sampled independently following the distribution $\mathbb{P}(\iota=\iota_i) = \frac{2^{-\iota_i}}{2-2^{-q}}$, for $l=0,\dots,q$. To ensure convergence, Wang et al. (2023) demonstrates that by choosing hyper-parameters $n_q^\circ = \mathcal{O}(1), q = \mathcal{O}(\log(\varepsilon^{-1}))$, RT-MLMC estimators ensures SGD achieving $\tilde{\mathcal{O}}(\varepsilon^{-2})$ near-optimal convergence when loss function $\ell(x;\xi)$ is convex and bounded. However, due to the nested structure of proposed Sinkhorn DRO dual formulation (3)*, we find that the RT-MLMC estimator is not well-suited for our setting, as* (14) *requires storing $q=\mathcal{O}(\log(\varepsilon^{-1}))$ samples of ζ. This, in turn, simultaneously increases the number of calls to Algorithm 1 per iteration within Algorithm 2 to obtain corresponding $\eta_{x_t}^{\tilde{d}}(\zeta^{\iota_i})$. The nested structure of the problem limits the effectiveness of variance reduction typically offered by the multilevel Monte Carlo method and fails to significantly improve the convergence order. Furthermore, it may even introduce additional challenges in sampling and computational efficiency, particularly when applying estimator* (14) *within the backpropagation process of deep neural networks. As we show later in Theorem 4, our proposed Algorithm 2, which employs gradient estimator* (13) *and requires only a single sample of ζ and ξ (namely $B=1$) per iteration, achieves a convergence rate of $T=\mathcal{O}(\varepsilon^{-4})$ in standard nonconvex setting.*

At each iteration, Algorithm 2 then uses SGD-algorithm with respect to ξ to update x_t iteratively. For simplicity, in the following article, we use the simplified notation $\hat{g}_t^B$ to denote expression (13) when the dependence on x_t, ζ and ξ_B is clear from the context.

Next, we analyze the convergence of Algorithm 2. We first study the smoothness property of the objective function $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$. We note that the bi-variate function $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\mathcal{L}_\zeta(x,\eta)\big]$ has been

Algorithm 2 Nested SGD for optimizing x

Input: Initialization x_0; number of iteration T; learning rate γ_t; batch size B.
while $t < T$ **do**
  Draw samples $\zeta \sim \mathbb{P}$ and $\{\xi\}_B \sim \nu$ with batch size B independently.
  Apply Algorithm 1 to compute estimator $\eta_x^{\tilde{d}}(\zeta)$.
  Compute $\hat{g}_t^B$ via (13).
  Update $x_{t+1} = x_t - \gamma_t \hat{g}_t^B$.
end while
Output: $x_{\tilde{t}}$, where $\tilde{t}$ sampled uniformly from $\{0, \dots, T-1\}$.

shown to satisfy a generalized-smooth condition (Zhang et al., 2020; Chen et al., 2023c). However, if η is chosen to be the minimizer $\eta_x^*(\zeta)$, we show that the objective function satisfies the following directional smoothness property (Mishkin et al., 2024) (See Appendix K for proof details).

Lemma 4 (Directional Smoothness) *Let Assumption 1 hold. For any x and x', we have*

$$\mathbb{E}_{\zeta\sim\mathbb{P}}\big\|\nabla\Psi_\zeta(x) - \nabla_1\mathcal{L}_\zeta(x', \eta_x^*(\zeta))\big\|^2 \le K^2\big\|x - x'\big\|^2, \text{ where } K = G^2(\lambda\beta)^{-1}M + L. \tag{15}$$

Recall that $\eta_x^{\tilde{d}}(\zeta)$ is inexact estimation of $\eta_x^*(\zeta)$ obtained from Algorithm 1. To analyze convergence of Algorithm 2, we obtain the upper bound of the second moment for $\hat{g}_t^B$, which is stated in next lemma (See Appendix L for proof details).

Lemma 5 (Second moment bound of $\hat{g}_t^B$) *The second moment of the mini-batch gradient estimator $\hat{g}_t^B$ with batch size B defined in (13) is upper bounded as follows*

$$\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta),\xi_B\sim\nu}\big\|\hat{g}_t^B\big\|^2 \le \frac{R_1 + 10\varepsilon^2}{B} + \big\|\nabla_1\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta)}\big[\mathcal{L}_\zeta(x_t, \eta_{x_t}^{\tilde{d}}(\zeta))\big]\big\|^2, \tag{16}$$

where $R_1 = 8G^2 + 24G^2M^2(\lambda\beta)^{-2}\sigma^2 + 24G^2M^2\beta^{-2}\delta^2$.

Based on the above lemma, we obtain the following convergence result of Nested-SGD for minimizing $\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x)]$ (See Appendix M for proof details).

Theorem 4 (Convergence of Algorithm 2) *Let Assumptions 1 and 2 hold. Denote $\Delta = \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_0)\big] - \inf_x \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$, apply Algorithm 2 to solve the outer objective in (3) using a constant learning rate $\gamma_t = \gamma = \min\{\frac{1}{24K}, \frac{\varepsilon^2}{2KR_1}\}$, and set the batch size $B = 1$. At each iteration, query Algorithm 1 to obtain an estimator $\eta_x^{\tilde{d}}(\zeta)$ for sampled ζ. Then, the output $x_{\tilde{t}}$ of Algorithm 2 satisfies*

$$\mathbb{E}_{x_{\tilde{t}}}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{x_{\tilde{t}}})\big]\big\|^2 \le 7\varepsilon^2, \tag{17}$$

after $T \ge \max\{96\Delta K\varepsilon^{-2}, 8\Delta KR_1\varepsilon^{-4}\} = \mathcal{O}(\Delta KR_1\varepsilon^{-4})$ number of iterations.

This theorem indicates Algorithm 2 takes $\mathcal{O}(\varepsilon^{-4})$ iterations to obtain an $\mathcal{O}(\varepsilon)$-stationary point, which implies the inexact minimizer $\eta_x^{\tilde{d}}(\zeta)$ has minor effects on the worst-case sample complexity of Algorithm 2 for solving proposed Sinkhorn DRO dual formulation (3). Next, we summarize the overall sample and iteration complexity, as well as the per-iteration sample and memory complexities of Algorithms 1 and 2 (See Appendix N for proof details).

Corollary 5 (Complexity Bound for $\min_x \mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x)]$) *Let Assumptions 1 and 2 hold. Then, the Nested-SGD algorithm (Algorithm 2) returns an ε-stationary point with a total sample complexity of $\mathcal{O}(\varepsilon^{-8})$ for sampling ξ and ζ. Furthermore, by setting the batch sizes $B, \tilde{B} \sim \Theta(\varepsilon^{-2})$, the total iteration complexity becomes $T \times D \sim \mathcal{O}(\varepsilon^{-4})$. At each iteration, Algorithms 1 and 2 incur memory complexities of $\mathcal{O}(1)$ and $\mathcal{O}(d)$, respectively.*

Compared with existing stochastic bilevel methods—e.g., Hu et al. (2023); Kwon et al. (2023); Chen et al. (2023a); Huang (2023)—that study bilevel optimization framework, our nested-SGD Algorithm 2 attains a moderate overall sample-complexity bound. However, the tighter bounds in those works highly rely on additional regularity assumptions that are incompatible with proposed dual formulation (3). Specifically, Hu et al. (2023) establish a sample complexity of $\tilde{\mathcal{O}}(\varepsilon^{-6})$ for mini-batch SGD under a strongly convex lower-level problem. Huang (2023); Chen et al. (2023b) obtain sample complexity of $\tilde{\mathcal{O}}(\varepsilon^{-4})$ and $\tilde{\mathcal{O}}(\varepsilon^{-6})$, respectively, by assuming the lower-level objective satisfies the Polyak–Łojasiewicz (PL) condition. Finally, Chen et al. (2023a); Kwon et al. (2023) report sample complexity of $\mathcal{O}(\varepsilon^{-4})$ and $\mathcal{O}(\varepsilon^{-7})$ for (i) a jointly convex upper–lower structure and (ii) a non-convex setting satisfying small-error proximal error-bound (EB) condition. Except Hu et al. (2023), none of these studies incorporate an in-context variable in the lower-level problem. Later through numerical experiments in section 7, We show that Algorithm 2 enables proposed dual formulation (3) to achieve performance comparable to that of Wang et al. (2023), while requiring only a small number of queries to ζ and ξ under the same iteration budget T and a small inner-loop depth D. Moreover, it remains scalable to large-scale problems, such as training robust deep neural networks under distribution shifts.

6.3 Proof Sketch of Theorem 4

Based on directional smooth property stated in Lemma 4, we obtain the similar descent lemma as L-smooth function along the direction $x_{t+1} - x_t$ when $\eta^*_{x_t}(\zeta)$ is fixed (See (57) in Appendix M.1). By replacing $x_{t+1} - x_t$ with biased gradient estimator, $\hat{g}_t$, one can obtain

$$\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{t+1})\big] \le \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big] - \gamma_t\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], \hat{g}_t\rangle + \frac{K\gamma_t^2}{2}\big\|\hat{g}_t\big\|^2.$$

When $\hat{g}_t^B$ is used, it induces randomness from x_t, ζ, $\{\xi\}_{\tilde{B}}$ and $\eta_x^{\tilde{d}}(\zeta)$. Taking expectation conditioned on x_t over $\zeta, \eta_x^{\tilde{d}}(\zeta), \{\xi\}_B$ on both sides of above inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\Psi_\zeta(x_{t+1})|x_t\big] \le \mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\Psi_\zeta(x_t)|x_t\big] \\
&\qquad - \mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], \gamma_t\hat{g}_t^B\rangle|x_t\big] + \underbrace{\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\frac{K\gamma_t^2}{2}\big\|\hat{g}_t^B\big\|^2|x_t\big]}_{\textit{second moment}}.
\end{aligned}$$

For upper bounding term *"second moment"*, one can utilize (16) stated in Lemma 5, which leads to

$$\begin{aligned}
&\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\Psi_\zeta(x_{t+1})|x_t\big] \le \mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\Psi_\zeta(x_t)|x_t\big] + \frac{K\gamma_t^2(R_1 + 10\varepsilon^2)}{2} \\
&\qquad \underbrace{-\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\,\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], \gamma_t\hat{g}_t^B\rangle\,|x_t\big]}_{\textit{Term 1}} + \underbrace{\frac{K\gamma_t^2}{2}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta)}\big[\mathcal{L}_\zeta(x_t, \eta^{\tilde{d}}_{x_t}(\zeta))\big]\big\|^2}_{\textit{Term 2}}
\end{aligned}$$

For *"Term 1"*, we expand this term as $-\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\gamma_t\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x_t)],\hat{g}^B_t-g^B_t+g^B_t\rangle|x_t\big]$. Since $\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\Psi(x_t)$ is independent from randomness induced by ξ_B and $\eta^{\tilde{d}}_{x_t}(\zeta)$, we move expectation over $\xi_B\sim\nu$ and $\eta^{\tilde{d}}_{x_t}(\zeta)$ into inner product. For $-\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\gamma_t\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x_t)],\mathbb{E}_{\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}[\hat{g}^B_t-g^B_t]\rangle|x_t\big]$, we first apply Cauchy-Scharwarz inequality and then use condition (7) stated in Theorem 2 to obtain upper bound $\gamma_t\varepsilon\mathbb{E}_{x_t}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|$. For $-\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\gamma_t\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x_t)],g^B_t\rangle|x_t\big]$, it is equivalent to rewrite as $-\gamma_t\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x_t)]\|^2$.

Similarly, For *"Term2"*, by plus and minus an additional term $\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x_t)]$ in squared norm, utilizing inequality $(a+b)^2\leq 2a^2+2b^2$ and condition (8) stated in Theorem 2, we can upper bound "Term2" by $K\gamma_t^2\varepsilon^2+K\gamma_t^2\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x_t)]\|^2$. Re-arranging above inequality and applying fact $\gamma_t=\gamma=\min\{\frac{1}{24K},\frac{\varepsilon^2}{2KR_1}\}<\frac{1}{2K}$, we have

$$\begin{aligned}&\frac{\gamma_t}{2}\mathbb{E}_{x_t}\Big(\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2-2\varepsilon\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|\Big)\\&\leq\mathbb{E}_{x_t,\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\big[\Psi_\zeta(x_t)-\Psi_\zeta(x_{t+1})\big]+\frac{KR_1\gamma_t^2}{2}+6K\varepsilon^2\gamma_t^2.\end{aligned}$$

Summing above inequality from 0 to $T-1$, applying $\gamma_t=\gamma=\min\{\frac{1}{24K},\frac{\varepsilon^2}{2KR_1}\}$ and $T\geq\max\{96\Delta K\varepsilon^{-2},8\Delta KR_1\varepsilon^{-4}\}$, we further conclude

$$\mathbb{E}_{x_{\tilde{t}}}\big[\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{\tilde{t}})\big]\big\|^2\big]\leq\frac{4\Delta}{T\gamma}+\frac{24K\varepsilon^2\gamma^2T}{T\gamma}+\frac{2KR_1\gamma^2T}{T\gamma}+\frac{4\varepsilon^2T\gamma}{T\gamma}\leq 7\varepsilon^2.$$

7 Experiments

In this section, we evaluate the performance of proposed Sinkhorn DRO dual formulation (3) in comparison to other baselines, including (constrained) Sinkhorn DRO dual formulation (4) from Wang et al. (2023), regularized f-divergence DRO from Jin et al. (2021); Duchi and Namkoong (2020), and empirical risk minimization (Vapnik and Chervonenkis, 2015) under distribution shifts[1]. To elaborate, we train logistic regression and LeNet (LeCun et al., 1998) for classification tasks on real-world datasets, where we simulate distribution shifts by applying adversarial attacks on test dataset. To align with assumption 1, for proposed Sinkhorn DRO dual formulation (3), we choose f^* to be the conjugate dual of the χ^2-divergence, which satisfies the M-smoothness property. We unify the training procedure across all formulations using vanilla SGD (Ghadimi and Lan, 2013), implemented via PyTorch's autograd toolbox (Paszke et al., 2019), *without* employing additional heuristics such as random shuffling, learning rate scheduling, or weight decay. Test results are reported using the model parameters obtained at the last epoch, rather than the uniform averaging over iterates used in theoretical analysis. Additionally, for gradient estimator (13), we evaluate the expression using the sampled ξ and $\eta^{\tilde{d}}_x(\zeta)$ obtained from Algorithm 1 to reduce computational overhead. To make a fair comparison between proposed Sinkhorn DRO dual formulation (3) and dual formulation from Wang et al. (2023), we generate sample ξ following same distribution. Due to page limitation, we refer readers to check Appendix B and C for more details on model initialization and algorithm hyper-parameter settings. We also conduct linear regression experiment over synthetic data and ablation studies of proposed Sinkhorn DRO dual formulation (3), where we present the

1. Code available at: https://github.com/ynyang94/GeneralSinkhorn-Regularized-DRO

corresponding results in Appendix A, D respectively. All the previously mentioned experiments were conducted on a PC computer with 32GB memory, 24 cores CPU running Python 3.8.

7.1 Logistic Regression on CIFAR-10

FGSM (Logistic)	Sinkhorn DRO[1]	f-DRO	ERM	Sinkhorn DRO[2]
$\epsilon_{\text{FGSM}} = 0.00$	**77.07%**	77.00%	73.26%	75.27%
$\epsilon_{\text{FGSM}} = 0.01$	65.60%	63.80%	60.73%	**67.13%**
$\epsilon_{\text{FGSM}} = 0.02$	54.0%	46.07%	45.67%	**58.13%**
ℓ_∞-PGD (Logistic)	Sinkhorn DRO[1]	f-DRO	ERM	Sinkhorn DRO[2]
$\epsilon_{\text{PGD}} = 0.01, iter = 20$	66.93%	66.0%	62.2%	**68.2%**
$\epsilon_{\text{PGD}} = 0.02, iter = 20$	57.93%	53.27%	51.0%	**60.47%**
$\epsilon_{\text{PGD}} = 0.03, iter = 20$	45.47%	37.93%	38.07%	**50.60%**

Table 1: Test classification accuracy of logistic regression.

In this section, we apply the proposed Sinkhorn DRO dual formulation (3) and other baselines on logistic regression over CIFAR-10 (Krizhevsky, 2009), and test the classification accuracy on adversarial samples generated by the fast gradient sign method (FGSM) (Goodfellow et al., 2015) and ℓ_∞-PGD (Madry et al., 2018) attacks utilizing model parameters obtained at *last* epoch. Table 1 reports the test accuracy under different perturbation magnitudes, where Sinkhorn DRO[1] refers to formulations (4) from Wang et al. (2023) and Sinkhorn DRO[2] refers to proposed Sinkhorn DRO dual formulation (3). The corresponding test loss curves are plotted in Figure 2. We found that when the test data are clean, the proposed Sinkhorn DRO dual formulation (3) achieves performance comparable to other baselines. However, as attack level increases, the model obtained via proposed Sinkhorn DRO formulation (3) is more robust than others, which demonstrates the advantage and effectiveness of Sinkhorn DRO and proposed Nested SGD Algorithm 2.

7.2 LeNet Classification on MNIST

In this section, we apply proposed Sinkhorn DRO dual formulation (3) and other baselines to train a LeNet (LeCun et al., 1998) over MNIST (Deng, 2012), and test the classification accuracy on adversarial samples generated by FGSM (Goodfellow et al., 2015), ℓ_∞, ℓ_2-PGD (Madry et al., 2018) and momentum iterative method (MIM) (Dong et al., 2018) attacks utilizing model parameters obtained at *last* epoch. Table 2 reports the test accuracies under different perturbation magnitudes and the corresponding test loss curves are plotted in Figure 3, 4, 5 and 6. Specially, we found f-DRO is vulnerable against FGSM and PGD attacks. This might be due to the generalized-smoothness property of f-DRO objective (Jin et al., 2021; Chen et al., 2023c), which makes it hard to optimize using vanilla SGD. As for the proposed Sinkhorn DRO dual formulation (3), we find that it achieves higher classification accuracy than the Sinkhorn DRO dual formulation (4) from Wang et al. (2023) across most attack magnitudes, and exhibits a smaller drop in accuracy as the attack strength increases. This demonstrates the effectiveness of our proposed Sinkhorn DRO dual formulation (3) and supports the validity of our algorithmic analysis.

FGSM (LeNet)	Sinkhorn DRO[1]	f-DRO	ERM	Sinkhorn DRO[2]
$\epsilon_{\text{FGSM}} = 0.00$	95.89%	94.60%	95.50%	**96.80%**
$\epsilon_{\text{FGSM}} = 0.02$	83.10%	60.30%	79.20%	**89.80%**
$\epsilon_{\text{FGSM}} = 0.05$	52.50%	16.90%	40.0%	**65.60%**
PGD ℓ_∞ (LeNet)	Sinkhorn DRO[1]	f-DRO	ERM	Sinkhorn DRO[2]
$\epsilon_{\text{PGD}_\infty} = 0.01, iter = 20$	91.50%	84.0%	90.80%	**94.30%**
$\epsilon_{\text{PGD}_\infty} = 0.02, iter = 20$	85.10%	62.30%	81.20%	**91.30%**
$\epsilon_{\text{PGD}_\infty} = 0.05, iter = 20$	46.60%	7.90%	34.30%	**66.20%**
PGD ℓ_2 (LeNet)	Sinkhorn DRO[1]	f-DRO	ERM	Sinkhorn DRO[2]
$\epsilon_{\text{PGD}_2} = 0.5, iter = 30$	87.10%	71.39%	85.30%	**92.50%**
$\epsilon_{\text{PGD}_2} = 0.8, iter = 30$	74.20%	46.20%	73.50%	**87.10%**
$\epsilon_{\text{PGD}_2} = 1.2, iter = 30$	55.50%	18.90%	51.40%	**74.00%**
MIM (LeNet)	Sinkhorn DRO[1]	f-DRO	ERM	Sinkhorn DRO[2]
$\epsilon_{\text{MIM}} = 0.01, iter = 30$	**83.20%**	42.50%	20.20%	80.60%
$\epsilon_{\text{MIM}} = 0.02, iter = 30$	**76.60%**	31.90%	15.20%	76.20%
$\epsilon_{\text{MIM}} = 0.05, iter = 30$	52.10%	10.60%	4.80%	**57.99%**

Table 2: Test classification accuracy of LeNet under different adversarial attack methods.

8 Conclusion

In this paper, we investigate generalized Sinkhorn distance-regularized distributionally robust optimization. By deriving a new dual formulation with strong duality guarantee, we show that the resultant Sinkhorn DRO problem has nested stochastic optimization structure, which enables us to design a Nested SGD algorithm with convergence guarantee under mild assumptions. Numerical studies demonstrate that our Sinkhorn DRO formulation is applicable to large-scale problems and can attain stronger robustness against distribution shifts through multiple datasets and tasks.

9 Acknowledgements

Yufeng Yang and Yi Zhou's work is supported by the National Science Foundation under grants DMS-2134223, ECCS-2237830. Zhaosong Lu's work is partially supported by the Office of Naval Research under grant N00014-24-1-2702, the Air Force Office of Scientific Research under grant FA9550-24-1-0343, and the National Science Foundation under grant IIS-2211491.

References

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, 2018.

Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable sinkhorn distances via the nyström method. In *Advances in Neural Information Processing Systems*, 2019.

Genevay Aude, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Adavances in Neural Information Processing Systems*, 2016.

Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Regularization for wasserstein distributionally robust optimization. *ESAIM: Control, Optimisation and Calculus of Variations*, 29:33, 2023.

Khanh Do Ba, Huy L Nguyen, Huy N Nguyen, and Ronitt Rubinfeld. Sublinear time algorithms for earth mover's distance. *Theory of Computing Systems*, 48:428–442, 2010.

Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *The operations research revolution*, pages 1–19. INFORMS, 2015.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0-387-31073-2.

Jose Blanchet, Daniel Kuhn, Jiajin Li, and Bahar Taskesen. Unifying distributionally robust optimization via optimal transport theory. *arXiv preprint arXiv:2308.05414*, 2023.

Lesi Chen, Jing Xu, and Jingzhao Zhang. Bilevel optimization without lower-level strong convexity from the hyper-objective perspective. *preprint*, 2023a.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. *37th Annual Conference on Learning Theory*, 2023b.

Ruidi Chen and Ioannis Ch. Paschalidis. A distributionally robust optimization approach for outlier detection. In *IEEE Conference on Decision and Control (CDC)*, pages 352–357, 2018a.

Ruidi Chen and Ioannis Ch. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19:1–48, 2018b.

Xi Chen, Simai He, Bo Jiang, Christopher Thomas Ryan, and Teng Zhang. The discrete moment problem with nonconvex shape constraints. *Operations Research*, 69:279–296, 2021.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, volume 202, 2023c.

Meysam Cheramin, Jianqiang Cheng, Ruiwei Jiang, and Kai Pan. Computationally efficient approximations for distributionally robust optimization under moment and wasserstein ambiguity. *INFORMS Journal on Computing*, 34:1768–1794, 2022.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268. PMLR, 13–18 Jul 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58:595–612, 2010.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *the Annals of Statistics*, 49:1378–1406, 2020.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1568–1577. PMLR, 10–15 Jul 2018.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48:603–655, 2023.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 2022.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617, 09–11 Apr 2018.

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

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Scarf Herbert. A min-max solution of an inventory problem. in: Studies in the mathematical theory of inventory and production, 1957.

Yifan Hu, Jie Wang, Yao Xie, Andreas Krause, and Daniel Kuhn. Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36:78412–78434, 2023.

Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Optimization Online*, 1:9, 2013.

Feihu Huang. On momentum-based gradient methods for bilevel optimization with nonconvex lower-level. *arXiv preprint arXiv:2303.03944*, 2023.

Jikai Jin, Bohang Zhang, Haiyang Wang, and Liwei Wang. Non-convex distributionally robust optimization: Non-asymptotic analysis. In *Advances in Neural Information Processing Systems*, 2021.

Burak Kocuk. Conic reformulations for kullback-leibler divergence constrained distributionally robust optimization and applications. *arXiv preprint arXiv:2007.05966*, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*, chapter 6, pages 130–166. Informs, 2019.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. *arXiv preprint arXiv:2309.01753*, 2023.

Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67:1090–1105, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Daniel Levy, Yair Carmon, John C. Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems*, 2020.

Jiashuo Liu, Jiayun Wu, Bo Li, and Peng Cui. Distributionally robust optimization with data geometry. In *Advances in Neural Information Processing Systems*, volume 35, pages 33689–33701. Curran Associates, Inc., 2022.

Zhenyuan Liu, Bart PG Van Parys, and Henry Lam. Smoothed f-divergence distributionally robust optimization. *arXiv preprint arXiv:2306.14041*, 2023.

Fengqiao Luo and Sanjay Mehrotra. Decomposition algorithm for distributionally robust optimization using wasserstein metric with an application to a class of regression models. *European Journal of Operational Research*, 278:20–35, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert Gower. Directional smoothness and gradient methods: Convergence and adaptivity. *Advances in Neural Information Processing Systems*, 37:14810–14848, 2024.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in Neural Information Processing Systems*, 29, 2016.

Adam Paszke, Sam Gross, Francesco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

Giorgio Patrini, Rianne Van den Berg, Patrick Forre, Marcello Carioni, Samarth Bhargav, Max Welling, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, pages 733–743, 2020.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *IEEE International Conference on Computer Vision*, pages 460–467, 2009.

Georg Pflug and David Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7:435–442, 2007.

Qi Qi, Jiameng Lyu, Er Wei Bai, Tianbao Yang, et al. Stochastic constrained dro with a complexity independent of sample size. *Transactions on Machine Learning Research*, 2023.

Wei Qian, Bin Hong, Deng Cai, Xiaofei He, Xuelong Li, et al. Non-negative matrix factorization with sinkhorn distance. In *International Joint Conference on Artificial Intelligence*, pages 1960–1966, 2016.

Zi-Hao Qiu, Siqi Guo, Mao Xu, Tuo Zhao, Lijun Zhang, and Tianbao Yang. To cool or not to cool? temperature network meets large foundation models via dro. *The Fortieth-First International Conference on Machine Learning*, 2024.

Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 256–269, 2015.

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

Hitesh Sapkota, Yiming Ying, Feng Chen, and Qi Yu. Distributionally robust optimization for deep kernel multiple instance learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2188–2196. PMLR, 2021.

Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

Alexander Shapiro, Enlu Zhou, and Yifan Lin. Bayesian distributionally robust optimization. *SIAM Journal on Optimization*, 33:1279–1304, 2023.

Evgeny Tenetov, Gershon Wolansky, and Ron Kimmel. Fast entropic regularized optimal transport using semidiscrete cost approximation. *SIAM Journal on Scientific Computing*, 40:A3400–A3422, 2018.

Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for alexey chervonenkis*, pages 11–30. Springer, 2015.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv: 2109.11926*, 2023.

Kaixin Wang, Navdeep Kumar, Kuangqi Zhou, Bryan Hooi, Jiashi Feng, and Shie Mannor. The geometry of robust value functions. In *International Conference on Machine Learning*, pages 22727–22751. PMLR, 2022.

Zifan Wang, Yi Shen, Michael Zavlanos, and Karl H Johansson. Outlier-robust distributionally robust optimization via unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 37:52189–52214, 2024.

Xiyuan Wei, Ming Lin, Fanjiang Ye, Fengguang Song, Liangliang Cao, My T That, and Tianbao Yang. Model steering: Learning with a reference model improves generalization bounds and scaling laws. *The Forty-Second International Conference on Machine Learning*, 2025.

David Wozabal. A framework for optimization under ambiguity. *Annals of Operations Research*, 193:21–47, 2012.

Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Xiang Wang, and Xiangnan He. Understanding contrastive learning via distributionally robust optimization. In *Adavances in Neural Information Processing Systems*, 2023.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

Qi Zhang, Yi Zhou, Ashley Prater-Bennette, Lixin Shen, and Shaofeng Zou. Large-scale non-convex stochastic constrained distributionally robust optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8217–8225, 2024.

Qi Zhang, Yi Zhou, Simon Khan, Ashley Prater-Bennette, Lixin Shen, and Shaofeng Zou. Revisiting large-scale non-convex distributionally robust optimization. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Dixian Zhu, Gang Li, Bokun Wang, Xiaodong Wu, and Tianbao Yang. When AUC meets DRO: Optimizing partial AUC for deep learning with non-convex convergence guarantee. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27548–27573. PMLR, 17–23 Jul 2022.

Qing Zhu, Xian Yu, and Guzin Bayraksan. Residuals-based contextual distributionally robust optimization with decision-dependent uncertainty. *arXiv*, 2024.

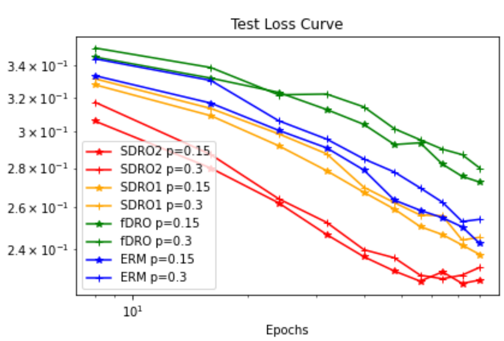
Appendix

Appendix A. Regression over synthetic data

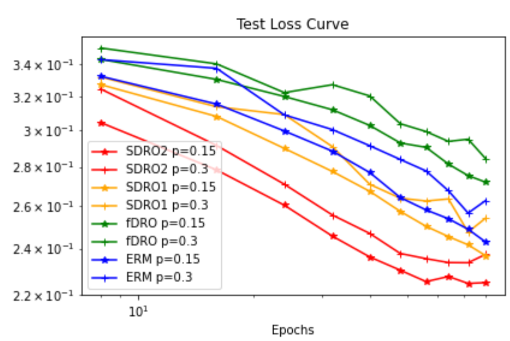
Through this section, we use synthetic training and test data. We generate the input samples with $3k$ measurements and dimension $d = 10$ from a multivariate normal distribution, where the mean vector and covariance matrix are $0.5\mathbf{e}$ and $0.1\mathbf{I}$, respectively. Ground-truth model parameters x^* are sampled from $\mathcal{N}(0, 9e^{-2})$, and the corresponding output data $\zeta_{\text{output}} \in \mathbf{R}^{3k\times 10}$ follows the rule $\zeta_{\text{output}} = \zeta_{\text{train}} \cdot x^* + \epsilon_{\text{noise}}$, where $\epsilon_{\text{noise}} \sim \mathcal{N}(0, 2.5e^{-2})$. For synthetic test data, we generate 500 measurements following the same way as training data, we normalize all the data, apply Gaussian and Laplacian attack over test data to compare their performances. For primal parameters, we initialize them as $x_0 \sim \mathcal{N}(x^* + 5e^{-2}, 1e^{-2})$. For dual variable η used in proposed Sinkhorn DRO dual formulation (3) and f-DRO, we initialize them as $\eta_0 \in \mathbf{R}^{3k} \sim \mathcal{N}(5, 2.25)$, $\eta_0 = 0.8$ respectively.

We fine-tuned the hyper-parameters for all models. The detailed formulation and algorithm settings are as follows. The loss function $\ell(.)$ is set to quadratic loss through all formulations. For proposed Sinkhorn DRO dual formulation (3) and Sinkhorn DRO dual formulation (4) from Wang et al. (2023), we set the reference measure ν as Gaussian measure following $\xi_{\text{train}} \sim \mathcal{N}(\zeta_{\text{train}}, 4e^{-2})$ for every ζ. The cost metric and f^* are set as $c(\zeta, \xi) = ||\zeta - \xi||_2^2$ and $f^*(t) = \frac{1}{4}(t+2)_+^2 - 1$, which corresponds to the dual function of χ^2-divergence. For regularization parameter λ and β used in generalized Sinkhorn distance (see Definition 1), we set them as $\lambda = 0.8, \beta = 1.0$. We trained all formulations using vanilla stochastic gradient descent (SGD) (Ghadimi and Lan, 2013). For f-DRO and ERM, we set the learning rates as $5e^{-4}$, $1e^{-3}$ respectively, and we optimize primal and dual variable of f-DRO in parallel following conclusion drawn from Jin et al. (2021). For Sinkhorn DRO formulation (4) from Wang et al. (2023), we set the learning rate as $1e^{-3}$ and subsample $\xi_{\tilde{B}}$ with $\tilde{B} = 8$ at each iteration. For proposed Sinkhorn DRO formulation (3), we subsample $\xi_{\tilde{B}}$ with $\tilde{B} = 8$, run algorithm 1 with 5 steps to minimize (5) at each iteration, and set the learning rates for algorithm 2 and 1 as $5e^{-2}$, $8e^{-2}$ respectively. For all algorithms (except inner SGD algorithm 1), we set batch size for sub-sampled ζ as 32 and ran SGD for 80 epochs. Figure 1 plots the training and test loss according to recorded checkpoints every 8 epochs.

We evaluate our proposed Sinkhorn DRO dual formulation (3), f-DRO and ERM on regression task over synthetic data. We plot test (quadratic) loss on the test data in log-log scale at Figure 1, where the left shows the loss value obtained from the test data under gaussian attack; the right shows loss value obtained from the test data under Laplacian attack. *SDRO*1 refers to the Sinkhorn DRO dual formulation from Wang et al. (2023) and *SDRO*2 refers to proposed Sinkhorn DRO dual formulation (3). Different marks represent different perturbation magnitudes p. As we can see, our proposed Sinkhorn DRO dual formulation (3) attains comparable performance under distribution shifts with Sinkhorn DRO dual formulation from (Wang et al., 2023). Additionally, we found f-DRO difficult to optimize using SGD with sample-average approximation. This observation is consistent with the findings of Jin et al. (2021); Chen et al. (2023c), where the f-DRO formulation satisfies a generalized smoothness condition and requires advanced optimization algorithms (Zhang et al., 2025; Cutkosky and Mehta, 2020) to ensure convergence.



(a) Gaussian attack

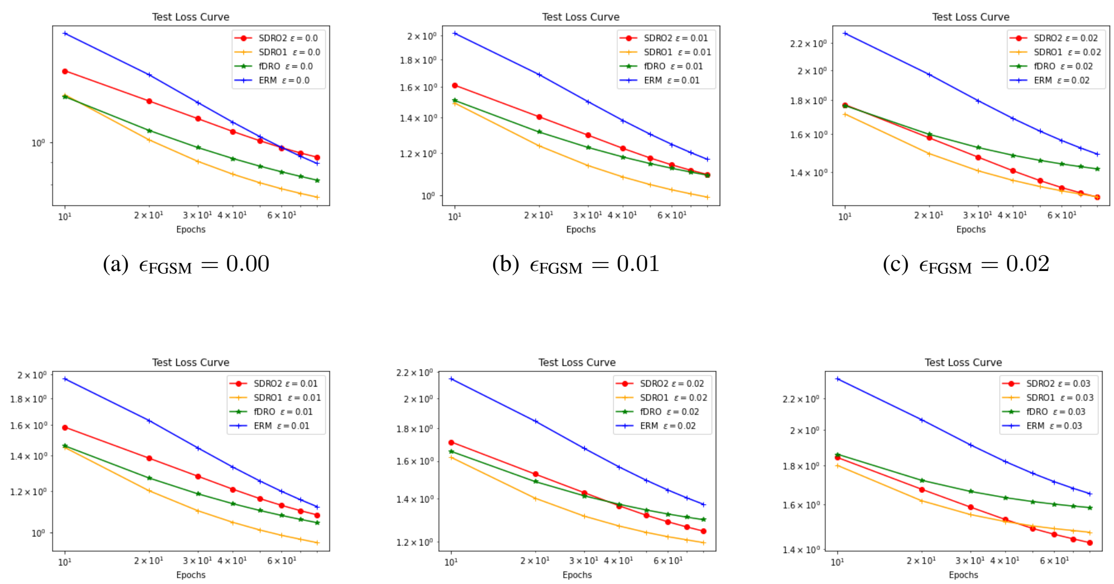


(b) Laplacian attack

Figure 1: Test performance of linear regression under Gaussian and Laplacian attack.

Appendix B. Detailed settings for training logistic regression over compressed CIFAR-10

Through this section, we use CIFAR-10 (Krizhevsky, 2009) as our train and test data. We preprocess the dataset by resizing images, normalizing and utilizing pre-trained ResNet-50 (He et al., 2016) over ImageNet (Deng et al., 2009) to compress each image into a vector with feature dimension $d = 250$. For test data, we subsampled 1500 samples from compressed CIFAR-10 test data, generating adversarial examples utilizing model parameters obtained at last epoch to evaluate test performances through all methods. For proposed Sinkhorn DRO dual formulation (3), we initialize the primal and dual parameters as $x_0 \sim \mathcal{N}(0, 4e^{-2})$, $\eta_0 \in \mathbf{R}^{50k} \sim \mathcal{N}(1, 1e^{-2})$ respectively. For Sinkhorn DRO dual formulation from Wang et al. (2023), f-DRO and ERM, we adopt same initialization for primal parameters and set the dual variable η for f-DRO as $\eta_0 = 1.5$.



(a) $\epsilon_{\text{FGSM}} = 0.00$ (b) $\epsilon_{\text{FGSM}} = 0.01$ (c) $\epsilon_{\text{FGSM}} = 0.02$

(d) $\epsilon_{\text{PGD}} = 0.01$ (e) $\epsilon_{\text{PGD}} = 0.02$ (f) $\epsilon_{\text{PGD}} = 0.03$

Figure 2: Logistic Regression Test Loss under FGSM (top) and PGD (bottom) attack

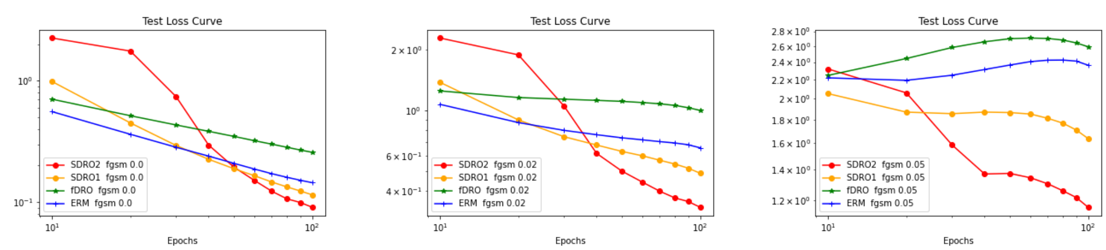
We fine-tuned all hyper-parameters for each baseline methods. The detailed formulation and algorithm setting are as follows. The loss function $\ell(.)$ is set to cross-entropy (CE) loss through all formulations. For proposed Sinkhorn DRO dual formulation (3) and formulation (4) from Wang et al. (2023), we set reference measure ν as Gaussian measure following $\xi_{\text{train}} \sim \mathcal{N}(\zeta_{\text{train}}, 4e^{-2})$ and keep $c(\cdot,\cdot)$, $f^*(\cdot)$ to be ℓ_2-norm and conjugate dual of χ^2-divergence. We trained all formulations using vanilla stochastic gradient descent (SGD) (Ghadimi and Lan, 2013). For f-DRO and ERM, we set their learning rates as $8e^{-2}$, $3e^{-2}$ respectively, and we optimize the primal and dual variable of f-DRO in parallel following conclusion drawn from Jin et al. (2021). For Sinkhorn DRO dual formulation (4) from Wang et al. (2023), we subsample $\xi_{\tilde{B}}$ with $\tilde{B} = 2$ and set learning rate as $8e^{-2}$. For proposed Sinkhorn DRO dual formulation (3), we also subsample $\xi_{\tilde{B}}$ with $\tilde{B} = 2$, run algorithm 1 with 5 steps to minimize inner objective (5) at each iteration and we set the learning rates for algorithm 2 and 1 as $8e^{-2}$,$1e^{-1}$ respectively. For all SGD algorithms (except inner Algorithm

1), we set batch size for sub-sampled ζ as 64, ran algorithms for 80 epochs and plot the test CE loss every 10 epochs. Figure 2 plots the test loss in log-log scale, where the first row represents the CE loss of test data under FGSM attack and the second row plots the CE loss of test data under ℓ_∞-PGD (Madry et al., 2018), where we set ℓ_∞-PGD attack iterations to be 20 and step size $\alpha = \epsilon_{\text{PGD}}/4$ through all perturbation magnitudes.

Combined test accuracy reported in Table 2, although every model's accuracy is affected when varying perturbation, we found our proposed Sinkhorn DRO dual formulation (3) achieves highest test classification accuracies across most scenarios, and exhibits smallest accuracy drop the attack strength increases, which demonstrates the effectiveness of proposed Sinkhorn formulation (3) and Nested-SGD algorithm (Algorithm 2).

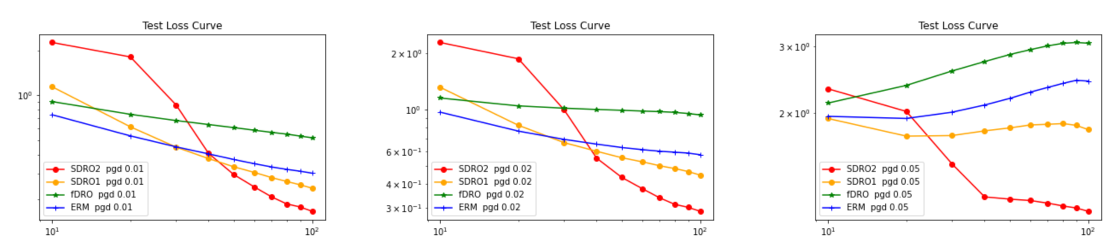
Appendix C. Detailed settings for training LeNet over MNIST data

Through this section, we use MNIST (Deng, 2012) as our train and test data. We preprocess them by resizing images into 32×32, normalizing them with mean and standard derivation, all equal to 0.5. For the test data, we randomly subsampled 1000 samples from MNIST test data, and generate adversarial test samples utilizing model parameters obtained at the last epoch. For proposed Sinkhorn DRO dual formulation (3) and other baseline methods, we initialize the primal parameters using kaiming initialization (He et al., 2015). For the dual parameters utilized in proposed Sinkhorn DRO dual formulation (3) and f-DRO, we initialize them as $\eta_0 \in \mathbf{R}^{60k} \sim \mathcal{N}(0.5, 1e^{-2})$ and $\eta_0 = 1.0$ respectively.



(a) $\epsilon_{\text{FGSM}} = 0.00$ (b) $\epsilon_{\text{FGSM}} = 0.02$ (c) $\epsilon_{\text{FGSM}} = 0.05$

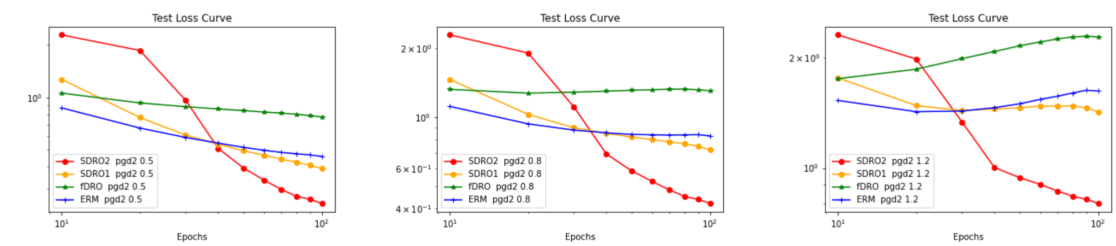
Figure 3: Test loss curves of LeNet under FGSM attacks with different perturbation levels.



(a) $\epsilon_{\text{PGD}} = 0.01$ (b) $\epsilon_{\text{PGD}} = 0.02$ (c) $\epsilon_{\text{PGD}} = 0.05$

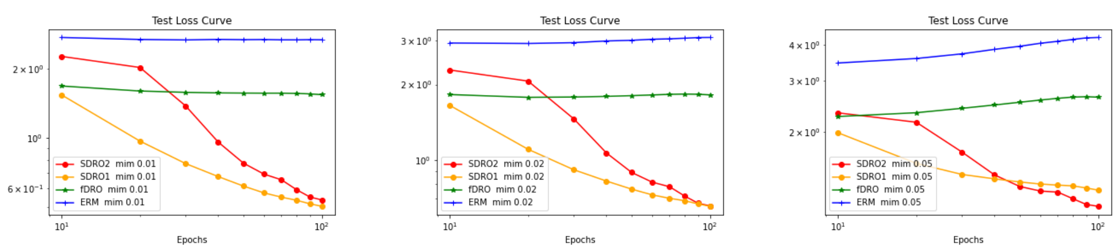
Figure 4: Test loss curves of LeNet under ℓ_∞-PGD attacks with different perturbation levels.

We fine-tuned all hyper-parameters for each method. The detailed formulation and algorithm settings are as follows. The loss function $\ell(.)$ is set to cross-entropy (CE) loss through all formulations. For proposed Sinkhorn DRO dual formulation (3) and formulation (4) from Wang et al. (2023), we set their reference measure ν as Gaussian measure following $\xi_{\text{train}} \sim \mathcal{N}(\zeta_{\text{train}}, 2.25e^{-2})$ and keep $c(\cdot,\cdot)$, $f^*(\cdot)$ to be ℓ_2-norm, conjugate dual of χ^2-divergence. For regularization parameter λ, β used in objective formulation and Sinkhorn distance, we set them as $\lambda = 0.5$, $\beta = 0.8$ respectively. We trained all formulations using vanilla stochastic gradient descent (SGD) (Ghadimi and Lan, 2013). For f-DRO and ERM, we set their learning rate to be $1e^{-3}$, optimize primal and dual variables of f in parallel according to conclusion drawn from Jin et al. (2021). For Sinkhorn DRO dual formulation (4) from Wang et al. (2023), we subsample $\xi_{\tilde{B}}$ with $\tilde{B} = 5$ and set learning rate as $1e^{-3}$. For proposed Sinkhorn DRO dual formulation (3), we subsample $\xi_{\tilde{B}}$ with $\tilde{B} = 4$, run algorithm 1 with 20 steps for minimizing inner objective (5), and set the learning rates for algorithm 2 and 1 as $5e^{-3}$ and$1e^{-1}$ respectively. We set the batch size for sub-sampled ζ as 128 and ran SGD for 100 epochs and record the loss every 10 epochs.



(a) $\epsilon_{\text{PGD2}} = 0.5$ (b) $\epsilon_{\text{PGD2}} = 0.8$ (c) $\epsilon_{\text{PGD2}} = 1.2$

Figure 5: Test loss curves of LeNet under ℓ_2-PGD attacks with different perturbation levels.



(a) $\epsilon_{\text{MIM}} = 0.01$ (b) $\epsilon_{\text{MIM}} = 0.02$ (c) $\epsilon_{\text{MIM}} = 0.05$

Figure 6: Test loss curves of LeNet under MIM attacks with different perturbation levels.

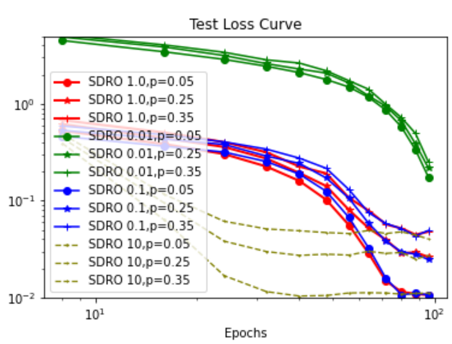
Figure 3, 4, 5 and 6 plot the test loss in log-log scale, where the first to last row represents test CE-loss under FGSM, ℓ_∞, ℓ_2-PGD attack (Madry et al., 2018) and MIM attack (Dong et al., 2018) respectively. For ℓ_∞, ℓ_2-PGD attack, we set their learning rates as $\alpha_{\text{PGD}_\infty} = \epsilon_{\text{PGD}_\infty}/10, \alpha_{\text{PGD}_2} = \epsilon_{\text{PGD}_2}/10$; For MIM attack, we set the moving-average parameter 1.0 and its learning rate $\alpha_{\text{mim}} = \epsilon_{\text{mim}}/15$. From above results, we conclude Sinkhorn DRO dual formulation in general is more robust against distribution shifts. Compared with Sinkhorn DRO dual formulation (4), our method attains smaller accuracy drop and better classification accuracy in most scenarios, which demonstrates the effectiveness of proposed Sinkhorn DRO dual formulation (3) and its convergence analysis.

Appendix D. Ablation Study

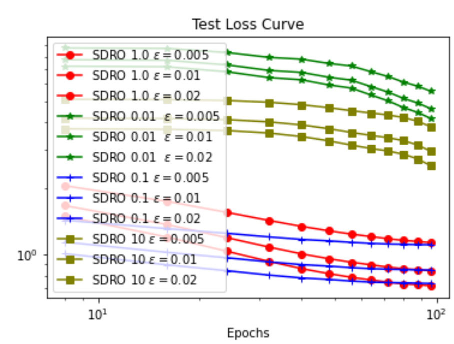
In this section, we conduct ablation studies for our proposed Sinkhorn DRO dual formulation (3) over linear and logistic regression, where we focus on components of (3) having potential effects of model robustness performance, including regularization parameter λ, the cost metric $c(\cdot,\cdot)$ and the choices of information divergence conjugate dual $f^*(\cdot)$. For linear regression, we slightly modify the training data generation and model initialization procedures as follows. The optimal model parameters x^* are sampled from $\mathcal{N}(0, 2.25e^{-2})$, and the corresponding output training data are generated according to $\zeta_{\text{output}} = \zeta_{\text{train}} \cdot x^* + \epsilon_{\text{noise}}$, where $\epsilon_{\text{noise}} \sim \mathcal{N}(0, 1e^{-2})$. In addition to normalizing the training and test data as described in Appendix A, we also normalize the initial parameters after sampling them from $x_0 \sim \mathcal{N}(x^*, 1)$. We apply gaussian attack on subsampled test dataset to evaluate the robust performance of linear regression trained by proposed Sinkhorn DRO dual formulation (3). For logistic regression, we adopt the same setup as described in Appendix B, except that we increase the iteration number of Algorithm 1 from 5 to 8. And we vary FGSM attack (Goodfellow et al., 2015) strength ϵ_{fgsm} over subsampled test data to evaluate the robust performance of logistic regression trained via proposed Sinkhorn DRO dual formulation (3).

D.0.1 Effects of regularization λ

In this section, we vary λ over the set $\{0.01, 0.1, 1.0, 10\}$. We fine-tune the learning rates for algorithm 2 and algorithm 1 to be $1e^{-2}$ and $1e^{-1}$, respectively, for all models. Figure 7 (left) plots the test (quadratic) loss for linear regression under different gaussian attack levels, where the numbers in the legend indicate the corresponding value of λ. We observe that using a smaller regularization parameter improves the model's robustness against distributional shifts. In particular, when $\lambda = 0.01$, the green curves exhibit smaller shifts relative to others, which aligns with classical insights on the effect of regularization (Bishop, 2006). However, from optimization perspective, a small λ makes the proposed dual formulation (3) more difficult to train with promising accuracy guarantee. We also conduct the same experiments for logistic regression, where the learning rates of



(a) linear regression

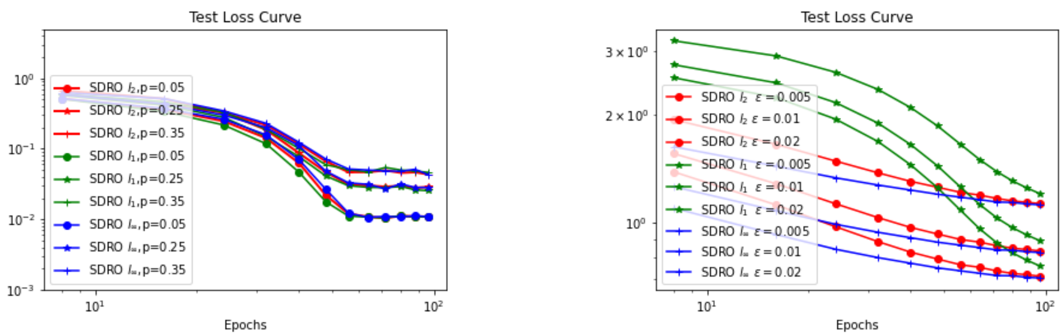


(b) logistic regression

Figure 7: Effects of Regularization λ

algorithm 2 and algorithm 1 are fine-tuned to $8e^{-2}$ and $1e^{-1}$, respectively. Figure 7 (right) plots the test loss curves for logistic regression under different FGSM attack levels. We observe that when $\lambda = 0.01$ or 10, the proposed dual formulation (3) fails to train a valid model, with the resulting classification accuracies dropping below 10%. In contrast, for $\lambda = 0.1$ and 1.0, the test losses converge to similar scales across different attack levels. These results suggest that a proper choice of the regularization parameter lies in the range around $[0.1, 1.0]$.

D.0.2 EFFECTS OF COST METRIC $c(\cdot, \cdot)$



(a) linear regression

(b) logistic regression

Figure 8: Effects of cost metric $c(\cdot, \cdot)$

In this section, we vary the cost metric $c(\cdot, \cdot)$ among the $\{\ell_1, \ell_2, \ell_\infty\}$-norms to examine the impact of cost metric choices. Specifically, for the ℓ_2-norm, we set $c(\zeta, \xi) = \|\zeta - \xi\|_2$; for the ℓ_1-norm, we set $c(\zeta, \xi) = 0.2 \cdot |\zeta - \xi|_1$; and for the $\ell\infty$-norm, we set $c(\zeta, \xi) = 2 \cdot \|\zeta - \xi\|_\infty$. To ensure a fair comparison, we fix the regularization parameter at $\lambda = 0.8$ and retain the same learning rates for Algorithm 2 and Algorithm 1 as used in Section D.0.1. Figure 8 (left) plots the test loss for linear regression trained with different cost metrics under the proposed dual formulation (3). As shown in the figure, the choice of cost metric has a marginal effect on linear regression, as the test curves across different norms are nearly indistinguishable. However, for logistic regression, we observe that using the ℓ_1-norm makes the proposed dual formulation (3) less effective in learning a robust model compared to the ℓ_2 and ℓ_∞ norms. This suggests that ℓ_2 and ℓ_∞ norms are more reliable choices for $c(\cdot, \cdot)$ in practice.

D.0.3 EFFECTS OF CHOICES OF f^*

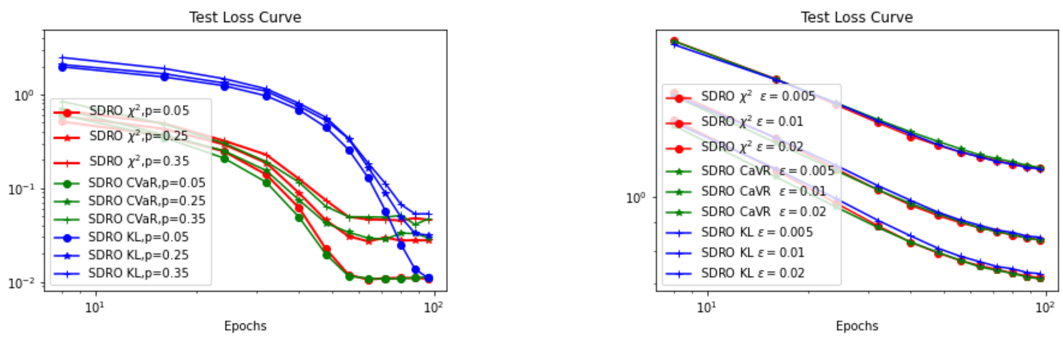
In this section, we test the effects of conjugate functions $f^*(\cdot)$. We select three classical divergence measures, including χ^2-divergence, KL-divergence and smoothed CVaR divergence (Jin et al., 2021). To elaborate, we list the primal and conjugate dual expressions in following Table 3.

For the linear regression task, we vary the conjugate functions corresponding to different information divergences, while keeping the same learning rates for Algorithm 2 and Algorithm 1 as in Section D.0.1, and fix the regularization parameter at $\lambda = 0.8$. Specifically, for conjugate dual of smoothed-CVaR, we set $\alpha = 0.5$. Figure 9 (left) plots the test loss for linear regression trained using the proposed dual formulation (3). We observe that using the conjugate dual of KL-divergence hinders fast convergence compared with other f^* satisfying M-smoothness property. However, the

Table 3: Primal and dual expressions for different divergence measures

Divergence	$f(t)$	$f^*(t)$
χ^2	$\frac{1}{2}(t-1)^2$	$-1+\frac{1}{4}(t+2)_+^2$
KL	$t\log t - t + 1$	$\exp(t)-1$
smoothed CVaR	$f_\alpha^{\text{smo}}(t) = \begin{cases} t\log t + \frac{1-\alpha t}{\alpha}\log\frac{1-\alpha t}{1-\alpha}, & t\in[0,1/\alpha) \\ +\infty, & \textit{otherwise} \end{cases}$	$\frac{1}{\alpha}\log(1-\alpha+\alpha\exp(t))$

test accuracy evaluated at last epoch reachs same level under the existence of gaussian attack. For



(a) linear regression (b) logistic regression

Figure 9: Effects of conjugate function f^*

logistic regression, we observe that the robust performance achieved using the conjugate dual of the KL-divergence is comparable to that obtained with a conjugate dual f^* satisfying the M-smoothness property. The performance gap across different models may be attributed to the sensitivity of the KL-divergence to model initialization, as the local M-smooth constant can vary depending on the starting point. Nevertheless, empirical results on both linear and logistic regression suggest that enforcing the M-smoothness assumption does not hinder practical applicability. As long as the model is not initialized in an ill-conditioned region, choosing f^* as the conjugate dual of the KL-divergence yields similar convergence, supporting the practical validity of the M-smoothness assumption.

Appendix E. Proof of Theorem 1

Theorem 1 (Dual formulation) *The DRO problem (2) has the following equivalent dual formulation*

$$\min_{x\in\mathbf{R^d}} \mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x)], \textit{ where } \Psi_\zeta(x) = \min_{\eta\in\mathbf{R}} \mathbb{E}_{\xi\sim\nu}\underbrace{\left[\lambda\beta f^*\left(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta}{\lambda\beta}\right)+\eta\right]}_{\mathcal{L}_{\xi,\zeta}(x,\eta)}, \tag{3}$$

and f^* denotes the conjugate function of f.

Proof Merging the infimum operator $\inf_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})}$ with the supremum operator $\sup_{\mathbb{Q}}$ in (2), we obtain the following equivalent form.

$$\min_{x\in\mathbf{R}^d}\sup_{\mathbb{Q},\gamma\in\Gamma(\mathbb{P},\mathbb{Q})}\left\{\mathbb{E}_{\xi\sim\mathbb{Q}}\Big[\ell(x;\xi)\Big]-\Big[\mathbb{E}_{(\zeta,\xi)\sim\gamma}\big[\lambda c(\zeta,\xi)\big]+\lambda\beta D_f(\gamma|\mathbb{P}\otimes\nu)\Big]\right\}. \tag{18}$$

Regarding the joint distribution $\gamma(\zeta,\xi)$, we decompose it as $\gamma(\zeta,\xi)=\gamma(\xi|\zeta)\mathbb{P}(\zeta)$, where $\mathbb{P}(\zeta)$ denotes the marginal distribution and $\gamma(\xi|\zeta)$ corresponds to the conditional distribution given ζ. Then, the constraint $\gamma\in\Gamma(\mathbb{P},\mathbb{Q})$ is equivalent to $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\gamma(\xi|\zeta)\big]=\mathbb{Q}$, and hence (18) reduces to

$$\min_{x\in\mathbf{R}^d}\sup_{\mathbb{Q}\text{ s.t. }\mathbb{E}_{\zeta\sim\mathbb{P}}\left[\gamma(\xi|\zeta)\right]=\mathbb{Q}}\left\{\mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\mathbb{E}_{\xi\sim\gamma(\cdot|\zeta)}\big[\ell(x;\xi)-\lambda c(\zeta,\xi)\big]-\lambda\beta D_f\big(\gamma(\xi|\zeta)|\nu(\xi)\big)\Big]\right\}. \tag{19}$$

We claim that the optimal value of (19) equals that of the following problem, which takes supremum over all possible conditional distributions $\gamma(\xi|\zeta)$.

$$\min_{x\in\mathbf{R}^d}\sup_{\gamma(\xi|\zeta)}\left\{\mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\mathbb{E}_{\xi\sim\gamma(\cdot|\zeta)}\big[\ell(x;\xi)-\lambda c(\zeta,\xi)\big]-\lambda\beta D_f\big(\gamma(\xi|\zeta)|\nu(\xi)\big)\Big]\right\}. \tag{20}$$

To show this, for any fixed x, suppose the supremum of (19) is achieved by a certain conditional distribution $\gamma(\xi|\zeta)$ that satisfies $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\gamma(\xi|\zeta)\big]=\mathbb{Q}$, and such $\gamma(\xi|\zeta)$ is feasible for the supremum of (20). Thus, the supremum of (19) is lower than the supremum of (20). On the other hand, for any fixed x, suppose the supremum of (20) is achieved by a certain conditional distribution $\gamma(\xi|\zeta)$. Then, the distribution $\mathbb{Q}$ given by $\mathbb{Q}=\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\gamma(\xi|\zeta)\big]$ is feasible for the supremum of (19). Consequently, the supremum of (19) is higher than that of (20). In summary, (19) and (20) are equivalent.

Furthermore, by principle of interchangeability (Theorem 7.92, Chapter 7.3.2 from Shapiro et al. (2021)), (20) can be rewritten as

$$\min_{x\in\mathbf{R}^d}\mathbb{E}_{\zeta\sim\mathbb{P}}\Bigg[\underbrace{\sup_{\gamma(\xi|\zeta)}\left\{\mathbb{E}_{\xi\sim\gamma(\cdot|\zeta)}\big[\ell(x;\xi)-\lambda c(\zeta,\xi)\big]-\lambda\beta D_f\big(\gamma(\xi|\zeta)|\nu(\xi)\big)\right\}}_{\Psi_\zeta(x)}\Bigg]. \tag{21}$$

Next, for every fixed ζ and x, denote $\mu_{\gamma|\zeta}$ and μ_ν as the distributions of the scalar random variable $\ell(x;\xi)-\lambda c(\zeta,\xi)$ under $\gamma(\xi|\zeta)$ and $\nu(\xi)$, respectively. we show that the $\Psi_\zeta(x)$ defined above is equivalent to the following auxiliary function

$$\widetilde{\Psi}_\zeta(x)=\sup_{\mu_{\gamma|\zeta}}\left\{\mathbb{E}_{\mu_{\gamma|\zeta}}\big[\ell(x;\xi)-\lambda c(\zeta,\xi)\big]-\lambda\beta D_f(\mu_{\gamma|\zeta}|\mu_\nu)\right\}, \tag{22}$$

where $\sup_{\mu_{\gamma|\zeta}}$ corresponds to the supremum over all possible distributions $\mu_{\gamma|\zeta}$ induced by $\gamma(\xi|\zeta)$. To show this, for any fixed x, suppose the supremum in $\Psi_\zeta(x)$ is achieved by a certain $\gamma(\xi|\zeta)$, and denote the induced distribution of $\ell(x;\xi)-\lambda c(\zeta,\xi)$ as $\mu_{\gamma|\zeta}$. It is straightforward to show that

$$\mathbb{E}_{\xi\sim\gamma(.|\zeta)}\big[\ell(x;\xi)-\lambda c(\zeta,\xi)\big]=\mathbb{E}_{\mu_{\gamma|\zeta}}\big[\ell(x;\xi)-\lambda c(\zeta,\xi)\big].$$

Moreover, by the data processing inequality, it holds that $D_f\big(\mu_{\gamma|\zeta}|\mu_\nu\big) \le D_f\big(\gamma(\xi|\zeta)|\nu(\xi)\big)$. Since $\mu_{\gamma|\zeta}$ is feasible for the supremum of (22), we conclude that $\Psi_\zeta(x) \le \widetilde{\Psi}_\zeta(x)$. Conversely, for any fixed x, suppose the supremum in $\widetilde{\Psi}_\zeta(x)$ is achieved by a certain $\mu_{\gamma|\zeta}$, then the corresponding $\gamma(\zeta|\xi)$ (which induces $\mu_{\gamma|\zeta}$) is feasible for the supremum in $\Psi_\zeta(x)$, and hence we have that $\widetilde{\Psi}_\zeta(x) \le \Psi_\zeta(x)$. Finally, we conclude that $\Psi_\zeta(x) = \widetilde{\Psi}_\zeta(x)$.

Using inverse c.d.f. sampling based on the cumulative distribution function over μ_ν, the f-divergence between $\mu_{\gamma|\zeta}$ and μ_ν can be rewritten as

$$D_f(\mu_{\gamma|\zeta}|\mu_\nu) = \int f\Big(\frac{\mathrm{d}\mu_{\gamma|\zeta}}{\mathrm{d}\mu_\nu}\Big)\mathrm{d}\mu_\nu = \int f\Big(\frac{\mathrm{d}\mu'}{\mathrm{d}\mathit{Unif}([0,1])}\Big)\mathrm{d}\mathit{Unif}([0,1]) = D_f\big(\mu' \mid \mathit{Unif}([0,1])\big),$$

where $\mathit{Unif}([0,1])$ represents the uniform distribution over $[0,1]$, $\mu' = \mu_\nu^{-1}\circ\mu_{\gamma|\zeta}$. Moreover, for fixed x and ζ, denote the cumulative distribution function of the scalar random variable $\ell(x;\xi) - \lambda c(\zeta,\xi)$ as $F(t) = \mathbb{P}\big(\ell(x;\xi) - \lambda c(\zeta,\xi) \le t\big)$. We can further transform $\Psi_\zeta(x)$ into

$$\Psi_\zeta(x) = \sup_{\mu'}\left\{\mathbb{E}_{F^{-1}(u)\sim\mu'}\big[F^{-1}(u)\big] - \lambda\beta D_f\big(\mu'|\mathit{Unif}([0,1])\big)\right\} \tag{23}$$

$$= \sup_{\mu'}\int F^{-1}(u)\mathrm{d}\mu'(u) - \lambda\beta\int f\Big(\frac{\mathrm{d}\mu'(u)}{\mathrm{d}\mathit{Unif}([0,1])(u)}\Big)\mathrm{d}\mathit{Unif}([0,1])(u) \tag{24}$$

$$= \sup_{\mu'}\int \Big(F^{-1}(u)\frac{\mathrm{d}\mu'(u)}{\mathrm{d}\mathit{Unif}([0,1])(u)} - \lambda\beta f\Big(\frac{\mathrm{d}\mu'(u)}{\mathrm{d}\mathit{Unif}([0,1])(u)}\Big)\Big)\mathrm{d}\mathit{Unif}([0,1])(u) \tag{25}$$

$$= \sup_{r\in\mathcal{R}}\int_0^1 \big[r(u)F^{-1}(u) - \lambda\beta f(r(u))\big]\mathrm{d}u, \tag{26}$$

where $r(u) = \frac{\mathrm{d}\mu'}{\mathrm{d}\mathit{Unif}}(u)$ and $\mathcal{R} = \{r : [0,1] \to \mathbf{R}_+ \mid \int_0^1 r(u)\mathrm{d}u = 1\}$. Introduce a dual variable η for the constraint $\mathcal{R}$, the Lagrange dual formulation of (26) can be written as

$$\Psi_\zeta(x) = \min_{\eta\in\mathbf{R}}\mathcal{L}_\zeta(x,\eta), \text{ where} \tag{27}$$

$$\mathcal{L}_\zeta(x,\eta) = \sup_r\int_0^1 \big[r(u)F^{-1}(u) - \eta(r(u)-1) - \lambda\beta f\big(r(u)\big)\big]\mathrm{d}u. \tag{28}$$

Since for fixed η, $\mathcal{L}_\zeta(x,\eta)$ can be denoted as $\sup_r \mathbb{E}_{u\sim \mathit{Unif}[0,1]}\big[\vartheta_\eta(r(u),u)\big]$, where $\vartheta_\eta(r(u),u) = r(u)F^{-1}(u) - \eta(r(u)-1) - \lambda\beta f\big(r(u)\big)$ is a continuous function, by principle of interchangeability stated in Theorem 7.92 (Shapiro et al., 2021), the order of sup and integral can be swapped, which yields that

$$\Psi_\zeta(x) = \min_{\eta\in\mathbf{R}}\underbrace{\int_0^1 \sup_{r\in\mathbf{R}_+}\big[rF^{-1}(u) - \eta(r-1) - \lambda\beta f(r)\big]\mathrm{d}u}_{\mathcal{L}_\zeta(x,\eta)}. \tag{29}$$

Define the conjugate function $f^*(v) = \sup_{r\in\mathbf{R}_+}\big\{vr - f(r)\big\}$, we further obtain that

$$\Psi_\zeta(x) = \min_{\eta\in\mathbf{R}}\int_0^1 \lambda\beta f^*\Big(\frac{F^{-1}(u)-\eta}{\lambda\beta}\Big)\mathrm{d}u + \eta = \min_{\eta\in\mathbf{R}}\mathbb{E}_{\xi\sim\nu}\Big[\lambda\beta f^*\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta}{\lambda\beta}\Big)\Big] + \eta. \tag{30}$$

Finally, the dual formulation of (21) is expressed as

$$\min_{x\in\mathbf{R^d}} \mathbb{E}_{\zeta\sim\mathbb{P}}\Bigg[\underbrace{\min_{\eta\in\mathbf{R}} \mathbb{E}_{\xi\sim\nu}\Big[\lambda\beta f^*\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta}{\lambda\beta}\Big)\Big]+\eta}_{\Psi_\zeta(x)}\Bigg],$$

which gives the desired result. ■

Appendix F. Proof of Lemma 1

Lemma 1 (Computation of $\nabla\Psi_\zeta(x)$ (Jin et al., 2021)) *Let Assumption 1 hold and consider fixed x and given ζ. Then, the function $\Psi_\zeta(x)$ is differentiable and satisfies $\nabla\Psi_\zeta(x)=\nabla_1\mathcal{L}_\zeta(x,\eta_x^*(\zeta))$, where $\eta_x^*(\zeta)\in\arg\min_\eta \mathcal{L}_\zeta(x,\eta)$.*

Proof We refer to Lemma 2.6 in Jin et al. (2021) for the detailed proof. ■

Appendix G. Proof of Theorem 2

Theorem 2 (Gradient approximation error bound) *Consider a stochastic algorithm minimizing (5). If the stochastic oracle outputs an $\eta_x^{\tilde d}(\zeta)$ converging to $\nabla_2\mathcal{L}_\zeta(x_t,\eta_x^*(\zeta))$ with scaled small target error $\tilde\varepsilon=\varepsilon/G$, i.e.,*

$$\mathbb{E}_{\eta_x^{\tilde d}(\zeta)}\big|\nabla_2\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))\big|^2\le\tilde\varepsilon^2, \tag{6}$$

then the gradient $\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))$ approximates full gradient $\nabla\Psi(x)$ with error up to ε, i.e.,

$$\big\|\nabla\Psi_\zeta(x)-\mathbb{E}_{\eta_x^{\tilde d}(\zeta)}[\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))]\big\|^2\le\varepsilon^2,\forall\zeta\sim\mathbb{P}. \tag{7}$$

Proof First, taking square root and applying Jensen's inequality on both sides, (6) implies

$$\mathbb{E}_{\eta_x^{\tilde d}(\zeta)}|\nabla_2\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))|\le\tilde\varepsilon. \tag{31}$$

Since at $\eta_x^*(\zeta)$, by optimality condition, $\nabla_2\mathcal{L}_\zeta(x,\eta_x^*(\zeta))=0$, we have

$$\begin{aligned}
&\mathbb{E}_{\eta_x^{\tilde d}(\zeta)}\big|\nabla_2\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))\big|\\
=&\mathbb{E}_{\eta_x^{\tilde d}(\zeta)}\big|\nabla_2\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))-\nabla_2\mathcal{L}_\zeta(x,\eta_x^*(\zeta))\big|\\
=&\mathbb{E}_{\eta_x^{\tilde d}(\zeta)}\Big|\mathbb{E}_{\xi\sim\nu}\Big[(f^*)'\Big(\frac{\ell(x;\xi)-c(\zeta,\xi)-\eta_x^{\tilde d}(\zeta)}{\lambda\beta}\Big)\Big]-\mathbb{E}_{\xi\sim\nu}\Big[(f^*)'\Big(\frac{\ell(x;\xi)-c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\Big]\Big|\\
\le&\tilde\varepsilon.
\end{aligned} \tag{32}$$

Applying expression (32), for given x and every ζ, the approximation error between $\nabla\Psi(x)$ and $\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}[\nabla_1\mathcal{L}(x,\eta_x^{\tilde{d}}(\zeta))]$ can be upper bounded as follows

$$
\begin{aligned}
&\left\|\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}[\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))]-\nabla\Psi_\zeta(x)\right\|\\
=&\left\|\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}[\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))-\nabla\Psi_\zeta(x)]\right\|\\
\overset{(i)}{\leq}&\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}\left\|\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))-\nabla\Psi_\zeta(x)\right\|\\
=&\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}\left\|\mathbb{E}_{\xi\sim\nu}\left[(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)\nabla\ell(x;\xi)-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\nabla\ell(x;\xi)\right]\right\|\\
=&\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}\left\|\mathbb{E}_{\xi\sim\nu}\left[(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\right]\cdot\nabla\ell(x;\xi)\right\|\\
\overset{(ii)}{\leq}&\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta),\xi\sim\nu}\left\|\left[(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\right]\cdot\nabla\ell(x;\xi)\right\|\\
\overset{(iii)}{=}&\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta),\xi\sim\nu}\left|\left[(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\right]\right|\cdot\left\|\nabla\ell(x;\xi)\right\|\\
\overset{(iv)}{\leq}&G\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta),\xi\sim\nu}\left\|(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\right\|\\
\overset{(v)}{=}&G\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}\left\|\mathbb{E}_{\xi\sim\nu}\left[(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\right]\right\|\\
=&G\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}\left\|\nabla_2\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))-\nabla_2\mathcal{L}_\zeta(x,\eta_x^*(\zeta))\right\|\\
\overset{(vi)}{\leq}&G\tilde{\varepsilon}=\varepsilon,
\end{aligned}
\tag{33}
$$

where (i) applies Jensen's inequality to extract out expectation over $\eta_x^{\tilde{d}}(\zeta)$; (ii) applies Jensen's inequality to extract out expectation over ξ; (iii) extracts scalar $(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)$ out; (iv) applies G-Lipschitz assumption of $\ell(x;\xi)$ stated at Assumption 1; (v) applies monotonicity property of $(f^*)'$ with regard to η given x and fixed ζ (the sign of $(f^*)'(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta})-(f^*)'(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta})$ is fixed regardless of ξ, which enables to move the expectation over ξ into the norm without changing its value); (vi) applies condition (32). Squaring both sides and re-arranging LHS, we conclude that

$$\left\|\nabla\Psi_\zeta(x)-\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}[\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))]\right\|^2\leq\varepsilon^2,\forall\zeta\sim\mathbb{P},\tag{34}$$

which gives the desired result. ■

Appendix H. Proof of Lemma 2

Lemma 2 (**K'-smoothness of inner objective** (5)) *Let Assumption 1 hold and denote* $\mathcal{L}_{\zeta,\xi}(x,\eta)=\lambda\beta\mathbb{E}_{\xi\sim\nu}\left[f^*(\frac{\ell(x;\xi)-c(\zeta,\xi)-\eta}{\lambda\beta})\right]+\eta$. *Then, for any* η *and* η', *we have*

$$\mathbb{E}_{\xi\sim\nu}\left\|\nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta)-\nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta')\right\|^2\leq(K')^2\left\|\eta-\eta'\right\|^2,$$

where $K' = M(\lambda\beta)^{-1}$.

Proof Notice that the gradient of $\mathcal{L}_{\zeta,\xi}(x,\eta)$ takes the form

$$\nabla_2 \mathcal{L}_{\zeta,\xi}(x,\eta) = 1 - \mathbb{E}_{\xi\sim\nu}\Big[(f^*)'\Big(\frac{\ell(x;\xi) - c(\zeta,\xi) - \eta}{\lambda\beta}\Big)\Big].$$

Then, we obtain that

$$\begin{aligned}
&\mathbb{E}_{\xi\sim\nu}\big\|\nabla_2 \mathcal{L}_{\zeta,\xi}(x,\eta) - \nabla_2 \mathcal{L}_{\zeta,\xi}(x,\eta')\big\|^2 \\
=&\mathbb{E}_{\xi\sim\nu}\big\|(f^*)'\Big(\frac{\ell(x;\xi) - c(\zeta,\xi) - \eta}{\lambda\beta}\Big) - (f^*)'\Big(\frac{\ell(x;\xi) - c(\zeta,\xi) - \eta'}{\lambda\beta}\Big)\big\|^2 \\
\le&(M(\lambda\beta)^{-1})^2\big\|\eta - \eta'\big\|^2,
\end{aligned}$$

where the inequality is due to the M-smoothness assumption of f^* stated in assumption 1. ■

Appendix I. Proof of Lemma 3

Through this work, we utilize the following proposition for variance computing.

Proposition 1 (Variance computing) *Given two i.i.d. random variables X_1 and X_2, the variance can be calculated as*

$$2Var(X_1) = 2Var(X_2) = \mathbb{E}\big\|X_1 - X_2\big\|^2. \tag{35}$$

Proof The proof simply extends from variance definition,

$$\begin{aligned}
2Var(X) =&2\mathbb{E}\big\|X_1 - \bar{X}\big\|^2 \\
=&2\mathbb{E}\|X_1 - \bar{X}\|^2 + 2\mathbb{E}(X_1 - \bar{X})(X_2 - \bar{X}) \\
=&\mathbb{E}\|X_1 - \bar{X}\|^2 + 2\mathbb{E}(X_1 - \bar{X})(X_2 - \bar{X}) + \mathbb{E}\|X_2 - \bar{X}\|^2 \\
=&\mathbb{E}\|X_1 - \bar{X}\|^2 + 2\mathbb{E}(X_1 - \bar{X})(\bar{X} - X_2) + \mathbb{E}\|\bar{X} - X_2\|^2 \\
=&\mathbb{E}\big\|X_1 - \bar{X} + \bar{X} - X_2\big\|^2 \\
=&\mathbb{E}\big\|X_1 - X_2\big\|^2,
\end{aligned}$$

where $\bar{X}$ denotes the mean and we use the facts that (i) $\mathbb{E}\|X_1 - \bar{X}\|^2 = \mathbb{E}\|X_2 - \bar{X}\|^2 = Var(X)$; (ii) $\mathbb{E}(X_1 - \bar{X})(X_2 - \bar{X}) = 0$ for i.i.d random variables. ■

Lemma 3 (Second moment bound for $\nabla_2 \mathcal{L}_{\zeta,\xi}(x,\eta)$) *Let Assumption 2 hold. Then, $v^{\tilde{B}}(\eta)$ is the unbiased estimator for $\nabla_2 \mathcal{L}_\zeta(x,\eta)$, and the second moment of $\nabla_2 \mathcal{L}_{\zeta,\xi}(x,\eta)$ satisfies*

$$\mathbb{E}_{\xi\sim\nu}\left\|\nabla_2 \mathcal{L}_{\zeta,\xi}(x,\eta)\right\|^2 \le R_2 + \left\|\nabla_2 \mathcal{L}_\zeta(x,\eta)\right\|^2, \tag{11}$$

where $R_2 = 2M^2(\lambda\beta)^{-2}(\sigma^2 + \lambda^2\delta^2)$.

Proof Utilizing proposition 1, we have

$$
\begin{aligned}
&Var_\xi\big(\nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta)\big)\\
=&\frac{1}{2}\mathbb{E}_{\xi_1,\xi_2}\Big[(f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x;\xi_2)-\lambda c(\zeta,\xi_2)-\eta}{\lambda\beta}\Big)\Big]^2\\
\leq&\frac{1}{2}M^2(\lambda\beta)^{-2}\mathbb{E}_{\xi_1,\xi_2}\big\|\ell(x;\xi_1)-\ell(x;\xi_2)-(\lambda c(\zeta,\xi_1)-\lambda c(\zeta,\xi_2))\big\|^2\\
\overset{(i)}{\leq}&\frac{1}{2}M^2(\lambda\beta)^{-2}2\cdot(\mathbb{E}_{\xi_1,\xi_2}\|\ell(x;\xi_1)-\ell(x;\xi_2)\|^2+\lambda^2\mathbb{E}_{\xi_1,\xi_2}\|c(\zeta,\xi_1)-c(\zeta,\xi_2)\|^2)\\
\overset{(ii)}{\leq}&\frac{1}{2}M^2(\lambda\beta)^{-2}2(2Var\big(\ell(x;\xi)\big)+2Var_\xi\big(c(\zeta,\xi)\big))\\
\overset{(iii)}{\leq}&2M^2(\lambda\beta)^{-2}(\sigma^2+\lambda^2\delta^2),
\end{aligned}
$$

where (i) applies the fact that given any vectors a,b, we have $\|a-b\|^2\leq 2\|a\|^2+2\|b\|^2$, (ii) applies proposition 1 and (iii) applies bounded variance assumptions 2 for $\ell(\cdot,\xi)$ and $c(\zeta,\cdot)$. ■

Appendix J. Proof of Corollary 3

Corollary 1 (Convergence of Algorithm 1) *Let Assumptions 1 and 2 hold, denote* $\widehat{\Delta}=\sup_\zeta\{L_\zeta(x_t,\eta^0)$ $-\Psi_\zeta(x_t)\}$. *Apply Algorithm 1 to solve the inner problem (5) with learning rate* $\alpha_d=\min\{\frac{1}{K'},\frac{\tilde{\varepsilon}^2}{R_2K'}\}$ *and batch size* $\tilde{B}=1$*, then Algorithm 1 outputs an* $\eta_x^{\tilde{d}}(\zeta)$ *satisfying*

$$
\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}\big|\nabla_2\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))\big|^2\leq\tilde{\varepsilon}^2. \tag{12}
$$

In particular, it takes $D=\mathcal{O}(\hat{\Delta}K'R_2\tilde{\varepsilon}^{-4})=\mathcal{O}(\hat{\Delta}K'R_2G^4\varepsilon^{-4})$ *number of iterations to obtain an* $\tilde{\varepsilon}$*-stationary point, and the stochastic gradient oracle complexity is* $\mathcal{O}(\widehat{\Delta}K'R_2G^4\varepsilon^{-4})$.

Proof Notice that the objective function is K'-smooth, we have

$$
\begin{aligned}
\mathcal{L}_\zeta(x_t,\eta_{x_t}^{d+1}(\zeta))&\overset{(i)}{\leq}\mathcal{L}_\zeta(x_t,\eta_{x_t}^d(\zeta))-\langle\nabla_2\mathcal{L}(x_t,\eta_{x_t}^d(\zeta)),\eta_{x_t}^d(\zeta)-\eta_{x_t}^{d+1}(\zeta)\rangle+\frac{K'}{2}|\eta_{x_t}^{d+1}(\zeta)-\eta_{x_t}^d(\zeta)|^2\\
&\overset{(ii)}{=}\mathcal{L}_\zeta(x_t,\eta_{x_t}^d(\zeta))-\langle\nabla_2\mathcal{L}(x_t,\eta_{x_t}^d(\zeta)),\alpha_d v^{\tilde{B}}(\eta_{x_t}^d(\zeta))\rangle+\frac{K'\alpha_d^2}{2}\big|v^{\tilde{B}}(\eta_{x_t}^d(\zeta))\big|^2,
\end{aligned}
$$

where (i) applies descent lemma for K'-smooth function; (ii) applies update rule $\eta_{x_t}^{d+1}(\zeta)=\eta_{x_t}^d(\zeta)-\alpha_d v^{\tilde{B}}(\eta_{x_t}^d(\zeta))$. Taking expectation over ξ on both sides, we further obtain

$$
\begin{aligned}
&\mathbb{E}_{\xi_{\tilde{B}}\sim\nu}[\mathcal{L}_\zeta(x_t,\eta_{x_t}^{d+1}(\zeta))]\\
&\leq\mathbb{E}_{\xi_{\tilde{B}}\sim\nu}[\mathcal{L}_\zeta(x_t,\eta_{x_t}^d(\zeta))]-\alpha_d\mathbb{E}_{\xi_{\tilde{B}}\sim\nu}|\nabla_2\mathcal{L}_\zeta(x_t,\eta_{x_t}^d(\zeta))|^2+\frac{K'\alpha_d^2}{2}\mathbb{E}_{\xi_{\tilde{B}}\sim\nu}|v^{\tilde{B}}(\eta_{x_t}^d(\zeta))|^2\\
&\overset{(i)}{\leq}\mathbb{E}_{\xi_{\tilde{B}}\sim\nu}[\mathcal{L}_\zeta(x_t,\eta_{x_t}^d(\zeta))]-\alpha_d\mathbb{E}_{\xi_{\tilde{B}}\sim\nu}|\nabla_2\mathcal{L}_\zeta(x_t,\eta_{x_t}^d(\zeta))|^2+\frac{K'R_2\alpha_d^2}{2\tilde{B}}+\frac{K'\alpha_d^2}{2}|\nabla_2\mathcal{L}_\zeta(x_t,\eta_{x_t}^d(\zeta))|^2\\
&=\mathbb{E}_{\xi_{\tilde{B}}\sim\nu}[\mathcal{L}_\zeta(x_t,\eta_{x_t}^d(\zeta))]-\alpha_d(1-\frac{K'\alpha_d}{2})\mathbb{E}_{\xi_{\tilde{B}}\sim\nu}|\nabla_2\mathcal{L}_\zeta(x_t,\eta_{x_t}^d(\zeta))|^2+\frac{K'\alpha_d^2R_2}{2}
\end{aligned}
$$

$$\overset{(ii)}{\leq} \mathbb{E}_{\xi_{\tilde{B}}\sim\nu}[\mathcal{L}_\zeta(x_t,\eta^d_{x_t}(\zeta))] - \frac{\alpha_d}{2}\mathbb{E}_{\xi_{\tilde{B}}\sim\nu}|\nabla_2\mathcal{L}_\zeta(x_t,\eta^d_{x_t}(\zeta))|^2 + \frac{K'\alpha_d^2R_2}{2} \tag{36}$$

where (i) utilizes second moment upper bound stated in Lemma 3; (ii) utilizes the fact $\alpha_d \leq \frac{1}{K'}$ and $\tilde{B} = 1$. Since $\mathcal{L}_\zeta(x_t,\eta^d_{x_t}(\zeta))$ does not include randomness from ξ, re-organizing above inequality gives

$$\frac{\alpha_d}{2}|\nabla_2\mathcal{L}_\zeta(x_t,\eta^d_{x_t}(\zeta))|^2 \leq \mathbb{E}_{\xi\sim\nu}[\mathcal{L}_\zeta(x_t,\eta^d_{x_t}(\zeta)) - \mathcal{L}_\zeta(x_{t+1},\eta^d_{x_t}(\zeta))] + \frac{K'\alpha_d^2R_2}{2} \tag{37}$$

Summing above equation through $d \in \{0....D-1\}$, we have

$$\frac{\alpha_d}{2D}\sum_{d=0}^{D-1}|\nabla_2\mathcal{L}_\zeta(x_t,\eta^d_{x_t}(\zeta))|^2 \leq \frac{\hat{\Delta}}{D} + \frac{K'\alpha_d^2R_2}{2} \tag{38}$$

To make sure the RHS smaller than or equal to $\tilde{\varepsilon}^2$, we need $\alpha_d \leq \frac{\tilde{\varepsilon}^2}{2K'R_2}$, and $D \geq \frac{8\hat{\Delta}K'R_2}{\tilde{\varepsilon}^4}$. Then, applying average argument, above inequality implies for any $\tilde{d}$ uniformly sampled from $\{0,....D-1\}$, we have

$$\mathbb{E}_{\eta^{\tilde{d}}_x(\zeta)}|\nabla_2\mathcal{L}_\zeta(x_t,\eta^{\tilde{d}}_x(\zeta))|^2 = \frac{1}{D}\sum_{d=0}^{D-1}\mathbb{E}_{\eta^d_{x_t}(\zeta)}|\nabla_2\mathcal{L}_\zeta(x_t,\eta^d_{x_t}(\zeta))|^2 \leq \tilde{\varepsilon}^2, \tag{39}$$

which gives the desired result. ■

Appendix K. Proof of Lemma 4

Lemma 4 (Directional Smoothness) *Let Assumption 1 hold. For any x and x', we have*

$$\mathbb{E}_{\zeta\sim\mathbb{P}}\left\|\nabla\Psi_\zeta(x) - \nabla_1\mathcal{L}_\zeta(x',\eta^*_x(\zeta))\right\|^2 \leq K^2\left\|x-x'\right\|^2, \text{ where } K = G^2(\lambda\beta)^{-1}M + L. \tag{15}$$

Proof For any fixed ζ, define the following two quantities.

$$A = \mathbb{E}_{\xi\sim\nu}\Big[(f^*)'\Big(\frac{\ell(x;\xi)-c(\xi;\zeta)-\eta^*_x(\zeta)}{\lambda\beta}\Big)\nabla\ell(x;\xi) - (f^*)'\Big(\frac{\ell(x;\xi)-c(\xi;\zeta)-\eta^*_x(\zeta)}{\lambda\beta}\Big)\nabla\ell(x';\xi)\Big],$$
$$B = \mathbb{E}_{\xi\sim\nu}\Big[(f^*)'\Big(\frac{\ell(x;\xi)-c(\xi;\zeta)-\eta^*_x(\zeta)}{\lambda\beta}\Big)\nabla\ell(x';\xi) - (f^*)'\Big(\frac{\ell(x';\xi)-c(\xi;\zeta)-\eta^*_x(\zeta)}{\lambda\beta}\Big)\nabla\ell(x';\xi)\Big].$$

It's easy to show that $A + B = \nabla\Psi_\zeta(x) - \nabla_1\mathcal{L}_\zeta(x',\eta^*_x(\zeta))$. Our proof strategy is to bound A and B separately and combine the bounds together to give the results.
We first proceed upper bound of A. Note that $\eta^*_x(\zeta) \in \arg\min_\eta \mathcal{L}_\zeta(x,\eta)$, and the first-order optimality condition over η gives that

$$1 - \mathbb{E}_{\xi\sim\nu}\Big[(f^*)'\Big(\frac{\ell(x;\xi)-c(\xi;\zeta))-\eta^*_x(\zeta)}{\lambda\beta}\Big)\Big] = 0. \tag{40}$$

Moreover, note that the derivative of the conjugate function f^* satisfies $(f^*)'(v) = r^*(v) = \arg\max_{r\in\mathbf{R}_+}\langle r, v\rangle - f(r)$. Thus, we must have $(f^*)'(v) \geq 0$, and the above equation further implies that

$$\mathbb{E}_{\xi\sim\nu}\Big|(f^*)'\Big(\frac{\ell(x;\xi)-c(\xi;\zeta)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\Big| = \Big|\mathbb{E}_{\xi\sim\nu}\Big[(f^*)'\Big(\frac{\ell(x;\xi)-c(\xi;\zeta)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\Big]\Big| = 1. \tag{41}$$

Consequently, using L-smoothness of $\ell(x;\xi)$ and the equality (41), we conclude that

$$\|A\| \leq L\mathbb{E}_\xi\Big|f^{*'}\Big(\frac{\ell(x;\xi)-c(\xi;\zeta)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\Big|\|x-x'\| \leq L\|x-x'\|. \tag{42}$$

Next, we bound B. By G-Lipschitz continuity of $\ell(x;\xi)$ and M-smoothness of $(f^*)'$, we have

$$\begin{aligned}
&\|B\| \\
\leq&\mathbb{E}_\xi\Big\|\Big[(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x';\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\Big]\nabla\ell(x';\xi)\Big\| \\
\leq&\mathbb{E}_\xi\Big\|(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x';\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\Big\|\|\nabla\ell(x';\xi)\| \\
\leq&G\mathbb{E}_\xi\Big\|(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)-(f^*)'\Big(\frac{\ell(x';\xi)-\lambda c(\zeta,\xi)-\eta_x^*(\zeta)}{\lambda\beta}\Big)\Big\| \\
\leq&GM(\lambda\beta)^{-1}\mathbb{E}_\xi\|\ell(x;\xi)-\ell(x';\xi)\| \\
\leq&(\lambda\beta)^{-1}MG^2\|x-x'\|.
\end{aligned} \tag{43}$$

Combining (42) and (43), we obtain that

$$\|\nabla\Psi_\zeta(x)-\nabla\mathcal{L}_\zeta(x',\eta_x^*(\zeta))\| \leq \|A\|+\|B\| \leq (G^2(\lambda\beta)^{-1}M+L)\|x-x'\|.$$

Squaring both sides and taking expectation over ζ gives the claimed result. ■

Appendix L. Proof of Lemma 5

Lemma 5 (Second moment bound of $\hat{g}_t^B$) *The second moment of the mini-batch gradient estimator $\hat{g}_t^B$ with batch size B defined in (13) is upper bounded as follows*

$$\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta),\xi_B\sim\nu}\|\hat{g}_t^B\|^2 \leq \frac{R_1+10\varepsilon^2}{B}+\|\nabla_1\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta)}[\mathcal{L}_\zeta(x_t,\eta_{x_t}^{\tilde{d}}(\zeta))]\|^2, \tag{16}$$

where $R_1 = 8G^2+24G^2M^2(\lambda\beta)^{-2}\sigma^2+24G^2M^2\beta^{-2}\delta^2$.

Proof Throughout, given x, η, we denote $\mathcal{L}_{\zeta,\xi}(x,\eta) = f^*(\frac{\ell(x;\xi)-c(\zeta,\xi)-\eta}{\lambda\beta})$ to simplify notation. Notice that the outputs from Algorithm 1, $\tilde{d}$ also contains randomness, we use $\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta)}(\cdot)$ to denote expectation over $\eta_x^{\tilde{d}}(\zeta)$ through the proof. Then, for $\mathbb{E}_{\zeta\sim\mathbb{P},\xi\sim\nu,\eta_x^{\tilde{d}}(\zeta)}\|\nabla_1\mathcal{L}_{\zeta,\xi}(x,\eta_x^{\tilde{d}}(\zeta)\|^2$, we decompose it as follows

$$\mathbb{E}_{(\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta)),\xi\sim\nu}\|\nabla_1\mathcal{L}_{\zeta,\xi}(x,\eta_x^{\tilde{d}}(\zeta))\|^2$$

$$
\begin{aligned}
&= \mathbb{E}_{(\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta))}\big[\mathbb{E}_{\xi\sim\nu}\big\|\nabla_1\mathcal{L}_{\zeta,\xi}(x,\eta_x^{\tilde{d}}(\zeta))\big\|^2\big] \\
&\overset{(i)}{=} \mathbb{E}_{(\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta))}\big[Var_\xi\big(\nabla_1\mathcal{L}_{\xi,\zeta}(x,\eta_x^{\tilde{d}}(\zeta))\big) + \big\|\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))\big\|^2\big] \\
&= \mathbb{E}_{(\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta))}\big[Var_\xi\big(\nabla_1\mathcal{L}_{\xi,\zeta}(x,\eta_x^{\tilde{d}}(\zeta))\big)\big] + \mathbb{E}_{(\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta))}\big\|\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\eta_x^{\tilde{d}}(\zeta)}(\zeta))\big\|^2 \\
&\overset{(ii)}{=} \mathbb{E}_{(\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta))}\big[Var_\xi\big(\nabla_1\mathcal{L}_{\xi,\zeta}(x,\eta_x^{\tilde{d}}(\zeta))\big)\big] + \big[Var_{(\zeta,\eta_x^{\tilde{d}}(\zeta))}\big(\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))\big)\big] \\
&\quad + \big[\big\|\nabla_1\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta)}\big[\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))\big]\big\|^2\big],
\end{aligned}
$$

where (i) and (ii) apply equation $\mathbb{E}\big[\rho(\varpi)^2\big] = Var_\varpi(\rho(\varpi)) + (\mathbb{E}\big[\rho(\varpi)\big])^2$, which holds for any random variable ϖ and $\rho(\cdot):\mathbf{R}^d\to\mathbf{R}$.
Next, we bound $\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta)}\big[Var_\xi\big(\nabla_1\mathcal{L}_{\xi,\zeta}(x,\eta_x^{\tilde{d}}(\zeta))\big)\big]$ and $Var_{(\zeta,\eta_x^{\tilde{d}}(\zeta))}\big(\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))\big)$ separately. For $Var_\xi\big(\nabla_1\mathcal{L}_{\xi,\zeta}(x,\eta_x^{\tilde{d}}(\zeta))\big)$, we have

$$
\begin{aligned}
& Var_\xi\big(\nabla_1\mathcal{L}_{\zeta,\xi}(x,\eta_x^{\tilde{d}}(\zeta))\big) \\
&\overset{(i)}{=} \frac{1}{2}\mathbb{E}_{\xi_1,\xi_2}\Big\|(f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)\nabla\ell(x;\xi_1) \\
&\qquad - (f^*)'\Big(\frac{\ell(x;\xi_2)-\lambda c(\zeta,\xi_2)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)\nabla\ell(x;\xi_2)\Big\|^2 \\
&= \frac{1}{2}\mathbb{E}_{\xi_1,\xi_2}\Big\|(f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)\nabla\ell(x;\xi_1) \\
&\qquad - (f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)\nabla\ell(x;\xi_2) \\
&\qquad + (f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)\nabla\ell(x;\xi_2) \\
&\qquad - (f^*)'\Big(\frac{\ell(x;\xi_2)-\lambda c(\zeta,\xi_2)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)\nabla\ell(x;\xi_2)\Big\|^2 \\
&\overset{(ii)}{\leq} \mathbb{E}_{\xi_1,\xi_2}\Big[(f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)^2\big\|\nabla\ell(x;\xi_1)-\nabla\ell(x;\xi_2)\big\|^2\Big] \\
&\qquad + \mathbb{E}_{\xi_1,\xi_2}\Big[\big\|\nabla\ell(x;\xi_2)\big\|^2\cdot\Big((f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big) \\
&\qquad\quad - (f^*)'\Big(\frac{\ell(x;\xi_2)-\lambda c(\zeta,\xi_2)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)\Big)^2\Big] \\
&\overset{(iii)}{\leq} 4G^2\mathbb{E}_{\xi_1}\Big[(f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)^2\Big] \\
&\qquad + G^2M^2(\lambda\beta)^{-2}\mathbb{E}_{\xi_1,\xi_2}\big[\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\ell(x;\xi_2)+\lambda c(\zeta,\xi_2)\big]^2 \\
&\overset{(iv)}{\leq} 4G^2\mathbb{E}_{\xi_1}\Big[(f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)^2\Big] + 2G^2M^2(\lambda\beta)^{-2}(2\sigma^2+2\lambda^2\delta^2),
\end{aligned}
\tag{44}
$$

where (i) applies proposition 1; (ii) applies the fact $(a+b)^2\leq 2(a^2+b^2)$; (iii) applies the assumption G-Lipschitz continuity of $\ell(x;\xi)$ function and M-smoothness of f^*; (iv) applies bounded variance

assumption of $\ell(x,\xi)$ and $c(\zeta,\xi)$.
Applying inequality $a^2 \le 2(a-1)^2+2$, the term in (44), $\mathbb{E}_{\xi_1}\Big[(f^*)'\big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde d}(\zeta)}{\lambda\beta}\big)^2\Big]$ can be further upper bounded as

$$\begin{aligned}
&\mathbb{E}_{\xi_1}\Big[(f^*)'\Big(\frac{\ell(x;\xi_1)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde d}(\zeta)}{\lambda\beta}\Big)^2\Big]\\
&\le 2+2\mathbb{E}_{\xi_1}\Big[1-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta,\xi_1)-\eta_x^{\tilde d}(\zeta)}{\lambda\beta}\Big)\Big]^2\\
&\le 2\Big(1+\big|\nabla_2\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))\big|^2+Var_\xi\big(\nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta_x^{\tilde d}(\zeta))\big)\Big).
\end{aligned}\tag{45}$$

Since at $\eta_x^{\tilde d}(\zeta)$, by (6) stated in Theorem 3, we conclude $\mathbb{E}_{\eta_x^{\tilde d}(\zeta)}|\nabla_2\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))|^2 \le \tilde\varepsilon^2=\varepsilon^2/G^2$. Moreover, from lemma 3, we have

$$Var_\xi\big(\nabla_2\mathcal{L}_{\zeta,\xi}(x,\eta_x^{\tilde d}(\zeta))\big)\le 2M^2(\lambda\beta)^{-2}(\sigma^2+\lambda^2\delta^2).\tag{46}$$

Thus, combining inequalities (44) (45) (46), we have

$$\begin{aligned}
&\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde d}(\zeta)}\big[Var_\xi\big(\nabla_1 L_{\zeta,\xi}(x,\eta_x^{\tilde d}(\zeta))\big)\big]\\
\le&\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde d}(\zeta)}\big[4G^2\cdot 2(1+\big|\nabla_2 L_\zeta(x,\eta_x^{\tilde d}(\zeta))\big|^2+2M^2(\lambda\beta)^{-2}(\sigma^2+\lambda^2\delta^2))\\
&\quad+2G^2M^2(\lambda\beta)^{-2}(2\sigma^2+2\lambda^2\delta^2)\big]\\
\le&8G^2+8G^2\tilde\varepsilon^2+20G^2M^2(\lambda\beta)^{-2}\sigma^2+20G^2M^2\beta^{-2}\delta^2\\
=&8G^2+8\varepsilon^2+20G^2M^2(\lambda\beta)^{-2}\sigma^2+20G^2M^2\beta^{-2}\delta^2.
\end{aligned}\tag{47}$$

For $Var_{(\zeta,\eta_{x_t}^{\tilde d}(\zeta))}\big(\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))\big)$, we have

$$\begin{aligned}
&Var_{(\zeta,\eta_x^{\tilde d}(\zeta))}\big(\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde d}(\zeta))\big)\\
=&\frac{1}{2}\mathbb{E}_{(\zeta_1,\eta_x^{\tilde d}(\zeta_1)),(\zeta_2,\eta_x^{\tilde d}(\zeta_2))}\Big\|\mathbb{E}_\xi\Big[\Big((f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta_1,\xi)-\eta_x^{\tilde d}(\zeta_1)}{\lambda\beta}\Big)\\
&\quad-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta_2,\xi)-\eta_x^{\tilde d}(\zeta_2)}{\lambda\beta}\Big)\Big)\cdot\nabla\ell(x;\xi)\Big]\Big\|^2\\
\overset{(i)}{\le}&\frac{1}{2}\mathbb{E}_{(\zeta_1,\eta_x^{\tilde d}(\zeta_1)),(\zeta_2,\eta_x^{\tilde d}(\zeta_2)),\xi}\Big\|\Big[(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta_1,\xi)-\eta_x^{\tilde d}(\zeta_1)}{\lambda\beta}\Big)\\
&\quad-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta_2,\xi)-\eta_x^{\tilde d}(\zeta_2)}{\lambda\beta}\Big)\Big]\cdot\nabla\ell(x;\xi)\Big\|^2\\
\overset{(ii)}{=}&\frac{1}{2}\mathbb{E}_{(\zeta_1,\eta_x^{\tilde d}(\zeta_1)),(\zeta_2,\eta_x^{\tilde d}(\zeta_2)),\xi}\Big|(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta_1,\xi)-\eta_x^{\tilde d}(\zeta_1)}{\lambda\beta}\Big)\\
&\quad-(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta_2,\xi)-\eta_x^{\tilde d}(\zeta_2)}{\lambda\beta}\Big)\Big|^2\big\|\nabla\ell(x;\xi)\big\|^2\\
\overset{(iii)}{\le}&\frac{1}{2}G^2\mathbb{E}_{(\zeta_1,\eta_x^{\tilde d}(\zeta_1)),(\zeta_2,\eta_x^{\tilde d}(\zeta_2)),\xi}\Big|(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta_1,\xi)-\eta_x^{\tilde d}(\zeta_1)}{\lambda\beta}\Big)-1
\end{aligned}$$

$$
\begin{aligned}
&\quad - (f^*)'\Big(\frac{\ell(x;\xi) - \lambda c(\zeta_2,\xi) - \eta_x^{\tilde{d}}(\zeta_2)}{\lambda\beta}\Big) + 1|^2 \\
&\overset{(iv)}{\le} G^2 \mathbb{E}_{(\zeta_1,\eta_x^{\tilde{d}}(\zeta_1)),(\zeta_2,\eta_x^{\tilde{d}}(\zeta_2))}\Big(\mathbb{E}_\xi|\nabla_2\mathcal{L}_{\zeta_1,\xi}(x;\eta_x^{\tilde{d}}(\zeta_1))|^2 + \mathbb{E}_\xi|\nabla_2\mathcal{L}_{\zeta_2,\xi}(x;\eta_x^{\tilde{d}}(\zeta_2))|^2\Big) \\
&\overset{(v)}{=} G^2 \mathbb{E}_{(\zeta_1,\eta_x^{\tilde{d}}(\zeta_1)),(\zeta_2,\eta_x^{\tilde{d}}(\zeta_2))}\Big(Var_\xi(\nabla_2\mathcal{L}_{\zeta_1,\xi}(x;\eta_x^{\tilde{d}}(\zeta_1))) + |\nabla_2\mathcal{L}_{\zeta_1}(x;\eta_x^{\tilde{d}}(\zeta_1))|^2 \\
&\quad + Var_\xi(\nabla_2\mathcal{L}_{\zeta_2,\xi}(x;\eta_x^{\tilde{d}}(\zeta_2))) + |\nabla_2\mathcal{L}_{\zeta_2}(x;\eta_x^{\tilde{d}}(\zeta_2))|^2\Big) \\
&\overset{(vi)}{\le} G^2\Big(4M^2(\lambda\beta)^{-2}(\sigma^2+\lambda\delta^2) + \mathbb{E}_{\zeta_1,\eta_x^{\tilde{d}}(\zeta_1)}|\nabla_2\mathcal{L}_{\zeta_2}(x;\eta_x^{\tilde{d}}(\zeta_1))|^2 + \mathbb{E}_{\zeta_2,\eta_x^{\tilde{d}}(\zeta_2)}|\nabla_2\mathcal{L}_{\zeta_2}(x;\eta_x^{\tilde{d}}(\zeta_2))|^2\Big) \\
&\overset{(vii)}{\le} 4G^2M^2(\lambda\beta)^{-2}(\sigma^2+\lambda\delta^2) + 2G^2\tilde{\varepsilon}^2 \\
&= 4G^2M^2(\lambda\beta)^{-2}\sigma^2 + 4G^2M^2\beta^{-2}\delta^2 + 2\varepsilon^2
\end{aligned} \tag{48}
$$

where (i) applies Jensen's inequality move expectation over ξ out squared-norm; (ii) extracts scalar $|(f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta_1,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big) - (f^*)'\Big(\frac{\ell(x;\xi)-\lambda c(\zeta_2,\xi)-\eta_x^{\tilde{d}}(\zeta)}{\lambda\beta}\Big)|$ out; (iii) applies G-Lipschitz assumption of $\ell(\cdot,\xi)$ stated at assumption 1. (iv) applies inequality $(a+b)^2 \le 2a^2+2b^2$ to decouple $|\nabla_2\mathcal{L}_{\zeta_1,\xi}(x;\eta_x^{\tilde{d}}(\zeta_1))|$ and $|\nabla_2\mathcal{L}_{\zeta_2,\xi}(x;\eta_x^{\tilde{d}}(\zeta_2))|$; (v) applies equality $\mathbb{E}\big[\rho(\varpi)^2\big] = Var_\varpi(\rho(\varpi)) + (\mathbb{E}\big[\rho(\varpi)\big])^2$, which holds for any random variable ϖ and $\rho(\cdot) : \mathbf{R}^d \to \mathbf{R}$; (v) and (vi) apply Lemma 3 to upper bound $Var_\xi(\nabla_2\mathcal{L}_{\zeta_1,\xi}(x;\eta_x^{\tilde{d}}(\zeta_1)))$, $Var_\xi(\nabla_2\mathcal{L}_{\zeta_2,\xi}(x;\eta_x^{\tilde{d}}(\zeta_2)))$; (vii) applies (6) stated in Theorem 3 to upper bound $\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta_1)}|\nabla_2\mathcal{L}_{\zeta_2}(x;\eta_x^{\tilde{d}}(\zeta_1))|^2$ and $\mathbb{E}_{\eta_x^{\tilde{d}}(\zeta_2)}|\nabla_2\mathcal{L}_{\zeta_2}(x;\eta_x^{\tilde{d}}(\zeta_2))|^2$ by $\tilde{\varepsilon}^2 = \varepsilon^2/G^2$.

Combining (47) (48), we have

$$
\begin{aligned}
&\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta),\xi\sim\nu}\big\|\nabla_1\mathcal{L}_{\zeta,\xi}(x,\eta_x^{\tilde{d}}(\zeta))\big\|^2 \\
\le&\mathbb{E}_{(\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta))}\big[Var_\xi\big(\nabla_1\mathcal{L}_{\xi,\zeta}(x,\eta_x^{\tilde{d}}(\zeta))\big)\big] + \big[Var_{(\zeta,\eta_x^{\tilde{d}}(\zeta))}\big(\nabla_1\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))\big)\big] \\
&\quad + \big[\big\|\nabla_1\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta)}\big[\mathcal{L}_\zeta(x,\eta_x^{\tilde{d}}(\zeta))\big]\big\|^2\big], \\
\le&8G^2 + 10\varepsilon^2 + 24G^2M^2(\lambda\beta)^{-2}\sigma^2 + 24G^2M^2\beta^{-2}\delta^2 + \big\|\nabla_1\mathbb{E}_{\zeta,\eta_x^{\tilde{d}}(\zeta)}\mathcal{L}(x,\eta_x^{\tilde{d}}(\zeta))\big\|^2.
\end{aligned} \tag{49}
$$

For mini-batch stochastic gradient estimator (13), the RHS of (49) becomes

$$
\begin{aligned}
&\mathbb{E}_{\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta),\xi_B\sim\nu}\big\|\nabla_1\mathcal{L}_{\zeta,\xi}(x,\eta_x^{\tilde{d}}(\zeta))\big\|^2 \\
\le&\frac{8G^2 + 24G^2M^2(\lambda\beta)^{-2}\sigma^2 + 24G^2M^2\beta^{-2}\delta^2}{B} + \frac{10\varepsilon^2}{B} + \big\|\nabla_1\mathbb{E}_{\zeta,\eta_x^{\tilde{d}}(\zeta)}[\mathcal{L}(x,\eta_x^{\tilde{d}}(\zeta))]\big\|^2,
\end{aligned}
$$

which gives the desired result. ■

Appendix M. Proof of Theorem 4

Theorem 4 (Convergence of Algorithm 2) *Let Assumptions 1 and 2 hold. Denote* $\Delta = \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_0)\big] - \inf_x \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$, *apply Algorithm 2 to solve the outer objective in* (3) *using a constant learning* *rate* $\gamma_t = \gamma = \min\{\frac{1}{24K}, \frac{\varepsilon^2}{2KR_1}\}$, *and set the batch size* $B = 1$. *At each iteration, query Algorithm 1 to obtain an estimator* $\eta_x^{\tilde{d}}(\zeta)$ *for sampled* ζ. *Then, the output* $x_{\tilde{t}}$ *of Algorithm 2 satisfies*

$$\mathbb{E}_{x_{\tilde{t}}}\big\|\nabla \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{x_{\tilde{t}}})\big]\big\|^2 \le 7\varepsilon^2, \tag{17}$$

after $T \ge \max\{96\Delta K\varepsilon^{-2}, 8\Delta KR_1\varepsilon^{-4}\} = \mathcal{O}(\Delta KR_1\varepsilon^{-4})$ *number of iterations.*

Proof Regarding the objective function $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x)\big]$, we have the following descent lemma

$$\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{t+1})\big] \le \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big] + \langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], x_{t+1}-x_t\rangle + \frac{K\gamma_t^2}{2}\big\|x_{t+1}-x_t\big\|^2. \tag{50}$$

The proof of (50) is provided in Appendix M.1. Replace $x_{t+1} - x_t$ by $-\gamma_t\hat{g}_t^B$, above inequality implies

$$\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{t+1})\big] \le \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big] - \gamma_t\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], \hat{g}_t\rangle + \frac{K\gamma_t^2}{2}\big\|\hat{g}_t\big\|^2. \tag{51}$$

Since during parameters' update, the inexact mini-batch stochastic gradient, $\hat{g}_t^B$ is utilized, where its randomness comes from x_t, ζ, $\{\xi\}_{\tilde{B}}$ and $\eta_x^{\tilde{d}}(\zeta)$. Taking expectation over $\zeta, \eta_x^{\tilde{d}}(\zeta), \{\xi\}_B$ on both sides conditioned on x_t, we have

$$\begin{aligned}
&\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_{t+1})|x_t\Big]\\
\overset{(i)}{\le}&\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big] - \mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], \gamma_t\hat{g}_t^B\rangle|x_t\Big]\\
&\quad + \mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\frac{K\gamma_t^2}{2}\big\|\hat{g}_t^B\big\|^2|x_t\Big]\\
\overset{(ii)}{\le}&\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big] - \mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\big\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], \gamma_t\hat{g}_t^B\big\rangle\,|x_t\Big]\\
&\quad + \frac{K\gamma_t^2(R_1+10\varepsilon^2)}{2}\\
&\quad + \frac{K\gamma_t^2}{2}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta)}\big[\mathcal{L}_\zeta(x_t,\eta_{x_t}^{\tilde{d}}(\zeta))\big]\big\|^2\\
\overset{(iii)}{\le}&\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big] - \gamma_t\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], \hat{g}_t^B - g_t^B + g_t^B\rangle|x_t\Big]\\
&\quad + K\gamma_t^2\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta)}\big[\mathcal{L}_\zeta(x_t,\eta_{x_t}^{\tilde{d}}(\zeta))\big] - \nabla\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta)}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
&\quad + K\gamma_t^2\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta)}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
&\quad + \frac{KR_1\gamma_t^2}{2} + 5K\varepsilon^2\gamma_t^2\\
\overset{(iv)}{\le}&\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big] - \gamma_t\mathbb{E}_{\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], \hat{g}_t^B - g_t^B + g_t^B\rangle|x_t\Big]\\
&\quad + K\gamma_t^2\mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\big\|\nabla_1\mathbb{E}_{\eta_{x_t}^{\tilde{d}}(\zeta)}\big[\mathcal{L}_\zeta(x_t,\eta_{x_t}^{\tilde{d}}(\zeta))\big] - \nabla\Psi_\zeta(x_t)\big\|^2|x_t\Big]\\
&\quad + K\gamma_t^2\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2 + \frac{KR_1\gamma_t^2}{2} + 5K\varepsilon^2\gamma_t^2
\end{aligned}$$

$$\begin{aligned}
&\overset{(v)}{\leq}\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big]-\gamma_t\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big],\hat g^B_t-g^B_t\rangle|x_t\Big]\\
&\qquad-\gamma_t\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big],g^B_t\rangle|x_t\Big]\\
&\qquad+K\gamma_t^2\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
&\qquad+K\varepsilon^2\gamma_t^2+\frac{KR_1\gamma_t^2}{2}+5K\varepsilon^2\gamma_t^2\\
&=\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big]-\gamma_t\mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\big\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big],\mathbb{E}_{\eta^{\tilde d}_{x_t}(\zeta),\xi_B\sim\nu}[\hat g^B_t-g^B_t]\big\rangle|x_t\Big]\\
&\qquad-(\gamma_t-K\gamma_t^2)\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
&\qquad+\frac{KR_1\gamma_t^2}{2}+6K\varepsilon^2\gamma_t^2\\
&\overset{(vi)}{=}\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big]-\gamma_t\mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\big\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big],\nabla_1\mathbb{E}_{\eta^{\tilde d}_x(\zeta)}[\mathcal{L}_\zeta(x_t,\eta^{\tilde d}_x(\zeta))-\nabla\Psi_\zeta(x_t)]\big\rangle|x_t\Big]\\
&\qquad-(\gamma_t-K\gamma_t^2)\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
&\qquad+\frac{KR_1\gamma_t^2}{2}+6K\varepsilon^2\gamma_t^2\\
&\overset{(vii)}{\leq}\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big]+\gamma_t\mathbb{E}_{\zeta\sim\mathbb{P}}\Big[\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|\cdot\big\|\nabla_1\mathbb{E}_{\eta^{\tilde d}_{x_t}(\zeta)}[\mathcal{L}_\zeta(x_t,\eta^{\tilde d}_{x_t}(\zeta))]-\nabla\Psi_\zeta(x_t)\big]\big\||x_t\Big]\\
&\qquad-(\gamma_t-K\gamma_t^2)\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
&\qquad+\frac{KR_1\gamma_t^2}{2}+6K\varepsilon^2\gamma_t^2\\
&=\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big]+\gamma_t\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|\cdot\mathbb{E}_{\zeta\sim\mathbb{P}}\big\|\nabla_1\mathbb{E}_{\eta^{\tilde d}_{x_t}(\zeta)}[\mathcal{L}_\zeta(x_t,\eta^{\tilde d}_{x_t}(\zeta))]-\nabla\Psi_\zeta(x_t)\big]\big\|\\
&\qquad-(\gamma_t-K\gamma_t^2)\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
&\qquad+\frac{KR_1\gamma_t^2}{2}+6K\varepsilon^2\gamma_t^2\\
&\overset{(viii)}{\leq}\mathbb{E}_{\zeta\sim\mathbb{P}(\zeta),\xi_B\sim\nu,\eta^{\tilde d}_{x_t}}\Big[\Psi_\zeta(x_t)|x_t\Big]+\gamma_t\varepsilon\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|-(\gamma_t-K\gamma_t^2)\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
&\qquad+\frac{KR_1\gamma_t^2}{2}+6K\varepsilon^2\gamma_t^2,
\end{aligned}\tag{52}$$

where (i) applies descent lemma (50); (ii) applies (16) stated in lemma 5 with $B=1$; (iii) applies $(a+b)^2\leq 2a^2+2b^2$ to further upper bound $\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta)}[\mathcal{L}_\zeta(x_t,\eta^{\tilde d}_{x_t}(\zeta))]\|^2$ by $2\|\nabla_1\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta)}[\mathcal{L}_\zeta(x_t,\eta^{\tilde d}_{x_t}(\zeta))]-\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x_t)]\|^2+2\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x_t)]\|^2$ (the expectation over $\eta^{\tilde d}_{x_t}(\zeta)$ can be neglected as $\Psi_\zeta(x_t)$ doesn't contain randomness from $\eta^{\tilde d}_{x_t}(\zeta)$); (iv) applies Jensen's inequality to extract $\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde d}_{x_t}(\zeta)}$ out from squared norm; (v) applies condition (7) stated in Theorem 2; (vi) moves expectation over ξ_B inside inner product; (vii) applies Cauchy-Schwarz inequality; (viii) again applies condition (7) stated Theorem 2 as such relationship holds for every ζ.
Then, for any $\gamma_t<\frac{1}{2K}$ by choice, we have $(\gamma_t-K\gamma_t^2)>\frac{\gamma_t}{2}$. Taking expectation over x_t, the above inequality further transformed to

$$\mathbb{E}_{x_t,\zeta\sim\mathbb{P},\eta^{\tilde d}_x(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_{t+1})\Big]$$

$$\leq \mathbb{E}_{x_t,\zeta\sim\mathbb{P},\eta_x^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)\Big] + \gamma_t\varepsilon\mathbb{E}_{x_t}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\| - \frac{\gamma_t}{2}\mathbb{E}_{x_t}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2 + \frac{KR_1\gamma_t^2}{2} + 6K\varepsilon^2\gamma_t^2.$$

Re-arranging above terms, we have

$$\begin{aligned}
&\frac{\gamma_t}{2}\mathbb{E}_{x_t}\Big(\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2 - 2\varepsilon\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|\Big)\\
=&\frac{\gamma_t}{2}\mathbb{E}_{x_t}\Big(\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\| - \varepsilon\Big)^2 - \frac{\gamma_t\varepsilon^2}{2}\\
\leq&\mathbb{E}_{x_t,\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\big[\Psi_\zeta(x_t) - \Psi_\zeta(x_{t+1})\big] + \frac{KR_1\gamma_t^2}{2} + 6K\varepsilon^2\gamma_t^2.
\end{aligned}$$

Re-arranging above inequality, we have

$$\frac{\gamma_t}{2}\mathbb{E}_{x_t}\Big(\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\| - \varepsilon\Big)^2 \leq \mathbb{E}_{x_t,\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\big[\Psi_\zeta(x_t) - \Psi_\zeta(x_{t+1})\big] + \frac{KR_1\gamma_t^2}{2} + 6K\varepsilon^2\gamma_t^2 + \frac{\gamma_t\varepsilon^2}{2}.$$

Applying $\frac{(a+b)^2}{2} \leq a^2 + b^2$ and re-arranging above terms, we have

$$\begin{aligned}
&\frac{\gamma_t}{4}\mathbb{E}_{x_t}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
\leq&\frac{\gamma_t}{2}\mathbb{E}_{x_t}\Big[\Big(\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\| - \varepsilon\Big)^2\Big] + \frac{\gamma_t\varepsilon^2}{2}\\
\leq&\mathbb{E}_{x_t,\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t) - \Psi_\zeta(x_{t+1})\Big] + \frac{KR_1\gamma_t^2}{2} + 6K\varepsilon^2\gamma_t^2 + \varepsilon^2\gamma_t.
\end{aligned}$$

Summing the above inequality from $t = 0$ to $T - 1$ leads to

$$\begin{aligned}
&\sum_{t=0}^{T-1}\frac{\gamma_t}{4}\mathbb{E}_{x_t}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
\leq&\sum_{t=0}^{T-1}\mathbb{E}_{x_t,\zeta\sim\mathbb{P},\eta_{x_t}^{\tilde{d}}(\zeta),\xi_B\sim\nu}\big[\Psi_\zeta(x_t) - \Psi(x_{t+1})\big] + 6K\sum_{t=0}^{T-1}\gamma_t^2\varepsilon^2 + \frac{KR_1}{2}\sum_{t=0}^{T-1}\gamma_t^2 + \varepsilon^2\sum_{t=0}^{T-1}\gamma_t\\
\leq&\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_0) - \Psi_\zeta(x^*)\big] + 6K\varepsilon^2\sum_{t=0}^{T-1}\gamma_t^2 + \frac{KR_1}{2}\sum_{t=0}^{T-1}\gamma_t^2 + \varepsilon^2\sum_{t=0}^{T-1}\gamma_t.
\end{aligned}$$

Denoting $\Delta = \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_0) - \Psi_\zeta(x^*)\big]$ and choosing constant learning rate $\gamma_t = \gamma$, we have

$$\begin{aligned}
\mathbb{E}_{x_{\tilde{t}}}\big[\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{\tilde{t}})\big]\big\|^2\big] =&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{x_t}\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi(x_t)]\|^2\\
\leq&\frac{4\Delta}{T\gamma} + \frac{24K\varepsilon^2\gamma^2T}{T\gamma} + \frac{2KR_1\gamma^2T}{T\gamma} + \frac{4\varepsilon^2T\gamma}{T\gamma}. \qquad (53)
\end{aligned}$$

Then, choosing $\gamma = \min\{\frac{1}{24K}, \frac{\varepsilon^2}{2KR_1}\}$, we immediately have

$$\frac{24K\varepsilon^2\gamma^2 T}{T\gamma} \leq \varepsilon^2 \text{ and } \frac{2KR_1\gamma^2 T}{T\gamma} \leq \varepsilon^2. \tag{54}$$

To make $\frac{4\Delta}{T\gamma} = \varepsilon^2$, we have

$$T = \frac{4\Delta}{\gamma\varepsilon^2} \geq \max\{\frac{96\Delta K}{\varepsilon^2}, \frac{8\Delta K R_1}{\varepsilon^4}\} = \mathcal{O}(\Delta K R_1 \varepsilon^{-4}). \tag{55}$$

Combining all inequalities together, we have

$$\mathbb{E}_{x_{\tilde{t}}}\big[\|\nabla \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{\tilde{t}})\big]\|^2\big] \leq 7\varepsilon^2, \tag{56}$$

which gives the desired result.

■

M.1 Proof of Descent Lemma (50)

Lemma A.6 *Denote $\eta^*_{x_t}(\zeta) \in \arg\min_\eta \mathcal{L}_\zeta(x_t, \eta)$. Then, for $\mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi(x_t)]$, we have the following descent lemma*

$$\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{t+1})\big] \leq \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big] + \langle \nabla \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], x_{t+1} - x_t\rangle + \frac{K}{2}\|x_{t+1} - x_t\|^2, \tag{57}$$

where $K = G^2(\lambda\beta)^{-1}M + L$.

Proof Notice that, applying Jensen's inequality and taking square-root on both sides, Lemma 4 implies

$$\big\|\nabla \mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x)] - \nabla_1 \mathbb{E}_{\zeta\sim\mathbb{P}}[\mathcal{L}_\zeta(x', \eta^*_x(\zeta))]\big\| \leq K\big\|x - x'\big\|.$$

From fundamental theorem of calculus, we have

$$\begin{aligned}
&\Big|\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\mathcal{L}_\zeta(x_{t+1}, \eta^*_{x_t}(\zeta))\big] - \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big] - \langle \nabla \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], x_{t+1} - x_t\rangle\Big| \\
=&\Big|\int_0^1 \langle \nabla_1 \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\mathcal{L}_\zeta(x_t + t(x_{t+1} - x_t), \eta^*_{x_t}(\zeta))\big], x_{t+1} - x_t\rangle - \langle \nabla \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], x_{t+1} - x_t\rangle \mathrm{d}t\Big| \\
=&\Big|\int_0^1 \langle \nabla_1 \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\mathcal{L}_\zeta(x_t + t(x_{t+1} - x_t), \eta^*_{x_t}(\zeta))\big] - \nabla_1 \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\mathcal{L}_\zeta(x_t, \eta^*_{x_t}(\zeta))\big], x_{t+1} - x_t\rangle \mathrm{d}t\Big| \\
\overset{(i)}{\leq}&\Big|\int_0^1 \big\|\nabla_1 \mathbb{E}_\zeta\big[\mathcal{L}_\zeta(x_t + t(x_{t+1} - x_t), \eta^*_{x_t}(\zeta))\big] - \nabla_1 \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\mathcal{L}_\zeta(x_t, \eta^*_x(\zeta))\big]\big\| \|x_{t+1} - x_t\| \,\mathrm{d}t\Big| \\
\overset{(ii)}{\leq}&\Big|\int_0^1 tK\|x_{t+1} - x_t\|^2 \,\mathrm{d}t\Big| \\
\leq&\frac{K}{2}\big\|x_{t+1} - x_t\big\|^2,
\end{aligned}$$

where (i) applies Cauchy-Schwarz inequality; (ii) applies directional smoothness property stated at Lemma 4.
Re-arranging above inequality, we have

$$\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\mathcal{L}_\zeta(x_{t+1},\eta^*_{x_t}(\zeta))\big] \le \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big] + \langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], x_{t+1}-x_t\rangle + \frac{K}{2}\left\|x_{t+1}-x_t\right\|^2.$$

Since for each ζ, $\eta^*_{x_{t+1}}(\zeta)\in\arg\min_\eta \mathcal{L}_\zeta(x_{t+1},\eta)$, it holds that $\Psi_\zeta(x_{t+1}) = \mathcal{L}_\zeta(x_{t+1},\eta^*_{x_{t+1}}(\zeta)) \le \mathcal{L}_\zeta(x_{t+1},\eta^*_{x_t}(\zeta))$. Taking expectation over ζ, we have $\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{t+1})\big] \le \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\mathcal{L}_\zeta(x_{t+1},\eta^*_{x_t}(\zeta))\big]$. Combining this fact with above inequality gives us the desired result,

$$\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{t+1})\big] \le \mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big] + \langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big], x_{t+1}-x_t\rangle + \frac{K}{2}\left\|x_{t+1}-x_t\right\|^2.$$

■

Appendix N. Proof of Corollary 5

Corollary 2 (Complexity Bound for $\min_x \mathbb{E}_{\zeta\sim\mathbb{P}}[\Psi_\zeta(x)]$) *Let Assumptions 1 and 2 hold. Then, the Nested-SGD algorithm (Algorithm 2) returns an ε-stationary point with a total sample complexity of $\mathcal{O}(\varepsilon^{-8})$ for sampling ξ and ζ. Furthermore, by setting the batch sizes $B,\tilde{B}\sim\Theta(\varepsilon^{-2})$, the total iteration complexity becomes $T\times D\sim\mathcal{O}(\varepsilon^{-4})$. At each iteration, Algorithms 1 and 2 incur memory complexities of $\mathcal{O}(1)$ and $\mathcal{O}(d)$, respectively.*

Proof According to Theorem 3, for Algorithm 1 to output $\eta^{\tilde{d}}_x(\zeta)$ satisfying the condition (6) in Theorem 2, a sample complexity of $\mathcal{O}(\hat{\Delta}G^4K'R_2\varepsilon^{-4})$ is required. Furthermore, by Theorem 4, Algorithm 2 requires $\mathcal{O}(1)$ mini-batch of ζ and ξ per iteration and runs for $\mathcal{O}(\Delta K R_1\varepsilon^{-4})$ iterations. Combining both results, the total sample complexity is $\mathcal{O}(\Delta\hat{\Delta}R_1R_2KK'G^4\varepsilon^{-8})\sim\mathcal{O}(\varepsilon^{-8})$.

For the total iteration complexity, i.e., $T\times D$, it can be improved as follows. By setting $B=\Theta(\varepsilon^{-2})$ and applying the conclusion from Lemma 5 along with the descent inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_{t+1})|x_t\Big]\\
\le&\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)|x_t\Big] - \mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta),\xi_B\sim\nu}\Big[\big\langle\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big],\gamma_t\hat{g}^B_t\big\rangle\,|x_t\Big]\\
&\quad+\frac{K\gamma_t^2(R_1\varepsilon^2+10\varepsilon^4)}{2}\\
&\quad+\frac{K\gamma_t^2}{2}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P},\eta^{\tilde{d}}_{x_t}(\zeta)}\big[\mathcal{L}_\zeta(x_t,\eta^{\tilde{d}}_{x_t}(\zeta))\big]\big\|^2.
\end{aligned}$$

Following the same logic as Proof M, we then have

$$\begin{aligned}
&\mathbb{E}_{x_t,\zeta\sim\mathbb{P},\eta^{\tilde{d}}_x(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_{t+1})\Big]\\
\le&\mathbb{E}_{x_t,\zeta\sim\mathbb{P},\eta^{\tilde{d}}_x(\zeta),\xi_B\sim\nu}\Big[\Psi_\zeta(x_t)\Big] + \gamma_t\varepsilon\mathbb{E}_{x_t}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\| - \frac{\gamma_t}{2}\mathbb{E}_{x_t}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2\\
&\quad+\frac{KR_1\gamma_t^2\varepsilon^2}{2} + K\varepsilon^2\gamma_t^2 + 5K\gamma_t^2\varepsilon^4.
\end{aligned}$$

Let $\gamma_t = \gamma$, re-arranging above inequality, summing t from $0, \cdots, T-1$ and dividing by T, we have

$$\frac{\gamma}{4T}\sum_{t=0}^{T-1}\mathbb{E}_{x_t}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_t)\big]\big\|^2 \leq \frac{\Delta}{T} + \frac{KR_1\gamma^2\varepsilon^2}{2} + K\varepsilon^2\gamma^2 + 5K\varepsilon^4\gamma^2 + \varepsilon^2\gamma.$$

For $\gamma = \min\{\frac{1}{2KR_1}, \frac{1}{4K}\}$, after $T = \max\{8\Delta KR_1, 16\Delta K\}\varepsilon^{-2} = \mathcal{O}(\Delta KR_1\varepsilon^{-2})$ iterations, above inequality further implies

$$\mathbb{E}_{x_{\tilde{t}}}\big\|\nabla\mathbb{E}_{\zeta\sim\mathbb{P}}\big[\Psi_\zeta(x_{\tilde{t}})\big]\big\|^2 \leq 7\varepsilon^2 + 5\varepsilon^4 = \mathcal{O}(\varepsilon^2). \tag{58}$$

This concludes that by choosing $B, \tilde{B} \sim \Theta(\varepsilon^{-2})$ as suggested in Remark 5, we have total iteration complexity $T \times D = \mathcal{O}(\Delta\hat{\Delta}R_1R_2KK'G^4\varepsilon^{-4}) \sim \mathcal{O}(\varepsilon^{-4})$. Based on gradient dimension with respect to x and η, we conclude the per-iteration complexity of algorithm 1 and 2 are $\mathcal{O}(1), \mathcal{O}(d)$ respectively. ■