Learning Robust to Distributional Uncertainties and Adversarial Data

Alireza Sadeghi, *Member, IEEE*, Gang Wang, *Senior Member, IEEE*, and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—Successful training of data-intensive deep neural networks critically rely on vast, clean, and high-quality datasets. In practice however, their reliability diminishes, particularly with noisy, outlier-corrupted data samples encountered in testing. This challenge intensifies when dealing with anonymized, heterogeneous data sets stored across geographically distinct locations due to, e.g., privacy concerns. This present paper introduces robust learning frameworks tailored for *centralized* and *federated* learning scenarios. Our goal is to fortify model resilience with a focus that lies in (i) addressing distribution shifts from training to inference time; and, (ii) ensuring test-time robustness, when a trained model may encounter outliers or adversarially contaminated test data samples. To this aim, we start with a centralized setting where the true data distribution is considered unknown, but residing within a Wasserstein ball centered at the empirical distribution. We obtain robust models by minimizing the worst-case expected loss within this ball, yielding an intractable infinite-dimensional optimization problem. Upon leverage the strong duality condition, we arrive at a tractable surrogate learning problem. We develop two stochastic primal-dual algorithms to solve the resultant problem: one for ϵ-accurate convex sub-problems and another for a single gradient ascent step. We further develop a distributionally robust federated learning framework to learn robust model using heterogeneous data sets stored at distinct locations by solving per-learner's sub-problems locally, offering robustness with modest computational overhead and considering data distribution. Numerical tests corroborate merits of our training algorithms against distributional uncertainties and adversarially corrupted test data samples.**

***Index Terms*—Wasserstein distance, distributionally robust optimization, minimax, primal-dual, federated learning.**

I. Introduction

THE RELIABLE performance of contemporary machine learning models largely hinges on their training algorithms. In practice, their generalization is often compromised during inference, especially when confronted with noisy data tainted by outliers, as cited in previous research [1]. These issues primarily stem from imperfect data acquisition and distribution shifts from training to inference phases. Furthermore, massive datasets are typically not confined to a single device, instead they are generated and distributed across multiple locations (referred to as workers) with limited computing and storage capabilities. Distribution shifts among workers, along with privacy and confidentiality concerns, pose additional challenges in training reliable models from distributed datasets [2], [3].

Recent research efforts have been devoted to providing reliable models with enhanced generalization guarantees, particularly by addressing the susceptibility of models to adversarial data samples. These strategies fall into two categories: detection approaches, which determine if the model input is anomalous [4], [5], and robust training methods that endow the model with resilience against attacks [6], [7].

While detection-based approaches have shown some success in ensuring trustworthy inference [8], [9], [10], adversarial training remains a promising approach for robust model training [11], [12]. By carefully maximizing the model's loss, adversarial training methods introduce imperceptible noise to clean input training data to enhance robustness [13]. This approach typically involves solving an optimization problem and relies on the trained model or a pre-trained one [14]. However, solving the resultant optimization problem is often challenging, especially when tuning hyperparameters, which can be cumbersome in certain cases [15], [16]. Moreover, these methods often focus on per-datum processing to generate adversarial noise, simply disregarding the underlying data-generating process or distribution. Additionally, quantifying the robustness gained through such training methods remains a challenge [17], and is less explored in research.

Additional challenges arise from the prevalent assumption in many methods that data is confined to a single device. In reality, data is often generated and stored across distributed locations, each having subsets of data with potentially different distributions. While data localization is essential for privacy among other reasons, the federated learning (FL) paradigm aims to train a global model by coordinating multiple devices through a central parameter server [2], [18]. Existing FL approaches have primarily focused on the trade-offs between communication and computation when aggregating model updates from learners [19], [20], [21], with limited attention to robust FL methods. Among those addressing robust FL, some focus on robust aggregation strategies [22], [23], or a blend of periodic averaging with adaptive sampling [24], [25], while others target robust models and assume affine distribution shifts across workers' data see, e.g., [26]. However, these

Manuscript received 25 October 2023; revised 3 February 2024; accepted 19 March 2024. Date of publication 26 March 2024; date of current version 16 April 2024. The work of Alireza Sadeghi and Georgios B. Giannakis was supported in part by NSF under Grant 1901134, Grant 2128593, Grant 2126052, Grant 2212318, Grant 2220292, Grant 2312547, Grant 2103256, and Grant 2212318. *(Corresponding author: Alireza Sadeghi.)*

The authors are with the Department of ECE, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: sadeghi@umn.edu; gangwang@bit.edu.cn; georgios@umn.edu).

Digital Object Identifier 10.1109/JSAIT.2024.3381869

methods have limited scopes and do not fully account for underlying data distributions when designing training algorithms. In a recent effort to deal with discrepancies in data distributions across workers, [27] has extended standard results of optimal transport theory to the multi-marginal optimal transport problem and studied personalized federated learning. Albeit interesting, the proposed method requires an additional model per worker and does not target robustness at all. From a theoretical perspective, the optimal transport theory has further been investigated to quantify adversarial risk of classifiers on adversarially perturbed data [28], [29].

Our work leverages the distributionally robust optimization framework to develop procedures for training robust models against distributional uncertainties and adversarial data. It builds on the implicit assumption that training data samples are clean or nominal, but obtained from an "unknown" data distribution. The objective is to target distribution shifts of data from training to inference, and ensure robustness to adversarial inputs during the inference phase. Our algorithms enable learning robust models from multiple heterogeneous datasets with potentially different distributions stored at different locations. Unlike most existing works, our proposed algorithms account for the data distribution when creating adversarial perturbations. We obtain robust models by minimizing the worst-case expected loss over all distributions residing within a Wasserstein ball centered at the empirical data distribution. The resulting formulation leads to a challenging and intractable infinite-dimensional functional optimization problem, which we address by exploiting the strong duality to arrive at a tractable, and under certain conditions, an equivalent unconstrained and affordable minimization problem. To solve this robust surrogate problem, we develop a stochastic proximal gradient descent (SPGD) algorithm based on an ϵ-accurate oracle, along with its lightweight stochastic proximal gradient descent-ascent (SPGDA) iterations. The former algorithm uses an oracle to solve the involved convex sub-problems to ϵ-accuracy, while the latter approximates the solution through a single gradient ascent step. To further accommodate learning from distributed datasets, we develop a distributionally robust federated learning (DRFL) framework. In summary, the main contributions of this paper are as follows.

- A generic distributionally robust learning framework, designed to endow machine learning models with resilience against distributional uncertainties and adversarial input data encountered during inference.
- Two scalable distributionally robust learning algorithms, with rigorous theoretical convergence guarantees.
- A DRFL framework, enabling the training of robust models from multiple sources with potentially different underlying data distributions.

The rest of this paper is structured as follows. Problem formulation and its robust surrogate are the subjects of Section II. The proposed SPGD with ϵ-accurate oracle and SPGDA algorithms with their convergence analyses are presented in Sections III and IV, respectively. The DRFL implementation is discussed in Section V. Numerical tests are given in Section VI with conclusions drawn in Section VII. Technical proofs are deferred to the Appendix.

II. Problem Statement

Consider the standard regularized statistical learning task

$$\min_{\boldsymbol{\theta}\in\Theta} \ \mathbb{E}_{z\sim P_0}\big[\ell(\boldsymbol{\theta};z)\big]+r(\boldsymbol{\theta}) \tag{1}$$

where $\ell(\boldsymbol{\theta};z)$ denotes the loss of a model parameterized by the unknown parameter vector $\boldsymbol{\theta}$ on a datum $z=(\boldsymbol{x},y)\sim P_0$, with feature $\boldsymbol{x}$ and label y, drawn from some nominal distribution P_0. Here, Θ denotes the feasible set for model parameters. To prevent over fitting or incorporate prior information, regularization term $r(\boldsymbol{\theta})$ is oftentimes added to the expected loss. Popular regularizers include $r(\boldsymbol{\theta}):=\beta\|\boldsymbol{\theta}\|_1^2$ or $\beta\|\boldsymbol{\theta}\|_2^2$, where $\beta\geq 0$ is a hyper-parameter controlling the importance of the regularization term relative to the expected loss.

In practice, the nominal distribution P_0 is typically unknown. Instead, we are given with independently and identically distributed training data samples $\{z_n\}_{n=1}^N\sim P_0$. Leveraging these samples, we form the empirical data distribution $\widehat{P}_0^{(N)}$ to replace P_0 in (1) in *empirical* loss minimization

$$\min_{\boldsymbol{\theta}\in\Theta} \ \bar{\mathbb{E}}_{z\sim\widehat{P}_0^{(N)}}\big[\ell(\boldsymbol{\theta};z)\big]+r(\boldsymbol{\theta}) \tag{2}$$

where $\bar{\mathbb{E}}_{z\sim\widehat{P}_0^{(N)}}[\ell(\boldsymbol{\theta};z)]=N^{-1}\sum_{n=1}^N\ell(\boldsymbol{\theta};z_n)$. Indeed, a variety of machine learning tasks can be cast as (2), including, e.g., ridge and Lasso regression, logistic regression, and reinforcement learning. The resultant models obtained by solving (2) however, have been shown vulnerable to adversarially corrupted data in $\widehat{P}_0^{(N)}$. Further, the testing data distribution often deviates from the available $\widehat{P}_0^{(N)}$. For this reason, targeting an adversarially robust model against a set of distributions corresponding to perturbations of the underlying data distribution, has led to the formulation

$$\min_{\boldsymbol{\theta}\in\Theta} \ \sup_{P\in\mathcal{P}} \ \mathbb{E}_{z\sim P}[\ell(\boldsymbol{\theta};z)]+r(\boldsymbol{\theta}) \tag{3}$$

where $\mathcal{P}$ represents a set of distributions centered around the data generating distribution $\widehat{P}_0^{(N)}$. Compared with (1), the worst-case formulation (3), yields models ensuring reasonable performance across a continuum of distributions characterized by $\mathcal{P}$. In practice, different types of ambiguity sets $\mathcal{P}$ can be considered, and they lead to different robustness guarantees and computational requirements. Popular choices of $\mathcal{P}$ include momentum [30], [31], KL divergence [32], statistical test [33], and Wasserstein distance-based ambiguity sets [33], [34]; see, e.g., [35] for a recent overview. Among all choices, it has been shown that the Wasserstein ambiguity set $\mathcal{P}$ results in a tractable realization of (3), thanks to the strong duality result of [33] and [34], which also motivates this work.

To formalize this, consider two probability measures P and Q supported on set $\mathcal{Z}$, and let $\Pi(P,Q)$ be the set of all joint measures supported on $\mathcal{Z}^2$, with marginals P and Q. Let $c:\mathcal{Z}\times\mathcal{Z}\to[0,\infty)$ measure the cost of transporting a unit of mass from z in P to another element z' in Q. The celebrated optimal transport problem is given by [36, p. 111]

$$W_c(P,Q):=\inf_{\pi\in\Pi}\mathbb{E}_\pi\big[c\big(z,z'\big)\big]. \tag{4}$$

Remark 1: If $c(\cdot,\cdot)$ satisfies the axioms of distance, then W_c defines a distance on the space of probability measures. For

instance, if P and Q are defined over a Polish space equipped with metric d, then choosing $c(z,z') = d^p(z,z')$ for some $p \in [1,\infty)$ asserts that $W_c^{1/p}(P,Q)$ is the well-known Wasserstein distance of order p between probability measures P and Q [36, Definition 6.1].

For a given empirical distribution $\widehat{P}_0^{(N)}$, define the uncertainty set $\mathcal{P} := \{P | W_c(P, \widehat{P}_0^{(N)}) \leq \rho\}$ to include all probability distributions having at most ρ-distance from $P_0^{(N)}$. Incorporating this ambiguity set into (3), yields the following reformulation

$$\min_{\boldsymbol{\theta}\in\Theta} \sup_{P} \mathbb{E}_{z\sim P}[\ell(\boldsymbol{\theta};z)] + r(\boldsymbol{\theta}) \tag{5a}$$

$$\text{s.t. } W_c\Big(P, \widehat{P}_0^{(N)}\Big) \leq \rho. \tag{5b}$$

Observe that the inner supremum in (5a) runs over all joint probability measures π on $\mathcal{Z}^2$ implicitly characterized by (5b). Intuitively, directly solving this optimization over the infinite-dimensional space of distribution functions is challenging, if not impossible. Fortunately, for a broad range of losses as well as transport costs, it has been shown that the inner maximization satisfies a strong duality condition [35]; that is, the optimal objective of this inner maximization and its Lagrangian dual optimal objective, are equal. In addition, the dual problem involves optimization over a one-dimensional dual variable. These two observations make it possible to solve (3) in the dual domain. To formally obtain a tractable surrogate to (5), we make the following assumptions.

Assumption 1: The transportation cost function $c : \mathcal{Z} \times \mathcal{Z} \to [0,\infty)$, is a lower semi-continuous function that to satisfy well-posedness of the transportation problem and the existence of optimal transport maps it satisfies the following conditions:

1) Non-Negativity of transportation $c(z,z') \geq 0 \quad \forall z, z' \in \mathcal{Z} \times \mathcal{Z}'$, and $c(z,z) = 0, \forall z \in \mathcal{Z} \times \mathcal{Z}'$.
2) Boundedness to ensuring there is a lower bound on the transportation cost, i.e.,
$$0 < \inf_{z,z'} c(z,z') \leq \sup_{z,z'} c(z,z') < \infty, \quad \forall z, z', z \neq z'.$$
3) Measurability of transportation cost in the product measure $P \times Q$ on the product space $\mathcal{Z} \times \mathcal{Z}'$, ensuring involved integrals in (4) are well-defined.
4) Strong convexity: Transportation cost is strongly convex, i.e., having a positive definite Hessian
$$H^1 := \begin{bmatrix} \frac{\partial^2 c}{\partial z^2} & \frac{\partial^2 c}{\partial z \partial z'} \\ \frac{\partial^2 c}{\partial z' \partial z} & \frac{\partial^2 c}{\partial z'^2} \end{bmatrix} \succ \mathbf{0}. \tag{6}$$

Remark 2: Unless otherwise noted, throughout this paper, a generic transportation cost $c(z,z') : \mathcal{Z} \times \mathcal{Z}' \to \mathbb{R}^+$ satisfying Assumption 1 is considered. Therefore, the Wasserstein distance (4) is for such generic cost accordingly. No additional details are considered for this distance except having power of the Wasserstein distance to be equal to 1 for the sake of simplicity of derivatives.

Assumption 2: The loss function $\ell : \Theta \times \mathcal{Z} \to [0,\infty)$, is upper semi-continuous, and integrable.

The following proposition provides a tractable surrogate for (5), whose proof can be found in [35, Th. 1].

Proposition 1: Let $\ell : \Theta \times \mathcal{Z} \to [0,\infty)$, and $c : \mathcal{Z} \times \mathcal{Z} \to [0,\infty)$ satisfy Assumptions 1 and 2, respectively. Then, for any given $\widehat{P}_0^{(N)}$, and $\rho > 0$, it holds that

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{z\sim P}[\ell(\boldsymbol{\theta};z)] = \inf_{\gamma\geq 0} \Big\{\bar{\mathbb{E}}_{z\sim\widehat{P}_0^{(N)}}\big[\sup_{\zeta\in\mathcal{Z}}\{\ell(\boldsymbol{\theta};\zeta) - \gamma(c(z,\zeta)-\rho)\}\big]\Big\} \tag{7}$$

where $\mathcal{P} := \{P | W_c(P, \widehat{P}_0^{(N)}) \leq \rho\}$.

Remark 3: Thanks to strong duality, the right-hand side in (7) simply is a uni-variate dual reformulation of the primal problem represented in the left-hand side. In sharp contrast with the primal formulation, the expectation in the dual domain is taken only over the empirical $\widehat{P}_0^{(N)}$ rather than any $P \in \mathcal{P}$. In addition, since this reformulation circumvents the need for finding the optimal $\pi \in \Pi$ to form $\mathcal{P}$, and characterizing the primal objective $\forall P \in \mathcal{P}$, it is practically more convenient.

Upon relying on Proposition 1, the following distributionally robust surrogate is obtained

$$\min_{\boldsymbol{\theta}\in\Theta} \inf_{\gamma\geq 0} \Big\{\bar{\mathbb{E}}_{z\sim\widehat{P}_0^{(N)}}\big[\sup_{\zeta\in\mathcal{Z}}\{\ell(\boldsymbol{\theta};\zeta) + \gamma(\rho - c(z,\zeta))\} + r(\boldsymbol{\theta})\big]\Big\}. \tag{8}$$

Remark 4: The robust surrogate in (3) is reminiscent of minimax (saddle-point) optimization problems. Solvers for such problems have been recently investigated in several contributions see, e.g., [37] and references therein. While such solvers address the standard minimax problems, our formulated problem under (3), has its own specific challenges, posed by the structure of (8) to promote robustness, as well as taking into the account the data distribution when training the model. It is important to note that [37] and references therein offer valuable insights for conducting convergence analysis of primal-dual type solvers, as has been leveraged in this work.

It is worth noting that a relaxed (hence suboptimal) version of (8) with a fixed γ value has recently been studied in [34]. Unfortunately, one has to select an appropriate γ value using cross validation over a grid search that is also application dependent. Heuristically choosing a γ does not guarantee optimality in solving the distributionally robust surrogate (8). Clearly, the effect of heuristically selecting γ is more pronounced when training deep neural networks. Instead, we advocate algorithms that optimize γ and $\boldsymbol{\theta}$ simultaneously.

Our approach to addressing this, relies on the structure of (8) to *iteratively* update parameters $\bar{\boldsymbol{\theta}} := [\boldsymbol{\theta}^\top \ \gamma]^\top$ and ζ. To end up with a differentiable function of $\bar{\boldsymbol{\theta}}$ after maximizing over ζ, Danskin's theorem requires the sup-problem to have a unique solution [38]. For this reason, we design the inner maximization to involve a strongly concave objective function through the selection of a strongly convex transportation cost, such as $c(z,z') := \|z - z'\|_p^2$ for $p \geq 1$. For the maximization over ζ to rely on a strongly concave objective, we let

[1]Although not necessary in general, $c(\cdot,\cdot)$ is considered to be strongly convex. Note that all relevant derivatives discussed in subsequent sections are derived under the assumption that $c(\cdot,\cdot)$ represents a *generic transportation* cost, adhering to the specified Assumptions 1.

$\gamma \in \Gamma := \{\gamma | \gamma > \gamma_0\}$, where γ_0 is large enough. Since γ is the dual variable corresponding to the constraint in (5), having $\gamma \in \Gamma$ is tantamount to tuning ρ, which in turn *controls* the level of *robustness*. Replacing $\gamma \geq 0$ in (8) with $\gamma \in \Gamma$, our *robust learning model* is obtained as the solution of

$$\min_{\boldsymbol{\theta}\in\Theta} \inf_{\gamma\in\Gamma} \bar{\mathbb{E}}_{z\sim\widehat{P}_0^{(T)}}\Big[\sup_{\zeta\in\mathcal{Z}} \psi\Big(\bar{\boldsymbol{\theta}}, \zeta; z\Big)\Big] + r\big(\bar{\boldsymbol{\theta}}\big) \tag{9}$$

where

$$\psi\Big(\bar{\boldsymbol{\theta}}, \zeta; z\Big) := \ell(\boldsymbol{\theta}; \zeta) + \gamma(\rho - c(z, \zeta)). \tag{10}$$

Intuitively, input z in (9) is pre-processed by maximizing ψ accounting for the adversarial perturbation. To iteratively solve our objective in (9), the ensuing sections provide efficient solvers under some mild conditions. Those include cases, every inner maximization (supremum) can be solved to ϵ-optimality by an oracle.

Before developing our algorithms, we start by making several standard assumptions.

Assumption 3: Function $c(z, \cdot)$ is L_c-Lipschitz and μ-strongly convex for any given $z \in \mathcal{Z}$, with respect to the norm $\|\cdot\|$.

Assumption 4: The loss function $\ell(\boldsymbol{\theta}; z)$ satisfies the following Lipschitz smoothness conditions

$$\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}; z) - \nabla_{\boldsymbol{\theta}}\ell\big(\boldsymbol{\theta}'; z\big)\|_* \leq L_{\boldsymbol{\theta}\boldsymbol{\theta}}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \tag{11a}$$

$$\|\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}; z) - \nabla_{\boldsymbol{\theta}}\ell\big(\boldsymbol{\theta}; z'\big)\|_* \leq L_{\boldsymbol{\theta}z}\|z - z'\| \tag{11b}$$

$$\|\nabla_{z}\ell(\boldsymbol{\theta}; z) - \nabla_{z}\ell\big(\boldsymbol{\theta}; z'\big)\|_* \leq L_{zz}\|z - z'\| \tag{11c}$$

$$\|\nabla_{z}\ell(\boldsymbol{\theta}; z) - \nabla_{z}\ell\big(\boldsymbol{\theta}'; z\big)\|_* \leq L_{z\boldsymbol{\theta}}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \tag{11d}$$

and it is continuously differentiable with respect to $\boldsymbol{\theta}$.

Assumption (4) guarantees that the supremum in (8) results in a smooth function of $\bar{\boldsymbol{\theta}}$; thus, one can execute gradient descent to update $\boldsymbol{\theta}$ upon solving the supremum. This will further help to provide convergence analysis of our proposed algorithms. To elaborate more on this, the following lemma characterizes the smoothness and gradient Lipschitz properties obtained upon solving the maximization problem in (9).

Lemma 1: For each $z \in \mathcal{Z}$, let us define $\bar{\psi}(\bar{\boldsymbol{\theta}}; z) = \sup_{\zeta} \psi(\bar{\boldsymbol{\theta}}, \zeta; z)$ with $\zeta_*(\bar{\boldsymbol{\theta}}; z) := \arg\max_{\zeta\in\mathcal{Z}} \psi(\bar{\boldsymbol{\theta}}, \zeta; z)$. Then $\bar{\psi}(\cdot)$ is differentiable, and its gradient is $\nabla_{\bar{\boldsymbol{\theta}}}\bar{\psi}(\bar{\boldsymbol{\theta}}; z) = \nabla_{\bar{\boldsymbol{\theta}}}\psi(\bar{\boldsymbol{\theta}}, \zeta_*(\bar{\boldsymbol{\theta}}; z); z)$. Moreover, the following conditions hold

$$\big\|\zeta_*\big(\bar{\boldsymbol{\theta}}_1; z\big) - \zeta_*\big(\bar{\boldsymbol{\theta}}_2; z\big)\big\| \leq \frac{L_{z\boldsymbol{\theta}}}{\lambda}\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\| + \frac{L_c}{\lambda}\|\gamma_2 - \gamma_1\| \tag{12a}$$

and

$$\big\|\nabla_{\bar{\boldsymbol{\theta}}}\bar{\psi}\big(\bar{\boldsymbol{\theta}}_1; z\big) - \nabla_{\bar{\boldsymbol{\theta}}}\bar{\psi}\big(\bar{\boldsymbol{\theta}}_2; z\big)\big\| \leq \frac{L_{\boldsymbol{\theta}z}L_c + L_c^2}{\lambda}\|\gamma_2 - \gamma_1\| + \left(L_{\boldsymbol{\theta}\boldsymbol{\theta}} + \frac{L_{\boldsymbol{\theta}z}L_{z\boldsymbol{\theta}} + L_cL_{z\boldsymbol{\theta}}}{\lambda}\right)\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|. \tag{12b}$$

where $\gamma^{1,2} \in \Gamma$, and $\psi(\bar{\boldsymbol{\theta}}, \cdot; z)$ is λ-strongly concave.

Proof: See Appendix-A for the proof.

Lemma 1 paves the way for iteratively solving the surrogate optimization (9), intuitively because it guarantees a differentiable and smooth objective upon solving the inner supremum to its optimum.

Remark 5: Equation (12a) is appealing in practice. Indeed, if $\bar{\boldsymbol{\theta}}^t = [\boldsymbol{\theta}^t, \gamma^t]$ is updated with a small enough step size, the corresponding $\zeta_*(\boldsymbol{\theta}^{t+1}; z)$ is close enough to $\zeta_*(\boldsymbol{\theta}^t; z)$. Building on this observation, instead of using an oracle to find the optimum $\zeta_*(\boldsymbol{\theta}^{t+1}; z)$, an ϵ-accurate solution $\zeta_\epsilon(\boldsymbol{\theta}^{t+1}; z)$ suffices to obtain comparable performance. This also circumvents the need to find the optimum for the inner maximization per iteration, which could be computationally demanding.

III. Stochastic Proximal Gradient Descent With ϵ-Accurate Oracle

A standard solver of regularized optimization problems is the proximal gradient algorithm. In this section, we develop a variant of it to tackle the robust surrogate (9). For convenience, let us define

$$f(\boldsymbol{\theta}, \gamma) := \mathbb{E}\left[\sup_{\zeta\in\mathcal{Z}}\{\ell(\boldsymbol{\theta}; \zeta) + \gamma(\rho - c(z, \zeta))\}\right] \tag{13}$$

and rewrite our objective as

$$\min_{\boldsymbol{\theta}\in\Theta} \inf_{\gamma\in\Gamma} F(\boldsymbol{\theta}, \gamma) := f(\boldsymbol{\theta}, \gamma) + r(\boldsymbol{\theta}) \tag{14}$$

where $f(\boldsymbol{\theta}, \gamma)$ is the smooth function in (13), and $r(\cdot)$ is a non-smooth and convex regularizer, such as the ℓ_1-norm. With a slight abuse of notation, upon introducing $\bar{\boldsymbol{\theta}} := [\boldsymbol{\theta}\ \gamma]$, we define $f(\bar{\boldsymbol{\theta}}) := f(\boldsymbol{\theta}, \gamma)$ and $F(\bar{\boldsymbol{\theta}}) := F(\boldsymbol{\theta}, \gamma)$.

The proximal gradient algorithm the updates $\bar{\boldsymbol{\theta}}^t$ as

$$\bar{\boldsymbol{\theta}}^{t+1} = \arg\min_{\boldsymbol{\theta}} \alpha_t r(\boldsymbol{\theta}) + \alpha_t\langle\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^t, g\big(\bar{\boldsymbol{\theta}}^t\big)\rangle + \frac{1}{2}\big\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^t\big\|^2$$

where $g(\bar{\boldsymbol{\theta}}^t) := \nabla f(\bar{\boldsymbol{\theta}})|_{\bar{\boldsymbol{\theta}}=\bar{\boldsymbol{\theta}}^t}$, and $\alpha_t > 0$ is some step size. The last update is expressed in the compact form

$$\bar{\boldsymbol{\theta}}^{t+1} = \text{prox}_{\alpha_t r}\big[\bar{\boldsymbol{\theta}}^t - \alpha_t g\big(\bar{\boldsymbol{\theta}}^t\big)\big] \tag{15}$$

where the proximal gradient operator is given by

$$\text{prox}_{\alpha r}[\boldsymbol{v}] := \arg\min_{\boldsymbol{\theta}} \alpha r(\boldsymbol{\theta}) + \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{v}\|^2. \tag{16}$$

The working assumption is that this optimization problem can be solved efficiently using off-the-shelf solvers.

Starting from the guess $\bar{\boldsymbol{\theta}}^0$, the proposed SPGD with ϵ-accurate oracle executes two steps per iteration $t = 1, 2, \ldots$ First, it relies on an ϵ-accurate maximum oracle to solve the inner problem $\sup_{\zeta\in\mathcal{Z}}\{\ell(\boldsymbol{\theta}^t; \zeta) - \gamma^t c(z, \zeta)\}$ for randomly drawn samples $\{z_n\}_{n=1}^N$ to yield ϵ-optimal $\zeta_\epsilon(\bar{\boldsymbol{\theta}}^t, z_n)$ with the corresponding objective values $\psi(\bar{\boldsymbol{\theta}}^t, \zeta_\epsilon(\bar{\boldsymbol{\theta}}^t, z_n); z_n)$. Next, $\bar{\boldsymbol{\theta}}^t$ is updated using a stochastic proximal gradient step as

$$\bar{\boldsymbol{\theta}}^{t+1} = \text{prox}_{\alpha_t r}\Big[\bar{\boldsymbol{\theta}}^t - \frac{\alpha_t}{N}\sum_{n=1}^{N}\nabla_{\bar{\boldsymbol{\theta}}}\psi\Big(\bar{\boldsymbol{\theta}}, \zeta_\epsilon(\bar{\boldsymbol{\theta}}^t; z_n); z_n\Big)\Big].$$

For implementation, the proposed SPGD algorithm with ϵ-accurate oracle is summarized in Alg. 1. Convergence performance of this algorithm is analyzed in the ensuing subsection.

Algorithm 1: SPGD With ϵ-Accurate Oracle

Input : Initial guess $\bar{\theta}^0$, step size sequence $\{\alpha_t > 0\}_{t=0}^T$, ϵ-accurate oracle

for $t = 1, \ldots, T$ **do**
- Draw i.i.d samples $\{z_n\}_{n=1}^N$
- Find ϵ-optimizer $\zeta_\epsilon(\bar{\theta}^t; z_n)$ via the oracle
- Update:

$$\bar{\theta}^{t+1} = \text{prox}_{\alpha_t r}\Big[\bar{\theta}^t - \tfrac{\alpha_t}{N}\sum_{n=1}^N \nabla_{\bar{\theta}}\psi(\bar{\theta}, \zeta_\epsilon(\bar{\theta}^t; z_n); z_n)\big|_{\bar{\theta}=\bar{\theta}^t}\Big]$$

end

A. Convergence of SPGD With ϵ-Accurate Oracle

In general, the postulated model is nonlinear, and the robust surrogate $F(\bar{\theta})$ is nonconvex. In this section, we characterize the convergence performance of Alg. 1 to a stationary point. However, lack of convexity and smoothness implies that stationary points must be understood in the sense of the Frèchet subgradient. Specifically, the Frèchet subgradient $\partial F(\check{\theta})$ for the composite optimization in (14), is the set [39]

$$\partial F\Big(\check{\theta}\Big) := \left\{ v \,\Big|\, \liminf_{\bar{\theta}\to\check{\theta}} \frac{F(\bar{\theta}) - F\Big(\check{\theta}\Big) - v^\top\Big(\bar{\theta} - \check{\theta}\Big)\Big)}{\|\bar{\theta} - \check{\theta}\|} \geq 0 \right\}.$$

Consequently, the distance between vector $\mathbf{0}$ and the set $\partial F(\check{\theta})$ is a measure characterizing whether a point is stationary or not. To this end, define the distance between a vector v and a set $\mathcal{S}$ as $\text{dist}(v, \mathcal{S}) := \min_{s\in\mathcal{S}} \|v - s\|$, and the notion of δ-stationary points as defined next.

Definition 1: Given a small $\delta > 0$, we call vector $\check{\theta}$ a δ-stationary point if and only if $\text{dist}(\mathbf{0}, \partial F(\check{\theta})) \leq \delta$.

Since $f(\cdot)$ in (13) is smooth, we have that $\partial F(\bar{\theta}) = \nabla f(\bar{\theta}) + \partial r(\bar{\theta})$ [39]. Hence, it suffices to prove that the algorithm converges to a δ-stationary point $\check{\theta}$ satisfying

$$\text{dist}\Big(\mathbf{0}, \nabla f\Big(\check{\theta}\Big) + \partial r\Big(\check{\theta}\Big)\Big) \leq \delta. \tag{17}$$

We further adopt the following assumption that is standard in stochastic optimization.

Assumption 5: Function f satisfies the next two conditions.

1) Gradient estimates are unbiased and have a bounded variance, i.e., $\mathbb{E}[g^*(\bar{\theta}^t) - \nabla f(\bar{\theta}^t)] = \mathbf{0}$, and there is a constant $\sigma^2 < \infty$, so that $\mathbb{E}[\|\nabla f(\bar{\theta}^t) - g^*(\bar{\theta}^t)\|_2^2] \leq \sigma^2$.
2) Function $f(\bar{\theta})$ is smooth with L_f-Lipschitz continuous gradient, i.e., $\|\nabla f(\bar{\theta}_1) - \nabla f(\bar{\theta}_2)\| \leq L_f\|\bar{\theta}_1 - \bar{\theta}_2\|$.

We are now ready to claim the convergence guarantees for Alg. 1; see Appendix-B for the proof.

Theorem 1: Let Alg. 1 run for T iterations with constant step sizes $\alpha, \eta > 0$. Under Assumptions 1–5, Alg. 1 generates a sequence of $\{\bar{\theta}^t\}$ that satisfies

$$\mathbb{E}\left[\text{dist}\Big(\mathbf{0}, \partial F(\bar{\theta}^{t'})\Big)^2\right] \leq \left(\frac{2}{\alpha} + \beta\right)\frac{\Delta_F}{T} + \left(\frac{\beta}{\eta} + 2\right)\sigma^2 + \frac{(\beta+2)L_{\bar{\theta}z}^2\epsilon}{\lambda_0} \tag{18}$$

where t' is uniformly sampled from $\{1, \ldots, T\}$; here, $\Delta_F := F(\bar{\theta}^0) - F(\bar{\theta}^{T+1})$; $L_{\bar{\theta}z}^2 := L_{\bar{\theta}z}^2 + \lambda_0 L_c$, and β, $\lambda_0 > 0$ are some constants.

Theorem 1 asserts that $\{\bar{\theta}^t\}_{t=1}^T$ generated by Alg. 1 converges to a stationary point on average. The upper bound here is characterized by the initial error Δ_F, which decays at the rate of $\mathcal{O}(1/T)$; and, the constant bias terms induced by the gradient estimate variance σ^2 as well as the oracle accuracy ϵ.

Remark 6: . Note in Theorem 1 the established convergence rate of $O(1/T)$ to a stationary point is for the squared error distance, which, upon taking the square root, yields a rate of $O(1/\sqrt{T})$, which relies on a *constant* stepsize. By carefully designing the step size, it may be possible to obtain even a tighter upper bound. In addition, it is important to note that this Theorem demonstrates the attainment of a $\delta-$stationary point, where parameter δ depends on the accuracy of the oracle captured via ϵ, and the variance of the gradient estimates σ^2. Please observe that σ^2 depends on the batch size M, as is evident from equation (22). While we have not explicitly stated the dependency of σ^2 on M in Assumption 5, yet it is evident from equation (22), that a larger batch size should yield to a decrease in gradient estimate variance. All in all, intuitively having a larger batch size will yield a smaller variance for gradient estimates, i.e., $\delta \propto \sigma^2 \propto \frac{1}{M}$, and it also holds that $\delta \propto \epsilon$. Thus, having an accurate oracle with small $\epsilon \to 0$ with a larger batch size $M \to \infty$ results in obtaining a stationary point (i.e., $\delta \to 0$). These details give an intuitive explanation why the convergence guarantee comes with a bias term.

The computational complexity of Alg. 1 can grow prohibitively when dealing with large-size datasets and complex models. This motivates lightweight, scalable, yet efficient methods. To this end, we introduce next a stochastic proximal gradient descent-ascent (SPGDA) algorithm.

IV. Stochastic Proximal Gradient Descent-Ascent

Leveraging the strong concavity of the inner maximization problem and Lemma 1, a lightweight variant of the SPGD with ϵ-accurate oracle is developed here. Instead of optimizing the inner maximization problem to ϵ-accuracy by an oracle, we approximate its solution after only a *single* gradient ascent step. Specifically, for a batch of data $\{z_m^t\}_{m=1}^M$ per iteration t, our SPGDA algorithm first perturbs each datum via a gradient ascent step

$$\zeta_m^t = z_m^t + \eta_t \nabla_\zeta \psi\Big(\bar{\theta}^t, \zeta; z_m^t\Big)\big|_{\zeta=z_m^t}, \ \forall m = 1, \ldots, M \tag{19}$$

and then forms

$$g^t\Big(\bar{\theta}^t\Big) := \frac{1}{M}\sum_{m=1}^M \nabla_{\bar{\theta}}\psi\big(\bar{\theta}, \zeta_m^t; z_m^t\big)\big|_{\bar{\theta}=\bar{\theta}^t}. \tag{20}$$

Using (20), an extra proximal gradient step is taken to obtain

$$\bar{\theta}^{t+1} = \text{prox}_{\alpha_t r}\Big[\bar{\theta}^t - \alpha_t g^t\Big(\bar{\theta}^t\Big)\Big]. \tag{21}$$

The SPGDA steps are summarized in Alg. 2. Besides its simplicity and scalability, SPGDA enjoys convergence to a stationary point as elaborated next.

Algorithm 2: SPGDA

Input : Initial guess $\bar{\boldsymbol{\theta}}^0$, step size sequence $\{\alpha_t, \eta_t > 0\}_{t=0}^T$, batch size M

for $t = 1, \ldots, T$ **do**

  Draw a batch of i.i.d samples $\{z_m\}_{m=1}^M$

  Find $\{\boldsymbol{\zeta}_m^t\}_{m=1}^M$ via gradient ascent:

  $\boldsymbol{\zeta}_m^t = z_m^t + \eta_t \nabla_{\boldsymbol{\zeta}} \psi(\bar{\boldsymbol{\theta}}^t, \boldsymbol{\zeta}; z_m^t)\big|_{\boldsymbol{\zeta}=z_m^t}, \quad m = 1, \ldots, M$

  Update:

  $\bar{\boldsymbol{\theta}}^{t+1} = \text{prox}_{\alpha_t r}\Big[\bar{\boldsymbol{\theta}}^t - \frac{\alpha_t}{M} \sum_{m=1}^M \nabla_{\bar{\boldsymbol{\theta}}} \psi(\bar{\boldsymbol{\theta}}^t, \boldsymbol{\zeta}_m^t; z_m^t)\big|_{\bar{\boldsymbol{\theta}}=\bar{\boldsymbol{\theta}}^t}\Big]$

end

A. Convergence of SPGDA

To prove convergence of Alg. 2, let us start by defining

$$g^*\left(\bar{\boldsymbol{\theta}}^t\right) := \frac{1}{M} \sum_{m=1}^M \nabla_{\bar{\boldsymbol{\theta}}} \boldsymbol{\psi}^*\left(\bar{\boldsymbol{\theta}}^t, \boldsymbol{\zeta}_m^*; z_m^t\right). \tag{22}$$

Different from (20), the gradient here is obtained at the optimum $\boldsymbol{\zeta}_m^* = \boldsymbol{\zeta}_m^*(\bar{\boldsymbol{\theta}}; z) := \arg\max_{\boldsymbol{\zeta} \in \mathcal{Z}} \psi(\bar{\boldsymbol{\theta}}, \boldsymbol{\zeta}; z_m)$.

To establish convergence, one more assumption is needed.

Assumption 6: Function f satisfies the following conditions.

1) Gradient estimates $\nabla_{\bar{\boldsymbol{\theta}}} \boldsymbol{\psi}^*(\bar{\boldsymbol{\theta}}^t, \boldsymbol{\zeta}_m^*; z_m)$ at $\boldsymbol{\zeta}_m^*$ are unbiased and have bounded variance. That is, for $m = 1 \cdots M$, we have $\mathbb{E}\left[\nabla_{\bar{\boldsymbol{\theta}}} \boldsymbol{\psi}^*(\boldsymbol{\theta}, \boldsymbol{\zeta}_m^*; z_m) - \nabla_{\bar{\boldsymbol{\theta}}} f(\boldsymbol{\theta})\right] = \mathbf{0}$ and $\mathbb{E}\left[\|\nabla_{\bar{\boldsymbol{\theta}}} \boldsymbol{\psi}^*(\boldsymbol{\theta}, \boldsymbol{\zeta}_m^*; z_m) - \nabla f(\boldsymbol{\theta})\|^2\right] \leq \sigma^2$.
2) The expected norm of $g^t(\bar{\boldsymbol{\theta}})$ is bounded, that is, $\mathbb{E}\|g^t(\bar{\boldsymbol{\theta}})\|^2 \leq B^2$.[2]

We can now present a theorem on the convergence of Alg. 2; see Appendix-C for the proof.

Theorem 2 (Convergence of Alg. 2): Let $\Delta_F := F(\bar{\boldsymbol{\theta}}^0) - \inf_{\bar{\boldsymbol{\theta}}} F(\bar{\boldsymbol{\theta}})$, and D denote the diameter of the feasible set Θ. Under As. 1–4 and 6, for a constant step size $\alpha > 0$, and a fixed batch size $M > 0$, after T iterations, Alg. 2 satisfies

$$\mathbb{E}\left[\text{dist}\left(\mathbf{0}, \partial F(\bar{\boldsymbol{\theta}}^T)\right)^2\right] \leq \frac{\upsilon}{T+1}\Delta_F + \frac{2L_{\theta z}^2 \nu}{M}\left[(1-\alpha\mu)D^2 + \alpha^2 B^2\right] + \frac{4\sigma^2}{M} \tag{23}$$

where υ, ν, and $\mu = \gamma_0 - L_{zz}$ are some positive constants.

Theorem 2 implies that the sequence $\{\bar{\boldsymbol{\theta}}^t\}_{t=1}^T$ generated by Alg. 2 converges to a stationary point. The upper bound in (23) is characterized by a vanishing term induced by initial error Δ_F, and constant bias terms.

V. Distributionally Robust Federated Learning

In practice, massive datasets are distributed geographically across multiple sites, where scalability, data privacy and integrity, as well as bandwidth scarcity typically discourage uploading them to a central server. This has propelled the so-called federated learning framework, where multiple workers exchange information with a server to learn a centralized model using data locally generated and/or stored across workers [2]. Workers in this learning framework communicate *iteratively* with the server. Albeit appealing for its scalability, one needs to carefully address the bandwidth bottleneck associated with server-worker links. Furthermore, the workers' data may have (slightly) different distributions, which further challenges the learning task. To seek a model robust to distribution drifts across workers, we will adapt our novel SPGDA approach to design a privacy-respecting and robust algorithm.

To that end, consider K workers with each worker $k \in \mathcal{K}$ collecting samples $\{z_n(k)\}_{n=1}^N$. A globally shared model parameterized by $\boldsymbol{\theta}$ is to be updated at the server by aggregating gradients computed locally per worker. For simplicity, we consider workers having the same number of samples N. The goal is to learn a single global model from stored data at all workers by minimizing the following objective function

$$\min_{\boldsymbol{\theta} \in \Theta} \bar{\mathbb{E}}_{z \sim \widehat{P}}[\ell(\boldsymbol{\theta}; z)] + r(\boldsymbol{\theta}) \tag{24}$$

where $\bar{\mathbb{E}}_{z \sim \widehat{P}}[\ell(\boldsymbol{\theta}; z)] := \frac{1}{NK} \sum_{n=1}^N \sum_{k=1}^K \ell(\boldsymbol{\theta}, z_n(k))$. To endow the learned model with robustness against distributional uncertainties, our novel formulation will solve the following problem in a distributed fashion

$$\min_{\boldsymbol{\theta} \in \Theta} \sup_{P \in \mathcal{P}} \mathbb{E}_{z \sim P}[\ell(\boldsymbol{\theta}; z)] + r(\boldsymbol{\theta}) \quad \text{s. to. } \mathcal{P} := \left\{P \Big| \sum_{k=1}^K W_c\left(P, \widehat{P}^{(N)}(k)\right) \leq \rho\right\} \tag{25}$$

where $W_c(P, \widehat{P}^{(N)}(k))$ denotes the Wasserstein distance between distribution P and the local $\widehat{P}^{(N)}(k)$, per worker k.

Clearly, the constraint $P \in \mathcal{P}$ couples the optimization in (25) across all workers. To offer distributed implementations, we resort to Prop. 1, to arrive at the equivalent reformulation

$$\min_{\boldsymbol{\theta} \in \Theta} \inf_{\gamma \in \Gamma} \sum_{k=1}^K \left\{\bar{\mathbb{E}}_{z(k) \sim \widehat{P}^{(N)}(k)}\left[\sup_{\zeta \in \mathcal{Z}} \{\ell(\boldsymbol{\theta}; \zeta) + \gamma(\rho - c(z(k), \zeta))\}\right]\right\} + r(\boldsymbol{\theta}). \tag{26}$$

Next, we present our communication- and computation-efficient DRFL that builds on the SPGDA scheme in Section IV.

Specifically, our DRFL hinges on the fact that with fixed server parameters $\bar{\boldsymbol{\theta}}^t := [\boldsymbol{\theta}^{t\top}, \gamma^t]^\top$ per iteration t, the optimization problem becomes *separable* across all workers. Hence, upon receiving $\bar{\boldsymbol{\theta}}^t$ from the server, each worker $k \in \mathcal{K}$: i) samples a minibatch $\mathcal{B}^t(k)$ of data from $\widehat{P}^{(N)}(k)$; ii) forms the *perturbed* loss $\psi_k(\bar{\boldsymbol{\theta}}^t, \boldsymbol{\zeta}; z) := \ell(\boldsymbol{\theta}^t; \boldsymbol{\zeta}) + \gamma^t(\rho - c(z, \boldsymbol{\zeta}))$ for each $z \in \mathcal{B}^t(k)$; iii) lazily maximizes $\psi_k(\bar{\boldsymbol{\theta}}^t, \boldsymbol{\zeta}; z)$ over $\boldsymbol{\zeta}$ using a single gradient ascent step to yield $\boldsymbol{\zeta}(\bar{\boldsymbol{\theta}}^t; z) = z + \eta_t \nabla_{\boldsymbol{\zeta}} \psi_k(\bar{\boldsymbol{\theta}}^t, \boldsymbol{\zeta}; z)|_{\boldsymbol{\zeta}=z}$; and, iv) sends the stochastic gradient $|\mathcal{B}^t(k)|^{-1} \sum_{z \in \mathcal{B}^t(k)} \nabla_{\bar{\boldsymbol{\theta}}} \psi_k(\bar{\boldsymbol{\theta}}^t, \boldsymbol{\zeta}(\bar{\boldsymbol{\theta}}^t; z); z)|_{\bar{\boldsymbol{\theta}}=\bar{\boldsymbol{\theta}}^t}$ back to the

[2]Further details regarding this assumption are provided in Section VII-D.

Algorithm 3: DRFL

Input : Initial guess $\bar{\theta}^1$, a set of workers $\mathcal{K}$ with data samples $\{z_n(k)\}_{n=1}^N$ per worker $k \in \mathcal{K}$, step size sequence $\{\alpha_t, \eta_t > 0\}_{t=1}^T$

Output: $\bar{\theta}^{T+1}$

for $t = 1, \ldots, T$ **do**

  Each worker:

  Samples a minibatch $\mathcal{B}^t(k)$ of samples

  Given $\bar{\theta}^t$ and $z \in \mathcal{B}^t(k)$, forms local perturbed loss

$$\psi_k(\bar{\theta}^t, \zeta; z) := \ell(\bar{\theta}^t; \zeta) + \gamma^t(\rho - c(z, \zeta))$$

  Lazily maximizes $\psi_k(\bar{\theta}^t, \zeta; z)$ over ζ to find

$$\zeta(\bar{\theta}^t; z) = z + \eta_t \nabla_\zeta \psi_k(\bar{\theta}^t, \zeta; z)|_{\zeta=z}$$

  Computes stochastic gradient

$$\frac{1}{|\mathcal{B}^t(k)|} \sum_{z \in \mathcal{B}^t(k)} \nabla_{\bar{\theta}} \psi_k(\bar{\theta}^t, \zeta(\bar{\theta}^t; z); z)\big|_{\bar{\theta}=\bar{\theta}^t}$$

  and uploads to server

  Server:

  Updates $\bar{\theta}^t$ according to (27)

  Broadcasts $\bar{\theta}^{t+1}$ to workers

end

server. Upon receiving all local gradients, the server updates $\bar{\theta}^t$ using a proximal gradient descent step to find $\bar{\theta}^{t+1}$, that is

$$\bar{\theta}^{t+1} = \text{prox}_{\alpha_t r}\Bigg[\bar{\theta}^t - \frac{\alpha_t}{K}\sum_{k=1}^{K}\frac{1}{|\mathcal{B}^t(k)|} \times \sum_{z \in \mathcal{B}^t(k)} \nabla_{\bar{\theta}} \psi_k\Big(\bar{\theta}^t, \zeta(\bar{\theta}^t; z); z\Big)\big|_{\bar{\theta}=\bar{\theta}^t}\Bigg] \quad (27)$$

which is then broadcast to all workers to begin a new round of local updates. Our DRFL approach is tabulated in Alg. 3.

VI. Numerical Tests

To assess the performance in the presence of distribution drifts and adversarial perturbations, we will rely on empirical classification of standard MNIST and Fashion- (F-)MNIST datasets using standard cross-entropy loss. Specifically, we compare performance using models trained with empirical risk minimization (ERM), the fast-gradient method (FGSM) [13], its iterated variant (IFGM) [40], and the Wasserstein robust method (WRM) [34]. We further evaluate the testing performance using the projected gradient descent (PGD) attack [41]. We first test performance of SPGD with ϵ-accurate oracle, and SPGDA on these standard classification tasks.

A. SPGD With ϵ-Accurate Oracle and SPGDA

The FGSM attack performs one step gradient update along the direction of the gradient's sign to find an adversarial sample; that is,

$$x_{\text{adv}} = \text{Clip}_{[-1,1]}\{x + \epsilon_{\text{adv}}\text{sign}(\nabla \ell_x(\theta; (x, y)))\} \quad (28)$$

where ϵ_{avd} controls the maximum ℓ_∞ perturbation of adversarial samples. The element-wise $\text{Clip}_{[a,b]}\{\}$ operator forces its input to reside in the prescribed range $[-1, 1]$. By running T_{adv} iterations of (28) iterative (I) FGSM attack samples are generated [13]. Starting with an initialization $x_{\text{adv}}^0 = x$, and considering the ℓ_∞ norm, the PGD attack iterates [41]

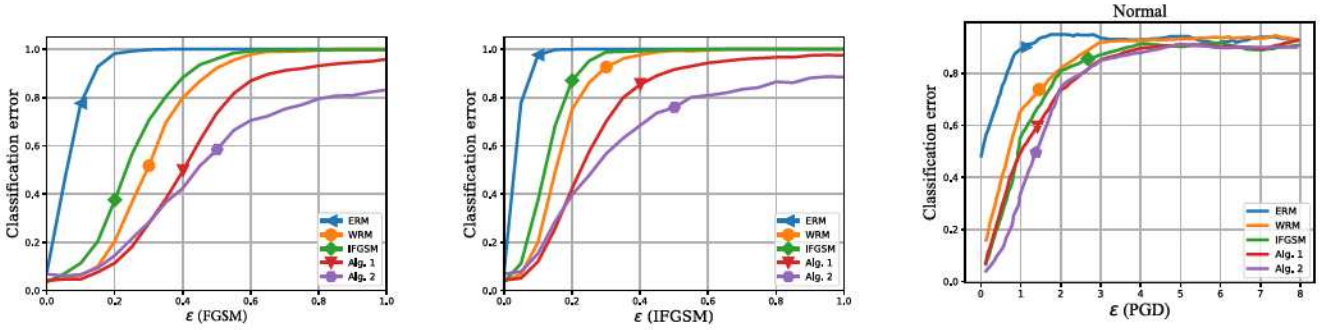
$$x_{\text{adv}}^{t+1} = \Pi_{\mathcal{B}_\epsilon(x_{\text{adv}}^t)}\big\{x_{\text{adv}}^t + \alpha\text{sign}\big(\nabla \ell_x(\theta; (x_{\text{adv}}^t, y))\big)\big\} \quad (29)$$

for T_{adv} steps, where Π denotes projection onto the ball $\mathcal{B}_\epsilon(x_{\text{adv}}^t) := \{x : \|x - x_{\text{adv}}^t\|_\infty \leq \epsilon_{\text{adv}}\}$, and $\alpha > 0$ is the stepsize set to 1 in our experiments. We use $T_{\text{adv}} = 10$ iterations for all iterative methods both in training and attack samples. The PGD can also be interpreted as an iterative algorithm that solves the optimization problem $\max_{x'} \ell(\theta; (x', y))$ subject to $\|x' - x\|_{\ell_\infty} \leq \alpha$. The Wasserstein attack on the other hand, generates adversarial samples by solving a perturbed training loss with an ℓ_2-based transportation cost associated with the Wasserstein distance between the training and adversarial data distributions [34].

Remark 7 (ϵ-accurate Oracle Implementation): To implement the ϵ-accurate oracle for Algorithm 1, several optimization algorithms can be employed. Here we rely on 10 iterations of gradient ascent with a constant step size $\eta = 0.001$. While this method provides only an approximate solution, in theory for large enough number of iterations it can generate ϵ-accurate solution. This holds specifically because the feasible set for γ is such that the function $\psi(\cdot)$ is guaranteed to be strongly concave, and thus with at least $O(1/(\epsilon\eta))$ gradient ascent steps an ϵ-accurate solution can be obtained.

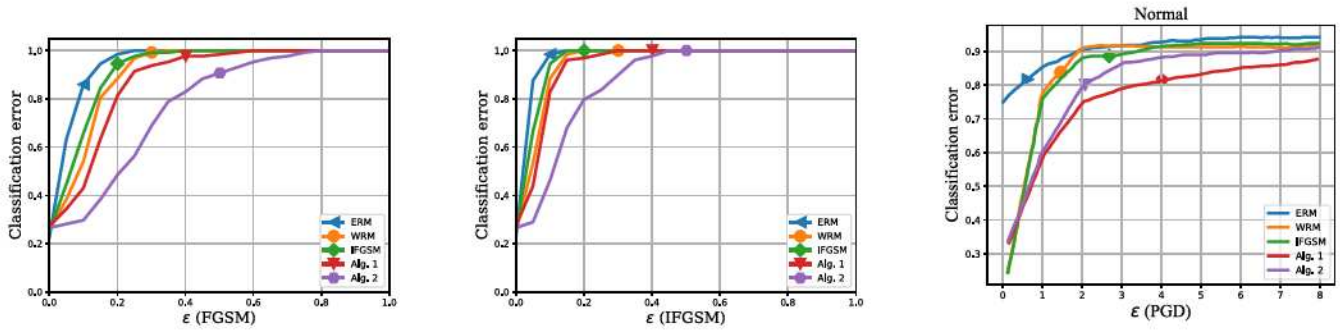
For the MNIST and F-MNIST datasets, a convolutional neural network (CNN) classifier consisting of 8×8, 6×6, and 5×5 filter layers with rectified linear units (ReLU) [42] and the same padding, is used. Its first, second, and third layers have 64, 128, and 128 channels, respectively, followed by a fully connected layer, and a softmax layer at the output.

CNNs with the same architecture are trained, using different adversarial samples. Specifically, to train a Wasserstein robust CNN model (WRM), $\gamma = 1$ was used to generate Wasserstein adversarial samples, ϵ_{adv} was set to 0.1 for the other two methods, and $\rho = 25$ was used to define the uncertainty set for both Algs. 1 and 2. Unless otherwise noted, we set the batch size to 128, the number of epochs to 30, the learning rates to $\alpha = 0.001$ and $\eta = 0.02$, and used the Adam optimizer [43]. Fig. 1(a) shows the classification error on the MNIST dataset. The error rates were obtained using testing samples generated according to the FGSM method with ϵ_{adv}. Clearly all training methods outperform ERM, and our proposed Algs. 1 and 2 offer improved performance over competing alternatives. The testing accuracy of all methods using samples generated according to an IFGSM attack is presented in Fig. 1(b). Likewise, Algs. 1 and 2 outperform other methods in this case. Fig. 1(c) depicts the testing accuracy of the considered methods under different levels of



(a) FGSM attack (b) IFGSM attack (c) PGD attack

Fig. 1. Misclassification error rate for different training methods using MNIST dataset.



(a) FGSM attack (b) IFGSM attack (c) PGD attack

Fig. 2. Misclassification error rate for different training methods using F-MNIST dataset.

a PGD attack. The plots in Fig. 1 showcase the improved performance obtained by CNNs trained using Algs. 1 and 2.

The F-MNIST article image dataset is adopted in the second experiment. Similar to the MNIST dataset, each example in F-MNIST is also a 28×28 gray-scale image, associated with a label from 10 classes. F-MNIST is a modern replacement for the original MNIST dataset for benchmarking machine learning algorithms. Using CNNs with similar architectures as before, the classification error is depicted for different training methods in Fig. 2. Three different attacks, namely FGSM, IFGSM, and PGD are used during testing. The resulting classification error rates are reported in Figs. 2(a), 2(b), and 2(c), respectively. The proposed SPGD and SPGDA algorithms outperform the other training methods, verifying the superiority of Algs. 1 and 2 in terms of yielding robust models.

B. Extended Numerical Tests on CIFAR-10 and CIFAR-100

In this section, we present an extended evaluation of our approach on more practical tasks using large-scale pre-trained CNNs for classification. Specifically, we leverage well-established models such as `ResNet-18` [44], `WideResNet-50-20` [45], and `VGG-11-BN` [46] for both 10-class and 100-class challenges on the CIFAR dataset. These models are initially pretrained on the ImageNet dataset [47] and fine-tuned on CIFAR-10 and CIFAR-100. CIFAR-10 comprises a total of $60,000$ color images of sizes 32×32 across 10 classes, with $50,000$ training images and $10,000$ for testing. On the other hand, CIFAR-100 features 600 images per class, with 500 for training and 100 for testing per class. To ensure fair comparisons, we kept hyperparameters consistent across experiments, utilizing SGD with a fixed learning rate of 0.001 and a momentum of 0.9 for fine-tuning over 15 epochs.

We conducted extensive adversarial testing using various attack strategies, including variance-tuning based (VNFIGSM) [48], ensemble adversarial (RFGSM) [49], transferable adversarial (TIFGSM) [50], transferable attack with input diversity (DIFGSM) [51], Nesterov accelerated scale invariance (NIFGSM) [52], Carlini and Wagner (CW) [53], KL-divergence based (TPGD) [54], and parameter-free attack (APGD) [55]. The hyperparameters used during adversarial testing in this case are reported in Table III in the Appendix.

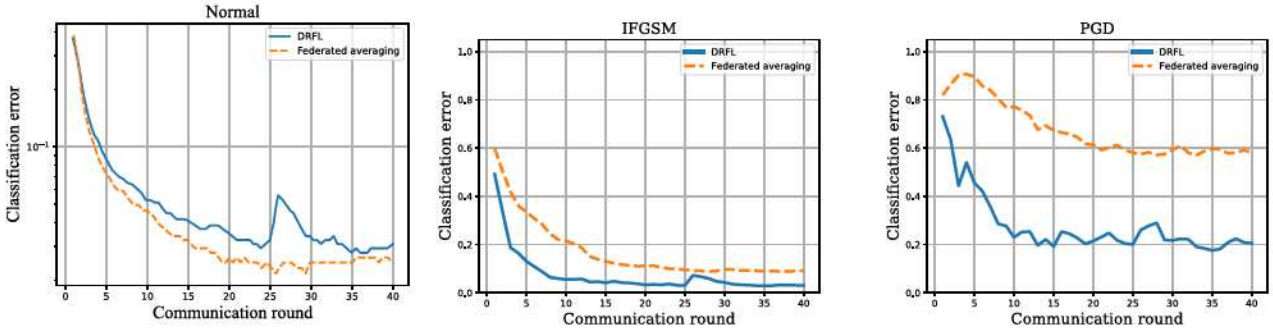
The results are presented in Tables I and II. The normal testing accuracies of all models for CIFAR-10 dataset were above 90% and for CIFAR-100 were above 78%, which shows that models were trained to classify normal input data accurately. The results in Tables I and II reveal that models trained using ERM demonstrate increased vulnerability to adversarial attacks compared to models trained using our robust method (Algorithm 1). This aligns with the expectation of the inherent vulnerability of conventionally trained models to adversarial perturbations. However, a notable finding is the consistent improvement in model performance when transitioning from ERM to robust models. This transition enhances performance across most adversarial testing scenarios while causing only a negligible drop in normal testing accuracy. Similar results were observed for the 100-class classification

TABLE I
TESTING ACCURACY RESULTS IN % FOR THE CIFAR-10 DATASET

Models	FGSM	TPGD	CW	PGD	APGD	TIFGSM	NIFGSM	DIFGSM	RFGSM	VNIFGSM
ResNet-18-**ERM**	37.03	27.57	0.45	0.57	0.68	2.30	0.55	1.29	0.54	0.88
ResNet-18-**Robust**	40.13	31.66	2.16	8.10	11.03	13.76	7.81	8.74	6.76	8.19
VGG-11-BN-**ERM**	28.75	24.72	0.06	0.16	0.24	1.16	0.16	0.33	0.16	0.71
VGG-11-BN-**Robust**	31.13	25.11	1.03	3.11	5.31	7.45	3.46	4.10	2.83	3.89
WideResNet-50-20-**ERM**	42.97	44.52	1.06	1.40	1.55	4.04	1.46	2.71	1.35	2.1
WideResNet-50-20-**Robust**	44.17	45.14	2.71	10.07	11.83	15.13	11.17	11.85	8.72	11.73

TABLE II
TESTING ACCURACY RESULTS IN % FOR THE CIFAR-100 DATASET

Models	FGSM	TPGD	CW	PGD	APGD	TIFGSM	NIFGSM	DIFGSM	RFGSM	VNIFGSM
ResNet-18-**ERM**	17.50	19.17	1.13	18.88	17.61	18.47	18.16	20.09	18.86	19.36
ResNet-18-**Robust**	21.13	24.34	4.54	22.97	21.69	21.84	21.53	24.88	21.93	24.52
VGG-11-BN-**ERM**	17.47	17.82	2.52	16.81	12.15	20.37	16.95	19.27	16.81	17.89
VGG-11-BN-**Robust**	21.02	21.92	6.26	18.84	15.31	20.70	19.65	21.70	19.79	21.79
WideResNet-50-20-**ERM**	20.05	21.44	5.87	21.02	17.78	22.87	21.24	22.69	21.00	21.66
WideResNet-50-20-**Robust**	23.54	25.06	9.45	23.83	18.36	26.71	24.67	26.52	23.63	23.92



(a) No attack (b) IFGSM attack (c) PGD attack

Fig. 3. Distributionally robust federated learning for image classification using the non-i.i.d. F-MNIST dataset.

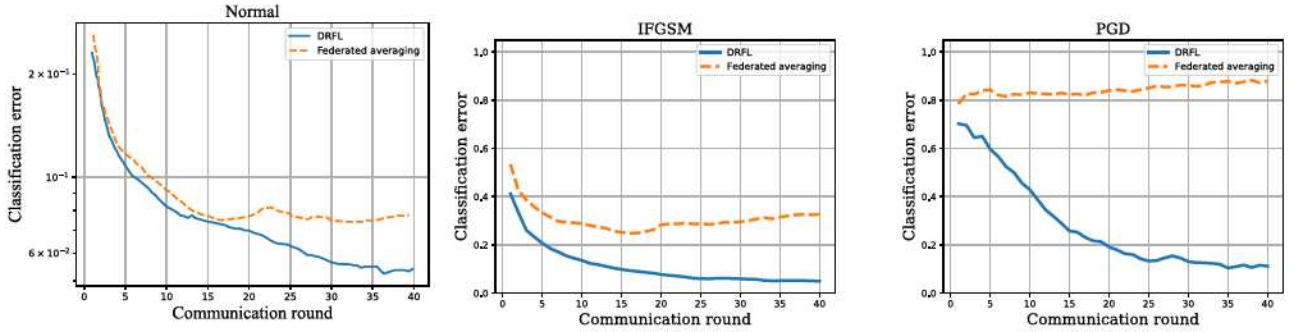
problem, which are reported in Table II. The hyperparameters used for testing are reported in Table IV in the Appendix.

These observations underscore the practical significance of designing and implementing distributionally robust algorithms. Furthermore, the consistency of these improvements across a range of models, including ResNet, VGG, and WideResNet, across different classification tasks, highlights the effectiveness and generalizability of our robust training algorithms in enhancing model reliability.

C. Distributionally Robust Federated Learning

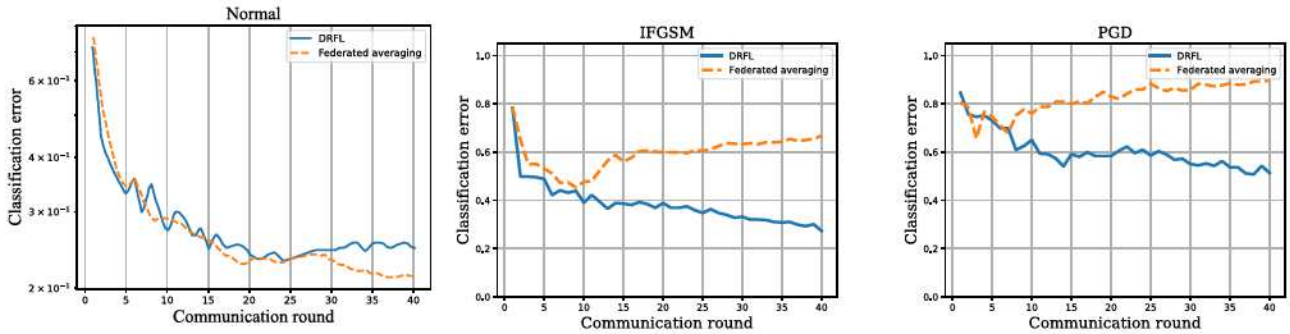
To validate the performance of our DRFL algorithm, we considered an FL environment consisting of a server and 10 workers, with local batch size 64, and assigned to every worker an equal-sized subset of training data containing i.i.d. samples from 10 different classes. All workers participated in each communication round. To benchmark the DRFL, we simulated the federated averaging method [56]. The testing accuracy on the MNIST dataset per communication round using clean (normal) images is depicted in Fig. 4(a). Clearly, both DRFL and federated averaging algorithms exhibit reasonable performance when the data is not corrupted. The performance is further tested against IFGSM and PGD attacks with a fixed $\epsilon_{\text{adv}} = 0.1$ during each communication round, and the corresponding misclassification error rates are shown in Figs. 4(b) and 4(c), respectively. The classification performance using federated averaging does not improve in Fig. 4(b), whereas the DRFL performance keeps improving across communication rounds. This is a direct consequence of accounting for the data uncertainties during the learning process. Moreover, Fig. 4(c) showcases that the federated averaging becomes even worse as the model gets progressively trained under the PGD attack. This indeed motivates our DRFL approach when data are from untrusted entities with possibly adversarial input perturbations. Similarly, Fig. 5 depicts the misclassification rate of the proposed DRFL method compared with federated averaging, when using the F-MNIST dataset.

As the distribution of data across devices may influence performance, we further considered a biased local data setting. In particular, each worker $k = 1, \ldots, 10$ has data from only one class, so the distributions at workers are highly perturbed, and data stored across workers are thus non-i.i.d. The testing error rate for normal inputs is reported in Fig. 3(a), while the test error against adversarial attacks is depicted in Figs. 3(b) and 3(c). This additional set of tests shows that having distributional shifts across workers can indeed enhance testing performance when the samples are adversarially manipulated.



(a) No attack (b) IFGSM attack (c) PGD attack

Fig. 4. Federated learning for image classification using the MNIST dataset.



(a) No attack (b) IFGSM attack (c) PGD attack

Fig. 5. Distributionally robust federated learning for image classification using F-MNIST dataset.

VII. Conclusion

A framework to robustify parametric machine learning models against distributional uncertainties was put forth in this paper. The learning task was cast as a distributionally robust optimization problem, for which two scalable stochastic optimization algorithms were developed. The first algorithm relies on an ϵ-accurate maximum-oracle to solve the inner convex subproblem, while the second approximates its solution via a single gradient ascent step. Convergence guarantees for both algorithms to a stationary point were obtained. The upshot of the proposed approach is that it is amenable to federated learning from unreliable datasets across multiple workers. The novel DRFL algorithm ensures data privacy and integrity, while offering robustness with minimal computational and communication overhead. Numerical tests for classifying standard real images showcased the merits of the proposed algorithms against distributional uncertainties and adversaries. This study uncovers promising avenues for future research, particularly in the domain of personalization. A key focus is on exploring the integration of distributionally robust learning with personalization. In this context, there is potential to extend this approach to heterogeneous datasets by incorporating the concept of multi-distribution Wasserstein distance [27]. This would involve utilizing *two* distinct models—one globally shared and another dedicated to aligning local data distributions within a common probability space. Further advancements in this research direction can be achieved by leveraging Bayesian and ensemble learning techniques to enhance both the robustness and personalization of federated learning methods.

Appendix

A. Proof of Lemma 1

Since function $\zeta \mapsto \psi(\bar{\theta}, \zeta; z)$ is λ-strongly concave, then $\zeta_*(\bar{\theta}) = \sup_{\zeta\in\mathcal{Z}} \psi(\bar{\theta}, \zeta; z)$ is unique. In addition, the first-order optimality condition gives $\langle \nabla_\zeta \psi(\bar{\theta}, \zeta_*(\bar{\theta}); z), \zeta - \zeta_*(\bar{\theta})\rangle \le 0$. Let us define $\zeta_*^1 = \zeta_*(\bar{\theta}_1)$, $\zeta_*^2 = \zeta_*(\bar{\theta}_2)$, and use the strong concavity for any $\bar{\theta}_1$ and $\bar{\theta}_2$, to write

$$\psi\left(\bar{\theta}_2, \zeta_*^2; z\right) \le \psi\left(\bar{\theta}_2, \zeta_*^1; z\right) + \left\langle \nabla_\zeta \psi\left(\bar{\theta}_2, \zeta_*^1; z\right), \zeta_*^2 - \zeta_*^1 \right\rangle - \frac{\lambda}{2}\|\zeta_*^2 - \zeta_*^1\|^2 \tag{30}$$

and

$$\begin{aligned}\psi\left(\bar{\theta}_2, \zeta_*^1; z\right) &\le \psi\left(\bar{\theta}_2, \zeta_*^2; z\right) + \left\langle \nabla_\zeta \psi\left(\bar{\theta}_2, \zeta_*^2; z\right), \zeta_*^1 - \zeta_*^2 \right\rangle - \frac{\lambda}{2}\|\zeta_*^2 - \zeta_*^1\|^2 \\ &\le \psi\left(\bar{\theta}_2, \zeta_*^2; z\right) - \frac{\lambda}{2}\|\zeta_*^2 - \zeta_*^1\|^2 \end{aligned} \tag{31}$$

where the last inequality is a consequence of the first-order optimality condition. Summing (30) and (31), we find that

$$\begin{aligned}\lambda\|\zeta_*^2 - \zeta_*^1\|^2 &\le \left\langle \nabla_\zeta \psi\left(\bar{\theta}_2, \zeta_*^1; z\right), \zeta_*^2 - \zeta_*^1 \right\rangle \\ &\le \left\langle \nabla_\zeta \psi\left(\bar{\theta}_2, \zeta_*^1; z\right), \zeta_*^2 - \zeta_*^1 \right\rangle - \left\langle \nabla_\zeta \psi\left(\bar{\theta}_1, \zeta_*^1; z\right), \zeta_*^2 - \zeta_*^1 \right\rangle \\ &= \left\langle \nabla_\zeta \psi\left(\bar{\theta}_2, \zeta_*^1; z\right) - \nabla_\zeta \psi\left(\bar{\theta}_1, \zeta_*^1; z\right), \zeta_*^2 - \zeta_*^1 \right\rangle. \end{aligned} \tag{32}$$

And using Hölder's inequality, we obtain that

$$\lambda\left\|\zeta_*^2-\zeta_*^1\right\|^2 \le \left\|\nabla_\zeta\psi\left(\bar\theta_2,\zeta_*^1;z\right)-\nabla_\zeta\psi\left(\bar\theta_1,\zeta_*^1;z\right)\right\|_\star\left\|\zeta_*^2-\zeta_*^1\right\| \quad (33)$$

from which we deduce

$$\left\|\zeta_*^2-\zeta_*^1\right\| \le \frac{1}{\lambda}\left\|\nabla_\zeta\psi\left(\bar\theta_2,\zeta_*^1;z\right)-\nabla_\zeta\psi\left(\bar\theta_1,\zeta_*^1;z\right)\right\|_\star. \quad (34)$$

Using $\psi(\bar\theta,\zeta;z) \coloneqq \ell(\theta;\zeta)+\gamma(\rho-c(z,\zeta))$, we have that

$$\begin{aligned}&\left\|\nabla_\zeta\psi\left(\bar\theta_2,\zeta_*^1;z\right)-\nabla_\zeta\psi\left(\bar\theta_1,\zeta_*^1;z\right)\right\|_\star\\ &=\Big\|\nabla_\zeta\ell\left(\theta_2;\zeta_*^1\right)-\nabla_\zeta\ell\left(\theta_1;\zeta_*^1\right)\\ &\quad+\gamma_1\nabla_\zeta c\left(z,\zeta_*^1\right)-\gamma_2\nabla_\zeta c\left(z,\zeta_*^1\right)\Big\|_\star\\ &\le\left\|\nabla_\zeta\ell\left(\theta_2;\zeta_*^1\right)-\nabla_\zeta\ell\left(\theta_1;\zeta_*^1\right)\right\|_\star\\ &\quad+\left\|\gamma_1\nabla_\zeta c\left(z,\zeta_*^1\right)-\gamma_2\nabla_\zeta c\left(z,\zeta_*^1\right)\right\|_\star\\ &\le L_{z\theta}\|\theta_2-\theta_1\|+\left\|\nabla_\zeta c\left(z,\zeta_*^1\right)\right\|_\star\|\gamma_2-\gamma_1\|. \quad (35)\end{aligned}$$

Substituting (35) into (34), yields

$$\begin{aligned}\|\zeta_*^2-\zeta_*^1\| &\le \frac{L_{z\theta}}{\lambda}\|\theta_2-\theta_1\|+\frac{1}{\lambda}\|\nabla_\zeta c\left(z,\zeta_*^1\right)\|_\star\|\gamma_2-\gamma_1\|\\ &\le \frac{L_{z\theta}}{\lambda}\|\theta_2-\theta_1\|+\frac{L_c}{\lambda}\|\gamma_2-\gamma_1\| \quad (36)\end{aligned}$$

where the last inequality holds because $\zeta\mapsto c(z,\zeta)$ is L_c-Lipschitz as per Assumption 3.

To obtain (12b), we first suppose without loss of generality that only a single datum z is given, and in order to prove existence of the gradient of $\bar\psi(\bar\theta,z)$ with respect to $\bar\theta$, we resort to the Danskin's theorem as follows.

Danskin's Theorem [57]: Consider the following minimax optimization problem

$$\min_{\theta\in\Theta}\max_{\zeta\in\mathcal{X}} f(\theta,\zeta) \quad (37)$$

where $\mathcal{X}$ is a nonempty compact set, and $f:\Theta\times\mathcal{X}\to[0,\infty)$ is such that $f(\cdot,\zeta)$ is differentiable for any $\zeta\in\mathcal{X}$, and $\nabla_\theta f(\theta,\zeta)$ is continuous on $\Theta\times\mathcal{X}$. With $\mathcal{S}(\theta)\coloneqq\{\zeta_*|\zeta_*=\arg\max_\zeta f(\theta,\zeta)\}$, the function

$$\bar f(\theta)\coloneqq\max_{\zeta\in\mathcal{Z}} f(\theta,\zeta)$$

is locally Lipschitz and directionally differentiable, where the directional derivatives satisfy

$$\bar f(\theta,d)=\sup_{\zeta\in\mathcal{S}(\theta)}\langle d,\nabla_\theta f(\theta,\zeta)\rangle. \quad (38)$$

For a given θ, if the set $\mathcal{S}(\theta)$ is a singleton, then the function $\bar f(\theta)$ is differentiable at θ with gradient

$$\nabla_\theta\bar f(\theta)=\nabla_\theta f\left(\theta,\zeta_*(\theta)\right). \quad (39)$$

Given θ, and the μ-strongly convex $c(z,\cdot)$, function $\psi(\bar\theta,\cdot;z)$ is concave if $L_{zz}-\gamma\mu<0$, which holds true for $\gamma_0>L_{zz}/\mu$. Replacing $\bar f(\theta,\zeta)$ with $\psi(\bar\theta,\zeta;z)$, and given the concavity of $\zeta\mapsto\psi(\bar\theta,\zeta;z)$, we have that $\bar\psi(\bar\theta;z)$ is a continuous function with gradient

$$\nabla_{\bar\theta}\bar\psi\left(\bar\theta;z\right)=\nabla_{\bar\theta}\bar\psi\left(\bar\theta,\zeta_*(\bar\theta;z);z\right). \quad (40)$$

We can then obtain the second inequality, as

$$\begin{aligned}&\left\|\nabla_{\bar\theta}\psi\left(\bar\theta_1,\zeta_*^1;z\right)-\nabla_{\bar\theta}\psi\left(\bar\theta_2,\zeta_*^2;z\right)\right\|\\ &\le\left\|\nabla_{\bar\theta}\psi\left(\bar\theta_1,\zeta_*^1;z\right)-\nabla_{\bar\theta}\psi\left(\bar\theta_1,\zeta_*^2;z\right)\right\|\\ &\quad+\left\|\nabla_{\bar\theta}\psi\left(\bar\theta_1,\zeta_*^2;z\right)-\nabla_{\bar\theta}\psi\left(\bar\theta_2,\zeta_*^2;z\right)\right\|\\ &\le\left\|\begin{bmatrix}\nabla_\theta\ell\left(\theta_1,\zeta_*^1\right)-\nabla_\theta\ell\left(\theta_1,\zeta_*^2\right)\\ c\left(z,\zeta_*^2\right)-c\left(z,\zeta_*^1\right)\end{bmatrix}\right\|\\ &\quad+\left\|\begin{bmatrix}\nabla_\theta\ell\left(\theta_1,\zeta_*^2\right)-\nabla_\theta\ell\left(\theta_2,\zeta_*^2\right)\\ 0\end{bmatrix}\right\|\\ &\le L_{\theta z}\left\|\zeta_*^1-\zeta_*^2\right\|+L_c\left\|\zeta_*^1-\zeta_*^2\right\|+L_{\theta\theta}\|\theta_1-\theta_2\|\\ &\le\left(L_{\theta\theta}+\frac{L_{\theta z}L_{z\theta}+L_cL_{z\theta}}{\lambda}\right)\|\theta_2-\theta_1\|\\ &\quad+\frac{L_{\theta z}L_c+L_c^2}{\lambda}\|\gamma_2-\gamma_1\| \quad (41)\end{aligned}$$

where we again used inequality (36). As a technical note, if the considered model is a neural network with a non-smooth activation function, the loss will not be continuously differentiable. However, we will not encounter this challenge often in practice.

B. Proof of Theorem 1

With slight abuse of notation, define for convenience $F(\theta,\gamma)\coloneqq f(\theta,\gamma)+r(\theta)+h(\gamma)$, where $h(\gamma)$ is the indicator function

$$h(\gamma)=\begin{cases}0, & \text{if } \gamma\in\Gamma\\ \infty, & \text{if } \gamma\notin\Gamma\end{cases} \quad (42)$$

with $\Gamma\coloneqq\{\gamma|\gamma\ge\gamma_0\}$, and for ease of representation we use $\bar r(\bar\theta)\coloneqq r(\theta)+h(\gamma)$. Having an L_f–smooth function f, yields

$$f\left(\bar\theta^{t+1}\right)\le f\left(\bar\theta^t\right)+\left\langle\nabla f\left(\bar\theta^t\right),\bar\theta^{t+1}-\bar\theta^t\right\rangle+\frac{L_f}{2}\left\|\bar\theta^{t+1}-\bar\theta^t\right\|^2. \quad (43)$$

For a given z^t, the gradients are

$$g^*\left(\bar\theta^t\right)\coloneqq\begin{bmatrix}\nabla_\theta\psi\left(\theta^t,\gamma,\zeta_*(\bar\theta^t;z^t);z^t\right)\\ \partial_\gamma\psi\left(\theta^t,\gamma,\zeta_*(\bar\theta^t;z^t);z^t\right)\end{bmatrix}=\begin{bmatrix}\nabla_\theta\psi\left(\theta^t,\gamma,\zeta_*(\bar\theta^t;z^t);z^t\right)\\ \rho-c(z^t,\zeta_*\left(\bar\theta^t;z^t\right)\end{bmatrix}.$$

and

$$g^\epsilon\left(\bar\theta^t\right)\coloneqq\begin{bmatrix}\nabla_\theta\psi\left(\theta^t,\gamma,\zeta_\epsilon(\bar\theta^t;z^t);z^t\right)\\ \partial_\gamma\psi\left(\theta^t,\gamma,\zeta_\epsilon(\bar\theta^t;z^t);z^t\right)\end{bmatrix}=\begin{bmatrix}\nabla_\theta\psi\left(\theta^t,\gamma,\zeta_\epsilon(\bar\theta^t;z^t);z^t\right)\\ \rho-c(z^t,\zeta_\epsilon\left(\bar\theta^t;z^t\right)\end{bmatrix}$$

obtained by an oracle at the optimal ζ_* and the ϵ-optimal ζ_ϵ solvers, respectively. Now, we define the error vector $\delta(\bar{\theta}^t) \coloneqq \nabla f(\bar{\theta}^t) - g^\epsilon(\bar{\theta}^t)$, and replace this into (43), to obtain

$$f\left(\bar{\theta}^{t+1}\right) \leq f\left(\bar{\theta}^t\right) + \left\langle g^\epsilon\left(\bar{\theta}^t\right) + \delta\left(\bar{\theta}^t\right), \bar{\theta}^{t+1} - \bar{\theta}^t \right\rangle + \frac{L_f}{2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2. \tag{44}$$

The following properties hold equivalently for the proximal operator, and for any $\boldsymbol{x}, \boldsymbol{y}$

$$\boldsymbol{u} = \text{prox}_{\alpha r}(\boldsymbol{x}) \iff (\boldsymbol{x} - \boldsymbol{u})^T(\boldsymbol{y} - \boldsymbol{u}) \leq \alpha r(\boldsymbol{y}) - \alpha r(\boldsymbol{u}). \tag{45}$$

With $\boldsymbol{u} = \bar{\theta}^{t+1}$ and $\boldsymbol{x} = \bar{\theta}^t - \alpha_t g^\epsilon(\bar{\theta}^t)$ in (45), it holds that

$$\left\langle \bar{\theta}^t - \alpha_t g^\epsilon\left(\bar{\theta}^t\right) - \bar{\theta}^{t+1}, \bar{\theta}^t - \bar{\theta}^{t+1} \right\rangle \leq \alpha_t \bar{r}\left(\bar{\theta}^t\right) - \alpha_t \bar{r}\left(\bar{\theta}^{t+1}\right)$$

and upon rearranging, we obtain

$$\left\langle g^\epsilon\left(\bar{\theta}^t\right), \bar{\theta}^{t+1} - \bar{\theta}^t \right\rangle \leq \bar{r}\left(\bar{\theta}^t\right) - \bar{r}\left(\bar{\theta}^{t+1}\right) - \frac{1}{\alpha_t}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2. \tag{46}$$

Adding inequalities in (46) and (44) gives

$$f\left(\bar{\theta}^{t+1}\right) \leq f\left(\bar{\theta}^t\right) + \left\langle \delta\left(\bar{\theta}^t\right), \bar{\theta}^{t+1} - \bar{\theta}^t \right\rangle + \frac{L_f}{2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \bar{r}\left(\bar{\theta}^t\right) - \bar{r}\left(\bar{\theta}^{t+1}\right) - \frac{1}{\alpha_t}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2$$

and with $F(\bar{\theta}) \coloneqq f(\bar{\theta}) + \bar{r}(\bar{\theta})$, we can write

$$F\left(\bar{\theta}^{t+1}\right) - F\left(\bar{\theta}^t\right) \leq \left\langle \delta\left(\bar{\theta}^t\right), \bar{\theta}^{t+1} - \bar{\theta}^t \right\rangle + \left(\frac{L_f}{2} - \frac{1}{\alpha_t}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2. \tag{47}$$

Using Young's inequality for any $\eta > 0$ gives $\langle \delta(\bar{\theta}^t), \bar{\theta}^{t+1} - \bar{\theta}^t \rangle \leq \frac{\eta}{2}\|\bar{\theta}^{t+1} - \bar{\theta}^t\|^2 + \frac{1}{2\eta}\|\delta(\bar{\theta}^t)\|^2$, and hence

$$F\left(\bar{\theta}^{t+1}\right) - F\left(\bar{\theta}^t\right) \leq \left(\frac{L_f + \eta}{2} - \frac{1}{\alpha_t}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \frac{\left\|\delta\left(\bar{\theta}^t\right)\right\|^2}{2\eta}. \tag{48}$$

Next, we will bound $\delta(\bar{\theta}^t) \coloneqq \nabla f(\bar{\theta}^t) - g^\epsilon(\bar{\theta}^t)$. By adding and subtracting $g^*(\bar{\theta}^t)$ to the right hand side, we find

$$\left\|\delta\left(\bar{\theta}^t\right)\right\|^2 \leq 2\left\|\nabla f\left(\bar{\theta}^t\right) - g^*\left(\bar{\theta}^t\right)\right\|^2 + 2\left\|g^*\left(\bar{\theta}^t\right) - g^\epsilon\left(\bar{\theta}^t\right)\right\|^2. \tag{49}$$

The Lipschitz smoothness of the gradient, implies that

$$\begin{aligned}
&\left\|g^*\left(\bar{\theta}^t\right) - g^\epsilon\left(\bar{\theta}^t\right)\right\|^2 \\
&= \left\|\begin{bmatrix} \nabla_\theta \psi\left(\theta^t, \gamma, \zeta_*(\bar{\theta}^t; z^t); z^t\right) \\ \rho - cz^t, \zeta_*(\bar{\theta}^t; z^t) \end{bmatrix} - \begin{bmatrix} \nabla_\theta \psi(\theta^t, \gamma, \zeta_\epsilon(\bar{\theta}^t; z^t); z^t) \\ \rho - c(z^t, \zeta_\epsilon(\bar{\theta}^t; z^t) \end{bmatrix}\right\|^2 \\
&= \left\|\nabla_\theta \psi(\theta^t, \gamma, \zeta_*(\bar{\theta}^t; z^t); z^t) - \nabla_\theta \psi(\theta^t, \gamma, \zeta_\epsilon(\bar{\theta}^t; z^t); z^t)\right\|^2 + \left\|c(z^t, \zeta_*^t) - c(z^t, \zeta_\epsilon^t)\right\|^2 \\
&\overset{(a)}{\leq} \left(\frac{L_{\theta z}^2}{\lambda^t} + L_c\right)\|\zeta_*^t - \zeta_\epsilon^t\|^2 \\
&\overset{(b)}{\leq} \left(\frac{L_{\theta z}^2}{\lambda^t} + L_c\right)\epsilon \\
&\leq \left(\frac{L_{\theta z}^2}{\lambda_0} + L_c\right)\epsilon
\end{aligned} \tag{50}$$

where (a) uses the $\lambda^t = \mu\gamma^t - L_{zz}$ strong-concavity of $\zeta \mapsto \psi(\bar{\theta}, \gamma, \zeta; z)$, and the second term is bounded by $L_c\|\zeta_*^t - \zeta_\epsilon^t\|^2$ according to Assumption 3. The last inequality holds for $\lambda_0 \coloneqq \mu\gamma_0 - L_{zz}$, where we used (42) to bound $\gamma^t \geq \gamma_0 > L_{zz}$. So far, we have established that

$$\left\|g^*\left(\bar{\theta}^t\right) - g^\epsilon\left(\bar{\theta}^t\right)\right\|^2 \leq \frac{L_{\bar{\theta}z}^2 \epsilon}{\lambda_0} \tag{51}$$

where for notational convenience we let $L_{\bar{\theta}z}^2 \coloneqq L_{\theta z}^2 + \lambda_0 L_c$. Substituting (51) into (49), the error can be bounded as

$$\left\|\delta\left(\bar{\theta}^t\right)\right\|^2 \leq 2\left\|\nabla f\left(\bar{\theta}^t\right) - g^*\left(\bar{\theta}^t\right)\right\|^2 + \frac{2L_{\bar{\theta}z}^2 \epsilon}{\lambda_0}. \tag{52}$$

Combining (48) and (50) yields

$$F\left(\bar{\theta}^{t+1}\right) - F\left(\bar{\theta}^t\right) \leq \left(\frac{L_f + \eta}{2} - \frac{1}{\alpha_t}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \frac{1}{\eta}\left\|\nabla f\left(\bar{\theta}^t\right) - g^*\left(\bar{\theta}^t\right)\right\|^2 + \frac{L_{\bar{\theta}z}^2 \epsilon}{\eta\lambda_0}. \tag{53}$$

Considering a constant step size α and summing these inequalities over $t = 1, \ldots, T$ yields

$$\left(\frac{1}{\alpha} - \frac{L_f + \eta}{2}\right)\sum_{t=0}^{T}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \leq F\left(\bar{\theta}^0\right) - F\left(\bar{\theta}^T\right) + \frac{1}{\eta}\sum_{t=0}^{T}\left\|\nabla f\left(\bar{\theta}^t\right) - g^*\left(\bar{\theta}^t\right)\right\|^2 + \frac{(T+1)L_{\bar{\theta}z}^2 \epsilon}{\lambda_0}. \tag{54}$$

From the proximal gradient update

$$\bar{\theta}^{t+1} = \arg\min_\theta \alpha\bar{r}(\theta) + \alpha\langle\theta - \bar{\theta}^t, g^\epsilon\left(\bar{\theta}^t\right)\rangle + \frac{1}{2}\|\theta - \bar{\theta}^t\|^2 \tag{55}$$

the optimality of $\bar{\theta}^{t+1}$ in (55), implies that

$$\bar{r}\left(\bar{\theta}^{t+1}\right) + \langle\bar{\theta}^{t+1} - \bar{\theta}^t, g^\epsilon\left(\bar{\theta}^t\right)\rangle + \frac{1}{2\alpha}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \leq \bar{r}\left(\bar{\theta}^t\right)$$

which combined with the smoothness of f (c.f. (43)) yields

$$\left\langle \bar{\theta}^{t+1} - \bar{\theta}^t, g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right) \right\rangle + \left(\frac{1}{2\alpha} - \frac{L_f}{2}\right)\|\bar{\theta}^{t+1} - \bar{\theta}^t\|^2 \leq F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right) \tag{56}$$

Subtracting $\langle \bar{\theta}^{t+1} - \bar{\theta}^t, \nabla f(\bar{\theta}^{t+1})\rangle$ from both sides gives

$$\left\langle \bar{\theta}^{t+1} - \bar{\theta}^t, g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right)\right\rangle + \left(\frac{1}{2\alpha} - \frac{L_f}{2}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \le F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right) - \left\langle \bar{\theta}^{t+1} - \bar{\theta}^t, \nabla f\left(\bar{\theta}^{t+1}\right) - \nabla f\left(\bar{\theta}^t\right)\right\rangle.$$

Considering $\|g^\epsilon(\bar{\theta}^t) - \nabla f(\bar{\theta}^{t+1}) + \frac{1}{\alpha}(\bar{\theta}^{t+1} - \bar{\theta}^t)\|^2$ on the left hand side, and adding relevant terms to the right hand side, we arrive at

$$\begin{aligned}
&\left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) + \frac{1}{\alpha}(\bar{\theta}^{t+1} - \bar{\theta}^t)\right\|^2 \\
&\le \left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right)\right\|^2 + \frac{1}{\alpha^2}\|\bar{\theta}^{t+1} - \bar{\theta}^t\|^2 \\
&\quad + \left(\frac{L_f}{\alpha} - \frac{1}{\alpha^2}\right)\|\bar{\theta}^{t+1} - \bar{\theta}^t\|^2 + \frac{2}{\alpha}\left(F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right)\right) \\
&\quad - \frac{2}{\alpha}\left\langle \bar{\theta}^{t+1} - \bar{\theta}^t, \nabla f\left(\bar{\theta}^{t+1}\right) - \nabla f\left(\bar{\theta}^t\right)\right\rangle \\
&\le \left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right)\right\|^2 + \frac{1}{\alpha^2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \\
&\quad + \left(\frac{L_f}{\alpha} - \frac{1}{\alpha^2}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \frac{2}{\alpha}\left(F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right)\right) \\
&\quad - \frac{2}{\alpha}\left\langle \bar{\theta}^{t+1} - \bar{\theta}^t\right), \nabla f\left(\bar{\theta}^{t+1}\right) - \nabla f\left(\bar{\theta}^t\right)\Big\rangle \\
&\le \left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right)\right\|^2 + \frac{1}{\alpha^2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \\
&\quad + \left(\frac{L_f}{\alpha} - \frac{1}{\alpha^2}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \frac{2}{\alpha}\left(F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right)\right) \\
&\quad + \frac{\eta}{\alpha}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \frac{L_f^2}{\eta}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \qquad (57)
\end{aligned}$$

where the last inequality is obtained by applying Young's inequality, and then using the L_f-Lipschitz continuity of $f(\cdot)$. By simplifying the last inequality, we obtain

$$\begin{aligned}
&\left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) + \frac{1}{\alpha}\left(\bar{\theta}^{t+1} - \bar{\theta}^t\right)\right\|^2 \\
&\le \left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right)\right\|^2 + \frac{2}{\alpha}\left(F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right)\right) \\
&\quad + \left(\frac{L_f^2}{\eta} + \frac{L_f + \eta}{\alpha}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2. \qquad (58)
\end{aligned}$$

The first term in the right hand side can be bounded by adding and subtracting $g^*(\bar{\theta}^t)$ and using (51), to arrive at

$$\begin{aligned}
&\left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) + \frac{1}{\alpha}\left(\bar{\theta}^{t+1} - \bar{\theta}^t\right)\right\|^2 \\
&\le 2\left\|\nabla f\left(\bar{\theta}^t\right) - g^*\left(\bar{\theta}^t\right)\right\|^2 \frac{2L_{\bar{\theta}}^2\epsilon}{\lambda_0} + \frac{2}{\alpha}\left(F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right)\right) \\
&\quad + \left(\frac{L_f^2}{\eta} + \frac{L_f + \eta}{\alpha}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2. \qquad (59)
\end{aligned}$$

Summing these inequalities over $t = 1, \ldots, T$, we find

$$\begin{aligned}
&\sum_{t=0}^{T}\left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) + \frac{1}{\alpha}\left(\bar{\theta}^{t+1} - \bar{\theta}^t\right)\right\|^2 \\
&\le 2\sum_{t=0}^{T}\left\|\nabla f\left(\bar{\theta}^t\right) - g^*\left(\bar{\theta}^t\right)\right\|^2 + \frac{2(T+1)L_{\bar{\theta}}^2\epsilon}{\lambda_0} \\
&\quad + \frac{2}{\alpha}\left[F\left(\bar{\theta}^0\right) - F\left(\bar{\theta}^T\right)\right] \\
&\quad + \left(\frac{L_f^2}{\eta} + \frac{L_f + \eta}{\alpha}\right)\sum_{t=0}^{T}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2. \qquad (60)
\end{aligned}$$

Using (54) to bound the last term yields

$$\begin{aligned}
&\sum_{t=0}^{T}\left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) + \frac{1}{\alpha}\left(\bar{\theta}^{t+1} - \bar{\theta}^t\right)\right\|^2 \\
&\le 2\sum_{t=0}^{T}\left\|\nabla f\left(\bar{\theta}^t\right) - g^*\left(\bar{\theta}^t\right)\right\|^2 + \frac{2(T+1)L_{\bar{\theta}}^2\epsilon}{\lambda_0} \\
&\quad + \frac{2}{\alpha}\Delta_F + \beta\Delta_F + \frac{\beta}{\eta}\sum_{t=0}^{T}\left\|\nabla f\left(\bar{\theta}^t\right) - g^*\left(\bar{\theta}^t\right)\right\|^2 \\
&\quad + \frac{\beta(T+1)L_{\bar{\theta}_z}^2\epsilon}{\lambda_0} \qquad (61)
\end{aligned}$$

where $\beta = (\frac{L_f^2}{\eta} + \frac{L_f+\eta}{\alpha})\frac{2\alpha}{2-(L_f+\eta)\alpha}$. By taking expectation of both sides of this inequality, we obtain

$$\begin{aligned}
&\frac{1}{T+1}\mathbb{E}\left[\sum_{t=0}^{T}\left\|g^\epsilon\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) + \frac{1}{\alpha}\left(\bar{\theta}^{t+1} - \bar{\theta}^t\right)\right\|^2\right] \\
&\le \left(\frac{2}{\alpha} + \beta\right)\frac{\Delta_F}{T+1} + \left(\frac{\beta}{\eta} + 2\right)\sigma^2 + \frac{(\beta+2)L_{\bar{\theta}}^2\epsilon}{\lambda_0} \qquad (62)
\end{aligned}$$

where we have used $\mathbb{E}[\|\nabla f(\bar{\theta}^t) - g^*(\bar{\theta}^t)\|_2^2] \le \sigma^2$, which holds according to Assumption 5. By [39, Th. 10] and [58], we know that

$$-g^\epsilon\left(\bar{\theta}^t\right) - \frac{1}{\alpha}\left(\bar{\theta}^{t+1} - \bar{\theta}^t\right) \in \partial\bar{r}\left(\bar{\theta}^{t+1}\right) \qquad (63)$$

which gives

$$\begin{aligned}
&\nabla f\left(\bar{\theta}^{t+1}\right) - g^\epsilon\left(\bar{\theta}^t\right) - \frac{1}{\alpha}\left(\bar{\theta}^{t+1} - \bar{\theta}^t\right) \\
&\in \nabla f\left(\bar{\theta}^{t+1}\right) + \partial\bar{r}\left(\bar{\theta}^{t+1}\right) \\
&= \partial F\left(\bar{\theta}^{t+1}\right).
\end{aligned}$$

Upon replacing the latter in the left hand side of (62), and recalling the definition of distance, we deduce that

$$\begin{aligned}
\mathbb{E}\left[\operatorname{dist}(\mathbf{0}, \partial\hat{F}(\bar{\theta}^{t'}))\right] &\le \left(\frac{2}{\alpha} + \beta\right)\frac{\Delta_F}{T} + \left(\frac{\beta}{\eta} + 2\right)\sigma^2 \\
&\quad + \frac{(\beta+2)L_{\bar{\theta}_z}^2\epsilon}{\lambda_0}
\end{aligned}$$

where t' is randomly drawn from $t' \in \{1, 2, \ldots, T+1\}$, which concludes the proof.

C. Proof of Theorem 2

Instead of resorting to an oracle to obtain an ϵ-optimal solver for the surrogate loss, here we utilize a single step

stochastic gradient ascent with mini-batch size M to solve the maximization step. Consequently, the updates become

$$\bar{\theta}^{t+1} = \text{prox}_{\alpha_t r}\left(\bar{\theta}^t - \alpha_t g^t\left(\bar{\theta}^t\right)\right) \tag{64}$$

where $g^t(\bar{\theta}^t) := \frac{1}{M}\sum_{m=1}^{M} g(\bar{\theta}^t, \zeta_m^t; z_m)$. Letting $\delta(\bar{\theta}^t) := \nabla f(\bar{\theta}^t) - g^t(\bar{\theta}^t)$, and using the L_f-smoothness of $f(\bar{\theta})$, we obtain

$$\begin{aligned} f\left(\bar{\theta}^{t+1}\right) &\le f\left(\bar{\theta}^t\right) + \left\langle \nabla f\left(\bar{\theta}^t\right), \bar{\theta}^{t+1} - \bar{\theta}^t \right\rangle + \frac{L_f}{2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \\ &\le f\left(\bar{\theta}^t\right) + \left\langle g^t\left(\bar{\theta}^t\right) + \delta\left(\bar{\theta}^t\right), \bar{\theta}^{t+1} - \bar{\theta}^t \right\rangle \\ &\quad + \frac{L_f}{2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2. \end{aligned} \tag{65}$$

Next, we substitute $\bar{\theta}^{t+1} \to \boldsymbol{u}$, $\theta^t \to \boldsymbol{y}$, and $\bar{\theta}^t - \alpha_t g^t(\bar{\theta}^t) \to \boldsymbol{x}$ in (45), to arrive at

$$\left\langle \bar{\theta}^t - \alpha_t g^t\left(\bar{\theta}^t\right) - \bar{\theta}^{t+1}, \bar{\theta}^t - \bar{\theta}^{t+1} \right\rangle \le \alpha_t \bar{r}\left(\bar{\theta}^t\right) - \alpha_t \bar{r}\left(\bar{\theta}^{t+1}\right)$$

which leads to

$$\left\langle g^t\left(\bar{\theta}^t\right), \bar{\theta}^{t+1} - \bar{\theta}^t \right\rangle \le \bar{r}\left(\bar{\theta}^t\right) - \bar{r}\left(\bar{\theta}^{t+1}\right) - \frac{1}{\alpha_t}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2.$$

Substituting the latter into (65), gives

$$\begin{aligned} f\left(\bar{\theta}^{t+1}\right) &\le f\left(\bar{\theta}^t\right) + \left\langle \delta\left(\bar{\theta}^t\right), \bar{\theta}^{t+1} - \bar{\theta}^t \right\rangle + \frac{L_f}{2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \\ &\quad + \bar{r}\left(\bar{\theta}^t\right) - \bar{r}\left(\bar{\theta}^{t+1}\right) - \frac{1}{\alpha_t}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \end{aligned}$$

and with $F(\theta) := f(\theta) + \bar{r}(\theta)$, we have

$$\begin{aligned} F\left(\bar{\theta}^{t+1}\right) - F\left(\bar{\theta}^t\right) &\le \left\langle \delta\left(\bar{\theta}^t\right), \bar{\theta}^{t+1} - \bar{\theta}^t \right\rangle \\ &\quad + \left(\frac{L_f}{2} - \frac{1}{\alpha_t}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2. \end{aligned} \tag{66}$$

Using Young's inequality $\langle \delta(\bar{\theta}^t), \bar{\theta}^{t+1} - \bar{\theta}^t \rangle \le \frac{1}{2}\|\delta(\bar{\theta}^t)\|^2 + \frac{1}{2}\|\bar{\theta}^{t+1} - \bar{\theta}^t\|^2$ implies that

$$F\left(\bar{\theta}^{t+1}\right) - F\left(\bar{\theta}^t\right) \le \left(\frac{L_f + 1}{2} - \frac{1}{\alpha_t}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \frac{\|\delta\left(\bar{\theta}^t\right)\|^2}{2} \tag{67}$$

and after adding the term $\langle \bar{\theta}^{t+1} - \bar{\theta}^t, \nabla f(\bar{\theta}^{t+1}) \rangle$ to both sides in (67), and simplifying terms, yields

$$\begin{aligned} &\left\langle \bar{\theta}^{t+1} - \bar{\theta}^t, g^t\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) \right\rangle \\ &\le -\left(\frac{1}{2\alpha_t} - \frac{L_f}{2}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right) \\ &\quad - \left\langle \bar{\theta}^{t+1} - \bar{\theta}^t, \nabla f\left(\bar{\theta}^{t+1}\right) - \nabla f\left(\bar{\theta}^t\right) \right\rangle. \end{aligned} \tag{68}$$

Completing the square yields

$$\begin{aligned} &\left\| g^t\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) + \frac{1}{\alpha_t}\left(\bar{\theta}^{t+1} - \bar{\theta}^t\right) \right\|^2 \\ &\le \left\| g^t\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) \right\|^2 + \frac{1}{\alpha_t^2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \\ &\quad + \left(\frac{L_f}{\alpha_t} - \frac{1}{\alpha_t^2}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \frac{2\left(F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right)\right)}{\alpha_t} \\ &\quad - \frac{2}{\alpha_t}\left\langle \bar{\theta}^{t+1} - \bar{\theta}^t, \nabla f\left(\bar{\theta}^{t+1}\right) - \nabla f\left(\bar{\theta}^t\right) \right\rangle \\ &\le 2\left\| g^t\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right) \right\|^2 + 2\left\| \nabla f\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^{t+1}\right) \right\|^2 \\ &\quad + \frac{1}{\alpha_t^2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \left(\frac{L_f}{\alpha_t} - \frac{1}{\alpha_t^2}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \\ &\quad + \frac{2\left(F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right)\right)}{\alpha_t} \\ &\quad - \frac{2}{\alpha_t}\left\langle \bar{\theta}^{t+1} - \bar{\theta}^t, \nabla f\left(\bar{\theta}^{t+1}\right) - \nabla f\left(\bar{\theta}^t\right) \right\rangle \\ &\le 2\left\| g^t\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right) \right\|^2 + 2L_f^2\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \\ &\quad + \frac{1}{\alpha_t^2}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 + \left(\frac{L_f}{\alpha_t} - \frac{1}{\alpha_t^2}\right)\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \\ &\quad + \frac{2\left(F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right)\right)}{\alpha_t} + \frac{2L_f}{\alpha_t}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2 \\ &\le 2\left\| g^t\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right) \right\|^2 + \frac{2\left(F\left(\bar{\theta}^t\right) - F\left(\bar{\theta}^{t+1}\right)\right)}{\alpha_t} \\ &\quad + \frac{3L_f + 2L_f^2\alpha_t}{\alpha_t}\left\|\bar{\theta}^{t+1} - \bar{\theta}^t\right\|^2. \end{aligned} \tag{69}$$

Recalling that $\delta(\bar{\theta}^t) := \nabla f(\bar{\theta}^t) - g^t(\bar{\theta}^t)$, we can bound the first term as

$$\begin{aligned} &\mathbb{E}\left[\left\| g^t\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right) \right\|^2 | \theta^t\right] \\ &= \mathbb{E}\left[\left\| g^*\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right) + \delta^t \right\|^2 | \theta^t\right] \\ &= \left\| g^*\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right) \right\|^2 + \|\delta^t\|^2 \\ &\quad + 2\mathbb{E}\left[\left\langle g^*\left(\bar{\theta}^t\right) - \nabla f\left(\bar{\theta}^t\right), \delta^t \right\rangle | \theta^t\right] \end{aligned} \tag{70}$$

where the third equality is obtained by expanding the square term, and using $\mathbb{E}[\langle g^*(\bar{\theta}^t) - \nabla f(\bar{\theta}^t), \delta^t \rangle | \bar{\theta}^t] = \mathbf{0}$. We will further bound the right hand side here as follows. Recalling that $\delta^t = \frac{1}{M}\sum_{m=1}^{M} g(\bar{\theta}^t, \zeta_m^t; z_m) - g^*(\bar{\theta}^t)$, where $g^*(\theta^t) := \frac{1}{M}\sum_{m=1}^{M} \nabla_{\bar{\theta}} \psi(\bar{\theta}^t, \zeta_m^{*t}; z_m)$, it holds that

$$\begin{aligned} &\mathbb{E}\left[\|\delta^t\|^2 \middle| \bar{\theta}^t, \zeta_m^t\right] \\ &= \mathbb{E}\left[\left\| \frac{1}{M}\sum_{m=1}^{M}\left[g(\bar{\theta}^t, \zeta_m^t; z_m) - g^*\left(\bar{\theta}^t\right)\right]\right\|^2 \middle| \bar{\theta}^t, \zeta_m^t\right] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[\left\| \nabla_{\bar{\theta}}\psi(\bar{\theta}^t, \zeta_m^t; z_m) - \nabla_{\bar{\theta}}\psi(\bar{\theta}^t, \zeta_m^{*t}; z_m)\right\|^2 \middle| \bar{\theta}^t, \zeta_m^t\right] \\ &\le \frac{L_{\theta z}^2}{M^2}\sum_{m=1}^{M}\|\zeta_m^t - \zeta_m^{*t}\|^2 \end{aligned} \tag{71}$$

where the second equality is because the samples $\{z_m\}_{m=1}^M$ are i.i.d., and last inequality holds due to the Lipschitz smoothness

of $\psi(\cdot)$. Since ζ_m^t is obtained by a single gradient ascent update over a μ-strongly concave function, we have that

$$\frac{L_{\theta z}^2}{M^2}\sum_{m=1}^{M}\left\|\zeta_m^t-\zeta_m^{*t}\right\|^2\leq\frac{L_{\theta z}^2}{M}\Big[(1-\alpha_t\mu)D^2+\alpha_t^2B^2\Big] \quad (72)$$

where D is the diameter of the feasible set, and $\alpha_t>0$ is the step size. The following holds for the expected error term

$$\mathbb{E}\Big[\|\delta^t\|^2|\bar{\theta}^t,\zeta_m^t\Big]\leq\frac{L_{\theta z}^2}{M}\Big[(1-\alpha_t\mu)D^2+\alpha_t^2B^2\Big] \quad (73)$$

and using it in (70), we arrive at

$$\mathbb{E}\Big[\Big\|g^t\Big(\bar{\theta}^t\Big)-\nabla f\Big(\bar{\theta}^t\Big)\Big\|^2|\theta^t\Big]\leq 2\Big\|g^*\Big(\bar{\theta}^t\Big)-\nabla f\Big(\bar{\theta}^t\Big)\Big\|^2+\frac{L_{\bar{\theta}z}^2}{M}\Big[(1-\alpha_t\mu)D^2+\alpha_t^2B^2\Big]. \quad (74)$$

Substituting the last inequality into (69) boils down to

$$\begin{aligned}\mathbb{E}\Big[\Big\|g^t\Big(\bar{\theta}^t\Big)-\nabla f\Big(\bar{\theta}^{t+1}\Big)+\frac{1}{\alpha_t}\Big(\bar{\theta}^{t+1}-\bar{\theta}^t\Big)\Big\|^2|\bar{\theta}^t\Big]&\leq 4\Big\|g^*\Big(\bar{\theta}^t\Big)-\nabla f\Big(\bar{\theta}^t\Big)\Big\|^2\\&+\frac{3L_f+2L_f^2\alpha_t}{\alpha_t}\mathbb{E}\Big[\Big\|\bar{\theta}^{t+1}-\bar{\theta}^t\Big\|^2|\bar{\theta}^t\Big]\\&+\frac{2F\Big(\bar{\theta}^t\Big)-2\mathbb{E}\Big[F\Big(\bar{\theta}^{t+1}\Big)|\bar{\theta}^t\Big]}{\alpha_t}\\&+\frac{L_{\bar{\theta}z}^2}{M}\Big[(1-\alpha_t\mu)D^2+\alpha_t^2B^2\Big]. \quad (75)\end{aligned}$$

Taking again expectation over $\bar{\theta}^t$ on both sides, yields

$$\begin{aligned}&\mathbb{E}\Big\|g^t\Big(\bar{\theta}^t\Big)-\nabla f\Big(\bar{\theta}^{t+1}\Big)+\frac{1}{\alpha_t}\Big(\bar{\theta}^{t+1}-\bar{\theta}^t\Big)\Big\|^2\\&\leq 4\mathbb{E}\Big[\Big\|g^*\Big(\bar{\theta}^t\Big)-\nabla f\Big(\bar{\theta}^t\Big)\Big\|^2\Big]+\frac{L_{\bar{\theta}z}^2}{M}\Big[(1-\alpha_t\mu)D^2+\alpha_t^2B^2\Big]\\&+\mathbb{E}\left[\frac{2F\Big(\bar{\theta}^t\Big)-2F\Big(\bar{\theta}^{t+1}\Big)}{\alpha_t}+\frac{3L_f+2L_f^2\alpha_t}{\alpha_t}\Big\|\bar{\theta}^{t+1}-\bar{\theta}^t\Big\|^2\right]. \quad (76)\end{aligned}$$

Recalling that $\mathbb{E}[\|\psi^*(\bar{\theta}^t,\zeta_m^t;z_m)-\nabla f(\bar{\theta}^t)\|^2]\leq\sigma^2$, and that $g^*(\bar{\theta}^t)=\frac{1}{M}\sum_{m=1}^{M}\psi(\bar{\theta}^t,\zeta_m^{*t};z_m)$, the first term on the right hand side can be bounded by $\frac{4\sigma^2}{M}$. For a fixed learning rate $\alpha>0$, summing inequalities (76) from $t=0,\ldots,T$, yields

$$\begin{aligned}&\frac{1}{T+1}\mathbb{E}\Big[\sum_{t=0}^{T}\Big\|g^t\Big(\bar{\theta}^t\Big)-\nabla f\Big(\bar{\theta}^{t+1}\Big)+\frac{1}{\alpha_t}\Big(\bar{\theta}^{t+1}-\bar{\theta}^t\Big)\Big\|^2\Big]\\&\leq\frac{2}{\alpha(T+1)}\Big(F\Big(\theta^0\Big)-\mathbb{E}[F(\theta^T)]\Big)+\frac{3L_f+2L_f^2\alpha}{\alpha}\frac{1}{T+1}\mathbb{E}\Big[\sum_{t=0}^{T}\Big\|\bar{\theta}^{t+1}-\bar{\theta}^t\Big\|^2\Big]+\frac{4\sigma^2}{M}\\&+\frac{2L_{\bar{\theta}z}^2}{M}\Big[(1-\alpha\mu)D^2+\alpha^2B^2\Big]\\&\leq\frac{1}{T+1}\left\{\frac{2}{\alpha}+\frac{6L_f+4L_f^2\alpha}{[2-\alpha(L_f+\beta)]}\right\}\Big(F\Big(\bar{\theta}^0\Big)-\mathbb{E}\Big[F\Big(\bar{\theta}^T\Big)\Big]\Big)\\&+\frac{2L_{\bar{\theta}z}^2}{M}\left\{1+\frac{3L_f+2L_f^2\alpha}{2(2-\alpha(L_f+\beta))}\right\}\Big[(1-\alpha\mu)D^2+\alpha^2B^2\Big]+\frac{4\sigma^2}{M}. \quad (77)\end{aligned}$$

Consider now replacing $F(\bar{\theta}^0)-F(\bar{\theta}^T)$ with $\Delta_F=F(\bar{\theta}^0)-\inf_{\bar{\theta}}F(\bar{\theta})$, and note that $g^t(\bar{\theta}^t)-\nabla f(\bar{\theta}^{t+1})+\frac{1}{\alpha_t}(\bar{\theta}^{t+1}-\bar{\theta}^t)\in\partial F(\bar{\theta}^{t+1})$, where ∂F denotes the set of subgradients of F. It then becomes clear that

$$\begin{aligned}&\mathbb{E}[\mathrm{dist}(0,\partial F)^2]\\&\leq\frac{1}{T+1}\mathbb{E}\Big[\sum_{t=0}^{T}\Big\|g^t\Big(\bar{\theta}^t\Big)-\nabla f\Big(\bar{\theta}^{t+1}\Big)+\frac{1}{\alpha_t}\Big(\bar{\theta}^{t+1}-\bar{\theta}^t\Big)\Big\|^2\Big]\\&\leq\frac{\zeta}{T+1}\Delta_F+\frac{2L_{\bar{\theta}z}^2\nu}{N}\Big[(1-\alpha\mu)D^2+\alpha^2B^2\Big]+\frac{4\sigma^2}{M}\end{aligned}$$

where $\zeta=\frac{2}{\alpha}+\frac{6L_f+4L_f^2\alpha}{(2-\alpha(L_f+\beta))}$ and $\nu=1+\frac{3L_f+2L_f^2\alpha}{2(2-\alpha(L_f+\beta))}$, which concludes the proof.

D. Bounded gradient norm assumption.

Regarding a bounded gradient norm assumption, this indeed a consequence of the other assumption already made. Specifically, the condition $\mathbb{E}[\|\nabla_{\bar{\theta}}\psi^*(\theta,\zeta_m^*;z_m)-\nabla f(\theta)\|^2]\leq\sigma^2$ to hold, requires that the gradient of the loss be bounded, that is $\|\nabla_\theta\ell(\theta,z)\|_2\leq B_1\ \forall\theta,z$, where B_1 is a constant dependent on σ. A similar argument is presented in [34]). To formalize this, recall the definition

$$g^t\Big(\overline{\theta}^t\Big):=\frac{1}{M}\sum_{m=1}^{M}\nabla_{\overline{\theta}}\psi\Big(\overline{\theta},\zeta_m^t;z_m^t\Big)\Big|_{\overline{\theta}=\overline{\theta}^t}. \quad (78)$$

Using this, the gradient norm can be bounded as

$$\begin{aligned}\|g^t\Big(\overline{\theta}^t\Big)\|&=\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla_{\overline{\theta}}\psi\Big(\overline{\theta},\zeta_m^t;z_m^t\Big)\right\| \quad (79)\\&\overset{(a)}{\leq}\frac{1}{M}\sum_{m=1}^{M}\left\|\nabla_{\overline{\theta}}\psi\Big(\overline{\theta},\zeta_m^t;z_m^t\Big)\right\|\\&\overset{(b)}{\leq}\frac{1}{M}\sum_{m=1}^{M}\max_m\left\|\nabla_{\overline{\theta}}\psi\Big(\overline{\theta},\zeta_m^t;z_m^t\Big)\right\|\\&=\max_m\left\|\nabla_{\overline{\theta}}\psi\Big(\overline{\theta},\zeta_m^t;z_m^t\Big)\right\|\end{aligned}$$

where (a) is due to norm properties, and (b) is simply since each term is replaced by the maximum value it can take. For brevity, with a slight abuse of notation let us define

$$\left\|\nabla_{\overline{\theta}}\psi^\star\Big(\overline{\theta},\zeta;z\Big)\right\|:=\max_m\left\|\nabla_{\overline{\theta}}\psi\Big(\overline{\theta},\zeta_m^t;z_m^t\Big)\right\| \quad (80)$$

Upon invoking the definition of $\psi(\overline{\theta},\zeta;z):=\ell(\theta;\zeta)+\gamma(\rho-c(z,\zeta))$ (c.f., (10)), it holds that

$$\left\|\nabla_{\overline{\theta}}\psi^\star\Big(\overline{\theta},\zeta;z\Big)\right\|^2:=\left\|\begin{bmatrix}\nabla_\theta\ell(\theta;\zeta)\\\nabla_\gamma(\gamma(\rho-c(z,\zeta)))\end{bmatrix}\right\|^2$$

TABLE III
HYPERPARAMETERS USED FOR TESTING ON CIFAR-10

Testing Method	FGSM	TPGD	CW	PGD	DIFGSM
Hyperparameters	$\epsilon = 0.03$	$\epsilon = 0.001$ $\alpha = 0.0007$ steps $= 5$	$c = 1$ $\kappa = 0.01$ steps $= 50$ lr $= 0.01$	$\epsilon = 0.03$ $\alpha = 0.008$ steps $= 10$ random start = True	$\epsilon = 0.03$ $\alpha = 0.007$ steps $= 10$ decay $= 0.0$
Testing Method	TIFGSM	NIFGSM	RFGSM	VNIFGSM	APGD
Hyperparameters	$\epsilon = 0.03$ $\alpha = 0.007$ nsig $= 3$ steps $= 10$ decay $= 0.0$ len kernel $= 15$ resize rate $= 0.9$ kernel $=$ gaussian diversity prob $= 0.5$ random start $=$ False	$\epsilon = 0.03$ $\alpha = 0.007$ steps $= 10$ decay $= 0.0$ resize rate $= 0.9$ diversity prob $= 0.5$ random start $=$ False	$\epsilon = 0.03$ $\alpha = 0.007$ steps $= 10$	$\epsilon = 0.06$ $\alpha = 0.007$ steps $= 10$ decay $= 1.0$ N $= 5, \beta = 1.5$	$\epsilon = 0.03$ $\rho = 0.75$ steps $= 10$ norm $= \ell_\infty$

TABLE IV
HYPER-PARAMETERS USED FOR TESTING ON CIFAR-100

Testing Method	FGSM	TPGD	CW	PGD	DIFGSM
Hyperparameters	$\epsilon = 0.03$	$\epsilon = 0.001$ $\alpha = 0.005$ steps $= 2$	$c = 1$ $\kappa = 0.01$ steps $= 5$ lr $= 0.01$	$\epsilon = 0.001$ $\alpha = 0.05$ steps $= 10$ random start = True	$\epsilon = 0.001$ $\alpha = 0.01$ steps $= 5$ decay $= 0.001$
Testing Method	TIFGSM	NIFGSM	RFGSM	VNIFGSM	APGD
Hyperparameters	$\epsilon = 0.001$ $\alpha = 0.01$ nsig $= 3$ steps $= 10$ decay $= 0.0$ len kernel $= 15$ resize rate $= 0.9$ kernel $=$ gaussian diversity prob $= 0.5$ random start $=$ False	$\epsilon = 0.03$ $\alpha = 0.007$ steps $= 10$ decay $= 0.0$ resize rate $= 0.9$ diversity prob $= 0.5$ random start $=$ False	$\epsilon = 0.001$ $\alpha = 0.01$ steps $= 5$	$\epsilon = 0.001$ $\alpha = 0.01$ steps $= 10$ decay $= 1.0$ N $= 5, \beta = 1.5$	$\epsilon = 0.01$ $\rho = 0.8$ steps $= 10$ norm $= \ell_\infty$

$$\begin{aligned} &\leq \|\nabla_\theta \ell(\theta;\zeta)\|^2 + \left\|\nabla_\gamma(\gamma(\rho - c(z,\zeta)))\right\|^2 \\ &= \|\nabla_\theta \ell(\theta;\zeta)\|^2 + \|\rho - c(z,\zeta)\|^2 \\ &\leq B_1^2 + B_2^2 \\ &:= B^2 \end{aligned} \tag{81}$$

where $\|\nabla_\theta \ell(\theta;\zeta)\|^2 \leq B_1^2$ follows from discussion above, and $\|\rho - c(z,\zeta)\|^2 \leq B_2^2$ is a direct consequence of boundedness constraint under the Assumption 1. Hence it is clear that $\mathbb{E}\|g^t(\overline{\theta}^t)\|^2 \leq \mathbb{E}\|\nabla_{\overline{\theta}}\psi^\star(\overline{\theta},\zeta;z)\|^2 \leq B^2$.

E. Hyperparameters for CIFAR-10 and CIFAR-100 Tests

The numerical test results presented in Table I and Table II correspond to testing against adversarial samples generated using the trained model parameters and specific algorithms with preselected hyperparameters. For reference, Table III details the hyperparameters utilized for the CIFAR-10 classification problem. To address the observed drop in performance in the 100-class classification scenario, we modified the hyperparameters, which are outlined in Table IV. These adjustments were necessary to ensure consistent and reliable testing across both classification tasks.

REFERENCES

[1] B. Li et al., "Trustworthy AI: From principles to practices," *ACM Comput. Surveys*, vol. 55, no. 9, pp. 1–46, 2023.

[2] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[3] W. Zhang, G. Wang, J. Sun, Y. Yuan, and G. Huang, "STORM: Efficient stochastic transformer based world models for reinforcement learning," in *Proc. Conf. Neural Inf. Process. Syst.*, 2023, pp. 1–20.

[4] A. Abusnaina et al., "Adversarial example detection using latent neighborhood graph," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 87–96.

[5] Y.-C. Lin and F. Yu, "DeepSHAP summary for adversarial example detection," in *Proc. IEEE/ACM Int. Workshop Deep Learn. Test.*, 2023, pp. 17–24.

[6] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2019, pp. 310–320.

[7] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," in *Proc. Int. Conf. Learn. Rep.*, Apr. 2018, pp. 1–8.

[8] T. Pang et al., "Two coupled rejection metrics can tell adversarial examples apart," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 23–33.

[9] Z. Deng, X. Yang, S. Xu, H. Su, and J. Zhu, "LiBRe: A practical Bayesian approach to adversarial detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 72–82.

[10] F. Sheikholeslami, S. Jain, and G. B. Giannakis, "Minimum uncertainty based detection of adversaries in deep neural networks," in *Proc. Inf. Theory Appl. Workshop*, 2020, pp. 1–16.

[11] A. Liu, X. Liu, H. Yu, C. Zhang, Q. Liu, and D. Tao, "Training robust deep neural networks via adversarial noise propagation," *IEEE Trans. Image Process.*, vol. 30, pp. 69–81, 2021.
[12] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu, "On the convergence and robustness of adversarial training," in *Proc. Mach. Learn. Res.*, Jun. 2021, pp. 6586–6595.
[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Rep.*, Dec. 2015, pp. 1–9.
[14] C. Zhang, P. Benz, C. Lin, A. Karjauv, J. Wu, and I. S. Kweon, "A survey on universal adversarial attack," Jan. 2022, *arXiv:2103.01498*.
[15] X. Yang, J. Lin, H. Zhang, X. Yang, and P. Zhao, "Improving the transferability of adversarial examples via direction tuning," Aug. 2023, *arXiv:2303.15109*.
[16] G. Wang, H. Yan, Y. Guo, and X. Wei, "Improving adversarial transferability with gradient refining," Apr. 2022, *arXiv:2105.04834*.
[17] E. Dobriban, H. Hassani, D. Hong, and A. Robey, "Provable tradeoffs in adversarially robust classification," *IEEE Trans. Inf. Theory*, vol. 69, no. 12, pp. 7793–7822, Dec. 2023.
[18] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, vol. 2, 2020, pp. 429–450.
[19] R. Song, L. Zhou, L. Lyu, A. Festag, and A. Knoll, "ResFed: Communication efficient federated learning with deep compressed residuals," *IEEE Internet Things J.*, vol. 11, no. 6, pp. 9458–9472, Mar. 2024.
[20] X. Liu and T. Ratnarajah, "Computation and communication efficient federated learning over wireless networks," Sep. 2023, *arXiv:2309.01816*.
[21] B. Weng, J. Sun, G. Huang, F. Deng, G. Wang, and J. Chen, "Competitive meta-learning," *IEEE/CAA J. Automatica Sinica*, vol. 10, no. 9, pp. 1902–1904, Sep. 2023.
[22] K. Pillutla, S. M. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 42–54, 2022.
[23] D. Data and S. N. Diggavi, "Byzantine-resilient high-dimensional federated learning," *IEEE Trans. Inf. Theory*, vol. 69, no. 10, pp. 6639–6670, Oct. 2023.
[24] Y. Deng, M. M. Kamani, and M. Mahdavi, "Distributionally robust federated averaging," in *Proc. NeurIPS*, vol. 33, 2020, pp. 15111–15122.
[25] K. Pillutla, Y. Laguel, J. Malick, and Z. Harchaoui, "Federated learning with superquantile aggregation for heterogeneous data," *Mach. Learn.*, pp. 1–68, May 2023.
[26] A. Reisizadeh, F. Farnia, R. Pedarsani, and A. Jadbabaie, "Robust federated learning: The case of affine distribution shifts," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–6.
[27] F. Farnia, A. Reisizadeh, R. Pedarsani, and A. Jadbabaie, "An optimal transport approach to personalized federated learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 2, pp. 162–171, Jun. 2022.
[28] M. S. Pydi and V. Jog, "Adversarial risk via optimal transport and optimal couplings," in *Proc. Int. Conf. Mach. Learn*, 2020, pp. 7814–7823.
[29] M. S. Pydi and V. Jog, "The many faces of adversarial risk: An expanded study," *IEEE Trans. Inf. Theory*, vol. 70, no. 1, pp. 550–570, Jan. 2024.
[30] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Oper. Res.*, vol. 58, no. 3, pp. 595–612, 2010.
[31] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Oper. Res.*, vol. 62, no. 6, pp. 1358–1376, 2014.
[32] Z. Hu and L. J. Hong, "Kullback–Leibler divergence constrained distributionally robust optimization," *Optimization*, to be published.
[33] C. Bandi and D. Bertsimas, "Robust option pricing," *Eur. J. Oper. Res.*, vol. 239, no. 3, pp. 842–853, 2014.
[34] A. Sinha, H. Namkoong, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," in *Proc. Int. Conf. Learn. Rep.*, Apr. 2018, pp. 1–34.
[35] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *Math. Oper. Res.*, vol. 44, no. 2, pp. 565–600, 2019.
[36] C. Villani, *Optimal Transport: Old and New*, vol. 338. New York, NY, USA: Springer, 2008.
[37] T. Lin, C. Jin, and M. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *Proc. Int. Conf. Mach. Learn.*, May 2020, pp. 6083–6093.
[38] D. P. Bertsekas, "Nonlinear programming," *J. Oper. Res. Soc.*, vol. 48, no. 3, pp. 334–334, 1997.
[39] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317. New York, NY, USA: Springer, 2009.
[40] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Rep.*, Apr. 2017, pp. 1–9.
[41] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Rep.*, Apr. 2018, pp. 1–9.
[42] G. Wang, G. B. Giannakis, and J. Chen, "Learning ReLU networks on linearly separable data: Algorithm, optimality, and generalization," *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2357–2370, May 2019.
[43] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Rep.*, May 2015, pp. 1–8.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[45] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, Jun. 2016, pp. 1–9.
[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Intl. Conf. Learn. Rep.*, Apr. 2014, pp. 1–6.
[47] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, Apr. 2015.
[48] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 24–33.
[49] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Rep.*, Apr. 2018, pp. 1–9.
[50] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12–21.
[51] C. Xie et al., "Improving transferability of adversarial examples with input diversity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2019, pp. 30–39.
[52] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. Int. Conf. Learn. Rep.*, May 2019, pp. 1–6.
[53] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security Privacy*, 2017, pp. 39–57.
[54] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. E. Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 72–82.
[55] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. Int. Conf. Mach. Learn.*, May 2020, pp. 2206–2216.
[56] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficientlearning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat.*, vol. 54, Apr. 2017, pp. 1273–1282.
[57] J. M. Danskin, "The theory of max–min, with applications," *SIAM J. Appl. Math.*, vol. 14, no. 4, pp. 641–664, Jul. 1966.
[58] Y. Xu, R. Jin, and T. Yang, "Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2019, pp. 2626–2636.

Alireza Sadeghi (Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from the Iran University of Science and Technology, Tehran, Iran, in 2012, the M.Sc. degree in electrical engineering from the University of Tehran in 2015, and the Ph.D. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 2021.

He is currently a Postdoctoral Associate with the Department of Electrical and Computer Engineering, University of Minnesota. His research interests include machine learning, optimization, and signal processing with applications to data science, networking, and cyber-physical systems. He was a recipient of ADC Fellowship and DDF from the University of Minnesota, and the Student Travel Awards from the IEEE Communications Society and the National Science Foundation.

Gang Wang (Senior Member, IEEE) received the B.Eng. degree in automatic control and the Ph.D. degree in control science and engineering from the Beijing Institute of Technology, Beijing, China, in 2011 and 2018, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Minnesota, Minneapolis, MN, USA, in 2018.

Since July 2018, he has been a Postdoctoral Researcher with the University of Minnesota. Since August 2020, he has been a Professor with the School of Automation, Beijing Institute of Technology. His research interests include signal processing, control, and reinforcement learning with applications to cyber–physical systems and multiagent systems.

Dr. Wang was a recipient of the Best Paper Award from the FITEE in 2021, the Excellent Doctoral Dissertation Award from the Chinese Association of Automation in 2019, the Eusipco Best Student Paper Award in 2017, and the Best Conference Paper at the 2019 IEEE Power and Energy Society General Meeting. He is currently an Associate Editor of *Signal Processing* and the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS, and an Early Career Advisory Board Member of the IEEE/CAA JOURNAL OF AUTOMATICA SINICA.

Georgios B. Giannakis (Fellow, IEEE) received the Diploma degree in electrical engineering from the National Technical University of Athens, Greece, in 1981, and the first M.Sc. degree in electrical engineering, the second M.Sc. degree in mathematics, and the Ph.D. degree in electrical engineering from the University of the Southern California in 1983, 1986, and 1986, respectively.

He was with the University of Virginia from 1987 to 1998, and since 1999, he has been with the University of Minnesota (UMN), where he held an Endowed Chair of Telecommunications, served as the Director of the Digital Technology Center from 2008 to 2021, and since 2016, he has been a UMN Presidential Chair in ECE. He is the (co-) inventor of 36 issued patents. His interests span the areas of statistical learning, communications, and networking - subjects on which he has published more than 490 journal papers, 800 conference papers, 26 book chapters, two edited books and two research monographs. His current research focuses on data science with applications to IoT and power networks with renewables. He is the (co-)recipient of ten Best Journal Paper Awards from the IEEE Signal Processing and Communications Societies, including the G. Marconi Prize. He also received the IEEE-SPS 'Nobert Wiener' Society Award in 2019; the EURASIP's 'A. Papoulis' Society Award in 2020; the Tech. Achievement Awards from the IEEE-SPS in 2000 and from EURASIP in 2005; the IEEE ComSoc Education Award in 2019; and the IEEE Fourier Technical Field Award in 2015. He is a member of the Academia Europaea, the Academy of Athens, Greece, the Royal Academy of Engineering U.K., and a Fellow of the U.S. National Academy of Inventors, the European Academy of Sciences, and EURASIP. He has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SPS.