
Certifiably Robust Model Evaluation in Federated Learning under Meta-Distributional Shifts

Amir Najafi¹ Samin Mahdizadeh Sani² Farzan Farnia³

Abstract

We address the challenge of certifying the performance of a federated learning model on an unseen target network using only measurements from the source network that trained the model. Specifically, consider a source network “A” with K clients, each holding private, non-IID datasets drawn from heterogeneous distributions, modeled as samples from a broader meta-distribution μ . Our goal is to provide certified guarantees for the model’s performance on a different, unseen network “B”, governed by an unknown meta-distribution μ' , assuming the deviation between μ and μ' is bounded—either in *Wasserstein* distance or an *f-divergence*. We derive worst-case uniform guarantees for both the model’s average loss and its risk CDF, the latter corresponding to a novel, adversarially robust version of the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality. In addition, we show how the vanilla DKW bound enables principled certification of the model’s true performance on unseen clients within the same (source) network. Our bounds are efficiently computable, asymptotically minimax optimal, and preserve clients’ privacy. We also establish non-asymptotic generalization bounds that converge to zero as K grows and the minimum per-client sample size exceeds $\mathcal{O}(\log K)$. Empirical evaluations confirm the practical utility of our bounds across real-world tasks. The project code is available at: github.com/samin-mehdizadeh/Robust-Evaluation-DKW

¹Department of Computer Engineering, Sharif University of Technology, Tehran, Iran (Corresponding author) ²Department of Electrical and Computer Engineering, University of Tehran, Tehran, Iran ³Department of Computer Science and Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong. Correspondence to: Amir Najafi <amir.najafi@sharif.edu>, Samin Mahdizadeh Sani <samin.mahdizadeh@ut.ac.ir>, Farzan Farnia <farnia@cse.cuhk.edu.hk>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

1. Introduction

The distributed nature of modern learning environments, where local datasets are scattered across clients in a network, presents significant challenges for the machine learning community. Federated learning (FL) addresses some of these challenges by enabling clients to collaboratively train a decentralized model through communications with a central server (McMahan et al., 2017; Liu et al., 2024). A major obstacle in FL is the heterogeneity of data distributions among clients. This non-IID nature of client data not only impacts the training stage but also complicates the evaluation of trained models, especially when applied to unseen clients from the same or different networks (Ye et al., 2023; Zawad et al., 2021). In this paper, we focus on the latter problem.

Consider a network with K clients, each possessing their own private dataset. For $k \in [K]$, the k th client’s dataset is assumed to include $n_k \geq 1$ samples from a unique and unknown distribution P_k , which forms an empirical and private estimate \hat{P}_k . Note that the distributions P_1, \dots, P_K may be highly heterogeneous and distant from one another. In this setting, assume the server wants to assess the performance of a given machine learning model h on this network. Let the *risk* function $R(h, \hat{P}_k)$ denote the loss of h when evaluated over the private data samples of the k th client, i.e., \hat{P}_k . This quantity can be queried by the server by sending h to client k and requesting its local loss value, without directly accessing any of the private data samples in client k ’s dataset. Hence, the server can collect all K loss values and empirically approximate the average performance of h over the network, using various metrics. One common approach is the average loss: $\frac{1}{K} \sum_{k=1}^K R(h, \hat{P}_k)$, which measures the empirical average performance of the model in the network.

Another approach is the loss CDF, representing the percentage of clients whose loss exceeds a given threshold λ : $\frac{1}{K} \sum_{k=1}^K \mathbb{1}(R(h, \hat{P}_k) \geq \lambda)$, which is useful for estimating the quantiles of the risk (see (Laguel et al., 2021)). For a recent related work on the use of super-quantiles in FL with heterogeneous clients, see (Pillutla et al., 2024). While that study focuses on optimization and convergence aspects, our work provides a theoretical treatment of generalization.

A critical challenge arises when certifying the performance of h on a different and unseen network with completely different clients. Mathematically, assume a large group of $T \gg 1$ new and unseen clients with corresponding heterogeneous distributions Q_1, \dots, Q_T , which may differ substantially from one another and also from P_k 's. This corresponds to a *beta testing* scenario, where, for instance, a product owner may test a product on a small subset of a network before wider deployment (Reisizadeh et al., 2020; Ma et al., 2024). Our goal is to certify the same performance metrics over Q_i 's without accessing them.

Clearly, without a connection between P_k 's and Q_i 's, the certification task is fundamentally infeasible. To address this, we can rely on the following common assumption in many beta testing scenarios: $P_{1:K}$ and $Q_{1:T}$ are all independent samples from the same *meta-distribution* μ (a distribution over distributions), which governs higher-level factors influencing the clients, such as cultural or geographical attributes (Yuan et al., 2021; Ajay et al., 2022; Wu et al., 2024a; Chen et al., 2023; Patel et al., 2022). In this setting, the empirical average risk and empirical risk CDF over P_k 's provide unbiased estimates for those of the unseen clients, i.e., Q_i 's. Still, a natural question would be how large K or either of n_k 's need to be such that the mentioned estimates become statistically reliable?

In this work, we propose a novel approach by leveraging the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Dvoretzky et al., 1956) that uniformly bounds the CDF estimation error from empirical observed samples. Note that in our setting, the CDF is considered for the performance metric of clients drawn from meta distribution μ . Therefore, the observed clients' performance values lead to a proxy *empirical CDF*, and we want to provide an upper-bound on the *true CDF* to provide a performance guarantee for an unseen client $Q_i \sim \mu$. Our approach is to extend the DKW inequality to bound the gap between the empirical and population (true) CDFs. Figure 2 illustrates our DKW inequality-based approach and how it can provide a certified upper-bound on the true CDF of unseen clients' performance.

On the other hand, in real-world applications, target clients can belong to distinct populations with considerable cultural or behavioral differences compared to those observed in the evaluation phase. For example, an app developer might be limited to test a mobile application on users from Hong Kong, but ultimately hopes to release it for users from New York. This scenario corresponds to a meta-distributionally robust evaluation, where the objective is to evaluate the performance of h under worst-case shifts from μ to μ' . Here, μ and μ' represent unknown meta-distributions, with K empirical samples from μ corresponding to P_1, \dots, P_K . Unfortunately, μ' is completely unseen, however, it is assumed to deviate at most ε from μ , for some known $\varepsilon > 0$.

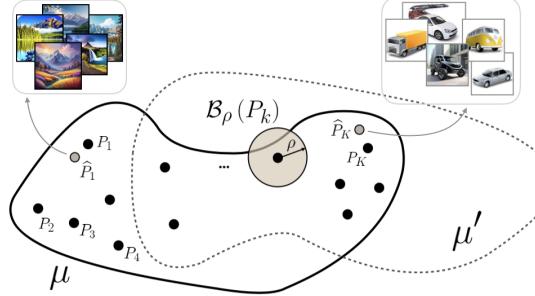


Figure 1. A graphical illustration of meta-distributional shifts between two client networks governed by μ and μ' . Meta-shifts shifts may only denote changes in densities (f -divergence) or can involve support as well (Wasserstein shifts). Each client P_k is a sample from μ , with \hat{P}_k s as its empirical counterpart based on a local datasets. The client distributions P_k can vary significantly, such as one having primarily landscape images while another contains mostly vehicle images.

Again, without this assumption certifying the performance of h under μ' is mathematically impossible. The degree of deviation (ε) between μ and μ' can be quantified using metrics such as Kullback-Leibler (KL) divergence or Wasserstein distance, both interpreted in the context of meta-distributions. KL, or more generally, an f -divergence captures situations where client types in one meta-distribution are reweighted in another (Mehta et al., 2024), whereas Wasserstein distance accounts for the emergence of entirely new client types, representing novel regions of the meta-distribution's support (Wang et al., 2022; Kuhn et al., 2019). For example, smartphone users in a coastal area may often take pictures of the sea, whereas residents in mountainous regions cannot do so unless they travel. See Figure 1 for more details. We are not aware of any prior theoretical bounds for meta-distributionally robust evaluation (or beta testing) in federated settings. In any case, a review of prior works on similar scenarios of non-IID federated learning and model evaluation is provided in Section A.

Our Contribution: Our problem, formally defined in Section 3, is to address the robust evaluation of any given h under the above-mentioned scenario. Section 4 presents some initial results for the non-robust evaluation when $\mu = \mu'$, i.e., standard beta testing. Theorem 4.1 proves non-asymptotic concentration bounds for both the empirical average and CDF of the risk, with exponential convergence as both K and $\min_k n_k$ increase. Notably, our convergence results for the loss CDF are *uniform* over thresholds λ , meaning they hold for all $\lambda \in \mathbb{R}$ simultaneously, via utilizing classical tools in statistics, including the DKW inequality and Glivenko-Cantelli theorem (Dvoretzky et al., 1956).

Next, we provide robust bounds on the average loss and/or risk CDF when target network is adversarially shifted from the source. We extend the concept of distributional robust-

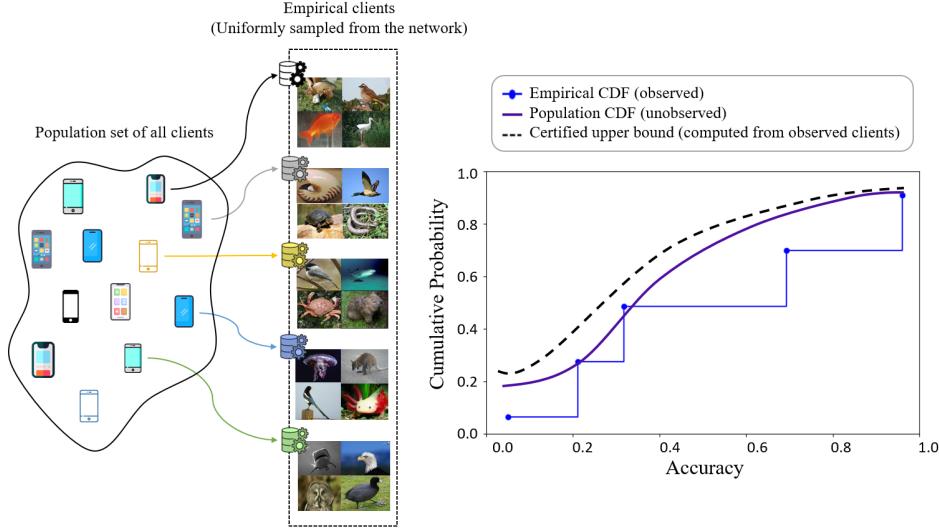


Figure 2. Our theoretical bounds are estimated from a set of seen users in a federated network governed by meta-distribution μ . We extend the application of DKW inequality to provide a certified upper-bound on the true CDF of the performance score of unseen clients following the same meta-distribution μ .

ness to meta-distributional robustness, considering both f -divergences and Wasserstein-type adversaries, respectively in Sections 5 and 6, to provide worst-case guarantees.

The majority of existing robust evaluation schemes require aggregating all private datasets in a central server for a collective adversarial manipulation of all the samples (Liu et al., 2024; Reisizadeh et al., 2020; Sinha et al., 2018), which severely violates user privacy. Another category of approaches simply compute the average ordinary risk of h in the source network, but are then forced to add a non-vanishing (ε -dependent) generalization gap that does not go to zero even when $K, n_{1:K}$ grow to infinity (see Section A and also (Rahimian and Mehrotra, 2019; Zeng and Lam, 2022)). The latter approaches fail to take into account the possible inherent wellness or robustness of h , thus usually result into excessively inflated risk certificates.

We prove that *globally* robust evaluations under both f -divergence and Wasserstein regimes can be done in a federated manner as long as: i) server can query the *local* ordinary or adversarial (only for Wasserstein attacks) loss of h from each client, and ii) server can repeat such queries with potentially various attack budgets for polynomially many times. Note that we do not need to directly access client's data. Our evaluation guarantees are asymptotically minimax optimal, and have vanishing generalization gaps which converge to zero as both K and $(\min_k n_k)/\log K$ grow. Our robust and uniform bounds on risk CDF in Section 5 are based on a novel extension of DKW inequality to adversarial scenarios, which is extensively detailed in Section B. The numerical results on standard datasets in Section 8 validate the tightness of our bounds.

2. Notations and Preliminaries

For $K \in \mathbb{N}$, $[K]$ denotes the set $\{1, 2, \dots, K\}$. Consider two measurable spaces \mathcal{X} and \mathcal{Y} , referred to as the *feature* and *label* spaces, respectively. Typically, we assume $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, while \mathcal{Y} may be $\{\pm 1\}$ for binary classification tasks or \mathbb{R} in regression problems. However, we make no extra assumptions regarding \mathcal{X} and \mathcal{Y} . We define the joint feature-label space as $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$. Let $\mathcal{M}(\mathcal{Z})$ denote the set of all probability measures supported on \mathcal{Z} . Each $P \in \mathcal{M}(\mathcal{Z})$ corresponds to a joint measure over a random feature vector $\mathbf{X} \in \mathcal{X}$ and its associated random label $y \in \mathcal{Y}$. The expectation operator with respect to a measure P is denoted by \mathbb{E}_P . We refer to μ as a meta-distribution over \mathcal{Z} , expressed as $\mu \sim \mathcal{M}^2(\mathcal{Z}) \triangleq \mathcal{M}(\mathcal{M}(\mathcal{Z}))$, where each sample from μ is itself a probability measure over \mathcal{Z} . In other words, a meta-distribution is a distribution over distributions supported on \mathcal{Z} . In our work, μ models the heterogeneity and non-IID-ness of the clients, since independent samples from μ represent different data distributions shifted from one another.

For simplicity, we abbreviate $\mathcal{M}(\mathcal{Z})$ as \mathcal{M} . For any two measures $P, Q \in \mathcal{M}$, let $\mathcal{D}(P, Q) \in \mathbb{R}_{\geq 0}$ denote any given distance or divergence between the two distributions such as KL or Wasserstein distance. In this context, let \mathcal{B}_ρ denote an ρ -distributional ambiguity ball for $\rho \geq 0$. Mathematically, for any measure $P \in \mathcal{M}$:

$$\mathcal{B}_\rho(P) \triangleq \{Q \in \mathcal{M} \mid \mathcal{D}(P, Q) \leq \rho\}, \quad (1)$$

which represents the set of distributions within ρ distance/divergence from P according to \mathcal{D} . Similarly, meta-distributional ambiguity balls $\mathcal{G}_\varepsilon(\mu)$ are defined as the set of meta-distributions over \mathcal{M}^2 within an ε distance

from a base meta-distribution μ . Mathematically, we have $\mathcal{G}_\varepsilon(\mu) \triangleq \left\{ \mu' \in \mathcal{M}^2 \mid \tilde{\mathcal{D}}(\mu, \mu') \leq \varepsilon \right\}$. Here, the *deviation*

$\tilde{\mathcal{D}}$ can be any properly defined f -divergence or a Wasserstein metric between the meta-distributions in \mathcal{M}^2 . In particular, the transportation cost in the Wasserstein distance can itself be a Wasserstein metric on \mathcal{M} in its ordinary sense.

Let $h : \mathcal{X} \rightarrow \mathcal{Y}$ represent a hypothesis (e.g., a classifier) that maps the feature space \mathcal{X} onto the label space \mathcal{Y} ¹. Additionally, assume a fixed loss function $\ell(y, \hat{y})$, which assigns a loss value to each pair of actual and predicted labels, y and \hat{y} , respectively. The expected loss, or *Risk*, of a hypothesis h w.r.t. a data distribution $P \in \mathcal{M}$ is defined as $R(h, P) \triangleq \mathbb{E}_P[\ell(y, h(\mathbf{X}))]$. Here, P could be the true distribution or an empirical approximation obtained from a dataset, usually denoted by \hat{P} . The *adversarial* risk of h w.r.t. a base measure P (or empirical \hat{P}) and a robustness radius $\varepsilon \geq 0$, is formulated as (Sinha et al., 2018):

$$R_{\varepsilon-\text{adv}}(h, P) \triangleq \sup_{Q \in \mathcal{B}_\varepsilon(P)} \mathbb{E}_Q[\ell(y, h(\mathbf{X}))], \quad (2)$$

and denotes the worst risk over all distributions in a ε -neighborhood of P according to \mathcal{D} . Similarly, the meta-distributionally robust loss of h with respect to a base meta-distribution μ is defined as:

$$\sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P[\ell(y, h(\mathbf{X}))]]. \quad (3)$$

The geometry of the ball \mathcal{G}_ε in \mathcal{M}^2 can be determined using various application-specific divergences or metrics over \mathcal{M}^2 . In Section 5, we deal with f -divergence balls, while Wasserstein meta-distributional balls (using the ordinary Wasserstein metric as their transportation cost) are utilized in Section 6. We also discuss how to practically implement such loss values in Section 7 and Appendix F.

3. Problem Definition

This section formally defines our problem. First, we outline the data generation process, privacy constraints, and specify the query policy that governs communication between clients and the server.

Data Generation: Consider $K \in \mathbb{N}$ clients connected to a central server, where client $k \in [K]$ has a unique data distribution $P_k \in \mathcal{M}(\mathcal{Z})$. We assume P_1, P_2, \dots, P_K are heterogeneous samples of an unknown meta-distribution μ over \mathcal{Z} . No one knows P_k s, however, client k has access to a dataset D_k of size $n_k \geq 1$, which contains independent samples from P_k , i.e.,

$$D_k \triangleq \left\{ \left(\mathbf{X}_i^{(k)}, y_i^{(k)} \right) \mid i \in [n_k] \right\} \sim P_k^{\otimes n_k}. \quad (4)$$

¹It should be noted that our work goes beyond this limitation and can be applied to any supervised or unsupervised machine learning task.

Let \hat{P}_k denote the empirical version of P_k based on the private samples in D_k , which are known only to client k .

Server-Client Query Policy: The server can query each of the K clients by sending a model h and a robustness radius $\rho \geq 0$ to client $k \in [K]$. In response, the client returns the adversarial loss of h around \hat{P}_k :

$$\widehat{\text{QV}}_k(h, \rho) \triangleq \sup_{Q \in \mathcal{B}_\rho(\hat{P}_k)} \mathbb{E}_Q[\ell(y, h(\mathbf{X}))]. \quad (5)$$

The type of distributional ball $\mathcal{B}_\rho(\cdot)$ can be defined using any user-defined divergence or metric over \mathcal{M} . When the robustness radius is unspecified, the client assumes it is zero, and returns the non-robust loss, i.e., $\widehat{\text{QV}}_k(h) = \widehat{\text{QV}}_k(h, 0)$. We later show that the local adversarial loss (when $\rho > 0$) is not needed for f -divergence-based guarantees and only appears in Section 6 which focuses on Wasserstein shifts. Each client k accepts a maximum number of queries, referred to as the *query budget*.

Our Problem Setup: Assume the server sends a model h to the clients and requests a number of robust or non-robust loss values for several arbitrary robustness radii. Server's ultimate goal is to use these values to provide a meta-distributionally robust upper bound for the average or CDF of the loss of h . Mathematically speaking, the objective is to *efficiently* compute empirical values $\widehat{E}_1(\varepsilon)$ and $\widehat{E}_2(\varepsilon, \lambda)$ such that the following bounds hold with high probability over the sampling of clients and datasets:

$$\sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P[\ell(y, h(\mathbf{X}))]] \leq \widehat{E}_1(\varepsilon) + \zeta, \quad (6)$$

$$\sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mu' \left(\mathbb{E}_P[\ell(y, h(\mathbf{X}))] \geq \lambda \right) \leq \widehat{E}_2(\varepsilon, \lambda) + \zeta'.$$

The generalization gaps ζ, ζ' should vanish as both K and $\min_k n_k$ increase asymptotically. Also, the second high-probability bound needs to hold for all threshold values λ , simultaneously (or *uniformly*). We consider both f -divergence and Wasserstein metrics to specify the geometry of the meta-distributional ambiguity ball $\mathcal{G}_\varepsilon(\mu)$.

From an algorithmic perspective, $\widehat{E}_{1:2}$ must be computable using only the private server-client query policy $\widehat{\text{QV}}_k(h, \rho)$ for various robustness radii ρ . Server decides the number of queries for each client, as well as the value of ρ for each query. However, the computational cost of evaluating each query at the client side, and the total number of queries per client should increase at most polynomially with parameters.

4. Non-Robust (Ordinary) Guarantees

In this section, we solve the problem setting of section 3 in the non-robust regime, i.e., when $\mu' = \mu$ or equivalently $\varepsilon = 0$. This scenario corresponds to the standard

beta testing, which has been extensively applied in practice (Chen et al., 2024). Simple statistical bounds can establish the high-probability concentration of the empirical average $\frac{1}{K} \sum_{k=1}^K R(h, \hat{P}_k)$ around its true mean w.r.t. μ , as long as the loss function $\ell(\cdot)$ is measurable and bounded (e.g., let ℓ be 1-bounded such as the 0-1 loss). However, the same argument does not directly extend to the loss CDF, as the bounds must hold uniformly for all threshold values $\lambda \geq 0$. Without uniformity, such guarantees are significantly weakened. By "uniform," we mean that, with high probability over the sampling of the K users, the worst-case deviation between the empirical and true tail probabilities is consistently bounded. This challenge is resolved using a classic result on the convergence of empirical CDFs, specifically the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality and the Glivenko–Cantelli theorem (see Section B, and in particular Lemma B.1). The following theorem summarizes our main findings in the non-robust setting.

Theorem 4.1. *Let μ be a meta-distribution, and assume P_1, \dots, P_K be K independent instances of μ , while \hat{P}_k for $k \in [K]$ represent their empirical counterparts, formed using n_k independent private samples from the k -th client. Let $h : \mathcal{X} \rightarrow \mathcal{Y}$ be any model and ℓ be any measurable and 1-bounded loss function. Then, for any $\delta > 0$, the following holds with probability at least $1 - \delta$:*

$$\begin{aligned} \mathbb{E}_{\mu} [\mathbb{E}_P (\ell(y, h(\mathbf{X})))] &\leq \frac{1}{K} \sum_{k=1}^K \widehat{\text{QV}}_k(h) + \\ &\quad \sqrt{\log(K\delta^{-1})} \mathcal{O} \left(\frac{1}{\sqrt{K}} + \frac{1}{K} \sum_{k=1}^K n_k^{-1/2} \right). \end{aligned} \quad (7)$$

For the loss CDF, the following bound holds with probability at least $1 - \delta$ uniformly for all $\lambda \in \mathbb{R}$:

$$\begin{aligned} \mu \left(\mathbb{E}_P [\ell(y, h(\mathbf{X}))] \geq \lambda \right) &\leq \\ &\quad \frac{1}{K} \sum_{k=1}^K \mathbb{1} \left(\widehat{\text{QV}}_k(h) \geq \lambda - \sqrt{\frac{\log \frac{(K+1)}{\delta}}{2n_k}} \right) + \sqrt{\frac{\log \frac{2(K+1)}{\delta}}{2K}}. \end{aligned} \quad (8)$$

The proof is given in Appendix C. By ignoring polylogarithmic terms, the empirical means over both the K users and the n_k samples from each client k converge at a rate of $(\min\{K, \min_k n_k / \log K\})^{-1/2}$. In both inequalities, the left-hand sides represent strong statistical quantities, while the right-hand sides consist of: i) vanishing generalization gaps, plus ii) empirical values that can be fully evaluated based on users' private datasets and using the private query policy described earlier (i.e., the $\widehat{\text{Q}}_i(h)$ values). Additionally, each client needs to be queried only once.

In Section 7, we discuss the computational complexity, certificates of privacy, and tightness of the all the bounds in

our work. Before that, Sections 5 and 6 extend Theorem 4.1 to the robust setting, where μ' can be adversarially perturbed from μ by either an f -divergence or a Wasserstein adversary.

5. f -Divergence Meta-Distributional Shifts

In this section, we provably bound the meta-distributionally robust performance of h , in terms of the average risk and risk CDF.

Definition 5.1. Consider two meta-distributions $\mu, \mu' \in \mathcal{M}^2$ where μ' is absolutely continuous with respect to μ . Let $f : [0, \infty) \rightarrow [-\infty, \infty]$ be a convex function such that $f(x)$ is finite for all $x > 0$, $f(1) = 0$, and $f(0) = \lim_{t \rightarrow 0^+} f(t)$ (which could be infinite). The f -divergence between μ and μ' is defined as:

$$\mathcal{D}_f(\mu' \| \mu) = \int_{P \in \mathcal{M}} f \left(\frac{d\mu'(P)}{d\mu(P)} \right) d\mu(P). \quad (9)$$

Also, for $\varepsilon \geq 0$, the f -divergence ball $\mathcal{G}_\varepsilon^{f\text{-div}}(\mu)$ is defined as $\{\mu' \mid \mathcal{D}_f(\mu' \| \mu) \leq \varepsilon\}$, which describes a neighborhood around μ where the divergence does not exceed ε .

Our goal is to provide a theoretical guarantee for the loss of h under the worst-case meta-distribution $\mu' \in \mathcal{G}_\varepsilon^{f\text{-div}}(\mu)$, using only the query values from client samples in the source network governed by μ . Since clients generated by μ' may follow different densities compared to μ , it is natural to reweight the robust loss to achieve a robust upper bound. The following theorem formalizes this idea by reweighting the query values using coefficients α_k , $k \in [K]$, and optimizing for the worst-case weights, provided they remain close to uniform weights. This approach yields a minimax-optimal bound with a vanishing generalization gap.

Theorem 5.2. *Assume an unknown meta-distributions $\mu \in \mathcal{M}^2$, let $\varepsilon \geq 0$, and consider $f(\cdot)$ to be as in Definition 5.1. Let $h : \mathcal{X} \rightarrow \mathcal{Y}$ be any model and ℓ be any measurable and 1-bounded loss function. Assume P_1, \dots, P_K represent K independent and unknown sample distributions from μ . Accordingly, let \hat{P}_k for $k \in [K]$ represent their empirical counterparts, formed using n_k independent private samples from the k -th client. Let $\Lambda \triangleq 1 + \kappa\varepsilon$. Define \widehat{B}^* as:*

$$\begin{aligned} \widehat{B}^*(\varepsilon) &\triangleq \sup_{0 \leq \alpha_1, \dots, \alpha_K \leq \Lambda} \frac{1}{K} \sum_{k=1}^K \alpha_k \widehat{\text{QV}}_k(h) \\ \text{subject to } &\quad \left| \frac{1}{K} \sum_{k=1}^K \alpha_k - 1 \right| \leq c_1 \sqrt{\log(\delta^{-1})/K}, \\ &\quad \frac{1}{K} \sum_{k=1}^K f(\alpha_k) \leq \varepsilon + c_2 \sqrt{\log(\delta^{-1})/K}, \end{aligned} \quad (10)$$

where constants $\kappa, c_1, c_2 \geq 0$ are known and depend on $f(\cdot)$. Then, for any $\delta > 0$, the following bound holds with

probability at least $1 - \delta$:

$$\sup_{\mu' \in \mathcal{G}_\varepsilon^{f-\text{div}}(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(\mathbf{X}))]] \leq \hat{B}^*(\varepsilon) \quad (11)$$

$$+ \sqrt{\log(K\delta^{-1})} \mathcal{O} \left[\frac{1}{\sqrt{K}} + \frac{1}{K} \sum_{k=1}^K n_k^{-1/2} \right].$$

The proof including the formulations for κ, c_1, c_2 are provided in Appendix D. This theorem establishes a robust bound on the expected loss under meta-shifts using $\hat{B}^*(\varepsilon)$, along with a vanishing generalization gap with a decay rate of $\tilde{\mathcal{O}}(\min\{K, n_{1:K}\}^{-1/2})$. The empirical quantity $\hat{B}^*(\varepsilon)$ is derived from a convex optimization problem performed server-side. In Section 7, we discuss all of aspects of computing this quantity. Also, it is worth noting that as both $\varepsilon, \frac{1}{K}$ tend to zero, the bound becomes increasingly similar to the non-robust case of Theorem 4.1, which should be expected. As discussed before, a key advantage of our result is that the generalization gap lacks any *non-vanishing* ε -dependent term. Next, we present our main result for robust CDF estimation of the loss under KL meta-shifts:

Theorem 5.3. Assume the setting of Theorem 5.2. For any $\lambda \in \mathbb{R}$, let us define the empirical value $\hat{J}^*(\varepsilon, \lambda)$ as

$$\begin{aligned} & \sup_{0 \leq \alpha_{1:K} \leq \Lambda} \frac{1}{K} \sum_{k=1}^K \alpha_k \mathbb{1} \left(\widehat{\text{QV}}_k(h) \geq \lambda - \sqrt{\frac{\log(\frac{K+2}{\delta})}{2n_k}} \right) \\ & \text{subject to } \left| \frac{1}{K} \sum_{k=1}^K \alpha_k - 1 \right| \leq c_1 \sqrt{\frac{1}{K} \log(K\delta^{-1})}, \\ & \quad \frac{1}{K} \sum_{k=1}^K f(\alpha_k) \leq \varepsilon + c_2 \sqrt{\frac{1}{K} \log(K\delta^{-1})}, \end{aligned}$$

Then, with probability at least $1 - \delta$, the following bound holds uniformly over all $\lambda \in \mathbb{R}$:

$$\sup_{\mu' \in \mathcal{G}_\varepsilon^{f-\text{div}}(\mu)} \mu' \left(\mathbb{E}_P [\ell(y, h(\mathbf{X}))] \geq \lambda \right) \leq \hat{J}^*(\varepsilon, \lambda) + \mathcal{O} \left(\sqrt{K^{-1} \log(K\delta^{-1})} \right). \quad (12)$$

Proof is in Appendix D, and extends the classical results from Glivenko-Cantelli theorem and DKW bound into an adversarial setting. We have detailed our new theoretical findings in Section B. The bound in Theorem 5.3 exhibits the following properties: i) It is forward-shifted with respect to the true CDF, meaning it exhibits a delayed reaction to increasing λ compared to the true CDF. The maximum delay is on the order of $\mathcal{O}(\sqrt{\log K / \min_k n_k})$. ii) The value of the CDF estimator also deviates from the true estimator by the amount $\mathcal{O}(\sqrt{\log K / K})$. Consequently, as K and $(\min_k n_k) / \log K$ tend to infinity, the generalization gap becomes zero. iii) The bound holds uniformly over all

$\lambda \in \mathbb{R}$, similar to the original DKW inequality. Again, the empirical value $\hat{J}^*(\varepsilon, \lambda)$ is the solution to a server-side convex and thus efficient program. More details will be given in Section 7.

6. Wasserstein Meta-Distributional Shifts

This section addresses the case of Wasserstein meta-shifts from μ to μ' . Such shifts present significant challenges, as the model h may encounter entirely unseen regions of the meta-distributional support. In practical terms, users may exhibit data distributions that are entirely novel compared to \hat{P}_k s of the evaluation phase in the source network. To mitigate this effect, as we will show later in this section, server needs to query out-of-domain loss values from each of the K clients. Here, we provide theoretical guarantees for the average risk, but not for the entire risk CDF. Addressing the full risk CDF requires a more advanced methodology, which falls beyond the scope of the current paper and is deferred to future work. We begin by introducing both standard and meta-distributional Wasserstein metrics, and then present our main result, an analog of Theorem 5.2 for Wasserstein-type shifts.

Definition 6.1. For any two measures $P, Q \in \mathcal{M}(\mathcal{Z})$ and a lower semi-continuous function $c : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}_{\geq 0}$, we define the *Wasserstein* distance between P and Q as

$$\mathcal{W}_c(P, Q) \triangleq \inf_{\nu \in \mathcal{C}(P, Q)} \mathbb{E}_{(\mathbf{Z}, \mathbf{Z}') \sim \nu} \{c(\mathbf{Z}, \mathbf{Z}')\}, \quad (13)$$

where $\mathcal{C}(P, Q)$ denotes the set of all couplings between P and Q , i.e., all joint probability measures in $\mathcal{M}(\mathcal{Z} \times \mathcal{Z})$ that have fixed P and Q as their respective marginals (Kuhn et al., 2019).

The function c is called the *transportation cost* and is user-defined. For example, $c(\mathbf{Z}, \mathbf{Z}') = \|\mathbf{Z} - \mathbf{Z}'\|_2 + \infty \cdot \mathbb{1}(y \neq y')$ corresponds to a typical *feature-shift* scenario in robust ML. $\mathcal{W}_c(P, Q)$ (which is a metric over \mathcal{M}) measures the minimum cost of transforming P into Q or vice versa according to the cost characterized by c . Unlike f -divergence, Wasserstein distance can stay bounded under support shifts. In fact, it is a powerful tool to model slight support changes between distributions and due to this property is widely used in adversarial robustness research.

In a similar fashion, one can define the Wasserstein distance between any two meta-distributions $\mu, \mu' \in \mathcal{M}^2(\mathcal{Z})$ with respect to any valid transportation cost over the space of measures $\mathcal{M}(\mathcal{Z})$, such as the ordinary Wasserstein distance.

Definition 6.2 (Wasserstein metric over \mathcal{M}^2). For any two meta-distributions $\mu, \mu' \in \mathcal{M}^2(\mathcal{Z})$, assume a distributional transportation cost such as ordinary Wasserstein distance $\mathcal{W}_c(\cdot, \cdot)$, where c is a bounded and proper (according to Definition 6.2) transportation cost on $\mathcal{Z} \times \mathcal{Z}$. In this regard,

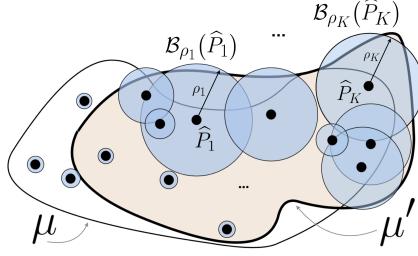


Figure 3. A graphical illustration of the core idea behind Theorem 6.3. Each tuple $\rho_{1:K}$ that satisfies the constraints corresponds to one possibility for μ' , via extending the local risk of \hat{P}_k s to their worst-case out-of-distribution values in $B_{\rho_k}(\hat{P}_k)$. Therefore, we maximize over all such $\rho_{1:K}$ to account for the worst μ' .

let us define

$$\|\mu - \mu'\|_{\mathcal{W}_c} \triangleq \inf_{\nu \in \mathcal{C}(\mu, \mu')} \mathbb{E}_{(P, Q) \sim \nu} \{ \mathcal{W}_c(P, Q) \} \quad (14)$$

as the Wasserstein distance between μ and μ' according to transportation cost \mathcal{W}_c .

Here, $\mathcal{C}(\mu, \mu')$ is the set of all couplings (joint measures in $\mathcal{M}(\mathcal{M}(\mathcal{Z}) \times \mathcal{M}(\mathcal{Z}))$) with μ and μ' as their respective marginals. Accordingly, for $\varepsilon \geq 0$ and $P \in \mathcal{M}$, we define the Wasserstein ball of radius ε around $P \in \mathcal{M}$ as $\mathcal{B}_\varepsilon^{\text{wass}}(P) \triangleq \{Q | \mathcal{W}_c(P, Q) \leq \varepsilon\}$, where the transportation cost c is hidden from formulation for the sake of simplicity. Similarly, one can define the Wasserstein ball around meta-distribution $\mu \in \mathcal{M}^2(\mathcal{Z})$ with radius ε as $\mathcal{G}_\varepsilon(\mu) \triangleq \{\mu' | \|\mu - \mu'\|_{\mathcal{W}_c} \leq \varepsilon\}$, which is the set of meta-distributions with a (meta-distributional) Wasserstein distance of at most ε from μ .

We now present our main result of this section: a quasi-convex (and thus polynomial-time) optimization problem that provides empirical meta-distributionally robust evaluation guarantees against Wasserstein shifts, with a asymptotically vanishing generalization gap.

Theorem 6.3 (Empirical Evaluation with Wasserstein Robustness). *Consider the same setting as in Theorem 5.2. Let c be a bounded and proper (according to Definition 6.2) transportation cost on $\mathcal{Z} \times \mathcal{Z}$. For any given $\varepsilon, \delta > 0$, consider the following constrained optimization problem:*

$$\widehat{U}^*(\varepsilon) \triangleq \sup_{\rho_1, \dots, \rho_K \geq \varepsilon/K} \frac{1}{K} \sum_{k=1}^K \widehat{\text{QV}}_k(h, \rho_k) \quad (15)$$

$$\text{subject to } \frac{1}{K} \sum_{k=1}^K \rho_k \leq \varepsilon \left(1 + \frac{1}{K}\right) + c_1 \sqrt{\frac{\log(\frac{K+2}{\delta})}{K}},$$

where c_1 is a universal constant. Then, the following bound holds with probability at least $1 - \delta$ for the meta-

distributionally robust loss of h around μ :

$$\sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} (\mathbb{E}_P [\ell(y, h(\mathbf{X}))]) \leq \widehat{U}^*(\varepsilon) + \mathcal{O} \left(\sqrt{\frac{1}{K} \log(K\delta^{-1})} + \frac{1}{K} \sum_{k=1}^K \sqrt{\frac{1}{n_k} \log \frac{Kn_k}{\varepsilon\delta}} \right).$$

The proof is provided in Appendix E. Similar to Theorem 5.2, Theorem 6.3 offers a robust bound on the expected loss under Wasserstein meta-distributional shifts using $\widehat{U}^*(\varepsilon)$. Once again, the generalization gap decreases asymptotically as both K and $(\min_{k \in [K]} n_k) / \log K$ increase with rate $\mathcal{O} \left(\sqrt{\log K / K} + \max_{k \in [K]} \sqrt{\frac{1}{n_k} \log(Kn_k)} \right)$. It is important to note that the inherent robustness of h against Wasserstein-type distributional shifts, if it exists, would be reflected in $\widehat{U}^*(\varepsilon)$, thereby reducing the bound.

A notable contribution of our work in this section is that we show robust evaluation against Wasserstein adversarial attacks to μ is possible using only $\widehat{\text{QV}}_i(h, \rho)$ values. In other words, it is not needed to move data points into a central server to perform collective attacks, and each client can independently assess the performance of h under local attacks. However, the budget values ρ_k for each client need to be carefully selected. In Section 7, we show this is possible at the server-side in polynomial time. Figure 3 gives a graphical illustration of the procedure in Theorem 6.3.

7. Asymptotic Minimax Optimality, Privacy, and Computational Efficiency

In this section, we address the computational complexity of our proposed empirical upper-bounds in Theorems 4.1, 5.2, 5.3 and 6.3 both the server and client sides. We also discuss their privacy w.r.t. local user data. Finally, we prove their asymptotic minimax optimality.

Remark 7.1 (Server-side Computational Complexity).

The server-side optimization problems in Theorems 5.2 and 5.3 are convex and efficiently (in polynomial-time w.r.t. K) solvable. Each client is queried only once for their non-robust (ordinary) loss value ($\rho_k = 0$). Robust local losses are not required for f -divergence shifts, since the reweighting attack can be carried out entirely at the server-side. On the other hand, the bound in Theorem 6.3, i.e., $\widehat{U}^*(\varepsilon)$, is the solution to a *quasi-convex* program, where a standard bisection method (e.g., Algorithm 1 in Appendix F), can approximate $\widehat{U}^*(\varepsilon)$ within an arbitrary error margin $\Delta > 0$, in polynomial time w.r.t. K and $\frac{1}{\Delta}$. The total query budget per client in this case is also polynomial with respect to both K and $\log \frac{1}{\Delta}$.

Remark 7.2 (Client-side Computational Complexity). For the case of f -divergence shifts in Theorems 5.2 and 5.3,

clients only return the ordinary loss values which can be computed in time $\mathcal{O}(n_k)$ at client k . In Theorem 6.3, assume the transportation cost c is convex with respect to its second argument² and is differentiable. Then, the client-side optimization problem to determine $\widehat{QV}_k(h, \rho)$ in equation 15, for any given h and ρ , is convex (Sinha et al., 2018). A standard stochastic gradient descent algorithm can approximate $\widehat{QV}_k(h, \rho)$ within an arbitrary error margin $\Delta > 0$, with polynomial time complexity relative to Δ^{-1} .

Proofs of Remarks 7.1 and 7.2, as well as the bisection method in Algorithm 1 can be found in Appendix F.

Remark 7.3 (Privacy). All empirical bounds in this work only use the local ordinary or adversarial loss values, without direct access to local private samples or model gradients. In particular, the empirical value $\widehat{U}^*(\varepsilon)$ in equation 15 relies solely on the Wasserstein adversarial loss of clients, i.e., $\widehat{QV}_k(h, \rho)$ for polynomially many values of ρ . To the best of our knowledge, there are no well-known privacy attacks capable of effectively recovering private data from this procedure. It should be noted that providing Differential Privacy (DP) certificates goes beyond the scope of our work.

Notably, working on novel attacks or providing information-theoretic analysis for the recoverability of private data from multiple local adversarial loss values can be a proper future research direction. Finally, the following remark (proved in Appendix F) provides tightness guarantees for our bounds.

Remark 7.4 (Asymptotic Minimax Optimality). All empirical upper-bounds in Theorems 4.1, 5.2, 5.3 and 6.3 are asymptotically minimax optimal, which means they can be achieved by a worst-case shift when K and all $n_k / \log K$ grow toward infinity. This means our bounds are tight and cannot be improved, at least when network and local dataset sizes become sufficiently large.

8. Experimental Results

Our work is mainly theoretical; In any case, we present a series of experiments on real-world datasets to show tightness and computability of our bounds in practice. First, we outline our client generation model and present a number of non-robust risk CDF guarantees. A more complete set of experiments with complementary explanations can be found in Appendix G. We simulated a federated learning scenario with $n = 1000$ nodes, where each node contains 1000 local samples. The experiments were conducted using four different datasets: CIFAR-10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), EMNIST (Cohen et al., 2017), and ImageNet (Russakovsky et al., 2015). To create each user’s data within the network, we applied three types of affine distribution shifts across users:

²This assumption is generally not restrictive, as any norm exhibits this property

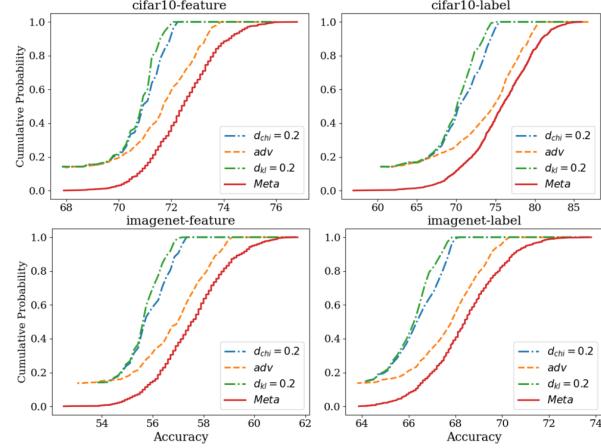


Figure 4. Non-robust Risk CDF bounds for unseen clients. Here, “Meta” refers to the main population with 1000 nodes. DKW-robust bounds are depicted only for tightness comparison.

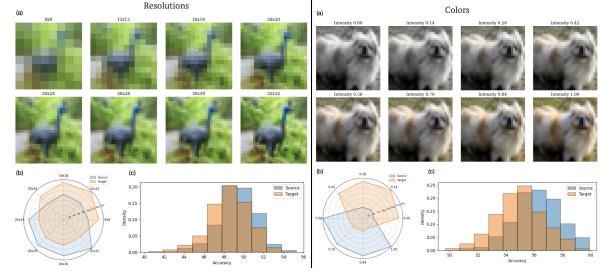


Figure 5. Meta-distributional shifts based on resolutions (left) and colors (right). (a) Sample images from each resolution/color. (b) Average number of samples per resolution/color within each network selected from the meta distributions. (c) Histogram of model accuracy densities for the two meta distributions.

Feature Distribution Shift: Each sample $X_i^{(k)}$ in the dataset is manipulated via a transformation chosen randomly for each node. Specifically, each user is assigned a unique matrix $\Lambda^{(k)}$ and shift vector $\delta^{(k)}$, and the data is modified as $\tilde{X}_i^{(k)} = (I + \Lambda^{(k)})X_i^{(k)} + \delta^{(k)}$, where $\Lambda^{(k)}$ and $\delta^{(k)}$ are respectively random matrices and vectors with i.i.d. zero-mean Gaussian entries. The standard deviation varies based on the dataset: 0.05 for CIFAR-10 and SVHN, 0.1 for EMNIST, and 0.01 for ImageNet.

Label Distribution Shift: Here, we assume that each meta-distribution is characterized by a specific α coefficient. To generate each user’s data under this shift, the number of samples per class is determined by a Dirichlet distribution with parameter α . In our experiments, we use $\alpha = 0.4$.

Feature & Label Distribution Shift: This shift combines both the feature and label distribution shifts described above to create a new distribution for each user.

Figure 4 illustrates our bounds on the risk CDF of unseen

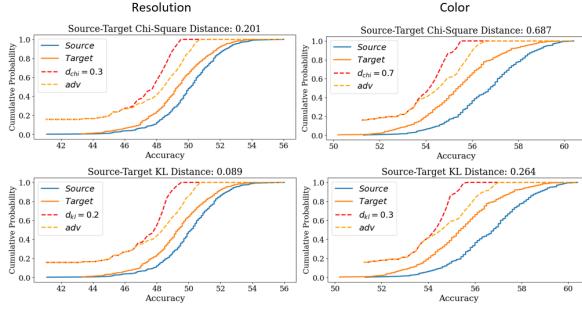


Figure 6. f -divergence certificates for two meta-distributions based on resolutions (Left) and colors (Right).

clients with no shifts. We selected 500 nodes from the population and considered 500 other nodes as unseen clients. We then plotted the CDFs based on 500 samples and confirmed that our bounds hold for the real population as well. Due to the standard DKW inequality, the empirical CDF is a good estimate for the test-time non-robust risk CDF.

8.1. f -Divergence Meta-Distributional Shifts

We assumed users belong to two distinct meta-distributions: the source and the target. A CNN-based model is initially trained on a network of clients sampled from the source. The resulting risk values are then fed into the optimization problems in Section 5 to obtain robust CDF bounds, considering both the Chi-Square and KL divergence as potential choices for f . Finally, we empirically estimate the risk CDF for users from the target meta-distribution and validate our bounds. Specifically, we tested our certificates in two distinct settings using the CIFAR-10 dataset (see Figure 5). We generated various image categories with differing resolutions or color schemes, and then sampled from these categories to create different distributions. More details of this procedure can be found in Section G.

Figure 6 verifies our CDF certificates based on both chi-square and KL-divergence (dotted curves) for the target meta distribution (orange curve). As can be seen, bounds have tightly captured the behavior of risk CDF in the target network. More detailed experiments are shown in Figure 9 in Appendix G.

We then used the above-mentioned affine distribution shifts to create new domains according to a Wasserstein metric. Figure 7 summarizes our numerical results in this scenario. To generate different networks within the meta-distribution, we applied the affine distribution shifts described in Section G.1. Once again, the results validate our certificates, this time for Wasserstein-type shifts. It is important to note that the bounds presented here remain tight, particularly under adversarial attacks as defined by a distributional adversary

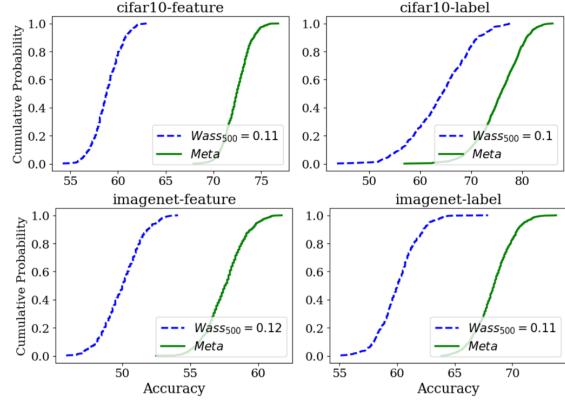


Figure 7. Wasserstein-based certificates for unseen clients. ‘‘Meta’’ refers to the main population with 1000 nodes. Dotted curves are based on 500 networks within the population.

in (Sinha et al., 2018). More detailed experiments with various levels of tightness are shown in Figure 10 in Appendix G. Although our theoretical findings in Section 6 focus on the average risk and not the risk CDF, we extended the same framework to the CDF in this experiment to explore whether the theory might also apply. The results were positive, suggesting potential for extending our theoretical findings in this area.

9. Conclusion

This work introduces new polynomial-time computable performance bounds to guarantee the performance of a model h on an unseen network B, using only data and queries from network A. The key assumption is that the meta-distributions behind user data in the two networks are ε -close, measured by f -divergence or Wasserstein shifts, which address diverse practical scenarios. The bounds, backed by rigorous proofs, achieve vanishing generalization gaps as network size K and normalized samples grow. Novel contributions include a robust GC theorem and DKW bound for f -divergence shifts. Experiments confirm the bounds’ tightness and efficiency.

Impact Statement

This paper proposes a certifiably robust evaluation framework for federated learning under meta-distributional shifts, providing theoretical guarantees on performance under deployment shifts. The proposed approach may benefit real-world federated systems by improving reliability and trust in evaluation across diverse clients. However, if misapplied without properly validating the assumptions on shift structure, the certificates may give a false sense of robustness. Care should be taken to ensure the theoretical assumptions hold approximately in practice, and to transparently communicate the limitations of the guarantees.

Acknowledgments

The work of Farzan Farnia is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and is partially supported by CUHK Direct Research Grants with CUHK Project No. 4055164 and 4937054.

References

- Anurag Ajay, Abhishek Gupta, Dibya Ghosh, Sergey Levine, and Pukit Agrawal. Distributionally adaptive meta reinforcement learning. *Advances in Neural Information Processing Systems*, 35:25856–25869, 2022.
- Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. *Advances in Neural Information Processing Systems*, 31, 2018.
- Adnan Ben Mansour, Gaia Carenini, Alexandre Duplessis, and David Naccache. Federated learning aggregation: New robust algorithms with guarantees. *arXiv preprint arXiv:2205.10864*, 2022.
- Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.
- Stephen Boyd. Convex optimization. *Cambridge UP*, 2004.
- Wei-Ning Chen, Graham Cormode, Akash Bharadwaj, Peter Romov, and Ayfer Ozgur. Federated experiment design under distributed differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2458–2466. PMLR, 2024.
- Yiqiang Chen, Wang Lu, Xin Qin, Jindong Wang, and Xing Xie. Metafed: Federated learning among federations with cyclic knowledge distillation for personalized healthcare. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- Shu-Ling Cheng, Chin-Yuan Yeh, Ting-An Chen, Eliana Pastor, and Ming-Syan Chen. Fedgcr: Achieving performance and fairness for federated learning with distinct client types via group customization and reweighting. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, 2024.
- Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.
- Theo Diamandis, Yonina Eldar, Alireza Fallah, Farzan Farnia, and Asuman Ozdaglar. A wasserstein minimax framework for mixed linear regression. In *International Conference on Machine Learning*, pages 2697–2706. PMLR, 2021.
- Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.
- Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.
- Farzan Farnia, Amirhossein Reisizadeh, Ramin Pedarsani, and Ali Jadbabaie. An optimal transport approach to personalized federated learning. *IEEE Journal on Selected Areas in Information Theory*, 3(2):162–171, 2022.
- Ruize Gao, Feng Liu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Maximum mean discrepancy test is aware of adversarial attacks. In *International Conference on Machine Learning*, pages 3564–3575. PMLR, 2021.
- Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.
- Peter W Glynn and Donald L Iglehart. Importance sampling for stochastic simulations. *Management science*, 35(11):1367–1392, 1989.
- Kai Hu, Yaogen Li, Shuai Zhang, Jiasheng Wu, Sheng Gong, Shanshan Jiang, and Liguo Weng. Fedmmd: A federated weighting algorithm considering non-iid and local model deviation. *Expert Systems with Applications*, 237:121463, 2024.
- Yongzhe Jia, Xuyun Zhang, Amin Beheshti, and Wanchun Dou. Fedlps: Heterogeneous federated learning for multiple tasks with local parameter sharing. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, 2024.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafeezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research*

- & management science in the age of analytics, pages 130–166. Informs, 2019.
- Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 29(4):967–996, 2021.
- Haoyu Lei, Shizhan Gong, Qi Dou, and Farzan Farnia. pfedfair: Towards optimal group fairness-accuracy trade-off in heterogeneous federated learning. *arXiv preprint arXiv:2503.14925*, 2025.
- Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning with matched averaging. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- Bingyan Liu, Nuoyan Lv, Yuanchun Guo, and Yawen Li. Recent advances on federated learning: A systematic survey. *Neurocomputing*, page 128019, 2024.
- Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.
- Mengmeng Ma, Tang Li, and Xi Peng. Beyond the federation: Topology-aware federated learning for generalization to unseen clients. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33794–33810. PMLR, 21–27 Jul 2024.
- Sarah Matta, Mathieu Lamard, Laurent Borderie, Alexandre Le Guilcher, Pascale Massin, Jean-Bernard Rottier, Beatrice Cochener, and Gwenolé Quellec. Domain generalization for multi-disease detection in fundus photographs. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4. IEEE, 2024.
- Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- Ronak Mehta, Vincent Roulet, Krishna Pillutla, and Zaid Harchaoui. Distributionally robust optimization with bias and variance reduction. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.
- Kumar Kshitij Patel, Lingxiao Wang, Blake E Woodworth, Brian Bullins, and Nati Srebro. Towards optimal communication complexity in distributed non-convex optimization. *Advances in Neural Information Processing Systems*, 35:13316–13328, 2022.
- Krishna Pillutla, Yassine Laguel, Jérôme Malick, and Zaid Harchaoui. Federated learning with superquantile aggregation for heterogeneous data. *Machine Learning*, 113(5):2955–3022, 2024.
- Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.
- Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. Robust federated learning: The case of affine distribution shifts. *Advances in Neural Information Processing Systems*, 33:21554–21565, 2020.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
- Seyed Amir Hossein Saberi, Amir Najafi, Abolfazl Motahari, and Babak Khalaj. Sample complexity bounds for learning high-dimensional simplices in noisy regimes. In *International Conference on Machine Learning*, pages 29514–29541. PMLR, 2023.
- Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.
- Behnaz Soltani, Yipeng Zhou, Venus Haghighi, and John Lui. A survey of federated evaluation in federated learning. *arXiv preprint arXiv:2305.08070*, 2023.
- Michel Talagrand. The glivenko-cantelli problem. *The Annals of Probability*, pages 837–870, 1987.
- Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022.
- Zhenyi Wang, Xiaoyang Wang, Li Shen, Qiuling Suo, Kaiqiang Song, Dong Yu, Yan Shen, and Mingchen

Gao. Meta-learning without data via wasserstein distributionally-robust model fusion. In *Uncertainty in Artificial Intelligence*, pages 2045–2055. PMLR, 2022.

Jun Wu, Jingrui He, and Hanghang Tong. Distributional network of networks for modeling data heterogeneity. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3379–3390, 2024a.

Xinghao Wu, Xuefeng Liu, Jianwei Niu, Haolin Wang, Shaojie Tang, and Guogang Zhu. Fedlora: When personalized federated learning meets low-rank adaptation. In *International Conference on Learning Representations (ICLR)*, 2024b.

Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

Honglin Yuan, Warren Morningstar, Lin Ning, and Karan Singh. What do we mean by generalization in federated learning? *arXiv preprint arXiv:2110.14216*, 2021.

Syed Zawad, Ahsan Ali, Pin-Yu Chen, Ali Anwar, Yi Zhou, Nathalie Baracaldo, Yuan Tian, and Feng Yan. Curse or redemption? how data heterogeneity affects the robustness of federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10807–10814, 2021.

Yan Zeng, Yuankai Mu, Junfeng Yuan, Siyuan Teng, Jilin Zhang, Jian Wan, Yongjian Ren, and Yunquan Zhang. Adaptive federated learning with non-iid data. *The Computer Journal*, 66(11):2758–2772, 2023.

Yibo Zeng and Henry Lam. Generalization bounds with minimal dependency on hypothesis class via distributionally robust optimization. *Advances in Neural Information Processing Systems*, 35:27576–27590, 2022.

Jie Zhang, Bo Li, Chen Chen, Lingjuan Lyu, Shuang Wu, Shouhong Ding, and Chao Wu. Delving into the adversarial robustness of federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Luhao Zhang, Jincheng Yang, and Rui Gao. A short and general duality proof for wasserstein distributionally robust optimization. *Operations Research*, 2024.

Yao Zhou, Jun Wu, Haixun Wang, and Jingrui He. Adversarial robustness through bias variance decomposition: A new perspective for federated learning. *arXiv preprint arXiv:2009.09026*, 2022.

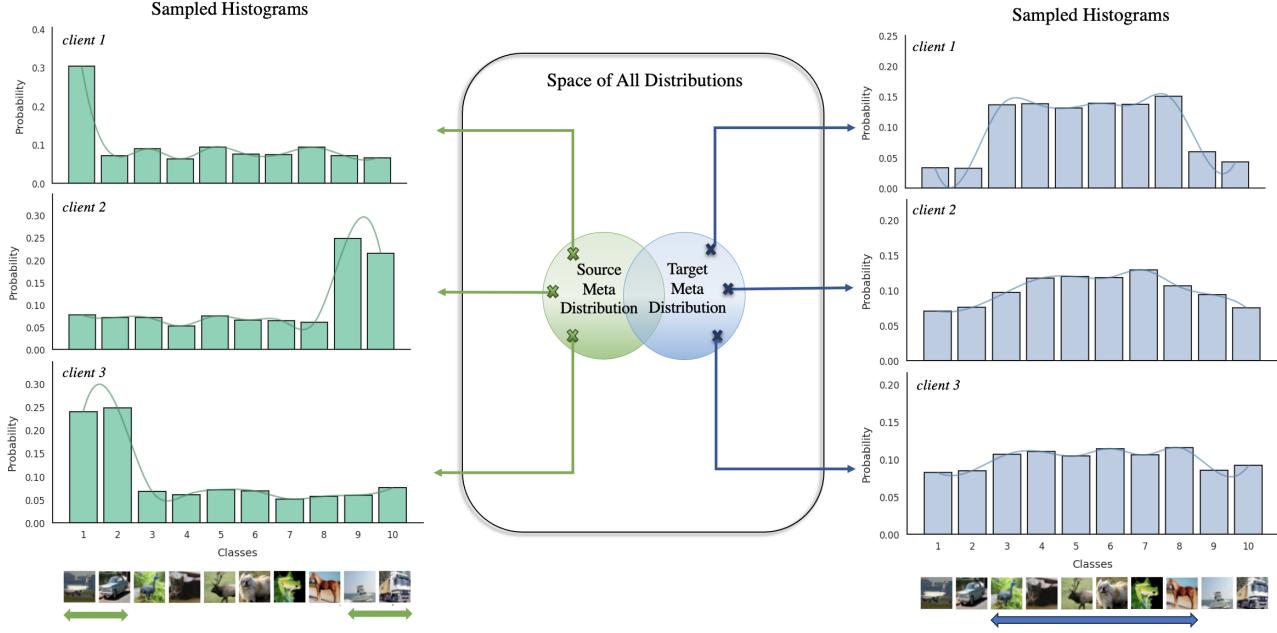


Figure 8. A graphical illustration of meta-distributional shifts between two different networks (or societies) of clients/users. Heterogeneous user data distributions can be considered as i.i.d. samples from a meta-distribution. In this example, clients from one meta-distribution primarily have vehicle images, while those from another meta-distribution mostly have animal images.

A. Complementary Review of Previous works

In this section, we review several related works in federated training and evaluation of models with non-IID clients. We then review a series of known (mostly theoretical) results regarding generalization gaps and shortcomings of current methodologies related to our problem set in this paper, i.e., beta testing in FL scenarios.

A.1. Generalization in FL

The challenge of generalizing FL models to unseen clients and distributions has been studied in several related works. Li et al. (2020) introduce a method to address the heterogeneity of client data, focusing on improving the generalization of FL models. Similarly, Ma et al. (2024) propose a topology-aware federated learning approach that leverages client relationships to enhance model robustness against out-of-federation data. Also, Zeng et al. (2023) explore adaptive federated learning techniques to dynamically adjust model parameters based on client data distributions. The robustness of FL models against distribution shifts and adversarial attacks has been the focus of several related references. Reisizadeh et al. (2020) propose a robust federated learning framework to handle affine distribution shifts across clients' data. Their proposed framework incorporates a Wasserstein-distance-based distribution shift model to account for device-dependent data perturbations. Also, Zhang et al. (2023) conduct comprehensive evaluations on the adversarial robustness of FL models, proposing the decision boundary-based FL Training algorithm to enhance the trained model's robustness. Zhou et al. (2022) gain insight from the bias-variance decomposition to improve adversarial robustness in FL. Also, Ben Mansour et al. (2022) propose a robust aggregation method to reduce the effect of adversarial clients.

A.2. Non-IID Federated Learning

The heterogeneous nature of FL, where clients have different data distributions has been a topic of great interest in the literature. In particular, modeling the heterogeneity of local data distributions/datasets via assuming a higher-level meta-distribution has been the center of several researches.

See, for example, Wu et al. (2024a); Chen et al. (2023); Patel et al. (2022); Matta et al. (2024). Please refer to Figures 1, 2 and 8 for a more illustrative definition of meta-distribution over local statistical or empirical distributions in a heterogeneous and non-IID network. In this regard, Fallah et al. (2020) introduce a personalized federated learning framework based on model-agnostic meta-learning, which provides performance guarantees by optimizing for data distribution heterogeneity. Farnia et al. (2022) propose an optimal transport approach to personalized federated learning by learning and inverting the optimal transport map between the distribution of clients, which has been further extended to fairness-aware personalized federated learning in (Lei et al., 2025). Diamandis et al. (2021) propose a Wasserstein-based framework to federated learning under the non-IID scenario by training a mixed linear regression model capturing the non-IID-ness of the clients' data. Luo et al. (2021) propose a classifier calibration method that adjusts for bias in heterogeneous data, offering improved performance guarantees in non-IID settings. Tan et al. (2022) develop FedProto, a framework that leverages prototype learning to improve convergence and robustness under non-convex objectives. Moreover, Wu et al. (2024b) propose FedLoRA, which adapts low-rank parameter sharing techniques to mitigate the effects of heterogeneity in personalized federated learning. Also, Cheng et al. (2024) and Jia et al. (2024) introduce group-based customization and local parameter sharing strategies, respectively, to provide fairness and efficiency guarantees for heterogeneous client types and multiple tasks in FL. A recent work on using super-quantiles in FL with heterogeneous clients is (Pillutla et al., 2024), which focuses on optimization and convergence aspects of the this problem.

A.3. Theory of Federated Model Evaluation

Evaluating models over networks of clients, also known as beta testing, or alpha/beta testing, or A/B testing, involves evaluating the performance of a model or application on a small, randomly selected sample of potential users before wider deployment (Soltani et al., 2023). Specifically, this process assumes that a central server has access to a limited subset of K clients from a large network and can request these clients to evaluate a given machine learning model h on their local data. The clients then report the results to the server. The objective is to estimate the *average* performance of the model across the entire network based on this limited sample (Liu et al., 2024).

Several critical challenges emerge in beta testing scenarios: i) **Data Privacy**: Can we reliably assess the average performance without direct access to clients' local data? ii) **Sample Efficiency**: How many randomly selected clients are necessary to achieve a reliable performance assessment? iii) **Meta-Distributional Robustness**: How can we guarantee a minimum level of performance for the model if the underlying meta-distribution governing the clients/users changes slightly? Other considerations include the computational efficiency of server-side and client-side algorithms required for such evaluations.

Questions (i) and (ii) above are relatively well-addressed in the literature, either through federated evaluation frameworks (see Sections A.1 and A.2), or via conventional concentration inequalities (see, for example, our Theorem 4.1). However, the challenge of *Meta-Distributionally Robust Assessment* (M-DRA) remains underexplored. This problem addresses the robustness of model evaluation when the meta-distribution governing client data (μ) shifts slightly to a new distribution (μ'). A straightforward yet naive approach to tackle this problem is to use the concept of **Maximum Mean Discrepancy (MMD)** (see Hu et al. (2024) or Gao et al. (2021)) in order to provide a bound on the performance of h on the target meta-distribution μ' , based on its performance on the source meta-distribution μ (Sinha et al., 2018). Mathematically, for any meta-distribution μ' which is in an ε -vicinity of μ , we have:

$$\begin{aligned} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P (\ell(y, h(\mathbf{X})))] &= \mathbb{E}_{P \sim \mu} [\mathbb{E}_P (\ell(y, h(\mathbf{X})))] + \mathbb{E}_{\mu' - \mu} [\mathbb{E}_P (\ell(y, h(\mathbf{X})))] \\ &\leq \mathbb{E}_{P \sim \mu} [\mathbb{E}_P (\ell(y, h(\mathbf{X})))] + \sup_{h \in \mathcal{H}} \mathbb{E}_{\mu' - \mu} [\mathbb{E}_P (\ell(y, h(\mathbf{X})))] \\ &= \mathbb{E}_{P \sim \mu} [\mathbb{E}_P (\ell(y, h(\mathbf{X})))] + \text{MMD}(\mu, \mu' | \mathcal{H}), \end{aligned} \quad (16)$$

where \mathcal{H} is a class of functions that h belongs to, and $\mathbb{E}_{\mu' - \mu} (\cdot)$ is defined as $\mathbb{E}_{\mu'} (\cdot) - \mathbb{E}_{\mu} (\cdot)$. This way, closeness of μ' to μ (ε -vicinity) is reflected into the MMD value, i.e., MMD is proportional to ε . Also, note that the first term, i.e., $\mathbb{E}_{P \sim \mu} [\mathbb{E}_P (\ell(y, h(\mathbf{X})))]$ can be estimated according to any unbiased averaging over the sample distributions P_1, \dots, P_K of μ (and not μ'), or their empirical counterparts $\widehat{P}_{1:K}$.

While MMD offers a promising starting point, however, the gap term $\text{MMD}(\mu, \mu' | \mathcal{H})$ does not shrink as K or either of n_k s increase asymptotically. It is an ε -dependent and non-vanishing error term that inflates the adversarial loss of h regardless of any possible inherent robustness in h around μ . To the best of our knowledge, there are no prior works that have already addressed this issue in federated beta testing (or model evaluation in general).

B. Uniform Convergence of Empirical CDFs from Adversarial Samples: GC Theory Revisited

Assume we are given n i.i.d. samples of a distribution P , denoted by $X_1, \dots, X_n \in \mathbb{R}$ and we aim to estimate the true cumulative distribution function (CDF) of $X \sim P$ based on these empirical observations. The Glivenko-Cantelli theorem provides strong guarantees on the asymptotic behavior of the worst-case error when estimating a CDF from a finite number of i.i.d. samples (Talagrand, 1987). The theorem states that the ℓ_∞ -norm of the difference between the true CDF of $X \sim P$, denoted by F , and its empirical version based on n i.i.d. samples, denoted by \hat{F}_n , which can be formulated as

$$\hat{F}_n(x) \triangleq \frac{1}{n} \sum_{i \in [n]} \mathbb{1}(X_i \leq x), \quad \forall x \in \mathbb{R},$$

almost surely converges to zero as n approaches infinity. Mathematically, this is expressed as:

$$\lim_{n \rightarrow \infty} \|F - \hat{F}_n\|_\infty \stackrel{a.s.}{=} 0. \quad (17)$$

Furthermore, the well-known Dvoretzky-Kiefer-Wolfowitz (DKW) theorem provides non-asymptotic bounds for this asymptotic behavior.

Lemma B.1 (DKW Inequality (Dvoretzky et al., 1956)). *Let \mathcal{X} be a measurable subset of \mathbb{R} , and let P be any probability measure supported on \mathcal{X} . Let $X_1, \dots, X_n \in \mathbb{R}$ be n i.i.d. samples from P . Then, the following bound holds for the ℓ_∞ -norm of the difference between the empirical and true CDF of P :*

$$\mathbb{P} \left(\sup_{\lambda \in \mathbb{R}} \left| P(X \geq \lambda) - \frac{1}{n} \sum_{i \in [n]} \mathbb{1}(X_i \geq \lambda) \right| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \right) \geq 1 - \delta, \quad (18)$$

for any $\delta > 0$.

The proof of this lemma can be found in the reference. This result allows us to provide uniform convergence guarantees on the tail probability of the loss of h , i.e., $\mathbb{P}(\ell(y, h(\mathbf{X})) \geq \lambda)$ for any $\lambda \in \mathbb{R}$, based on K i.i.d. observations of the loss across the network. By “uniform,” we mean that, with high probability over the sampling of the K users, the worst deviation between the empirical and statistical tail probability is consistently bounded.

For the case of f -divergence robustness, we show that it is also possible to derive an asymptotically consistent and *uniform* bound for the risk distribution (CDF) over the *target* network, which estimates

$$\mu'(\mathbb{E}_P[\ell(y, h(\mathbf{X}))] \geq \lambda),$$

for all $\lambda \in \mathbb{R}$, simultaneously. Here, μ' represents the unknown target meta-distribution, which is assumed to be within an ε -proximity of the source meta-distribution μ . Note that μ is also unknown, and our access to it is through K independent empirical realizations $\hat{P}_1, \dots, \hat{P}_K$, where each \hat{P}_k is known only to client k via a private sample set of size n_k . To achieve this, we first derive a robust version of the uniform convergence bound on empirical CDFs, i.e., the robust Dvoretzky-Kiefer-Wolfowitz (DKW) inequality:

Lemma B.2 (Robust Version of DKW Inequality). *Let μ and μ' be two probability measures on \mathbb{R} where μ' is absolutely continuous w.r.t. μ . Assume X_1, \dots, X_n for $n \in \mathbb{N}$ to be i.i.d. samples drawn from μ . Suppose we have $\mathcal{D}_f(\mu' \parallel \mu) \leq \varepsilon$ for some $\varepsilon \geq 0$ and a proper convex function $f(\cdot)$. For $\delta > 0$, let us define the set $\mathcal{A}_n(\varepsilon, \delta)$ as*

$$\mathcal{A}_n(\varepsilon, \delta) \triangleq \left\{ \boldsymbol{\alpha} \in \mathbb{R}_+^n \mid \left| \frac{1}{n} \sum_{i=1}^n \alpha_i - 1 \right| \leq c_1 \sqrt{n^{-1} \log \left(\frac{1}{\delta} \right)}, \frac{1}{n} \sum_{i=1}^n f(\alpha_i) \leq \varepsilon + c_2 \sqrt{n^{-1} \log \left(\frac{1}{\delta} \right)} \right\}, \quad (19)$$

where constants c_1 and c_2 depend only on $f(\cdot)$. Then, there exists $\boldsymbol{\alpha} \in \mathcal{A}_n(\varepsilon, \delta)$ such that the following uniform bound holds with probability at least $1 - \delta$:

$$\sup_{\lambda \in \mathbb{R}} \left| \mu'(\mathbf{X} \leq \lambda) - \frac{1}{n} \sum_{i=1}^n \alpha_i \mathbb{1}(X_i \leq \lambda) \right| \leq \mathcal{O} \left(\sqrt{n^{-1} \log \left(\frac{n}{\delta} \right)} \right). \quad (20)$$

Proof of Lemma B.2. The proof for most of its initial parts follows the same path as in the proof of Theorem 5.2. In particular, we use Lemmas D.1 and D.2 from the proof of Theorem 5.2 (see Section D) to show that the following events occur separately with probability at least $1 - \frac{\delta}{3}$, for any $\delta > 0$:

$$\left| \frac{1}{n} \sum_{i \in [n]} \frac{d\mu'(X_i)}{d\mu(X_i)} - 1 \right| \leq c_1 \sqrt{\frac{\log \frac{1}{\delta}}{n}}, \quad (21)$$

$$\left| \mathcal{D}_f(\mu' \| \mu) - \frac{1}{n} \sum_{i \in [n]} f\left(\frac{d\mu'(X_i)}{d\mu(X_i)}\right) \right| \leq c_2 \sqrt{\frac{\log \frac{1}{\delta}}{n}}, \quad (22)$$

where $c_1, c_2 > 0$ are constants depending on $f(\cdot)$. These probabilities are with respect to the randomness in drawing i.i.d. samples $X_1, \dots, X_n \sim \mu$. This is equivalent to the following statement:

$$\mathbb{P}\left[\left(\frac{d\mu'(X_i)}{d\mu(X_i)}\right)_{i \in [n]} \in \mathcal{A}_n\right] \geq 1 - \frac{2\delta}{3}. \quad (23)$$

Next, define $\hat{F}_n(\lambda)$ for $\lambda \in \mathbb{R}$ as

$$\hat{F}_n(\lambda) \triangleq \frac{1}{n} \sum_{i=1}^n \omega(X_i) \mathbb{1}(X_i \leq \lambda), \quad (24)$$

where the weight function $\omega(X) \geq 0$ is the unknown (bounded) density ratio between μ' and μ , i.e.,

$$\omega(X) \triangleq \frac{d\mu'(X)}{d\mu(X)}. \quad (25)$$

Since $\omega(\cdot)$ is non-negative, $\hat{F}_n(\lambda)$ is non-decreasing in λ , starting at 0 when $\lambda = -\infty$ and not exceeding an upper-bound $\Lambda < +\infty$ (due to absolute continuity) as $\lambda \rightarrow \infty$.

Consider the probability measure μ' . For $m \geq 2$, define $\lambda_0^*, \lambda_1^*, \dots, \lambda_m^*$ such that (i) $\lambda_0^* = -\infty$ and $\lambda_m^* = \infty$, and (ii) $\mu'(X \leq \lambda_i^*) = i/m$ for $i \in [m-1]$. These λ_i^* , $i \in [m] \cup \{0\}$ represent the m -quantiles of μ' . For any $\lambda \in \mathbb{R}$, let $i = i(\lambda) \in [m]$ be such that $\lambda \in [\lambda_{i-1}^*, \lambda_i^*)$. Then, the following chain of inequalities holds almost surely for all $\lambda \in \mathbb{R}$:

$$\begin{aligned} \hat{F}_n(\lambda) - \mu'(X \leq \lambda) &\leq \hat{F}_n(\lambda_i^*) - \mu'(X \leq \lambda_{i-1}^*) \leq \hat{F}_n(\lambda_i^*) - \mu'(X \leq \lambda_i^*) + \frac{1}{m}, \\ \hat{F}_n(\lambda) - \mu'(X \leq \lambda) &\geq \hat{F}_n(\lambda_{i-1}^*) - \mu'(X \leq \lambda_i^*) \geq \hat{F}_n(\lambda_{i-1}^*) - \mu'(X \leq \lambda_{i-1}^*) - \frac{1}{m}. \end{aligned} \quad (26)$$

Thus, the following bound holds for all $\lambda \in \mathbb{R}$:

$$\begin{aligned} \|\hat{F}_n - \mu'(X \leq \cdot)\|_\infty &= \sup_{\lambda \in \mathbb{R}} |\hat{F}_n(\lambda) - \mu'(X \leq \lambda)| \\ &\leq \max_{i \in \{0, 1, \dots, m\}} |\hat{F}_n(\lambda_i^*) - \mu'(X \leq \lambda_i^*)| + \frac{1}{m}. \end{aligned} \quad (27)$$

On the other hand, for any fixed $\lambda \in \mathbb{R}$, we have the following relation for the expectation of $\hat{F}_n(\lambda)$:

$$\begin{aligned} \mathbb{E}_\mu[\hat{F}_n(\lambda)] &= \mathbb{E}_\mu \left[\frac{1}{n} \sum_{i=1}^n \omega(X_i) \mathbb{1}(X_i \leq \lambda) \right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\mu [\omega(X_i) \mathbb{1}(X_i \leq \lambda)] \\ &= \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} \frac{d\mu'(X)}{d\mu(X)} \mathbb{1}(X \leq \lambda) d\mu(X) \\ &= \mu'(X \leq \lambda). \end{aligned} \quad (28)$$

Given that the weight functions $\omega(X_i)$ for $i \in [n]$ are bounded, and $\mathbb{1}(\cdot) \in \{0, 1\}$, McDiarmid's inequality states that for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\widehat{F}_n(\lambda_i^*) - \mu'(X \leq \lambda_i^*)\right| > \varepsilon\right) \leq 2e^{-2\mathcal{O}(n\varepsilon^2)}. \quad (29)$$

Therefore, using the union bound over all $i = 0, 1, \dots, m$, we obtain:

$$\mathbb{P}\left(\max_{i \in \{0, 1, \dots, m\}} \left|\widehat{F}_n(\lambda_i^*) - \mu'(X \leq \lambda_i^*)\right| > \varepsilon\right) \leq 2(m+1)e^{-2\mathcal{O}(n\varepsilon^2)}. \quad (30)$$

Equivalently, for any $\delta > 0$, the following bound holds with probability at least $1 - \delta/3$:

$$\max_{i \in \{0, 1, \dots, m\}} \left|\widehat{F}_n(\lambda_i^*) - \mu'(X \leq \lambda_i^*)\right| \leq \mathcal{O}\left(\sqrt{(2n)^{-1} \log\left(\frac{6(m+1)}{\delta}\right)}\right). \quad (31)$$

Using the preceding inequalities, in particular relations in equation 23, equation 27 and equation 31, we can say there exists $\alpha \in \mathcal{A}_n$ such that the following bounds for \widehat{F}_n hold with probability at least $1 - \delta$:

$$\begin{aligned} \left\|\widehat{F}_n - \mu'(X \leq \cdot)\right\|_\infty &\leq \max_{i \in \{0, 1, \dots, m\}} \left|\widehat{F}_n(\lambda_i^*) - \mu'(X \leq \lambda_i^*)\right| + \frac{1}{m} \\ &\leq \mathcal{O}\left(\inf_{m \in \mathbb{N}_{\geq 2}} \left\{\sqrt{(2n)^{-1} \log\left(\frac{6(m+1)}{\delta}\right)} + \frac{1}{m}\right\}\right) \\ &\leq \mathcal{O}\left(\sqrt{n^{-1} \log\left(\frac{n}{\delta}\right)}\right). \end{aligned} \quad (32)$$

Thus, the proof is complete. \square

A direct corollary of Lemma B.2 is the following robust (again with respect to f -divergence adversaries) of the well-known Gilivenko-Cantelli theorem:

Corollary B.3 (Robust Version of Glivenko-Cantelli Theorem). *Let μ and μ' be two probability measures on \mathbb{R} and let $\mathcal{S} \triangleq \{X_i\}_{i=1}^\infty$ be an i.i.d. sequence drawn from μ . Assume μ and μ' are absolutely continuous with respect to each other. Additionally, suppose $\mathcal{D}_f(\mu' \parallel \mu) \leq \varepsilon$ for some $\varepsilon \geq 0$ and a proper convex function $f(\cdot)$. Then, there exists a non-negative sequence $\{\alpha_i\}_{i=1}^\infty$ that can depend on \mathcal{S} , has the following properties:*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \alpha_i = 1 \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f(\alpha_i) \leq \varepsilon, \quad (33)$$

and also satisfies the following condition:

$$\lim_{n \rightarrow \infty} \sup_{\lambda \in \mathbb{R}} \left| \mu'(X \leq \lambda) - \frac{1}{n} \sum_{i=1}^n \alpha_i \mathbb{1}(X_i \leq \lambda) \right| \stackrel{a.s.}{=} 0. \quad (34)$$

Corollary B.3 follows directly from Lemma B.2.

C. Proofs of the Statements in Section 4

Proof of Theorem 4.1. The proof for the average risk is straightforward and combines several applications of McDiarmid's inequality (or, more simply in this case, Hoeffding's inequality). For any instance of distributions $P \sim \mu$, let us define the random variable $\zeta = \zeta(P)$ as follows:

$$\zeta(P) \triangleq \mathbb{E}_P[\ell(y, h(\mathbf{X}))] = \int_{\mathcal{Z}} \ell(y, h(\mathbf{X})) P(d\mathbf{Z}), \quad (35)$$

where ℓ denotes the loss function, and \mathcal{Z} is the space of possible outcomes. We omit the detailed proof that P being a random variable implies $\zeta(P)$ is also a random variable, as this follows from standard measurability arguments.

Now, for any $\delta > 0$, we apply McDiarmid's inequality to obtain:

$$\mathbb{P} \left(\zeta - \mathbb{E}_\mu[\zeta] \leq \sqrt{\frac{\log \frac{K+1}{\delta}}{2K}} \right) \geq 1 - \frac{\delta}{K+1}, \quad (36)$$

where the bound holds due to the one-sided version of McDiarmid's inequality, and the fact that $\zeta \in [0, 1]$ almost surely.

Next, for each $k \in [K]$, we similarly have:

$$\mathbb{P} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \ell \left(y_i^{(k)}, h \left(\mathbf{X}_i^{(k)} \right) \right) - \mathbb{E}_{P_i} [\ell(y, h(\mathbf{X}))] \leq \sqrt{\frac{\log \frac{K+1}{\delta}}{2n_k}} \right) \geq 1 - \frac{\delta}{K+1}. \quad (37)$$

This follows from Hoeffding's inequality, given that the data points in the local dataset of the k -th client are i.i.d. By the union bound, the above $K+1$ inequalities hold simultaneously with probability at least $1 - \delta$. Finally, combining these inequalities gives us the desired bound in the theorem, thus completing the proof.

For the bound concerning the empirical CDF, the proof follows more or less the same path. Let us define the *statistical* query value of the k th client as $\text{QV}_k(h)$, i.e.,

$$\text{QV}_k(h) \triangleq \mathbb{E}_{P_k} [\ell(y, h(\mathbf{X}))], \quad (38)$$

where P_k is the true (unknown) data generating distribution which is assigned to client $k \in [K]$. In this regard, according to Lemma B.1, for any $\delta > 0$ we have

$$\mathbb{P} \left(\sup_{\lambda \in \mathbb{R}} \left| \mu \left(\mathbb{E}_P [\ell(y, h(\mathbf{X}))] \geq \lambda \right) - \frac{1}{K} \sum_{k \in [K]} \mathbb{1} (\text{QV}_k(h) \geq \lambda) \right| \leq \sqrt{\frac{\log \frac{2(K+1)}{\delta}}{2K}} \right) \geq 1 - \frac{\delta}{K+1}.$$

Also, applying the McDiarmid's inequality for K times (once, with respect to each client $k \in [K]$), the following bounds also hold:

$$\mathbb{P} \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \ell \left(y_i^{(k)}, h \left(\mathbf{X}_i^{(k)} \right) \right) - \mathbb{E}_{P_i} [\ell(y, h(\mathbf{X}))] \leq \sqrt{\frac{\log \frac{K+1}{\delta}}{2n_k}} \right) \geq 1 - \frac{\delta}{K+1}, \quad (39)$$

which show the boundedness of the deviation between the empirical query values $\widehat{\text{QV}}_k(h)$ and statistical ones $\text{QV}_k(h)$. This is similar to the previous part of the proof. Again, applying union bound and combining all the inequalities mentioned so far in the proof, the final bound in the statement of the theorem holds with probability at least $1 - \delta$ and the proof is complete. \square

D. Proofs of the Statements in Section 5

Proof of Theorem 5.2. For each $k \in [K]$, let us define the event $\xi_1^{(k)}$ as follows:

$$\xi_1^{(k)} \equiv \mathbb{E}_{P_k} [\ell(y, h(\mathbf{X}))] \leq \frac{1}{n_k} \sum_{i \in [n_k]} \ell \left(y_i^{(k)}, h \left(\mathbf{X}_i^{(k)} \right) \right) + \sqrt{\frac{\log \left(\frac{K+3}{\delta} \right)}{2n_k}}, \quad (40)$$

where, since $\ell(\cdot)$ is assumed to be a 1-bounded loss function and the samples are drawn independently, McDiarmid's inequality tells us that $\mathbb{P} \left(\xi_1^{(k)} \right) \geq 1 - \frac{\delta}{K+3}$ (McDiarmid et al., 1989). For the rest of the proof, we assume the density ratio of the meta-distributions μ, μ' are bounded, i.e., we have

$$\frac{d\mu'}{d\mu}(P) \leq \Lambda, \quad \forall P \in \text{supp}(\mu), \quad (41)$$

for some positive Λ . We call this property the Λ -boundedness of density ratio. At the end of the proof, we show that in our problem we have $\Lambda \leq 1 + \kappa\varepsilon$. We now state and prove two essential lemmas which will be used in the subsequent arguments.

Lemma D.1. Consider two meta-distributions $\mu, \mu' \in \mathcal{M}^2$ which are absolutely continuous with respect to each other and have a Λ -bounded density ratio for some $\Lambda \geq 1$. For $K \in \mathbb{N}$, assume $P_1, \dots, P_K \in \mathcal{M}$ to be i.i.d. sample distributions sampled from μ . Then, for all $\epsilon > 0$, the following concentration bound holds:

$$\mathbb{P}\left(\left|\frac{1}{K} \sum_{k \in [K]} \frac{d\mu'(P_k)}{d\mu(P_k)} - 1\right| \geq \epsilon\right) \leq \exp\left(\frac{-2K\epsilon^2}{\Lambda^2(1-\Lambda^{-2})^2}\right). \quad (42)$$

Proof of Lemma D.1. Due to the assumed mutual absolute continuity, μ and μ' share the same support. Therefore, for $P \sim \mu$, we can define the scalar random variable

$$\zeta = \zeta(P) \triangleq \frac{d\mu'(P)}{d\mu(P)}. \quad (43)$$

This variable is bounded by $\Lambda^{-1} \stackrel{a.s.}{\leq} \zeta \stackrel{a.s.}{\leq} \Lambda$. Regarding the expected value of ζ , we have:

$$\mathbb{E}[\zeta] = \mathbb{E}_{P \sim \mu}\left[\frac{d\mu'(P)}{d\mu(P)}\right] = \int_{\mathcal{M}} \frac{d\mu'(P)}{d\mu(P)} d\mu(P) = 1. \quad (44)$$

Let $\zeta_k = \zeta(P_k)$. Since $\zeta(P_1), \dots, \zeta(P_K)$ represent i.i.d. instances of ζ , McDiarmid's inequality states that:

$$\mathbb{P}\left(\left|\frac{1}{K} \sum_{k \in [K]} \zeta_k - \mathbb{E}[\zeta]\right| \geq \epsilon\right) \leq \exp\left(-\frac{2K\epsilon^2}{(\Lambda - \Lambda^{-1})^2}\right), \quad (45)$$

which completes the proof. \square

Lemma D.2. For $\Lambda \geq 1$, assume two meta-distributions $\mu, \mu' \in \mathcal{M}^2$ are absolutely continuous with respect to each other and have a Λ -bounded density ratio. Let f be a convex function that satisfies the conditions described in Definition 5.1. For $K \in \mathbb{N}$, assume $P_1, \dots, P_K \in \mathcal{M}$ to be i.i.d. sample distributions sampled from μ . Then, the following concentration bound holds:

$$\mathbb{P}\left(\left|\mathcal{D}_f(\mu' \| \mu) - \frac{1}{K} \sum_{k \in [K]} f\left(\frac{d\mu'(P_k)}{d\mu(P_k)}\right)\right| \geq \epsilon\right) \leq \exp\left(\frac{-2K\epsilon^2}{\text{BW}^2(f(\cdot), \Lambda)}\right), \quad (46)$$

where $\text{BW}(f(\cdot), \Lambda)$ is defined as:

$$\text{BW}(f(\cdot), \Lambda) \triangleq \sup_{\Lambda^{-1} \leq u, v \leq \Lambda} f(u) - f(v). \quad (47)$$

Proof of Lemma D.2. The proof follows similarly to that of Lemma D.1. For $P \sim \mu$, let us define the random variable

$$\zeta(P) = f\left(\frac{d\mu'(P)}{d\mu(P)}\right). \quad (48)$$

Then, having defined $\zeta_k = \zeta(P_k)$, we know that ζ_1, \dots, ζ_K represent i.i.d. instances of ζ . Moreover, the expected value of ζ is the f -divergence between μ' and μ :

$$\mathbb{E}[\zeta] = \mathbb{E}_{P \sim \mu}\left[f\left(\frac{d\mu'(P)}{d\mu(P)}\right)\right] = \mathcal{D}_f(\mu' \| \mu). \quad (49)$$

Finally, since μ' and μ have the Λ -bounded density ratio property, the following bounds hold almost surely:

$$\zeta \stackrel{a.s.}{\leq} \sup_{\Lambda^{-1} \leq u \leq \Lambda} f(u) \quad , \quad \zeta \stackrel{a.s.}{\geq} \inf_{\Lambda^{-1} \leq v \leq \Lambda} f(v), \quad (50)$$

which means the range of ζ is almost surely equal to $\text{BW}(f(\cdot), \Lambda)$. Hence, again using McDiarmid's inequality, we get the bound:

$$\mathbb{P} \left(\left| \frac{1}{K} \sum_{k \in [K]} \zeta_k - \mathbb{E}[\zeta] \right| \geq \epsilon \right) \leq \exp \left(\frac{-2K\epsilon^2}{\text{BW}^2(f(\cdot), \Lambda)} \right), \quad (51)$$

and this completes the proof. \square

With the lemmas established, let us define additional events ξ_2 and ξ_3 based on the concentration bounds:

$$\xi_2 \equiv \frac{1}{K} \sum_{k \in [K]} \frac{d\mu'(P_k)}{d\mu(P_k)} \leq 1 \pm C_1 K^{-1/2}, \quad (52)$$

$$\xi_3 \equiv \frac{1}{K} \sum_{k \in [K]} f \left(\frac{d\mu'(P_k)}{d\mu(P_k)} \right) \leq \mathcal{D}_f(\mu' \| \mu) \pm C_2 K^{-1/2}, \quad (53)$$

where C_1, C_2 are constants which only depend on Λ and $f(\cdot)$, according to Lemmas D.1 and D.2. Based on the above arguments and the results of the mentioned lemmas, we have $\mathbb{P}(\xi_2), \mathbb{P}(\xi_3) \geq 1 - \frac{\delta}{K+3}$. Using the central idea for *importance sampling* (Glynn and Iglehart, 1989), the following equations hold for all μ, μ', ℓ and h :

$$\begin{aligned} \mathbb{E}_{P \sim \mu} \left[\left(\frac{d\mu'(P)}{d\mu(P)} \right) \mathbb{E}_P [\ell(y, h(\mathbf{X}))] \right] &= \int_{P \in \mathcal{M}} \left(\frac{d\mu'(P)}{d\mu(P)} \right) \mathbb{E}_P [\ell(y, h(\mathbf{X}))] d\mu(P) \\ &= \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(\mathbf{X}))]]. \end{aligned} \quad (54)$$

At this point, and similar to the idea of Lemma D.1, we define ξ_4 as the event of the empirical loss over meta-distribution μ' concentrates (with high probability) around its expected value, i.e.,

$$\begin{aligned} \xi_4 &\equiv \\ \mathbb{E}_{P \sim \mu} \left[\left(\frac{d\mu'(P)}{d\mu(P)} \right) \mathbb{E}_P [\ell(y, h(\mathbf{X}))] \right] &\leq \\ \frac{1}{K} \sum_{k \in [K]} \frac{d\mu'(P_k)}{d\mu(P_k)} \mathbb{E}_{P_k} [\ell(y, h(\mathbf{X}))] + \Lambda \sqrt{\frac{\log(\frac{K+3}{\delta})}{2K}}. \end{aligned} \quad (55)$$

Again, since

$$0 \stackrel{a.s.}{\leq} \frac{d\mu'(P_k)}{d\mu(P_k)} \mathbb{E}_{P_k} [\ell(y, h(\mathbf{X}))] \stackrel{a.s.}{\leq} \Lambda,$$

McDiarmid's inequality states that the probability bound $\mathbb{P}(\xi_4) \geq 1 - \frac{\delta}{K+3}$ holds. Our final definition in this proof is a random set of meta-distributions $\mathcal{G} \subseteq \mathcal{M}^2$ which represents an empirical candidate for the neighbors of μ . Mathematically speaking, let us define:

$$\begin{aligned} \mathcal{G} \triangleq \left\{ \nu \in \mathcal{M}^2 \mid \frac{1}{K} \sum_{k \in [K]} \frac{d\nu(P_k)}{d\mu(P_k)} \leq 1 \pm C_1 K^{-1/2}, \right. \\ \left. \frac{1}{K} \sum_{k \in [K]} f \left(\frac{d\nu(P_k)}{d\mu(P_k)} \right) \leq \varepsilon + C_2 K^{-1/2} \right\}, \end{aligned} \quad (56)$$

which depends on ε and has a random (empirical) nature since it also depends on sample distributions P_1, \dots, P_K . Based on prior discussions and lemmas, we have $\mu' \stackrel{a.s.}{\in} \mathcal{G}$ as long as the events ξ_2 and ξ_3 hold, simultaneously. By further assuming

that events ξ_4 and $\xi_1^{(k)}$ s for all $k \in [K]$ also hold, we can finally write the following chain of inequalities:

$$\begin{aligned} & \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(\mathbf{X}))]] \\ & \leq \frac{1}{K} \sum_{k \in [K]} \frac{d\mu'(P_k)}{d\mu(P_k)} \mathbb{E}_{P_k} [\ell(y, h(\mathbf{X}))] + \Lambda \sqrt{\frac{\log(\frac{K+3}{\delta})}{2K}} \\ & \leq \frac{1}{K} \sum_{k \in [K]} \frac{d\mu'(P_k)}{d\mu(P_k)} \widehat{\mathbb{E}}_{P_k} [\ell(y, h(\mathbf{X}))] + \Lambda \sqrt{\frac{\log(\frac{K+3}{\delta})}{2K}} + \frac{1}{K} \sum_{k \in [K]} \sqrt{\frac{\log(\frac{K+3}{\delta})}{2n_k}} \\ & \stackrel{a.s.}{\leq} \frac{1}{K} \sup_{\nu \in \mathcal{G}} \sum_{k \in [K]} \frac{d\nu(P_k)}{d\mu(P_k)} \widehat{\mathbb{E}}_{P_k} [\ell(y, h(\mathbf{X}))] + \sqrt{\log\left(\frac{K+3}{\delta}\right)} \left[\sqrt{\frac{\Lambda^2}{2K}} + \frac{1}{K} \sum_{k \in [K]} \sqrt{\frac{1}{2n_k}} \right]. \end{aligned} \quad (57)$$

It should be noted that the condition $\nu \in \mathcal{G}$ can be interpreted as introducing

$$\alpha_k \triangleq \frac{d\nu(P_k)}{d\mu(P_k)}, \quad \forall k \in [K],$$

and force $\alpha_1, \dots, \alpha_K$ to satisfy the constraints in the definition of $\widehat{B}^*(\varepsilon)$. Hence, this gives us the high probability bound claimed inside the statement of theorem. The only remaining part of the proof is to show events $\xi_1^{(k)}, \xi_2, \xi_3$ and ξ_4 for all $k \in [K]$ hold, simultaneously, with a probability at least $1 - \delta$.

For any event ξ , let ξ^c denote its complement. Then, we already have

$$\mathbb{P}(\xi_1^{(k)c}), \mathbb{P}(\xi_2^c), \dots, \mathbb{P}(\xi_4^c) \leq \frac{\delta}{K+3}, \quad \forall k \in [K].$$

In this regard, one can simply use the union bound and obtain the following chain of inequalities:

$$\mathbb{P}\left(\bigcup_{k \in [K]} \xi_1^{(k)c} \cup \xi_2 \cup \xi_3 \cup \xi_4\right) \leq \sum_{k \in [K]} \mathbb{P}(\xi_1^{(k)c}) + \sum_{i=2}^4 \mathbb{P}(\xi_i^c) = \frac{K\delta}{K+3} + 3\frac{\delta}{K+3} = \delta. \quad (58)$$

This means the bound in the statement of theorem holds with a probability at least $1 - \delta$, and thus completes the proof.

Finally, let us derive an upper-bound for Λ , a.k.a., the density ratio bound. According to the definition of f -divergence, we have:

$$\mathcal{D}_f(\mu' \parallel \mu) = \mathbb{E}_{P \sim \mu} \left[f\left(\frac{d\mu'}{d\mu}(P)\right) \right] \leq \varepsilon. \quad (59)$$

Assume there exists a region in the distributional space $\mathcal{S} \subseteq \mathcal{M}(\mathcal{Z})$, where for all $P \in \mathcal{S}$ the density ratio $d\mu'/d\mu(P)$ is at least Λ , for some $\Lambda \geq 1$. Then, we have $\mathcal{D}_f(\mu' \parallel \mu) \geq \mu(\mathcal{S}) f(\Lambda)$. Setting $\mu(\mathcal{S})$ equal to δ (our high-probability error margin in this problem), we get

$$f(\Lambda) \leq \varepsilon/\delta,$$

which means $\Lambda \leq f^{-1}(\varepsilon/\delta)$. On the other hand, we already know that f is a convex function (see Definition 5.1) which means f^{-1} must be concave. Additionally, we must have $f(1) = 0$. Therefore, the following inequality holds:

$$\Lambda \leq f^{-1}(\varepsilon/\delta) \leq 1 + \left(\delta^{-1} [f^{-1}]'(0)\right) \varepsilon \triangleq 1 + \kappa\varepsilon, \quad (60)$$

where κ does not depend on ε . \square

Proof of Theorem 5.3. Similar to the proof of Theorem 5.2, we begin by noting that, due to McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \frac{K\delta}{K+2}$, the following set of inequalities holds simultaneously for all $k \in [K]$:

$$\widehat{\text{QV}}(h) \leq \text{QV}(h) + \sqrt{\frac{\log(\frac{K+2}{\delta})}{2n_k}}. \quad (61)$$

Next, it can be readily verified that:

$$\mathbb{1}(\text{QV}(h) \geq \lambda) \leq \mathbb{1}\left(\widehat{\text{QV}}(h) \geq \lambda - \sqrt{\frac{\log\left(\frac{K+2}{\delta}\right)}{2n_k}}\right), \quad \forall k \in [K]. \quad (62)$$

Additionally, note that:

$$\mathbb{E}_{P \sim \mu'} [\mathbb{1}(\text{QV}(h) \geq \lambda)] = \mu'(\text{QV}(h) \geq \lambda). \quad (63)$$

The remainder of the proof simply involves applying the result of Lemma B.2 with a maximum error probability of $\frac{2\delta}{K+2}$. This concludes the proof. \square

E. Proofs of the Statements in Section 6

Proof of Theorem 6.3. Proof consists of two parts:

- Proving the statement of theorem for the statistical case, where $\min_{k \in [K]} n_k \rightarrow \infty$ and thus we have $\widehat{\text{QV}}_k(h, \rho) = \text{QV}_k(h, \rho)$ for all $h \in \mathcal{H}, \rho \geq 0$ and $k \in [K]$.
- Replacing the statistically exact adversarial loss $\text{QV}_k(h, \rho)$ which is based on the unknown distribution sample P_k with its empirically calculated counterpart $\widehat{\text{QV}}_k(h, \rho)$ which is computed based on the known (yet private) distribution \widehat{P}_k for all $k \in [K]$. This part of the proof requires establishing a uniform convergence bound over all values of $\rho \geq 0$.

Part I The core mathematical tool used throughout the proof is the following duality result from (Sinha et al., 2018) (originally derived in (Blanchet and Murthy, 2019)) which works for general Wasserstein-constrained optimization problems:

Lemma E.1 (Proposition 1 of Sinha et al. (2018)). *Let P be a probability measure defined over a measurable space Ω , $\ell(\cdot) : \Omega \rightarrow \mathbb{R}$ be any loss function, c denote a proper and lower semi-continuous transportation cost on $\Omega \times \Omega$, and assume $\varepsilon \geq 0$. Then, the following equality holds for the Wasserstein-constrained DRO around P :*

$$\sup_{Q \in \mathcal{B}_\varepsilon^{\text{wass}}(P)} \mathbb{E}_Q [\ell(Z)] = \inf_{\gamma \geq 0} \left\{ \gamma\varepsilon + \mathbb{E}_P \left[\sup_{Z' \in \Omega} \ell(Z') - \gamma c(Z', Z) \right] \right\}. \quad (64)$$

Proof can be found inside the reference. Also, (Blanchet and Murthy, 2019) and (Zhang et al., 2024) along with several other papers have theoretically analyzed alternative proofs. Based on the duality formulation in Lemma E.1, and considering the fact that meta-distribution μ is also a “distribution” over the measurable space \mathcal{M} , one can rewrite the original Wasserstein-constrained MDRO in the statement of the theorem in its dual form:

$$\begin{aligned} & \sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(X))]] \\ &= \inf_{\gamma \geq 0} \left\{ \gamma\varepsilon + \mathbb{E}_{P \sim \mu} \left[\sup_Q \mathbb{E}_Q [\ell(y, h(X))] - \gamma \mathcal{W}_c(P, Q) \right] \right\} \\ &= \inf_{\gamma \geq 0} \left\{ \mathbb{E}_\mu \left[\sup_Q \mathbb{E}_Q [\ell(y, h(X))] - \gamma (\mathcal{W}_c(P, Q) - \varepsilon) \right] \right\}. \end{aligned} \quad (65)$$

The main advantage achieved by this reformulation is the substitution of μ' with the fixed meta-distribution μ inside the expectation operators. Therefore, the optimization no longer has to be carried out in the \mathcal{M}^2 space. For the sake of simplicity in the proof, assume supreme value in equation 65 is attainable. This assumption is not necessary, and can be relaxed by using a more detailed mathematical analysis which is replacing the optimal distribution Q^* with a Cauchy series of distributions and proceed with similar arguments. However, we have decided to avoid this scenario in order to simplify the proof. In this regard, let us define:

$$Q^*(P, \gamma; \varepsilon) \triangleq \arg \max_Q \mathbb{E}_Q [\ell(y, h(X))] - \gamma (\mathcal{W}_c(P, Q) - \varepsilon), \quad \forall P \in \text{supp}(\mu). \quad (66)$$

Then, the following relation holds:

$$\begin{aligned} \sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(\mathbf{X}))]] &= \inf_{\gamma \geq 0} \{\mathbb{E}_\mu [\mathbb{E}_{Q^*(P, \gamma)} [\ell(y, h(\mathbf{X}))]] - \gamma (\mathcal{W}_c(P, Q^*(P, \gamma)) - \varepsilon)\} \\ &= \inf_{\gamma \geq 0} \{\mathbb{E}_{P \sim \mu} [\mathbb{E}_{Q^*(P, \gamma)} [\ell(y, h(\mathbf{X}))]] - \gamma \mathbb{E}_{P \sim \mu} [\mathcal{W}_c(P, Q^*(P, \gamma)) - \varepsilon]\}. \end{aligned} \quad (67)$$

which, can be simply rewritten as:

$$\begin{aligned} \sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(\mathbf{X}))]] &= \inf_{\gamma \geq 0} \mathbb{E}_\mu [\mathbb{E}_{Q^*(P, \gamma)} [\ell(y, h(\mathbf{X}))]] \\ \text{subject to } \mathbb{E}_{P \sim \mu} [\mathcal{W}_c(P, Q^*(P, \gamma))] &\leq \varepsilon. \end{aligned} \quad (68)$$

Using a similar argument as before, let us assume the $\inf_{\gamma \geq 0}$ in equation 68 is also attainable and denote the optimal value by $\gamma^* = \gamma^*(\mu, \varepsilon)$. Once again, this assumption is not necessary and can be relaxed at the expense of introducing more mathematical details and making the proof less readable. In this regard, we have:

$$\sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(\mathbf{X}))]] = \mathbb{E}_\mu [\mathbb{E}_{Q^*(P, \gamma^*)} [\ell(y, h(\mathbf{X}))]], \quad (69)$$

where it has been already guaranteed that the optimal parameter $\gamma^* \geq 0$ and optimal distribution $Q^*(P, \gamma^*)$, the following constraint holds:

$$\mathbb{E}_{P \sim \mu} [\mathcal{W}_c(P, Q^*(P, \gamma^*))] \leq \varepsilon. \quad (70)$$

For any $P \in \text{supp}(\mu) \subseteq \mathcal{M}$, let us define the following *optimal robustness radius* function

$$\rho^*(P; \varepsilon, \mu) \triangleq \mathcal{W}_c(P, Q^*(P, \gamma^*)). \quad (71)$$

Therefore, the original MDRO objective in the statement of the theorem can be readily upper-bounded using the following distributionally robust formulation:

$$\begin{aligned} \sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(\mathbf{X}))]] &= \mathbb{E}_{P \sim \mu} [\mathbb{E}_{Q^*(P, \gamma^*)} [\ell(y, h(\mathbf{X}))]] \\ &\leq \mathbb{E}_{P \sim \mu} \left[\sup_{Q \in \mathcal{B}_{\rho^*(P)}^{\text{wass}}(P)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] \right]. \end{aligned} \quad (72)$$

Using the upper-bound in equation 72 and the inequality condition on optimal Wasserstein radius functions $\rho^*(P)$ described in equation 70, we can proceed to the empirical stage of the proof. At this stage, the true expectation operators should be replaced by their empirical counterparts which are based on i.i.d. realizations of meta-distribution μ , i.e., unknown distributions P_1, \dots, P_K and their known yet private empirical realizations, i.e., \hat{P}_i for $i \in [K]$.

For $P \sim \mu$, let us define the following new and real-valued random variables $\psi(P)$ and $\zeta(P)$ as follows:

$$\begin{aligned} \psi(P) &\triangleq \sup_{Q \in \mathcal{B}_{\rho^*(P)}^{\text{wass}}(P)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))], \\ \zeta(P) &\triangleq \rho^*(P; \varepsilon, \mu). \end{aligned} \quad (73)$$

It should be noted that $\psi(P)$ is readily known to be (almost surely) bounded by 1, since $\ell(\cdot)$ is assumed to be 1-bounded. Additionally, the boundedness for $\zeta(P)$ directly results from the assumption that c is a bounded transportation cost.

Lemma E.2. *There exists $R < +\infty$ such that We have $\rho^*(P; \mu, \varepsilon) \stackrel{a.s.}{<} R$ for $P \sim \mu$.*

Proof. The proof is straightforward and directly results from the definition of Wasserstein distance:

$$\begin{aligned} \zeta(P) &\triangleq \rho^*(P; \varepsilon, \mu) = \mathcal{W}_c(P, Q^*(P, \gamma^*)) \\ &= \inf_{\nu \in \mathcal{C}(P, Q^*)} \mathbb{E}_\nu [c(\mathbf{Z}, \mathbf{Z}')] \\ &\stackrel{a.s.}{\leq} \sup_{\mathbf{Z}, \mathbf{Z}' \in \mathcal{Z}} c(\mathbf{Z}, \mathbf{Z}') < +\infty, \end{aligned} \quad (74)$$

which concludes the proof. \square

Using a similar series of arguments to the ones explained in Lemmas D.1 and D.2 (proof of Theorem 5.2), together with the fact that $\psi(P_k)$ s are all bounded by 1, one can directly apply the McDiarmid's inequality and show that the following bound holds with probability at least $1 - \frac{\delta}{K+2}$, for any $\delta > 0$:

$$\mathbb{E}_{P \sim \mu} \left[\sup_{Q \in \mathcal{B}_{\rho^*(P)}^{\text{wass}}(P)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] \right] \leq \frac{1}{K} \sum_{k \in [K]} \sup_{Q \in \mathcal{B}_{\rho^*(P_k)}^{\text{wass}}(P_k)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] + \sqrt{\frac{\log(\frac{K+2}{\delta})}{2K}}. \quad (75)$$

On the other hand, by using the boundedness property for $\zeta(P)$ proved in Lemma E.2 and applying McDiarmid's inequality once again, the following bound holds with probability $1 - \frac{\delta}{K+2}$ (for any $\delta > 0$) for the empirical mean of $\zeta(P)$ over true sample distributions P_1, \dots, P_k :

$$\frac{1}{K} \sum_{k \in [K]} \zeta(P_k) \leq \mathbb{E}_{P \sim \mu} [\zeta(P)] + c_1 \sqrt{\frac{\log(\frac{K+2}{\delta})}{K}} \leq \varepsilon + c_1 \sqrt{\frac{\log(\frac{K+2}{\delta})}{K}}, \quad (76)$$

where c_1 is a known universal constant depending only on the bound on transportation cost c . Here, the last inequality is a direct consequence of the property shown in equation 70.

Let $\mathcal{S} \subset \mathbb{R}_{\geq 0}^K$ be defined as the following subset:

$$\mathcal{S} \triangleq \left\{ (\zeta_1, \dots, \zeta_K) \in \mathbb{R}^K \mid \zeta_k \geq 0, \forall k \in [K], \frac{1}{K} \sum_{k \in [K]} \zeta_k \leq \varepsilon + c_1 \sqrt{\frac{\log(\frac{K+2}{\delta})}{K}} \right\}. \quad (77)$$

So far, we have shown that

$$\mathbb{P} \left(\{\zeta(P_k)\}_{k \in [K]} \in \mathcal{S} \right) \geq 1 - \frac{\delta}{K+2}. \quad (78)$$

In a similar procedure to the one used in the proof of Theorem 5.2, union bound ensures that the bound in equation 75 and the mathematical statement of $\{\zeta(P_k)\}_{k \in [K]} \in \mathcal{S}$ simultaneously hold with probability at least $1 - \frac{2\delta}{K+2}$. Then the following chain of bounds also hold with the same probability w.r.t. drawing of P_1, \dots, P_K from μ :

$$\begin{aligned} \sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(\mathbf{X}))]] &\leq \mathbb{E}_{P \sim \mu} \left[\sup_{Q \in \mathcal{B}_{\rho^*(P)}^{\text{wass}}(P)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] \right] \\ &\leq \frac{1}{K} \sum_{k \in [K]} \sup_{Q \in \mathcal{B}_{\rho^*(P_k)}^{\text{wass}}(P_k)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] + \sqrt{\frac{\log(\frac{K+2}{\delta})}{2K}} \\ &\leq \sup_{\underline{\rho} \in \mathcal{S}} \frac{1}{K} \sum_{k \in [K]} \sup_{Q \in \mathcal{B}_{\rho_k}^{\text{wass}}(P_k)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] + \sqrt{\frac{\log(\frac{K+2}{\delta})}{2K}} \\ &= \sup_{\underline{\rho} \in \mathcal{S}} \frac{1}{K} \sum_{k \in [K]} QV_k(h, \rho_k) + \sqrt{\frac{\log(\frac{K+2}{\delta})}{2K}}. \end{aligned} \quad (79)$$

For reasons that become clear in the final stages of the proof, we need to replace the set \mathcal{S} with a new one denoted by \mathcal{S}' which should be defined as:

$$\mathcal{S}' \triangleq \left\{ (\zeta_1, \dots, \zeta_K) \in \mathbb{R}^K \mid \zeta_k \geq \frac{\varepsilon}{K}, \forall k \in [K], \frac{1}{K} \sum_{k \in [K]} \zeta_k \leq \varepsilon \left(1 + \frac{1}{K} \right) + c_1 \sqrt{\frac{\log(\frac{K+2}{\delta})}{2K}} \right\}.$$

Evidently, replacing \mathcal{S}' with \mathcal{S} in the maximization step of equation 79, i.e., $\sup_{\underline{\rho} \in \mathcal{S}'}$, gives an upper bound for the original formulation of $\sup_{\underline{\rho} \in \mathcal{S}}$, since each member of \mathcal{S}' can be formed by taking a member from \mathcal{S} and add all radius values by a constant ε/K . Obviously, this procedure only makes the adversarial loss value larger and hence all the bounds still apply.

Part II: So far, we have managed to (partially) prove the proposed bound in the statement of the theorem in scenarios where $\min_k n_k \rightarrow \infty$ and thus we have $\widehat{P}_k \stackrel{a.s.}{\equiv} P_k$ for all $k \in [K]$. At this stage of the proof we focus on replacing

$$\text{QV}_k(h, \rho) = \sup_{Q \in \mathcal{B}_{\rho_k}^{\text{wass}}(P_k)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))],$$

for any $k \in [K]$ and arbitrary $\rho \geq 0$, with its empirical version $\widehat{\text{QV}}_k(h, \rho)$. Let us reiterate that we do not have any knowledge regarding P_k , and only client k has access to its empirical version \widehat{P}_k which is based on n_k i.i.d. samples. Therefore, $\widehat{\text{QV}}_k(h, \rho)$ is computable via the querying policy described in Section 3, while the true query value $\text{QV}(h, \rho)$ is always unknown.

To this aim, similar to (Sinha et al., 2018) first let us define the following $\phi_\gamma(\mathbf{Z})$ function for $\gamma \geq 0$ and $\mathbf{Z} \in \mathcal{Z}$:

$$\phi_\gamma(\mathbf{Z}) \triangleq \sup_{\mathbf{Z}' \in \mathcal{Z}} \ell(\mathbf{Z}') - \gamma c(\mathbf{Z}', \mathbf{Z}), \quad (80)$$

where $c(\cdot, \cdot)$ is the original transportation cost and $\ell(\mathbf{Z})$ abbreviates $\ell(y, h(\mathbf{X}))$ where we have omitted h for simplicity in notation. First, it can readily verified that if ℓ is bounded between 0 and 1, so does ϕ_γ for any $\gamma \geq 0$. Second, note that from Lemma E.1 we have the following duality formulation for QV_k and $\widehat{\text{QV}}_k$ for any $\rho_k \geq 0$:

$$\begin{aligned} \text{QV}_k(h, \rho_k) &\triangleq \sup_{Q \in \mathcal{B}_{\rho_k}^{\text{wass}}(P_k)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] = \inf_{\gamma \geq 0} \{\gamma \rho_k + \mathbb{E}_{P_k} [\phi_\gamma(\mathbf{Z})]\}, \\ \widehat{\text{QV}}_k(h, \rho_k) &\triangleq \sup_{Q \in \mathcal{B}_{\rho_k}^{\text{wass}}(\widehat{P}_k)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] = \inf_{\gamma \geq 0} \left\{ \gamma \rho_k + \frac{1}{n_k} \sum_{i \in [n_k]} \phi_\gamma(\mathbf{Z}_i^{(k)}) \right\}. \end{aligned} \quad (81)$$

In the following, first we show that $\phi_\gamma(\mathbf{Z})$, for any $\mathbf{Z} \in \mathcal{Z}$ is a Lipschitz function with respect to γ where the Lipschitz constant only depends on the way the transportation cost c is bounded, i.e., the inherent boundedness of c or the compactness of \mathcal{Z} . Additionally, we show that the optimal $\gamma \geq 0$ in both minimization problems on the right-hand sides of equation 81 is bounded by a known constant. The latter result is deduced from the fact that all robustness radii ρ_k , $k \in [K]$ in the statement of theorem has a known margin from zero. Finally, we show the above-mentioned properties can guarantee that $\mathbb{E}_{\widehat{P}_k} [\phi_\gamma(\mathbf{Z})]$ uniformly converges to its true expected value $\mathbb{E}_{P_k} [\phi_\gamma(\mathbf{Z})]$ for all relevant value of $\gamma \geq 0$. Hence, the empirical and statistical query values are always within a controlled and asymptotically small deviation from each other regardless of the robustness radius value ρ_k .

In this regard, the following lemma shows that $\phi_\gamma(\mathbf{Z})$ for any $\mathbf{Z} \in \mathcal{Z}$ is a Lipschitz function with respect to $\gamma \geq 0$:

Lemma E.3. *There exists a constant $R \geq 0$ which only depends on transportation cost c such that function $\phi_\gamma(\mathbf{Z})$ is R -Lipschitz with respect to $\gamma \geq 0$, for all $\mathbf{Z} \in \mathcal{Z}$.*

Proof. For any two distinct values $\gamma, \gamma' \geq 0$, let \mathbf{Z}_γ^* and $\mathbf{Z}_{\gamma'}^*$ denote the optimal values for which the sup in equation 80 is attained. Similar to several previous arguments, attainability of the sup in this case is not necessary again, and thus this assumption is made for the sake of simplifying the proof.

Then, for any $\mathbf{Z} \in \mathcal{Z}$ we have:

$$\begin{aligned} \phi_\gamma(\mathbf{Z}) &= \sup_{\mathbf{Z}' \in \mathcal{Z}} \ell(\mathbf{Z}') - \gamma c(\mathbf{Z}', \mathbf{Z}) \geq \ell(\mathbf{Z}_{\gamma'}^*) - \gamma c(\mathbf{Z}_{\gamma'}^*, \mathbf{Z}), \\ \phi_{\gamma'}(\mathbf{Z}) &= \ell(\mathbf{Z}_{\gamma'}^*) - \gamma' c(\mathbf{Z}_{\gamma'}^*, \mathbf{Z}), \end{aligned}$$

which directly gives us the following bound:

$$\phi_\gamma(\mathbf{Z}) - \phi_{\gamma'}(\mathbf{Z}) \geq -(\gamma - \gamma') c(\mathbf{Z}_{\gamma'}^*, \mathbf{Z}). \quad (82)$$

Through a set of similar arguments and replacing γ and γ' , the following complementary bound can be achieved as well:

$$\phi_\gamma(\mathbf{Z}) - \phi_{\gamma'}(\mathbf{Z}) \leq -(\gamma - \gamma') c(\mathbf{Z}_\gamma^*, \mathbf{Z}). \quad (83)$$

Therefore, the following inequality can be established according to the boundedness of c (or alternatively, compactness of \mathcal{Z}):

$$\begin{aligned} |\phi_\gamma(\mathbf{Z}) - \phi_{\gamma'}(\mathbf{Z})| &\leq |\gamma - \gamma'| \max_{r \in \{\gamma, \gamma'\}} \{c(\mathbf{Z}_r^*, \mathbf{Z})\} \\ &\leq |\gamma - \gamma'| \sup_{\mathbf{Z}' \in \mathcal{Z}} c(\mathbf{Z}', \mathbf{Z}) \\ &\leq R |\gamma - \gamma'|, \end{aligned} \quad (84)$$

which proves the Lipschitz-ness of ϕ_γ with respect to γ . \square

The following lemma shows that optimal values of γ in the right-hand side minimization of equation 81 (or the infimum-achieving sequence in case the infimum is not attainable) is bounded by a known constant:

Lemma E.4. *In both minimization problems on the right-hand side of equation 81, the optimal γ value denoted by γ^* (if attained), or the tail of its sequence in case the inf is not attainable, satisfies $0 \leq \gamma^* \leq \frac{1}{\rho_k}$.*

Proof. Proof directly results from the fact that $\ell(\cdot)$ is bounded between 0 and 1. Therefore, looking at the dual optimization problem in equation 81, increasing γ beyond $1/\rho_k$ results in $\gamma\rho_k > 1$ while the second term (i.e., the adversarial loss) is always lower-bounded by zero which makes the whole objective to become larger than 1. On the other hand, setting $\gamma = 0$ would (at worst) result in the objective to be 1. Therefore, the optimizer $\inf_{\gamma \geq 0}$ should not choose a γ value that is larger than $1/\rho_k$. \square

At this point, we can state the main lemma in the second part of the proof, which theoretically shows that empirical query values, i.e., $\widehat{\text{QV}}_k(h, \rho)$ for any fixed $h \in \mathcal{H}$ and uniformly all $\rho \geq 0$ converge to their true statistical expected values with a high probability.

Lemma E.5 (Uniform Convergence of Empirical Queries). *For $k \in [K]$, assume the unknown sample distribution P_k and let $\{\mathbf{Z}_i^{(k)} = (\mathbf{X}_i^{(k)}, y_i^{(k)})\}$ for $i \in [n_k]$ denote $n_k \in \mathbb{N}$ i.i.d. feature-label pairs drawn from P_k . The k th dataset is only known to client k . Then, for any fixed classifier h and any $\delta > 0$, the following bound holds with probability at least $1 - \delta/(K+2)$:*

$$\sup_{\rho \geq \varepsilon/K} \left| \widehat{\text{QV}}_k(h, \rho) - \text{QV}_k(h, \rho) \right| \leq \mathcal{O} \left(\sqrt{n_k^{-1} \log \left(\frac{(K+2)n_k}{\varepsilon\delta} \right)} \right). \quad (85)$$

Proof. Using the dual formulation of Lemma E.1, we can rewrite the main objective of the theorem as follows:

$$\begin{aligned} &\sup_{\rho \geq \varepsilon/K} \left| \widehat{\text{QV}}_k(h, \rho) - \text{QV}_k(h, \rho) \right| \\ &= \sup_{\rho_k \geq \varepsilon/K} \left\{ \inf_{\hat{\gamma} \geq 0} [\hat{\gamma}\rho_k + \mathbb{E}_{\hat{P}_k} [\phi_{\hat{\gamma}}(\mathbf{Z})]] - \inf_{\gamma \geq 0} [\gamma\rho_k + \mathbb{E}_{P_k} [\phi_\gamma(\mathbf{Z})]] \right\}. \end{aligned} \quad (86)$$

Again, for the sake of simplicity in the proof let us assume both optimal values γ^* and $\hat{\gamma}^*$ in the minimization problems on the right-hand side of equation 86 are attainable. It should be noted that this assumption is not necessary and can be relaxed by adding more mathematical work. Then, from Lemma E.4 we already know

$$0 \leq \gamma^*, \hat{\gamma}^* \leq \frac{1}{\rho_k} \leq \frac{K}{\varepsilon}.$$

Let us partition the feasible search set of $\gamma, \hat{\gamma} \geq 0$, i.e., $[0, K/\varepsilon]$ into $L \triangleq \lceil \frac{K^2 R}{\varepsilon\Delta} \rceil$ equal intervals, where $\Delta > 0$ is a small constant which should to be determined later in the proof. For each interval, let us choose a representative (for example, the value at the beginning of the interval) denoted by γ_i with $i = 1, \dots, L$. Then, based on Lemma E.3, for any $\rho_k \in [\varepsilon/K, 2K\varepsilon]$, any corresponding $\gamma \in [0, 1/\rho_k]$ and all $\mathbf{Z} \in \mathcal{Z}$ we have

$$|\phi_\gamma(\mathbf{Z}) - \phi_{\gamma_{i^*}}(\mathbf{Z})| \leq R |\gamma - \gamma_{i^*}| \leq R \cdot \frac{K}{\varepsilon} \cdot \frac{\varepsilon\Delta}{K^2 R} = \frac{\Delta}{K}, \quad (87)$$

where $i^* = \arg \min_{i \in [L]} |\gamma - \gamma_i|$. On the other hand, from the maximization problem that defines $\widehat{U}^*(\varepsilon)$ in equation 15, we have that

$$\rho_k \leq \mathcal{O} \left(K\varepsilon + \sqrt{K \log \left(\frac{K+2}{\delta} \right)} \right), \quad \forall k \in [K], \quad (88)$$

where we have omitted constants for the sake of readability. Therefore, the above discussions directly lead to the following bound in the statistical sense (i.e., with respect to P_k):

$$\begin{aligned} & \left| \inf_{\gamma \geq 0} [\gamma \rho_k + \mathbb{E}_{P_k} [\phi_\gamma(\mathbf{Z})]] - \min_{i \in [L]} [\gamma_i \rho_k + \mathbb{E}_{P_k} [\phi_{\gamma_i}(\mathbf{Z})]] \right| \\ & \leq \min_{i \in [L]} |\gamma^* - \gamma_i| (R + \sup \rho_k) \\ & \leq \frac{K}{\varepsilon} \cdot \frac{\varepsilon \Delta}{K^2 R} \cdot \mathcal{O} \left(R + K\varepsilon + \sqrt{K \log \left(\frac{(K+2)}{\delta} \right)} \right) \\ & \leq \Delta \cdot \mathcal{O} \left(\frac{1}{K} + \frac{\varepsilon}{R} + \frac{1}{R} \sqrt{\frac{\log \left(\frac{K+2}{\delta} \right)}{K}} \right). \end{aligned} \quad (89)$$

Through a similar procedure, the following bound also holds for the empirical case (i.e., \widehat{P}_k), but this time *almost surely*:

$$\begin{aligned} & \left| \inf_{\widehat{\gamma} \geq 0} [\widehat{\gamma} \rho_k + \mathbb{E}_{\widehat{P}_k} [\phi_{\widehat{\gamma}}(\mathbf{Z})]] - \min_{i \in [L]} [\gamma_i \rho_k + \mathbb{E}_{\widehat{P}_k} [\phi_{\gamma_i}(\mathbf{Z})]] \right| \\ & \stackrel{a.s.}{\leq} \Delta \cdot \mathcal{O} \left(\frac{1}{K} + \frac{\varepsilon}{R} + \frac{1}{R} \sqrt{\frac{\log \left(\frac{K+2}{\delta} \right)}{K}} \right). \end{aligned} \quad (90)$$

Therefore, according to the fact that \widehat{P}_k is an empirical estimate of P_k based on n_k i.i.d. samples, we have:

$$\begin{aligned} \sup_{\rho \geq \varepsilon/K} \left| \widehat{\text{QV}}_k(h, \rho) - \text{QV}_k(h, \rho) \right| & \stackrel{a.s.}{\leq} \Delta \cdot \mathcal{O} \left(\frac{1}{K} + \frac{\varepsilon}{R} + \frac{1}{R} \sqrt{\frac{\log \left(\frac{K+2}{\delta} \right)}{K}} \right) + \\ & \max_{i \in [L]} \left| \mathbb{E}_{\widehat{P}_k} [\phi_{\gamma_i}(\mathbf{Z})] - \mathbb{E}_{P_k} [\phi_{\gamma_i}(\mathbf{Z})] \right|. \end{aligned} \quad (91)$$

Since ϕ_{γ_i} for each $i \in [L]$ is a non-negative and 1-bounded (adversarial) loss function, simply applying McDiarmid's inequality and the union bound over all L values of γ_i s would give us the following bound which holds with probability at least $1 - \delta/(K+2)$ for any $\delta > 0$:

$$\max_{i \in [L]} \left| \mathbb{E}_{\widehat{P}_k} [\phi_{\gamma_i}(\mathbf{Z})] - \mathbb{E}_{P_k} [\phi_{\gamma_i}(\mathbf{Z})] \right| \leq \sqrt{\frac{\log \left[\frac{L(K+2)}{\delta} \right]}{2n_k}}, \quad (92)$$

which gives us the following bound for $\sup_{\rho \geq \varepsilon/K} \left| \widehat{\text{QV}}_k(h, \rho) - \text{QV}_k(h, \rho) \right|$ with the same high probability:

$$\sup_{\rho \geq \varepsilon/K} \left| \widehat{\text{QV}}_k(h, \rho) - \text{QV}_k(h, \rho) \right| \leq \mathcal{O} \left(\inf_{\Delta > 0} \left\{ \Delta \zeta(K, \varepsilon, \delta) + \sqrt{\frac{\log \left[\frac{RK^2(K+2)}{\varepsilon \delta \Delta} \right]}{n_k}} \right\} \right), \quad (93)$$

where function $\zeta(\cdot)$ is defined as

$$\zeta(K, \varepsilon, \delta) \triangleq \frac{1}{K} + \frac{\varepsilon}{R} + \frac{1}{R} \sqrt{\frac{\log \left(\frac{K+2}{\delta} \right)}{K}}.$$

It should be noted that R is a constant that does not depend on other parameters. Also, we have minimized over $\Delta > 0$ since the bound holds irrespective of Δ . Exact solution of the minimization problem in equation 93 is not needed, since choosing $\Delta = \mathcal{O}\left(K^{-1}n_k^{-1/2}\right)$ gives us the following bound:

$$\sup_{\rho \geq \varepsilon/K} \left| \widehat{\text{QV}}_k(h, \rho) - \text{QV}_k(h, \rho) \right| \leq \mathcal{O}\left(\sqrt{\frac{\log\left(\frac{(K+2)n_k}{\varepsilon\delta}\right)}{n_k}}\right), \quad (94)$$

and completes the proof. \square

By using the uniform convergence result from Lemma E.5 and applying it to all K clients simultaneously, we see that (via a union bound argument) with probability at least $1 - K\delta/(K+2)$ the empirical and statistical queries $\widehat{\text{QV}}_k(h, \rho_k)$ and $\text{QV}_k(h, \rho_k)$ are asymptotically close for all $k \in [K]$, and uniformly for all robustness radii

$$\rho_k \in \left[\frac{\varepsilon}{K}, \mathcal{O}\left(K\varepsilon + \sqrt{\frac{\log((K+2)/\delta)}{K}}\right) \right]$$

which are considered for the maximization problem of equation 15. Finally, using another union bound argument to incorporate the bounds from equation 75 and equation 78 in addition to the previous K events, we can say that with probability at least $1 - \delta$ the following bound holds:

$$\begin{aligned} & \sup_{\mu' \in \mathcal{G}_\varepsilon(\mu)} \mathbb{E}_{P \sim \mu'} [\mathbb{E}_P [\ell(y, h(\mathbf{X}))]] \\ & \leq \sup_{\rho \in \mathcal{S}'} \frac{1}{K} \sum_{k \in [K]} \sup_{Q \in \mathcal{B}_{\rho_k}^{\text{Wass}}(P_k)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] + c_2 \sqrt{\frac{\log\left(\frac{(K+2)n_k}{\varepsilon\delta}\right)}{n_k}} + \sqrt{\frac{\log\left(\frac{K+2}{\delta}\right)}{2K}} \\ & = \sup_{\rho \in \mathcal{S}'} \frac{1}{K} \sum_{k \in [K]} \widehat{\text{QV}}_k(h, \rho_k) + c_2 \sqrt{\frac{\log\left(\frac{(K+2)n_k}{\varepsilon\delta}\right)}{n_k}} + \sqrt{\frac{\log\left(\frac{K+2}{\delta}\right)}{2K}}, \end{aligned} \quad (95)$$

where c_2 is a constant that only depends on either the transportation cost c or the compactness of sample space \mathcal{Z} . This completes the proof. \square

F. Proofs of Remarks in Section 7

Proof of Remark 7.1. We begin by discussing the server-side programs in Section 5, which address f -divergence meta-robustness, followed by the analysis in Section 6, which focuses on Wasserstein meta-robustness.

Server-side Programs in Theorems 5.2 and 5.3 : The objective function in both programs is a linear function of the optimization variables $\alpha_1, \dots, \alpha_K$. Furthermore, the constraints define a convex feasible set:

- The constraint

$$\left| \frac{1}{K} \sum_{k=1}^K \alpha_k - 1 \right| \leq C,$$

for any positive constant C , is equivalent to the intersection of the two constraints: $\frac{1}{K} \sum_{k=1}^K \alpha_k \leq 1 + C$ and $\frac{1}{K} \sum_{k=1}^K \alpha_k \geq 1 - C$. These correspond to a band region bounded by two hyperplanes in \mathbb{R}^K , forming a convex set.

- The other constraint,

$$\frac{1}{K} \sum_{k=1}^K f(\alpha_k) \leq \varepsilon + C',$$

Algorithm 1 Server-side Bisection Algorithm

Require: K, ε, δ
input $h, \Delta > 0$, and $\text{poly}(K, \log \frac{1}{\Delta})$ query budget for $\widehat{\text{QV}}_k$, for all $k \in [K]$

 1: Initialize $a \leftarrow \min \ell(\cdot)$ (or 0), $b \leftarrow \max \ell(\cdot)$ (or 1)

 2: **while** $b - a > \Delta$ **do**

 3: $t \leftarrow (a + b)/2$

4: Solve convex feasibility problem:

 5: Find $\rho_1, \dots, \rho_K \geq \varepsilon/K$ such that

$$6: \quad \frac{1}{K} \sum_{k \in [K]} \rho_k \leq \varepsilon \left(1 + \frac{1}{K}\right) + c_1 \sqrt{\frac{\log(\frac{K+2}{\delta})}{K}}$$

$$7: \quad \frac{1}{K} \sum_{k \in [K]} \widehat{\text{QV}}_k(h, \rho_k) \geq t$$

 8: **if** problem is feasible **then**

 9: $a \leftarrow t$

 10: **else**

 11: $b \leftarrow t$

 12: **end if**

 13: **end while**
output upper-bound b

for a constant C' , represents the $(\varepsilon + C')$ -sublevel set of the convex function $\frac{1}{K} \sum_{k=1}^K f(\alpha_k)$. The convexity of this function follows directly from the convexity of $f(\cdot)$, as stated in Definition 5.1. Consequently, this constraint also defines a convex set.

Therefore, both programs feature linear objectives and feasible sets that are intersections of convex sets in \mathbb{R}^K , implying that both programs are convex and can be solved efficiently (Boyd, 2004; Ghadimi and Lan, 2013). This completes the proof.

Server-side Program in Theorem 6.3 : The query functions $\widehat{\text{QV}}_k(h, \rho)$, for any fixed $h \in \mathcal{H}$, are non-decreasing with respect to $\rho \geq 0$. Therefore, the following function:

$$\zeta(\rho_1, \dots, \rho_K) \triangleq \frac{1}{K} \sum_{k \in [K]} \widehat{\text{QV}}_k(h, \rho_k) \quad (96)$$

is a summation of K non-decreasing functions, where each function depends only on one of the ρ_k values. This function, $\zeta(\rho_1, \dots, \rho_K)$, is a *quasi-concave* function. Although quasi-concave functions are not generally concave, they do possess concave superlevel sets. Specifically, for each $t \in \mathbb{R}$, the following sets are convex in \mathbb{R}^K :

$$\mathcal{S}_t \triangleq \left\{ (\rho_1, \dots, \rho_K) \in \mathbb{R}^K \mid \frac{1}{K} \sum_{k \in [K]} \widehat{\text{QV}}_k(h, \rho_k) \geq t \right\}. \quad (97)$$

As a result, the original optimization problem in equation 15 can be decomposed into a sequence of convex optimization problems. Each sub-problem is essentially a feasibility problem that checks whether the set \mathcal{S}_t is feasible for a given $t \in \mathbb{R}$. Given that the original objective function is bounded between 0 and 1, a binary search can be employed to iteratively approximate the maximum attainable value of the objective within any desired error margin $\Delta > 0$. This process is detailed in Algorithm 1, which implements a *bisection* algorithm.

It is important to note that each feasibility check sub-problem in Algorithm 1 is a convex problem, and therefore can be solved in polynomial time with polynomial evaluations of the constraint functions, i.e., the $\widehat{\text{QV}}_k$ functions. Consequently, both the computational complexity and the required query budget remain polynomial. In particular, for a given $\Delta > 0$, the maximum of the objective in equation 96 can be approximated with an error of at most Δ , requiring at most $\log \frac{1}{\Delta}$ iterations (feasibility checks) in Algorithm 1 (Boyd, 2004). \square

Proof of Remark 7.2. We turn to client-side optimizations for this theorem, which take the form:

$$\sup_{Q \in \mathcal{B}_\rho^{\text{wass}}(P)} \mathbb{E}_Q [\ell(\mathbf{Z})],$$

where the notation $\ell(y, h(\mathbf{X}))$ has been simplified to $\ell(\mathbf{Z})$ (for $\mathbf{Z} = (\mathbf{X}, y)$) for clarity and ease of notation. For a client $k \in [K]$, the local distribution P corresponds to \widehat{P}_k , which is derived from n_k samples drawn from an arbitrary distribution P_k . This optimization program is initially defined within a restricted distributional space, $\mathcal{B}_\rho^{\text{wass}}$, and may appear intractable at first glance. However, by leveraging Lemma E.1 (see Appendix E), which relies on a core mathematical tool from the seminal works of (Sinha et al., 2018) (originally derived in (Blanchet and Murthy, 2019)), this formulation can be rewritten in its dual form. This dual reformulation is subsequently shown to be implementable in polynomial time (Sinha et al., 2018). The rest of the proof of Remark 7.2 is the summary of the method proposed and subsequently used in the above-mentioned reference (**not our work**).

Let P be a probability measure defined over a measurable space Ω , $\ell(\cdot) : \Omega \rightarrow \mathbb{R}$ be any loss function, c denote a proper and lower semi-continuous transportation cost on $\Omega \times \Omega$, and assume $\varepsilon \geq 0$. Then, the following equality holds for the Wasserstein-constrained DRO centered around P :

$$\sup_{Q \in \mathcal{B}_\rho^{\text{wass}}(P)} \mathbb{E}_Q [\ell(\mathbf{Z})] = \inf_{\gamma \geq 0} \left\{ \gamma\rho + \mathbb{E}_P \left[\sup_{\mathbf{Z}' \in \Omega} \ell(\mathbf{Z}') - \gamma c(\mathbf{Z}', \mathbf{Z}) \right] \right\}. \quad (98)$$

The optimization over γ is one-dimensional. On the other hand, the term $\mathbb{E}_P [\sup_{\mathbf{Z}' \in \Omega} \ell(\mathbf{Z}') - \gamma c(\mathbf{Z}', \mathbf{Z})]$ is non-increasing for $\gamma \geq 0$, while the other term, $\gamma\rho$, is affine. In fact, γ and ρ are *dual* counterparts, where increasing ρ decreases the optimal γ value and vice versa. Therefore, the outer minimization $\inf_{\gamma \geq 0}$ can be handled in polynomial time as long as the inner expected maximization can be numerically solved efficiently (in polynomial time).

For the inner maximization, we have

$$\mathbb{E}_{\widehat{P}_k} \left[\sup_{\mathbf{Z}' \in \Omega} \ell(\mathbf{Z}') - \gamma c(\mathbf{Z}', \mathbf{Z}) \right] = \frac{1}{n_k} \sum_{i \in [n_k]} \sup_{\mathbf{Z}' \in \Omega} \ell(\mathbf{Z}') - \gamma c(\mathbf{Z}', \mathbf{Z}_i),$$

for $\mathbf{Z}_i = (\mathbf{X}_i^{(k)}, y_i^{(k)})$, which follows from the fact that \widehat{P}_k is an atomic *empirical* measure. We follow the same procedure as in (Sinha et al., 2018):

- Assume ℓ is twice differentiable with a bounded-from-below (but potentially negative) curvature, i.e., the Hessian matrix satisfies $\nabla^2 \ell \succeq -\beta \mathbf{I}$ for some bounded $\beta \geq 0$.
- Assume c is 1-strongly convex in its first argument (same as Assumption A in Sinha et al. (2018)).

These assumptions are not overly restrictive, as most candidates for the loss function $\ell(\cdot)$, such as cross-entropy loss coupled with a deep neural network architecture for h , directly correspond to analytic and twice-differentiable functions with bounded curvature from below. Note that we **do not** assume convexity for ℓ , thus maintaining the generality of the problem. On the other hand, most choices for the transportation cost c are convex in their first argument (e.g., any valid norm).

Based on the above-mentioned two assumptions, it is straightforward to see that the term $\ell(\mathbf{Z}') - \gamma c(\mathbf{Z}', \mathbf{Z}_i)$ is *concave* as long as $\gamma \geq \beta$ (i.e., ε is not excessively large). According to standard convergence theorems in convex optimization theory (see, for example, Ghadimi and Lan (2013) or Boyd (2004)), any vanilla local first-order optimization algorithm with oracle access to gradients, such as stochastic gradient descent (SGD), can achieve an $\mathcal{O}(T^{-1/2})$ approximation of

$$\text{Obj}(\mathbf{Z}_1, \dots, \mathbf{Z}_{n_k}; \gamma) \triangleq \frac{1}{n_k} \sum_{i \in [n_k]} \sup_{\mathbf{Z}' \in \Omega} \ell(\mathbf{Z}') - \gamma c(\mathbf{Z}', \mathbf{Z}_i),$$

after $T \geq 2$ iterations. This completes the proof. \square

Proof of Remark 7.4. As $K, \min_k n_k \rightarrow \infty$, the bounds presented in Theorems 4.1, 5.2, 5.3, and 6.3 achieve minimax optimality, referred to as Asymptotic Minimax Optimality (AMO). Mathematically, this implies that the proposed empirical bounds on the right-hand sides become attainable by an adversary, rendering them unimprovable. This particular approach in demonstrating the tightness of such bounds is a standard practice in nearly all theoretical works within this field and related areas (see Sinha et al. (2018); Ashtiani et al. (2018); Saberi et al. (2023); etc.).

AMO for Theorem 4.1 : This theorem provides non-robust guarantees and thus there are no adversaries involved. Therefore, the proof for this part is straightforward since we only need to show the right-hand side of the inequalities equal the left-hand sides in the asymptotic regime. In this regard, we have

$$\begin{aligned} \lim_{K, n_{1:K} \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \widehat{\text{QV}}_k(h) &= \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \left(\lim_{n_k \rightarrow \infty} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(y_i^{(k)}, h(\mathbf{X}_i^{(k)})) \right) \\ &\stackrel{a.s.}{=} \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \mathbb{E}_{P_k} [\ell(y, h(\mathbf{X}))] \\ &\stackrel{a.s.}{=} \mathbb{E}_{P \sim \mu} (\mathbb{E}_{P_k} [\ell(y, h(\mathbf{X}))]). \end{aligned} \quad (99)$$

The first *almost sure* equality follows from the fact that that dataset $\left\{(\mathbf{X}_i^{(k)}, y_i^{(k)})\right\}_{i=1}^{n_k}$ consists of i.i.d. samples from P_k for each $k \in [K]$. The second one is a result of P_1, \dots, P_K being i.i.d. samples of μ . In both cases we have used the strong law of large numbers.

In a similar fashion, for the loss CDF and all subsequent values of $\lambda \in \mathbb{R}$, the following almost sure equalities hold:

$$\begin{aligned} \lim_{K, n_{1:K} \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \mathbb{1}\left(\widehat{\text{QV}}_k(h) \geq \lambda - \sqrt{\frac{\log \frac{(K+1)}{\delta}}{2n_k}}\right) &= \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \mathbb{1}\left(\lim_{n_k \rightarrow \infty} \widehat{\text{QV}}_k(h) + \sqrt{\frac{\log \frac{(K+1)}{\delta}}{2n_k}} \geq \lambda\right) \\ &= \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \mathbb{1}\left(\lim_{n_k \rightarrow \infty} \widehat{\text{QV}}_k(h) + \lim_{n_k \rightarrow \infty} \sqrt{\frac{\log \frac{(K+1)}{\delta}}{2n_k}} \geq \lambda\right) \\ &\stackrel{a.s.}{=} \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \mathbb{1}(\mathbb{E}_{P_k} [\ell(y, h(\mathbf{X}))] \geq \lambda) \\ &\stackrel{a.s.}{=} \mathbb{E}_{P \sim \mu} [\mathbb{1}(\mathbb{E}_P [\ell(y, h(\mathbf{X}))] \geq \lambda)] \\ &= \mu(\mathbb{E}_P [\ell(y, h(\mathbf{X}))] \geq \lambda). \end{aligned} \quad (100)$$

Again, the *almost sure* equalities are the result of applying law of large numbers in both distributional and meta-distributional levels. Therefore, the proof for this part is complete.

AMO for Theorems 5.2 and 5.3 : Assume two arbitrary meta-distributions μ' and μ and only assume $\mathcal{D}_f(\mu' \parallel \mu) \leq \varepsilon$ (note that for $\varepsilon < \infty$, this assumption automatically enforces relative continuity as well), and let P_1, P_2, \dots be an infinite i.i.d. sequence from μ . The only extra restriction over μ' is that it represents a *probability measure*. Combining all these natural restrictions we get

- Non-negativity of μ' , which means for any $\mathcal{S} \subseteq \mathcal{M}(\mathcal{Z})$ we have $\mu'(\mathcal{S}) \geq 0$.
- Normalization condition of probability measures which mathematically translates into the following equalities:

$$1 = \mu'(\mathcal{M}(\mathcal{Z})) = \int_{P \in \mathcal{M}} \mu'(dP) = \mathbb{E}_{P \sim \mu} \left(\frac{\mu'(dP)}{\mu(dP)} \right) \stackrel{a.s.}{=} \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \frac{d\mu'}{d\mu}(P_k).$$

- Boundedness of f -divergence, which can be written as follows:

$$\varepsilon \geq \mathcal{D}_f(\mu' \parallel \mu) = \mathbb{E}_{P \sim \mu} \left[f \left(\frac{d\mu'}{d\mu}(P) \right) \right] \stackrel{a.s.}{=} \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K f \left(\frac{d\mu'}{d\mu}(P_k) \right).$$

In all the above statements, the *almost sure* equality holds with respect to the randomness of drawing the i.i.d. sequence of distributions P_1, P_2, \dots from μ and utilizes the application of the law of large numbers. Based on the above discussions, any measure defined on a Borel σ -field over $\mathcal{M}(\mathcal{Z})$ that satisfies the above conditions is a potential candidate for the worst-case μ' in the statement of theorem.

On the other hand, the proposed bound in the right-hand side of the main result in Theorem 5.2 is computed via the following optimization problem:

$$\begin{aligned} \widehat{B}^*(\varepsilon) &\triangleq \sup_{0 \leq \alpha_1, \dots, \alpha_K \leq \Lambda} \frac{1}{K} \sum_{k=1}^K \alpha_k \widehat{\text{QV}}_k(h) \\ \text{subject to } & \left| \frac{1}{K} \sum_{k=1}^K \alpha_k - 1 \right| \leq c_1 \sqrt{\log(\delta^{-1})/K}, \\ & \frac{1}{K} \sum_{k=1}^K f(\alpha_k) \leq \varepsilon + c_2 \sqrt{\log(\delta^{-1})/K}. \end{aligned} \quad (101)$$

Since, for all $\delta > 0$ we have

$$\lim_{K \rightarrow \infty} c_1 \sqrt{\log(\delta^{-1})/K} = 0, \quad \text{and} \quad \lim_{K \rightarrow \infty} \varepsilon + c_2 \sqrt{\log(\delta^{-1})/K} = \varepsilon, \quad (102)$$

in the asymptotic regime of $K \rightarrow \infty$ the vector $(\alpha_1, \dots, \alpha_K)$ converges to the sequence $\{\alpha_i\}_{i \in \mathbb{N}}$, with the following properties:

$$\alpha_i \geq 0, \forall i \in \mathbb{N}, \quad \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{i=1}^K \alpha_i = 1 \quad \text{and} \quad \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{i=1}^K f(\alpha_i) \leq \varepsilon. \quad (103)$$

Therefore, it can be deduced that any $\{\alpha_i\}_{i \in \mathbb{N}}$ that satisfies such conditions corresponds to an achievable density ratio, i.e. $\alpha_k = d\mu'/d\mu(P_k)$, $k \in \mathbb{N}$, for some $\mu' \in \mathcal{M}^2(\mathcal{Z})$. In other words, the sup value is achievable. The same set of arguments hold for Theorem 5.3, and the proof for this part is complete.

AMO for Theorem 6.3 : The proof relies heavily on the theory developed in the proof of Theorem 6.3 (see Section E), and we adopt the same procedure and notations. Recall the bound proposed in Theorem 6.3: For any $\varepsilon, \delta > 0$, consider the following constrained optimization problem:

$$\begin{aligned} \widehat{U}^*(\varepsilon) &\triangleq \sup_{\rho_1, \dots, \rho_K \geq \varepsilon/K} \frac{1}{K} \sum_{k=1}^K \widehat{\text{QV}}_k(h, \rho_k) \\ \text{subject to } & \frac{1}{K} \sum_{k=1}^K \rho_k \leq \varepsilon \left(1 + \frac{1}{K} \right) + c_1 \sqrt{\frac{\log(\frac{K+2}{\delta})}{K}}. \end{aligned} \quad (104)$$

In the asymptotic regime (i.e., $K, \min_k n_k \rightarrow \infty$), we have

$$\lim_{n_k \rightarrow \infty} \widehat{\text{QV}}_k(h, \rho_k) \xrightarrow{a.s.} \sup_{Q \in \mathcal{B}_{\rho_k}(P_k)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))], \quad (105)$$

where P_1, P_2, \dots is an infinite (but countable) i.i.d. sequence from the meta-distribution μ . The almost sure convergence follows from the law of large numbers. Define $\rho : \mathcal{M}_{\mathcal{Z}} \rightarrow \mathbb{R}_+$. In the asymptotic regime ($K \rightarrow \infty$), the constraint becomes:

$$\begin{aligned} \mathbb{E}_{P \sim \mu} [\rho(P)] &\xrightarrow{a.s.} \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K \rho_k \\ &\leq \lim_{K \rightarrow \infty} \varepsilon \left(1 + \frac{1}{K} \right) + c_1 \sqrt{\frac{\log(\frac{K+2}{\delta})}{K}} \\ &= \varepsilon. \end{aligned} \quad (106)$$

Thus, the optimization problem defining \widehat{U}^* when $K, n_k \rightarrow \infty$ (almost surely) reduces to:

$$\begin{aligned} \widehat{U}^*(\varepsilon) &\xrightarrow{a.s.} \sup_{\rho: \mathcal{M}(\mathcal{Z}) \rightarrow \mathbb{R}_+} \mathbb{E}_{P \sim \mu} \left[\sup_{Q \in \mathcal{B}_{\rho_k}(P_k)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] \right] \\ \text{subject to } & \mathbb{E}_{P \sim \mu} [\rho(P)] \leq \varepsilon. \end{aligned} \quad (107)$$

Introducing a Lagrange multiplier $\gamma \geq 0$, and using strong duality, this problem becomes:

$$\begin{aligned} & \inf_{\gamma \geq 0} \left\{ \sup_{\rho: \mathcal{M}(\mathcal{Z}) \rightarrow \mathbb{R}_+} \mathbb{E}_{P \sim \mu} \left[\sup_{Q \in \mathcal{B}_{\rho}(P)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] \right] - \gamma (\mathbb{E}_{P \sim \mu} [\rho(P)] - \varepsilon) \right\} \\ &= \inf_{\gamma \geq 0} \left\{ \sup_{\rho: \mathcal{M}(\mathcal{Z}) \rightarrow \mathbb{R}_+} \mathbb{E}_{P \sim \mu} \left[\sup_{Q \in \mathcal{B}_{\rho}(P)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] - \gamma (\rho(P) - \varepsilon) \right] \right\} \\ &= \inf_{\gamma \geq 0} \left\{ \gamma \varepsilon + \sup_{\rho: \mathcal{M}(\mathcal{Z}) \rightarrow \mathbb{R}_+} \mathbb{E}_{P \sim \mu} \left[\sup_{Q \in \mathcal{B}_{\rho}(P)} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] - \gamma \rho(P) \right] \right\}. \end{aligned} \quad (108)$$

The combined effect of the two supremum operators ($\sup_{\rho: \mathcal{M}(\mathcal{Z}) \rightarrow \mathbb{R}_+}$ and $\sup_{Q \in \mathcal{B}_{\rho}(P)}$) yields the following min-max problem:

$$\inf_{\gamma \geq 0} \left\{ \gamma \varepsilon + \mathbb{E}_{P \sim \mu} \left[\sup_{Q \in \mathcal{M}_{\mathcal{Z}}} \mathbb{E}_Q [\ell(y, h(\mathbf{X}))] - \gamma \mathcal{W}_c(P, Q) \right] \right\}. \quad (109)$$

By Lemma E.1, this is equal to:

$$\sup_{\mu' \in \mathcal{G}_{\varepsilon}(\mu)} \mathbb{E}_{P \sim \mu'} (\mathbb{E}_P [\ell(y, h(\mathbf{X}))]). \quad (110)$$

Thus, the proposed bound $\widehat{\mathcal{U}}^*$ converges to the above expression as $K, \min_k n_k \rightarrow \infty$, which is exactly equal to the statistical quantity that it was supposed to upper-bound. This completes the proof. \square

G. Complementary Experimental Results

In this section, we present a complementary series of experiments on real-world datasets to demonstrate: (i) the validity and tightness of our bounds, and (ii) that our bounds are not only theoretically sound but also practical to compute. For clarity and readability, we repeat key explanations from Section 8 to maintain a cohesive narrative.

G.1. Client Generation and Risk CDF Certificates for Unseen Clients

In the first part of our experiments, we outline our client generation model and present a number of risk CDF guarantees. We simulated a federated learning scenario with $n = 1000$ nodes, where each node contains 1000 local samples. The experiments were conducted using four different datasets: CIFAR-10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), EMNIST (Cohen et al., 2017), and ImageNet (Russakovsky et al., 2015). To create each user’s data within the network, we applied three types of affine distribution shifts across users:

- **Feature Distribution Shift:** Each sample $\mathbf{X}_i^{(k)}$ in the dataset is manipulated via a transformation chosen randomly for each node. Specifically, each user is assigned a unique matrix $\Lambda^{(k)}$ and shift vector $\delta^{(k)}$, and the data is modified as follows:

$$\widetilde{\mathbf{X}}_i^{(k)} = (I + \Lambda^{(k)}) \mathbf{X}_i^{(k)} + \delta^{(k)}. \quad (111)$$

In our experiments, $\Lambda^{(k)}$ and $\delta^{(k)}$ are respectively random matrices and vectors with i.i.d. zero-mean Gaussian entries. The standard deviation varies based on the dataset: 0.05 for CIFAR-10 and SVHN, 0.1 for EMNIST, and 0.01 for ImageNet.

- **Label Distribution Shift:** Here, we assume that each meta-distribution is characterized by a specific α coefficient. To generate each user’s data under this shift, the number of samples per class is determined by a Dirichlet distribution with parameter α . In our experiments, we use $\alpha = 0.4$.
- **Feature & Label Distribution Shift:** As the name suggests, this shift combines both the feature and label distribution shifts described above to create a new distribution for each user.

Figure 4 illustrates our performance certificates (i.e., bounds on the risk CDF) for unseen clients when there are no shifts. We selected 100 nodes from the population and considered 400 other nodes as unseen clients. We then plotted the CDFs based on 100 samples and confirmed that our bounds hold for the real population as well. Due to the standard DKW inequality, the empirical CDF is a good estimate for the test-time non-robust risk CDF.

G.2. Certificates for f -Divergence Meta-Distributional Shifts

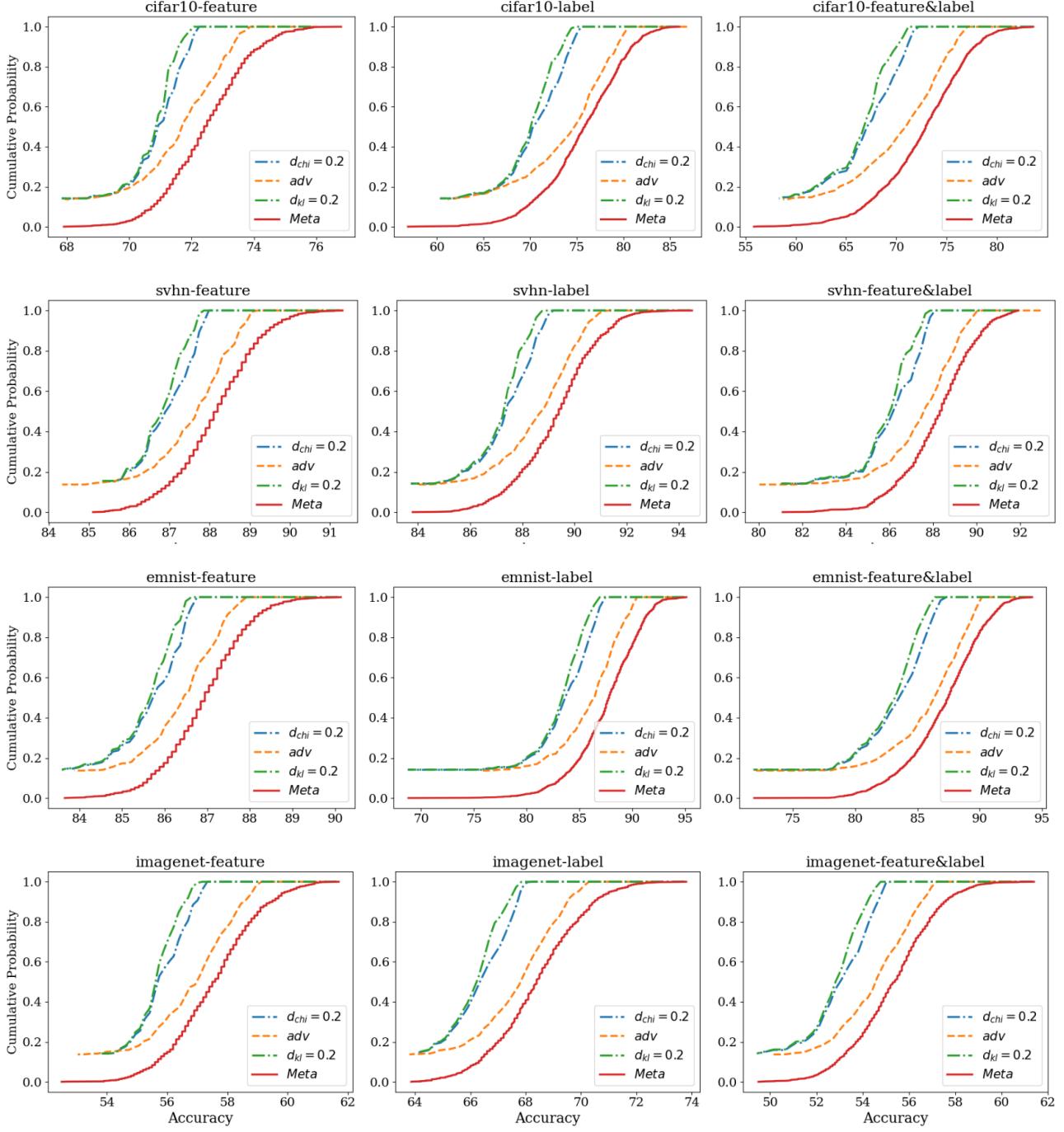


Figure 9. Extension of previous experiments to a broader range of adversarial budgets for f -divergence attacks. DKW-based certificates for unseen clients in our four examined datasets. *Meta* refers to the main population with 1000 nodes.

In this section, we examined scenarios where users belong to two distinct meta-distributions: the source and the target. A DNN-based model is initially trained on a network of clients sampled from the source. The resulting risk values are then fed into the optimization problems introduced in Section 5 to obtain robust CDF bounds, considering both the Chi-Square and KL divergence as potential choices for f . Finally, we empirically estimate the risk CDF for users from the target

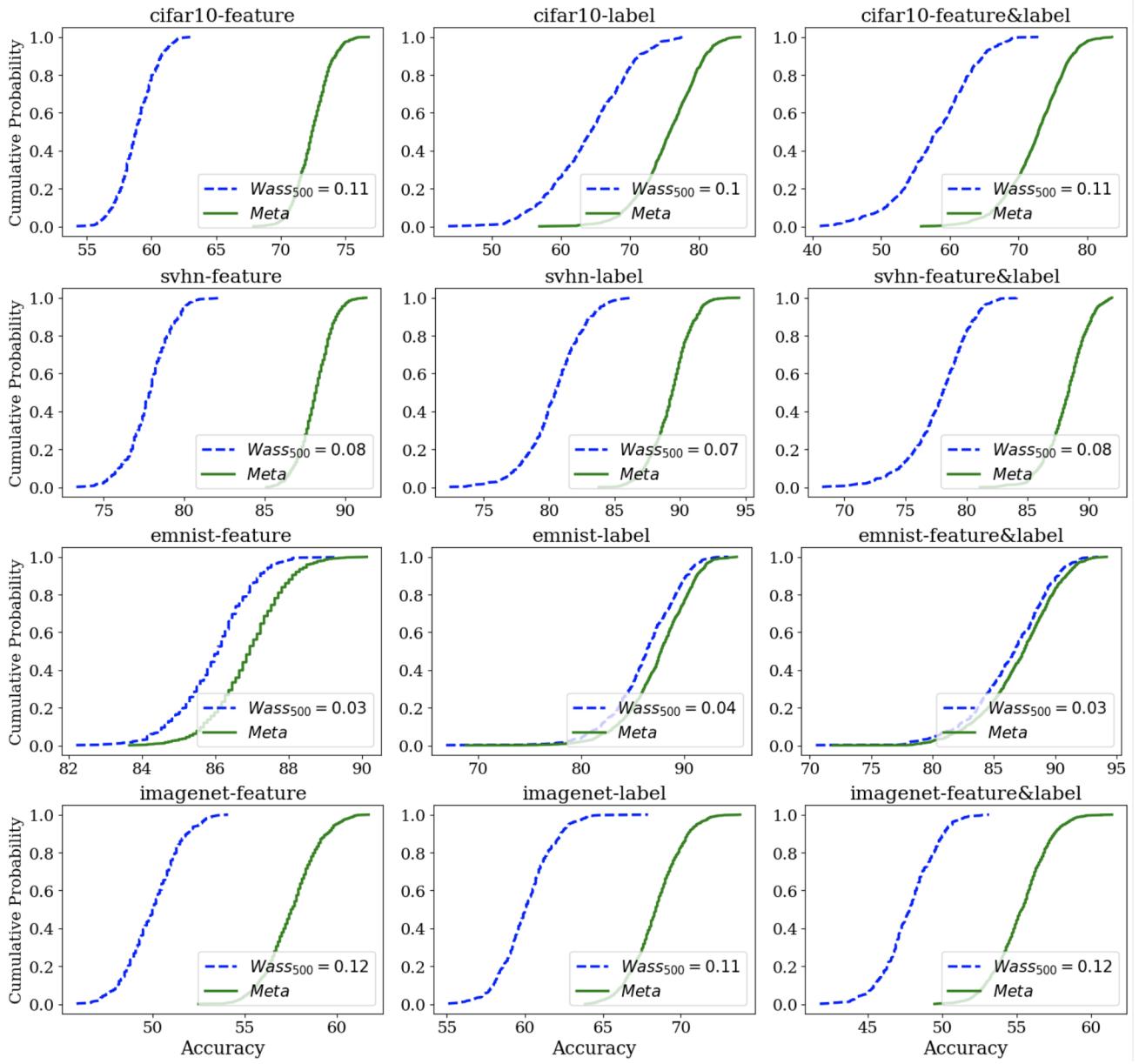


Figure 10. Wasserstein distance-based certificates for unseen clients in our four examined datasets. *Meta* refers to the main population with 1000 nodes. Dotted curves are based on 500 networks within the population.

meta-distribution and validate our bounds. Specifically, we tested our certificates in two distinct settings using the CIFAR-10 dataset (see Figure 5). We generated various image categories with differing resolutions or color schemes, and then sampled from these categories to create different distributions:

- **Resolutions:** Images were cropped and resized to create eight different resolutions. The Dirichlet α coefficients for the first (source) meta-distribution range from 0.4 to 0.7 for the four lower resolutions and from 0.7 to 1 for the four higher resolutions. For the second (target) meta-distribution, the ranges are reversed: 0.7 to 1 for the lower resolutions and 0.4 to 0.7 for the higher resolutions. The number of samples per resolution for each user is determined using a Dirichlet distribution, with α coefficients randomly selected from the specified range for the meta-distribution. As a result, users sampled from source will have more high-resolution images, while users from the target will have more low-resolution

samples.

- **Colors:** The color intensity of the images varies from 0.00 (gray-scale) to 1.00 (fully colored). For the source meta-distribution, the α coefficients range from 0 to 0.5 for images with color intensity below 0.5, and from 0.5 to 1 for images above 0.5. As with the resolution setting, the ranges are reversed for the target, and the number of samples per color intensity for each user is calculated similarly. Therefore, users sampled from source will have more colorful images, while those from the target will have more gray-scale images.

Figure 6 (left) verifies our CDF certificates based on both chi-square and KL-divergence (dotted curves) for the target meta distribution (orange curve). As can be seen, bounds have tightly captured the behavior of risk CDF in the target network. More detailed experiments are shown in Figure 9 in Appendix G.

G.3. Certificates for Wasserstein-based Meta-Distributional Shifts

In this experiment, as previously mentioned, we used affine distribution shifts to create new domains. Figure 6 (right) summarizes our numerical results in this scenario. To generate different networks within the meta-distribution, we applied the affine distribution shifts described in Section G.1. Once again, the results validate our certificates, this time for Wasserstein-type shifts. The blue curve, representing the real population, consistently falls within or beneath the blue shaded area. Regarding tightness, it is important to note that the bounds presented here remain tight, particularly under adversarial attacks as defined by a distributional adversary in (Sinha et al., 2018). More detailed experiments with various levels of tightness are shown in Figure 10 in Appendix G.

Although our theoretical findings in Section 6 focus solely on the average risk and not the risk CDF, we extended the same framework to the CDF in this experiment to explore whether the theory might also apply. The results were positive, suggesting potential for extending our theoretical findings in this area. A more detailed series of simulation results are presented via Figures 9 and 10, respectively.

Figure 9 illustrates complementary results for f -divergence bounds on risk CDF, which are robust extensions of the DKW bound. Simulations have been repeated for several different values of ε , where both KL and χ^2 (Chi-Square) divergences have been considered.

Figure 10 shows the robust CDF bounds for Wasserstein shifts. Again, different values of ε for Wasserstein distance have been considered which have resulted in several bounds with various levels of tightness. It should be noted that all bounds are asymptotically minimax optimal, meaning that an adversary can choose a distribution to perform exactly as bad as the bounds, at least in the asymptotic regime where both K and $\min_k n_k$ tend to infinity.