Distributionally and Adversarially Robust Logistic Regression via Intersecting Wasserstein Balls

Aras Selvi[*†]
Imperial College Business School
a.selvi19@imperial.ac.uk

Eleonora Kreačić [*]
JP Morgan AI Research
eleonora.kreacic@jpmorgan.com

Mohsen Ghassemi
JP Morgan AI Research
mohsen.ghassemi@jpmorgan.com

Vamsi K. Potluru
JP Morgan AI Research
vamsi.k.potluru@jpmchase.com

Tucker Balch
JP Morgan AI Research
tucker.balch@jpmchase.com

Manuela Veloso
JP Morgan AI Research
manuela.veloso@jpmchase.com

Abstract

Empirical risk minimization often fails to provide robustness against adversarial attacks in test data, causing poor out-of-sample performance. Adversarially robust optimization (ARO) has thus emerged as the *de facto* standard for obtaining models that hedge against such attacks. However, while these models are robust against adversarial attacks, they tend to suffer severely from overfitting. To address this issue for logistic regression, we study the Wasserstein distributionally robust (DR) counterpart of ARO and show that this problem admits a tractable reformulation. Furthermore, we develop a framework to reduce the conservatism of this problem by utilizing an auxiliary dataset (*e.g.*, synthetic, external, or out-of-domain data), whenever available, with instances independently sampled from a nonidentical but related ground truth. In particular, we intersect the ambiguity set of the DR problem with another Wasserstein ambiguity set that is built using the auxiliary dataset. We analyze the properties of the underlying optimization problem, develop efficient solution algorithms, and demonstrate that the proposed method consistently outperforms benchmark approaches on real-world datasets.

1 Introduction

Supervised learning traditionally involves access to a training dataset whose instances are assumed to be independently sampled from a true data-generating distribution [12, 32]. Optimizing an expected loss for the empirical distribution constructed from such a training set, also known as *empirical risk minimization* (ERM), enjoys several desirable properties in relatively generic settings, including convergence to the true risk minimization problem as the number of training samples increases [72, Chapter 2]. In practice, however, data is finite, and ERM suffers from the "optimism bias" that is also known as overfitting [43] or the optimizer's curse [20, 63], which causes deteriorated out-of-sample performance. A popular paradigm to prevent this phenomenon is *distributionally robust optimization* (DRO) [19] which optimizes the expected loss for the worst-case distribution that resides in an ambiguity set constructed from the empirical distribution.

[*]corresponding authors
[†]work completed during the internship at JPMorganChase

Preprint. Under review.

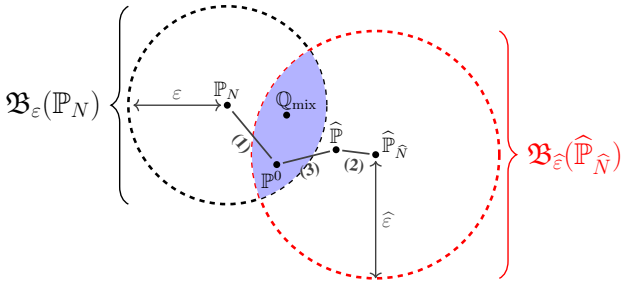

Figure 1: ARO optimizes the expected adversarial loss over the empirical distribution $\mathbb{P}_N$ that is constructed from N i.i.d. samples of the (unknown) true data-generating distribution $\mathbb{P}^0$. Replacing $\mathbb{P}_N$ with the worst-case distribution in a ball $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$ gives us its DR counterpart. To reduce the size of this ball while ensuring $\mathbb{P}^0$ is still included with high confidence, we intersect it with another ball $\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{P}_{\widehat{N}})$. The latter ball is built around an empirical distribution $\widehat{\mathbb{P}}_N$ that is constructed from $\widehat{N}$ i.i.d. samples of some auxiliary data distribution $\widehat{\mathbb{P}}$ (*cf.* Section 5). The intersection includes $\mathbb{P}^0$ if $\varepsilon \geq (1)$ and $\widehat{\varepsilon} \geq (2) + (3)$ (*cf.* Section 6). Recent works using auxiliary data in ARO propose optimizing the expected adversarial loss over a mixture $\mathbb{Q}_{\text{mix}}$ of $\mathbb{P}_N$ and $\widehat{P}_{\widehat{N}}$; we show that this distribution also resides in $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ under some conditions (Proposition 5.8).

Another actively studied real-world challenge causing poor out-of-sample performance of ERM is adversarial attacks, where an adversary perturbs the observed features in the testing or deployment phase [67, 28]. For neural networks, the paradigm of *adversarial training* (AT) [38] is therefore designed to provide adversarial robustness by simulating the attacks during the training stage. Many successful variants of AT specialized to different domains, losses, and attacks have been proposed in the literature to achieve adversarial robustness without significantly deteriorating the performance on training sets [56, 81, 45, 27]. While some works (*e.g.,* [14, 70]) examine adversarial robustness guarantees of various training algorithms, there is a recent stream of work (*e.g.,* [6, 76]) that studies properties of optimal solutions to the *adversarially robust optimization* (ARO) problems where one optimizes the empirical risk subject to worst-case adversarial attacks.

Recently, it has been observed that adversarially robust (AR) models may suffer from severe overfitting (*robust overfitting*, [49, 79, 37]), that is, AR models are not DR. Indeed, it is observed that robust overfitting is even more severe than traditional overfitting [51]. While some works address robust overfitting of AT through algorithmic adjustments [16, 36], a recent study [6, Thm 3.2] proves that robust overfitting is more severe than traditional overfitting via DRO theory. The authors of the latter work thus propose the simultaneous adoption of DR and AR.

In this paper, we adopt a Wasserstein DRO approach to address robust overfitting in the ℓ_p-attack setting [18] for logistic regression. We study both the traditional setting with an empirical dataset and an extension that incorporates an auxiliary dataset whose instances are sampled from a nonidentical but related distribution. Examples of auxiliary data include synthetic data generated from a generative model (*e.g.*, releasing portions of data under privacy constraints), data in the presence of distributional shifts (*e.g.*, different time period/geographic region), noisy data (*e.g.*, measurement errors), or out-of-domain data (*e.g.*, different source); any distribution is applicable as long as the Wasserstein distance between its underlying data-generating distribution and the true data-generating distribution is known or can be estimated (formal setup in Section 5). We propose a *distributionally and adversarially robust* model, constructing its ambiguity set from empirical and auxiliary datasets. Specifically, we first develop a Wasserstein DR counterpart of ARO without auxiliary data, which already improves the benchmark ARO methods. Our primary contribution, however, is intersecting this empirical Wasserstein ambiguity set (ball) with an additional ball formed around the auxiliary data. This method mitigates conservatism in DRO by refining its ambiguity set. We analyze the statistical properties and complexities attributed to this problem, and develop efficient approximation algorithms. Figure 1 illustrates the idea and Appendix A contains notation. Our contributions are:

- We show that ARO for logistic loss is equivalent to the ERM of a new loss function, which is convex and Lipschitz, allowing us to use recent Lipschitz DRO theory (*cf.* Section 4).

- We thus formulate distributionally *and* adversarially robust logistic regression (LR) and provide an *exact* tractable convex optimization reformulation (*cf.* Section 4).
- We utilize auxiliary data to reduce the conservatism of the aforementioned DRO problem in Section 5 (*cf.* Figure 1). We prove that the resulting optimization problem is NP-hard and develop a tractable approximation.
- We prove that Wasserstein finite sample guarantees are inherited by our optimization models and discuss how to set the radii of the Wasserstein balls (*cf.* Section 6).
- Experiments on UCI datasets and MNIST/EMNIST datasets demonstrate that our approach achieves better out-of-sample performance than benchmark algorithms with and without adversarial attacks, and scales graciously in practical settings (*cf.* Section 7).

2 Related work

Auxiliary data in ARO Despite the difference in motivation from ours, auxiliary data appears in the ARO literature. In particular, it is shown that additional unlabeled data sampled from the same [15, 75] or different [21] data-generating distributions could improve adversarial robustness. [54] shows adversarial robustness guarantees can be certified even when AT is done on a synthetic dataset if its generator's distance to the true distribution can be quantified. [30, 76] propose optimizing a weighted combination of ARO over empirical and synthetic datasets. We show that the latter approach is generalized by our model (*cf.* Proposition 5.8).

DRO-ARO interactions In our work, we solve ARO for the worst-case data distribution residing in a type-1 Wasserstein ball around the empirical distribution, since the type-1 Wasserstein metric is arguably the most common choice in machine learning (ML) with Lipschitz losses [59, 26]. In the literature, it is shown that the standard (non-DR) ARO is equivalent to the DRO of the original loss function with a type-∞ Wasserstein metric [65, 33, 50] (or a Lévy-Prokhorov metric [7]). Hence, our DR ARO approach can be interpreted as optimizing the logistic loss over the worst-case distribution whose 1-Wasserstein distance is bounded by a pre-specified radius from at least one distribution that resides in an ∞-Wasserstein ball around the empirical distribution. Conversely, [62] discusses that while DRO over Wasserstein balls is intractable for generic functions (*e.g.*, neural networks), its Lagrange relaxation resembles ARO and thus AT yields a certain degree of (relaxed) distributional robustness; this introduces a DRO perspective to AT algorithms [74, 13, 47]. However, to the best of our knowledge, there have not been works optimizing a pre-specified level of Wasserstein distributional robustness (that hedges against overfitting, [34]) and adversarial robustness (that hedges against adversarial attacks, [28]) *simultaneously*. To our knowledge, the only work that considers the DR counterpart of ARO is [6] where the distributional ambiguity is modeled with φ-divergences and the prediction model is a neural network.

Intersecting ambiguity sets in DRO Recent work started to explore the intersection of ambiguity sets for different contexts [2] or different metrics [82]. Our idea of intersecting Wasserstein balls is inspired by the "Surround, then Intersect" strategy [68, §5.2] to train linear regression under sequential domain adaptation in a non-adversarial setting (see [57] and [64] for robustness in domain adaptation/transfer learning). The aforementioned work focuses on a case where the loss function is the squared loss, and the metric is a variant of the Wasserstein metric developed for the first and second distributional moments.

Logistic Loss in DRO and ARO Our choice of LR aligns with the current directions and open questions in the relevant literature. In the ARO literature, there are recent theory developments on understanding the effect of auxiliary data (*e.g.*, [76]) for squared and logistic loss functions. In the DRO literature, even in the absence of adversarial attacks, the aforementioned work [68] on the intersection of Wasserstein ambiguity sets is restricted to linear regression. The authors show that this problem admits a tractable convex optimization reformulation, and the proof relies on the properties of the squared loss. We contribute to the DRO literature for adversarial and non-adversarial settings because we show that such a problem would be NP-hard for the logistic loss (*cf.*, Proposition 5.3), and develop specialized approximation techniques. Our problem recovers DR LR [60, 55] as a special case in the absence of adversarial attacks and auxiliary data. The theoretical challenges posed by the

Table 1: Comparison of the risks taken under various training paradigms.

	ERM	DRO	ARO
Training risk	$\mathbb{E}_{\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$	$\sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$	$\mathbb{E}_{\mathbb{P}_N}\big[\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p\leq\alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z}, y)\}\big]$
True risk	$\mathbb{E}_{\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$	$\mathbb{E}_{\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$	$\mathbb{E}_{\mathbb{P}^0}\big[\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p\leq\alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z}, y)\}\big]$

logistic loss have been a significant focus in DRO literature, with extensions such as DR LR [60] to DRO Lipschitz ML [59] and mixed-feature DR LR [55] to mixed-feature DR Lipschitz ML [3].

3 Problem setting and preliminaries

We consider a binary classification problem where an instance is modeled as $(\boldsymbol{x}, y) \in \Xi := \mathbb{R}^n \times \{-1, +1\}$ and the labels depend on the features probabilistically with $\text{Prob}[y \mid \boldsymbol{x}] = [1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x})]^{-1}$, for some $\boldsymbol{\beta} \in \mathbb{R}^n$; its associated loss is the *logloss* $\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y) := \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x}))$.

Distributional ambiguity and Wasserstein balls Let $\mathcal{P}(\Xi)$ denote the set of probability distributions on Ξ. We model distributional ambiguity via the *Wasserstein (Earth mover's) distance*.

Definition 3.1 (Feature-label metric)**.** The distance $d(\boldsymbol{\xi}, \boldsymbol{\xi'})$ between two instances $\boldsymbol{\xi} = (\boldsymbol{x}, y) \in \Xi$ and $\boldsymbol{\xi'} = (\boldsymbol{x'}, y') \in \Xi$ is

$$d(\boldsymbol{\xi}, \boldsymbol{\xi'}) = \|\boldsymbol{x} - \boldsymbol{x'}\|_q + \kappa \cdot \mathbb{1}[y \neq y'],$$

where $\kappa > 0$ controls the label weight and $q > 0$ specifies a rational norm on $\mathbb{R}^n$.

Definition 3.2. The type-1 Wasserstein distance between distributions $\mathbb{Q} \in \mathcal{P}(\Xi)$ and $\mathbb{Q}' \in \mathcal{P}(\Xi)$, with ground metric $d(\boldsymbol{\xi}, \boldsymbol{\xi'})$ on Ξ, is defined as

$$\mathrm{W}(\mathbb{Q}, \mathbb{Q}') = \inf_{\Pi\in\mathcal{C}(\mathbb{Q},\mathbb{Q}')} \left\{ \int_{\Xi\times\Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi'}) \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \right\}$$

where $\mathcal{C}(\mathbb{Q}, \mathbb{Q}') := \{\Pi \in \mathcal{P}(\Xi \times \Xi) : \Pi(\mathrm{d}\boldsymbol{\xi}, \Xi) = \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}), \Pi(\Xi, \mathrm{d}\boldsymbol{\xi'}) = \mathbb{Q}'(\mathrm{d}\boldsymbol{\xi'})\}$.

For $\varepsilon > 0$, the Wasserstein ball around $\mathbb{P} \in \mathcal{P}(\Xi)$ is defined $\mathfrak{B}_\varepsilon(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}(\Xi) : \mathrm{W}(\mathbb{Q}, \mathbb{P}) \leq \varepsilon\}$. We next review several training paradigms, see Table 1.

Empirical Risk Minimization Let $\mathbb{P}^0$ denote the true data-generating distribution. Ideally, one wants to minimize the expected loss over $\mathbb{P}^0$, or more precisely

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \quad \mathbb{E}_{\mathbb{P}^0}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]. \tag{RM}$$

In practice, $\mathbb{P}^0$ is hardly ever known, and one thus resorts to the empirical distribution $\mathbb{P}_N = \frac{1}{N}\sum_{i\in[N]} \delta_{\boldsymbol{\xi}^i}$ where $\{\boldsymbol{\xi}^i = (\boldsymbol{x}^i, y^i)\}_{i\in[N]}$ are i.i.d. samples from $\mathbb{P}^0$ and $\delta_{\boldsymbol{\xi}}$ denotes the Dirac distribution supported on $\boldsymbol{\xi}$. The empirical risk minimization (ERM) problem is thus given by

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \mathbb{E}_{\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)] \;=\; \inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \tfrac{1}{N}\textstyle\sum_{i\in[N]} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i, y^i). \tag{ERM}$$

Distributionally Robust Optimization As summarized in the introduction, DRO is motivated by the fact that in the finite-data setting, the distance between the true and empirical distributions is upper-bounded by some $\epsilon > 0$, that is, $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$. The goal in DRO is to optimize the expected loss over the worst possible realization of a distribution residing in $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$:

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \quad \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \quad \mathbb{E}_{\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]. \tag{DRO}$$

We refer to [40] and [34] for the generalization guarantees and ML applications of DRO.

Adversarial Robustness The goal of adversarial robustness is to provide robustness against adversarial attacks [28]. An adversarial attack, in the widely studied ℓ_p-noise setting [18], perturbs the features of the test instances $(\boldsymbol{x}, y)$ by adding additive noise $\boldsymbol{z}$ to $\boldsymbol{x}$. The adversary chooses the noise vector $\boldsymbol{z}$, subject to $\|\boldsymbol{z}\|_p \leq \alpha$, so as to maximize the loss $\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z}, y)$ associated with this perturbed test instance. Therefore, ARO solves the following optimization problem in the training stage to hedge against adversarial perturbations at the test stage:

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \quad \mathbb{E}_{\mathbb{P}_N}\left[\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p\leq\alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z}, y)\}\right]. \tag{ARO}$$

ARO reduces to ERM when $\alpha = 0$. Note that ARO is identical to feature robust training [9] which is not motivated by adversarial attacks, but the presence of noisy observations in the training set [4, 29].

4 Wasserstein adversarially robust optimization

ARO replaces the loss function of ERM with the worst-case loss (with respect to adversarial attacks). Here we show that ARO is equivalent to an ERM of a modified loss, which is convex and Lipschitz.

Proposition 4.1. *Let $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y) := \log(1+\exp(-y\cdot\boldsymbol{\beta}^\top\boldsymbol{x} + \alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$ denote the adversarial loss associated with the logloss, and $L^\alpha(z) := \log(1+\exp(-z+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$ its univariate counterpart. We have $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y) = \sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p\leq\alpha}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z}, y)\}$ and so ARO is identical to*

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \quad \mathbb{E}_{\mathbb{P}_N}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)].$$

Moreover, $\mathrm{Lip}(L^\alpha) = 1$ *for any* $\alpha \geq 0$.

The proof is in Appendix B.1. Proposition 4.1 tells us that *true* expected loss under adversarial attacks is $\mathbb{E}_{\mathbb{P}^0}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$. Therefore, instead of optimizing the empirical risk $\mathbb{E}_{\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$, ARO optimizes the empirical *adversarial* risk $\mathbb{E}_{\mathbb{P}_N}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$. This means that ARO calibrates the loss function so that we train and test with the *same* loss $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$. However, ARO still optimizes this loss for the empirical distribution $\mathbb{P}_N$ and is thus prone to overfitting due to the statistical error of estimating $\mathbb{P}^0$ with $\mathbb{P}_N$. To address overfitting in the adversarial setting (robust overfitting of ARO), we derive a Wasserstein DR counterpart of ARO. We start with the following assumption.

Assumption 4.2. We are given finite $\varepsilon > 0$ satisfying $\mathrm{W}(\mathbb{P}^0, \mathbb{P}_N) \leq \varepsilon$ (*i.e.*, $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$).

We discuss relaxing this assumption in Section 6. We now introduce the *distributionally and adversarially robust logistic regression* problem:

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p\leq\alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z}, y)\}]. \tag{DR-ARO}$$

The following result shows that, for a fixed ε, DR-ARO can be reformulated as a convex optimization problem. This is a direct corollary of Proposition 4.1 and Theorem 14 (ii) of [59]; see Appendix B.2.

Corollary 4.3. *Problem DR-ARO admits the following tractable convex optimization reformulation:*

$$\begin{array}{ll} \inf\limits_{\boldsymbol{\beta},\lambda,\boldsymbol{s}} & \varepsilon\lambda + \frac{1}{N}\sum_{i=1}^N s_i \\ \text{s.t.} & \ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}^i, y^i) \leq s_i,\ \ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}^i, -y^i) - \lambda\kappa \leq s_i \quad \forall i\in[N] \\ & \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda,\ \boldsymbol{\beta}\in\mathbb{R}^n,\ \lambda\geq 0,\ \boldsymbol{s}\in\mathbb{R}^N_+. \end{array}$$

The constraints of this problem are exponential cone representable (Appendix C), and for $q \in \{1, 2, \infty\}$, the yielding problem can be solved with the exponential cone solver of MOSEK [41] in polynomial time (with respect to their input size [44]). DR-ARO addresses the overfitting issue of ARO by solving its distributionally robust counterpart. However, the DRO approach of considering the worst-case distribution within a ball around the empirical distribution can be overly conservative [53]. Next, we explore how employing auxiliary data can reduce this conservatism.

5 Reducing conservatism of DR-ARO via intersection of Wasserstein balls

So far we have discussed the setting where we have access to an empirical distribution $\mathbb{P}_N$ that is constructed from N i.i.d. samples of the true distribution $\mathbb{P}^0$. Suppose that we have an auxiliary distribution $\widehat{\mathbb{P}}_{\widehat{N}}$ which is constructed from $\widehat{N}$ i.i.d. samples $\{\widehat{\boldsymbol{\xi}}^j = (\widehat{\boldsymbol{x}}^j, \widehat{y}^j)\}_{j\in[\widehat{N}]}$ of another distribution $\widehat{\mathbb{P}}$. In this section, we explore how auxiliary data can help us identify a *subset* of the Wasserstein ball $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$ in which $\mathbb{P}^0$ still resides. By shrinking the size of its ambiguity set, we expect to reduce the conservatism of DR-ARO. We start with the following assumption.

Assumption 5.1. We are given finite $\varepsilon, \widehat{\varepsilon} > 0$ such that $\mathrm{W}(\mathbb{P}^0, \mathbb{P}_N) \leq \varepsilon$ and $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}}_{\widehat{N}}) \leq \widehat{\varepsilon}$.

We relax this assumption in Section 6. Given Assumption 5.1, we want to solve the revised problem:

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)\cap\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})} \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]. \tag{Inter-ARO}$$

We first reformulate the intersected DR ARO problem (Inter-ARO) as a semi-infinite optimization problem with finite variables and then provide a complexity result.

Proposition 5.2. *Inter-ARO admits the following reformulation.*

$$\begin{aligned}
\inf_{\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}} \quad & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \tfrac{1}{N}\textstyle\sum_{i=1}^N s_i + \tfrac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \widehat{s}_j \\
\text{s.t.} \quad & \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, l) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq s_i + \frac{\kappa(1-ly^i)}{2}\lambda + \widehat{s}_j + \frac{\kappa(1-l\widehat{y}^j)}{2}\widehat{\lambda} \\
& \qquad\qquad \forall i \in [N],\ j \in [\widehat{N}],\ l \in \{-1, 1\} \\
& \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda} \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N_+,\ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+.
\end{aligned}$$

The proof is in Appendix B.3. Even though this problem recovers DR-ARO (hence admits tractable reformulations) when the radius $\widehat{\varepsilon}$ of the second ball satisfies $\widehat{\varepsilon} \to \infty$, Proposition 5.3 shows that it is NP-hard in the finite radius settings. We reformulate Inter-ARO as an adjustable robust optimization problem [5, 77], and borrow tools from this literature to obtain the following result.

Proposition 5.3. *Inter-ARO is equivalent to an adjustable robust optimization problem with $\mathcal{O}(N\cdot\widehat{N})$ two-stage robust constraints, which is NP-hard even when $N = \widehat{N} = 1$.*

The proof is in Appendix B.4. The adjustable robust optimization literature has developed a rich arsenal of relaxation techniques that can be leveraged for Inter-ARO. We adopt the 'static relaxation technique' [10] to restrict the feasible region of Inter-ARO and obtain a tractable approximation.

Theorem 5.4. *The following convex optimization problem is a feasible relaxation (safe approximation) of Inter-ARO:*

$$\begin{aligned}
\inf_{\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}},\boldsymbol{z}^+_{ij},\boldsymbol{z}^-_{ij}} \quad & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \tfrac{1}{N}\textstyle\sum_{i=1}^N s_i + \tfrac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \widehat{s}_j \qquad \text{(Inter-ARO}^\star\text{)} \\
\text{s.t.} \quad & \begin{bmatrix} L^\alpha(l\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \boldsymbol{z}^{l\top}_{ij}(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i)) \leq s_i + \frac{\kappa(1-ly^i)}{2}\lambda + \widehat{s}_j + \frac{\kappa(1-l\widehat{y}^j)}{2}\widehat{\lambda}, \\ \|l\boldsymbol{\beta} - \boldsymbol{z}^l_{ij}\|_{q^\star} \leq \lambda,\ \|\boldsymbol{z}^l_{ij}\|_{q^\star} \leq \widehat{\lambda} \end{bmatrix} \\
& \qquad\qquad \forall i \in [N],\ j \in [\widehat{N}],\ l \in \{-1, 1\} \\
& \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda} \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N_+,\ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+,\ \boldsymbol{z}^l_{ij} \in \mathbb{R}^n,
\end{aligned}$$

where $L^\alpha(z) := \log(1 + \exp(-z + \alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$ is the univariate representation of $\ell^\alpha_{\boldsymbol{\beta}}$.

The proof is in Appendix B.5. Inter-ARO* relaxes the NP-hard problem Inter-ARO so that it becomes efficiently solvable, and it enjoys similar tractable formulations to DR-ARO.

Remark 5.5. Inter-ARO* admits an exponential cone reformulation, analogously to Appendix C.

Recall that for $\widehat{\varepsilon}$ large enough so that $\mathfrak{B}_\varepsilon(\mathbb{P}_N)\cap\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}}) = \mathfrak{B}_\varepsilon(\mathbb{P}_N)$, Inter-ARO reduces to DR-ARO. The following corollary (proof in Appendix B.6) shows that a similar desired property holds for the relaxed problem Inter-ARO*. That is, "not learning anything from auxiliary data" remains feasible: the static relaxation does not force learning from $\widehat{\mathbb{P}}_{\widehat{N}}$, it learns from auxiliary data only if the objective improves. Moreover, we show that as $\widehat{\varepsilon} \to \infty$, Inter-ARO* converges to Inter-ARO.

Corollary 5.6. *Feasibility of disregarding auxiliary data: Any feasible solution $(\boldsymbol{\beta}, \lambda, \boldsymbol{s})$ of DR-ARO gives a feasible solution $(\boldsymbol{\beta}, \lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}, \boldsymbol{z}_{ij}^{+}, \boldsymbol{z}_{ij}^{-})$ for Inter-ARO* with $\widehat{\lambda} = 0$, $\widehat{\boldsymbol{s}} = \mathbf{0}$, $\boldsymbol{z}_{ij}^{+} = \boldsymbol{z}_{ij}^{-} = \mathbf{0}$. Convergence to Inter-ARO: As $\widehat{\varepsilon} \to \infty$, the optimal value of Inter-ARO* converges to the optimal value of Inter-ARO, with the same set of optimal $\boldsymbol{\beta}$ solutions.*

Inter-ARO and Related Problems Recall that Inter-ARO can simply ignore the auxiliary data once $\widehat{\varepsilon}$ is set large enough, reducing this problem to DR-ARO. Moreover, notice that $\alpha = 0$ reduces $\ell_{\boldsymbol{\beta}}^{\alpha}$ to $\ell_{\boldsymbol{\beta}}$, hence for $\alpha = 0$ and $\widehat{\varepsilon} = \infty$ Inter-ARO recovers the Wasserstein LR model of [60]. We next relate Inter-ARO to the problems in the ARO literature that use auxiliary data $\{(\widehat{\boldsymbol{x}}^j, \widehat{y}^j)\}_{j\in[\widehat{N}]}$. The works in this literature [30, 76] propose solving the following for some $w > 0$:

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{N + w\widehat{N}} \Big[\sum_{i\in[N]} \sup_{\boldsymbol{z}^i\in\mathcal{B}_p(\alpha)} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i + \boldsymbol{z}^i, y^i)\} + w \sum_{j\in[\widehat{N}]} \sup_{\boldsymbol{z}^j\in\mathcal{B}_p(\alpha)} \{\ell_{\boldsymbol{\beta}}(\widehat{\boldsymbol{x}}^j + \boldsymbol{z}^j, \widehat{y}^j)\} \Big], \quad (1)$$

where $\mathcal{B}_p(\alpha) := \{\boldsymbol{z} \in \mathbb{R}^n : \|\boldsymbol{z}\|_p \le \alpha\}$. We first observe that this resembles ARO, with the empirical distribution $\mathbb{P}_N$ being replaced with its mixture with $\widehat{\mathbb{P}}_{\widehat{N}}$:

Proposition 5.7. *Problem* (1) *is equivalent to:*

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \quad \mathbb{E}_{\mathbb{Q}_{\rm mix}}[\ell_{\boldsymbol{\beta}}^{\alpha}(\boldsymbol{x}, y)] \quad (2)$$

where $\mathbb{Q}_{\rm mix} := \lambda \cdot \mathbb{P}_N + (1-\lambda) \cdot \widehat{\mathbb{P}}_{\widehat{N}}$ for $\lambda = \frac{N}{N + w\widehat{N}}$.

The proof is in Appendix B.7. Next, we give a condition on ε and $\widehat{\varepsilon}$ to guarantee that the mixture distribution introduced in Proposition 5.7 lives in $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$, that is, the distribution $\mathbb{Q}_{\rm mix}$ will be feasible in the sup problem of Inter-ARO.

Proposition 5.8. *For any $\lambda \in (0,1)$ and $\mathbb{Q}_{\rm mix} := \lambda\cdot\mathbb{P}_N + (1-\lambda)\cdot\widehat{\mathbb{P}}_{\widehat{N}}$, whenever $\varepsilon + \widehat{\varepsilon} \ge \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$ and $\frac{\widehat{\varepsilon}}{\varepsilon} = \frac{\lambda}{1-\lambda}$ are satisfied, we have $\mathbb{Q}_{\rm mix} \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$.*

The proof is in Appendix B.8. For $\lambda = \frac{N}{N+\widehat{N}}$, if the intersection $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ is nonempty, Proposition 5.8 implies that a sufficient condition for this intersection to include the mixture $\mathbb{Q}_{\rm mix}$ is $\widehat{\varepsilon}/\varepsilon = N/\widehat{N}$, which is intuitive since the radii of the Wasserstein ambiguity sets are typically chosen inversely proportional to the number of samples [34, Theorem 18].

6 Selecting Wasserstein radii

Our analyses thus far have assumed knowledge of DRO ball radii ε and $\widehat{\varepsilon}$ (Assumptions 4.2 and 5.1). These are unrealistic in most real-world scenarios. Here we discuss how to set ε and $\widehat{\varepsilon}$ based on the data such that Problems DR-ARO and Inter-ARO remain well-defined. We consider two settings. First we discuss the case where $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ is known. Then, we discuss the most realistic scenario where this distance is unknown. To this end, we investigate the statistical properties of our distributionally and adversarially robust optimization models to be able to set ε and $\widehat{\varepsilon}$ values.

Choosing ϵ in DR-ARO. In order to relax Assumption 4.2 in Problem DR-ARO, one needs to infer ε value from the empirical data so that $\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$ with a pre-specified level of confidence. The following theorem presents tight characterizations for ε so that the ball $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$ includes the true distribution $\mathbb{P}^0$ with arbitrarily high confidence, and shows that for an ε chosen in such manner, Problem DR-ARO is well-defined. The detailed statement and the proof are in Appendix B.9.

Theorem 6.1 (abridged). *For light-tailed distribution $\mathbb{P}^0$ and $\varepsilon \ge \mathcal{O}(\frac{\log(\eta^{-1})}{N})^{1/n}$ for $\eta \in (0,1)$, we have:* (i) *$\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$ with $1-\eta$ confidence;* (ii) *DR-ARO overestimates true loss with $1-\eta$ confidence;* (iii) *DR-ARO is asymptotically consistent $\mathbb{P}^0$-a.s.;* (iv) *worst-case distributions for optimal solutions of DR-ARO are supported on at most $N+1$ outcomes.*

Choosing ϵ and ϵ' in Inter-ARO. Inter-ARO revises DR-ARO by intersecting $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$ with another ball $\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ centered at the auxiliary distribution. We need a nonempty intersection for Inter-ARO to be well-defined. A sufficient condition follows from the triangle inequality: $\varepsilon + \widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$. Moreover, provided that $\varepsilon \geq \mathrm{W}(\mathbb{P}_N, \mathbb{P}^0)$, a sufficient condition for $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}}) = \mathfrak{B}_\varepsilon(\mathbb{P}_N)$ is $\widehat{\varepsilon} \geq \varepsilon + \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}) + \mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \widehat{\mathbb{P}})$ (*cf.* Figure 1). While choosing such $\widehat{\varepsilon}$ to reduce the size of the ambiguity set of DR-ARO, we want this intersection to include $\mathbb{P}^0$, assuming ε is set in light of Theorem 6.1. The auxiliary data $\widehat{\mathbb{P}}_{\widehat{N}}$ is constructed from instances that are independently sampled from $\widehat{\mathbb{P}}$ and thus Wasserstein finite sample statistics can estimate $\mathrm{W}(\widehat{\mathbb{P}}, \widehat{\mathbb{P}}_{\widehat{N}})$. To have confidence guarantees on $\mathbb{P}^0 \in \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$, however, we must additionally know $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ which we use in the following result. Full statement of the theorem and its proof are in Appendix B.10.

Theorem 6.2 (abridged)**.** *For light-tailed $\mathbb{P}^0$ and $\widehat{\mathbb{P}}$, if $\varepsilon \geq \mathcal{O}(\frac{\log(\eta_1^{-1})}{N})^{1/n}$ and $\widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}}) + \mathcal{O}(\frac{\log(\eta_2^{-1})}{\widehat{N}})^{1/n}$ for $\eta_1, \eta_2 \in (0,1)$ with $\eta := \eta_1 + \eta_2 < 1$, we have:* (i) $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ *with $1-\eta$ confidence;* (ii) *Inter-ARO overestimates true loss with $1-\eta$ confidence.*

Remark 6.3. Inter-ARO is not asymptotically consistent, given that $\widehat{N} \to \infty$ will let $\widehat{\varepsilon} \to \mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ due to the non-zero distance between the true distribution $\mathbb{P}^0$ and the auxiliary distribution $\widehat{\mathbb{P}}$. Inter-ARO is thus not useful in asymptotic data regimes, which is not surprising given that we introduced it to reduce the conservatism of DR-ARO which by design arises in non-asymptotic settings.

Knowledge of $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$. In the above results, we assumed that $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ is known. However, this is challenging in most practical settings [48] and we estimate it via cross validation (as in the transfer learning and domain adaptation literature, [83]). For some special cases, we can use domain knowledge (*e.g.*, the "Uber vs Lyft" example of [68]). For example, in a differential privacy context, a data holder shares a subset of opt-in data to form $\mathbb{P}_N$, and generates a privacy-preserving synthetic dataset from the rest. Due to challenges in synthetic data generation under privacy constraints, the synthetic distribution approximates the true distribution, resulting in a nonzero Wasserstein distance [24, 71]. Using this distance will complete the above discussion. Another research direction relies on $\mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}})$ when it is known, especially when synthetic data generators are trained on the empirical dataset. By employing Wasserstein GANs, which minimize the Wasserstein-1 distance, the distance between the generated distribution and the training distribution is minimized. This ensures that the synthetic distribution remains within a small radius of the training distribution [1].

7 Experiments

We conduct a series of experiments to test the proposed DR ARO models using empirical and auxiliary datasets. We use the following abbreviations: ERM and ARO stand for solving problems ERM (*i.e.*, minimization of the empirical logistic loss) and ARO (*i.e.*, adversarial training for logistic loss), respectively. ARO+Aux refers to solving problem (1), that is, replacing the empirical distribution of ARO with its mixture with auxiliary data. DRO+ARO is solving DR-ARO, which is the Wasserstein DR counterpart of ARO. Finally, DRO+ARO+Aux refers to solving Inter-ARO, which revises DR-ARO by intersecting its ambiguity set with a Wasserstein ball built using auxiliary data. Note that, ERM, ARO, and DRO+ARO are oblivious to auxiliary data. Finally, recall that DRO+ARO and DRO+ARO+Aux are the models that we propose. All Wasserstein radii of DR models, as well as the weight parameters of ARO+Aux are cross-validated. Implementation details are in Appendix D.

7.1 UCI datasets (auxiliary data is synthetic)

We compare the out-of-sample error rates of each method on 5 UCI datasets for classification [22, 39].

For each dataset, we run 10 simulations as follows: *(i)* Select 40% of the data as a test set ($N_{\mathrm{te}} \propto 0.4$); *(ii)* Sample 25% of the remaining to form a training set ($N \propto 0.6 \cdot 0.25$); *(iii)* The rest ($\widehat{N} \propto 0.6 \cdot 0.75$) is used to fit a synthetic generator Gaussian Copula from the SDV package [46], to sample auxiliary data from. The mean errors on the test set are reported in Table 2 for ℓ_2-attacks of strength $\alpha = 0.05$. The best error is always achieved by DRO+ARO+Aux, followed by DRO+ARO, DRO+Aux, ARO, ERM, respectively. In Appendix D.1, we report similar results for 5 more UCI datasets along with attack strengths $\alpha \in \{0, 0.05, 0.2\}$, and share data preprocessing details and standard deviations.

Table 2: Out-of-sample errors of UCI experiments with ℓ_2-attacks of strength $\alpha = 0.05$.

Data	ERM	ARO	ARO+Aux	DRO+ARO	DRO+ARO+Aux
absent	44.02%	38.82%	35.95%	34.22%	**32.64%**
anneal	18.08%	16.61%	14.97%	13.50%	**12.78%**
audio	21.43%	21.54%	17.03%	11.76%	**9.01%**
breast-c	4.74%	4.93%	3.87%	3.06%	**2.52%**
contrac	44.14%	42.86%	40.98%	40.00%	**39.65%**

7.2 MNIST/EMNIST datasets (auxiliary data is out-of-domain)

We use the MNIST [35] digits dataset to classify whether a digit is 1 or 7. For an auxiliary dataset, we use the EMNIST [17] digits dataset, as the authors of [17] summarize that the EMNIST dataset has additional samples "collected from high school students and pose a more challenging problems". Since EMNIST digits include MNIST digits, we removed the latter from the EMNIST dataset. We simulated the following 20 times: *(i)* Sample $1,000$ instances from the MNIST dataset as a training set; *(ii)* The remaining instances in the MNIST dataset are our test set; *(iii)* Sample $1,000$ instances from the EMNIST dataset as an auxiliary dataset. Table 3 reports the mean test errors under various adversarial attack regimes. The results are analogous to UCI experiments.

Table 3: Out-of-sample errors of MNIST/EMNIST experiments with various attacks.

Attack	ERM	ARO	ARO+Aux	DRO+ARO	DRO+ARO+Aux
No attack ($\alpha = 0$)	1.55%	1.55%	1.19%	0.64%	**0.53%**
ℓ_1 ($\alpha = 68/255$)	2.17%	1.84%	1.33%	0.66%	**0.57%**
ℓ_2 ($\alpha = 128/255$)	99.93%	3.36%	2.54%	2.40%	**2.12%**
ℓ_∞ ($\alpha = 8/255$)	100.00%	2.60%	2.38%	2.20%	**1.95%**

7.3 Artificial experiments (auxiliary data is perturbed)

We generate empirical and auxiliary datasets by controlling their data-generating distributions in line with the standard practice (more details in Appendix D.3). We simulate 25 cases, each with $N = 100$ training, $\widehat{N} = 200$ auxiliary, and $N_{\mathrm{te}} = 10,000$ test instances and $n = 100$. The performance of benchmark models with varying ℓ_2-attacks is available in Figure 2 (left). ERM provides the worst performance, followed by ARO, and our DRO+ARO+Aux model gives the best performance. The relationship between DRO+ARO and ARO+Aux is not monotonic: the latter works better in larger attack regimes, conforming to the robust overfitting phenomenon. Finally, Adv+DRO+Aux always performs the best. We conduct a similar simulation for datasets with $n = 100$, and gradually increase $N = \widehat{N}$ to report median ($50\% \pm 15\%$ quantiles shaded) runtimes of each method (*cf.* Figure 2, left). The fastest methods is ARO, followed by ERM, ARO+Aux, DRO+ARO, and DRO+ARO+Aux. The slowest is DRO+ARO+Aux as expected, but the runtime still scales graciously.

8 Conclusions and future work

We formulate the distributionally robust counterpart of adversarially robust logistic regression. Additionally, we demonstrate how to effectively utilize appropriately curated auxiliary data (if available) to mitigate the inherent conservatism of distributional robustness. We illustrate the superiority of the proposed approach in terms of out-of-sample performance and confirm its scalability in practical settings.

It would be natural to extend our results to more loss functions as is typical for theoretical DRO studies stemming from logistic regression. Moreover, the recent breakthroughs in the area of foundation models naturally pose the question of whether the ideas presented in this work apply to these models. For example, [78] uses a pre-trained language model (PLM) to generate synthetic pairs of text sequences and labels which are then used to train a downstream model. It would be interesting to adapt our ideas to the text domain to explore robustness in the presence of two PLMs.

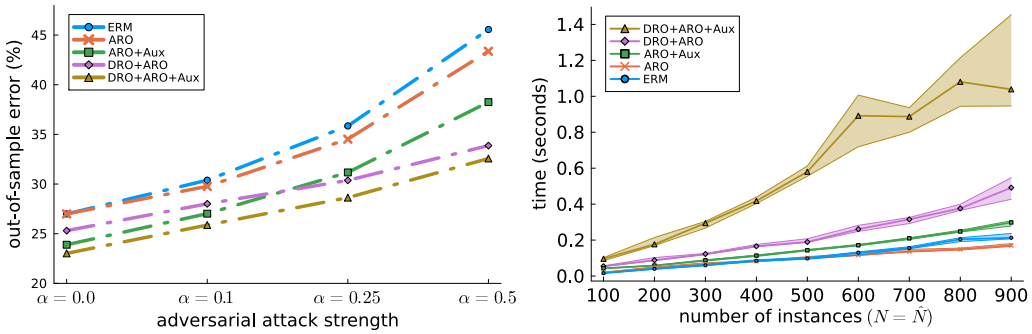

Figure 2: Out-of-sample errors under varying attack strengths (left) and runtimes under varying numbers of empirical and auxiliary instances (right) of artificial experiments.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2017.

[2] P. Awasthi, C. Jung, and J. Morgenstern. Distributionally robust data join. *arXiv:2202.05797*, 2022.

[3] R. Belbasi, A. Selvi, and W. Wiesemann. It's all in the mix: Wasserstein machine learning with mixed features. *arXiv:2312.12230*, 2023.

[4] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[5] A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99(2):351–376, 2004.

[6] A. Bennouna, R. Lucas, and B. Van Parys. Certified robust neural networks: Generalization and corruption resistance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 2023.

[7] A. Bennouna and B. Van Parys. Holistic robust data-driven decisions. *arXiv:2207.09560*, 2022.

[8] D. Bertsimas and D. Den Hertog. *Robust and Adaptive Optimization*. Dynamic Ideas, 2022.

[9] D. Bertsimas, J. Dunn, C. Pawlowski, and Y. D. Zhuo. Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34, 2019.

[10] D. Bertsimas, V. Goyal, and B. Y. Lu. A tight characterization of the performance of static solutions in two-stage adjustable robust linear optimization. *Mathematical Programming*, 150(2):281–319, 2015.

[11] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. `https://doi.org/10.1137/141000671`, 2014.

[12] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[13] T. A. Bui, T. Le, Q. Tran, H. Zhao, and D. Phung. A unified Wasserstein distributional robustness framework for adversarial training. *arXiv:2202.13437*, 2022.

[14] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. *arXiv:1902.06705*, 2019.

[15] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[16] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang. Robust overfitting may be mitigated by properly learned smoothening. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

[17] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. EMNIST: Extending MNIST to handwritten letters. In *International Joint Conference on Neural Networks*, pages 2921–2926, 2017.

[18] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv:2010.09670*, 2020.

[19] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):596–612, 2010.

[20] V. DeMiguel and F. J. Nogales. Portfolio selection with robust estimation. *Operations Research*, 57(3):560–577, 2009.

[21] Z. Deng, L. Zhang, A. Ghorbani, and J. Zou. Improving adversarial robustness via unlabeled out-of-domain data. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 2845–2853, 2021.

[22] D. Dua and C. Graff. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 1998.

[23] I. Dunning, J. Huchette, and M. Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.

[24] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[25] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[26] R. Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306, 2023.

[27] R. Gao, T. Cai, H. Li, C.-J. Hsieh, L. Wang, and J. D. Lee. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[28] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[29] B. L. Gorissen, İ. Yanıkoğlu, and D. Den Hertog. A practical guide to robust optimization. *Omega*, 53:124–137, 2015.

[30] S. Gowal, S.-A. Rebuffi, O. Wiles, F. Stimberg, D. A. Calian, and T. A. Mann. Improving robustness using generated data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[31] E. Guslitser. Uncertainty-immunized solutions in linear programming. Master's thesis, Technion – Israeli Institute of Technology, 2002.

[32] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical learning: Data mining, Inference, and Prediction*. Springer, 2009.

[33] J. Khim and P.-L. Loh. Adversarial risk bounds via function transformation. *arXiv:1810.09519*, 2018.

[34] D. Kuhn, P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research*, pages 130–169, 2019.

[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[36] B. Li and Y. Li. Why clean generalization and robust overfitting both happen in adversarial training. *arXiv:2306.01271*, 2023.

[37] L. Li and M. Spratling. Understanding and combating robust overfitting via input loss landscape analysis and regularization. *Pattern Recognition*, 136:1–11, 2023.

[38] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[39] K. N. Markelle Kelly, Rachel Longjohn. The UCI Machine Learning Repository. `https://archive.ics.uci.edu`.

[40] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):1–52, 2018.

[41] MOSEK ApS. Modeling cookbook. `https://docs.mosek.com/MOSEKModelingCookbook-letter.pdf`, 2023.

[42] MOSEK ApS. MOSEK Optimizer API for Julia 10.1.12. `https://docs.mosek.com/latest/juliaapi/index.html`, 2023.

[43] K. P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[44] Y. Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018.

[45] T. Pang, M. Lin, X. Yang, J. Zhu, and S. Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[46] N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics*, 2016.

[47] H. Phan, T. Le, T. Phung, A. T. Bui, N. Ho, and D. Phung. Global-local regularization via distributional robustness. In *International Conference on Artificial Intelligence and Statistics*, volume 206, pages 7644–7664, 2023.

[48] V. K. Potluru, D. Borrajo, A. Coletta, N. Dalmasso, Y. El-Laham, E. Fons, M. Ghassemi, S. Gopalakrishnan, V. Gosai, E. Kreačić, G. Mani, S. Obitayo, D. Paramanand, N. Raman, M. Solonin, S. Sood, S. Vyetrenko, H. Zhu, M. Veloso, and T. Balch. Synthetic data applications in finance, 2024.

[49] A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang. Adversarial training can hurt generalization. *arXiv:1906.06032*, 2019.

[50] C. Regniez, G. Gidel, and H. Berard. A distributional robustness perspective on adversarial training with the ∞-Wasserstein distance, 2022.

[51] L. Rice, E. Wong, and Z. Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[52] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

[53] E. Roos and D. den Hertog. Reducing conservatism in robust optimization. *INFORMS Journal on Computing*, 32(4):1109–1127, 2020.

[54] V. Sehwag, S. Mahloujifar, T. Handina, S. Dai, C. Xiang, M. Chiang, and P. Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[55] A. Selvi, M. R. Belbasi, M. Haugh, and W. Wiesemann. Wasserstein logistic regression with mixed features. *Advances in Neural Information Processing Systems*, 35, 2022.

[56] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

[57] A. Shafahi, P. Saadatpanah, C. Zhu, A. Ghiasi, C. Studer, D. W. Jacobs, and T. Goldstein. Adversarially robust transfer learning. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

[58] S. Shafieezadeh-Abadeh, L. Aolaritei, F. Dörfler, and D. Kuhn. New perspectives on regularization and computation in optimal transport-based distributionally robust optimization. *arXiv:2303.03900*, 2023.

[59] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[60] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[61] A. Shapiro. On duality theory of conic linear problems. *Nonconvex Optimization and its Applications*, 57:135–155, 2001.

[62] A. Sinha, H. Namkoong, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[63] J. E. Smith and R. L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

[64] C. Song, K. He, L. Wang, and J. E. Hopcroft. Improving the generalization of adversarial training with domain adaptation. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[65] M. Staib and S. Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 3, page 4, 2017.

[66] A. Subramanyam, C. E. Gounaris, and W. Wiesemann. K-adaptability in two-stage mixed-integer robust optimization. *Mathematical Programming Computation*, 12:193–224, 2020.

[67] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

[68] B. Taskesen, M.-C. Yue, J. Blanchet, D. Kuhn, and V. A. Nguyen. Sequential domain adaptation by synthesizing distributionally robust experts. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[69] J. F. Toland. Duality in nonconvex optimization. *Journal of Mathematical Analysis and Applications*, 66(2):399–415, 1978.

[70] J. Uesato, B. O'donoghue, P. Kohli, and A. Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[71] J. Ullman and S. Vadhan. PCPs and the hardness of generating synthetic data. *Journal of Cryptology*, 33(4):2078–2112, 2020.

[72] V. Vapnik. *The nature of statistical learning theory*. Springer, 1999.

[73] C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[74] D. Wu, S.-T. Xia, and Y. Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[75] Y. Xing, Q. Song, and G. Cheng. Unlabeled data help: Minimax analysis and adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, volume 151, pages 136–168, 2022.

[76] Y. Xing, Q. Song, and G. Cheng. Why do artificially generated data help adversarial robustness. *Advances in Neural Information Processing Systems*, 35:954–966, 2022.

[77] İ. Yanıkoğlu, B. L. Gorissen, and D. den Hertog. A survey of adjustable robust optimization. *European Journal of Operational Research*, 277(3):799–813, 2019.

[78] J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, and L. Kong. Zerogen: Efficient zero-shot learning via dataset generation. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

[79] C. Yu, B. Han, L. Shen, J. Yu, C. Gong, M. Gong, and T. Liu. Understanding robust overfitting of adversarial training and beyond. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[80] M. Yue, D. Kuhn, and W. Wiesemann. On linear optimization over Wasserstein balls. *Mathematical Programming*, 195(1-2):1107–1122, 2022.

[81] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[82] Y. Zhang, L. N. Steimle, and B. T. Denton. Data-driven distributionally robust optimization: Intersecting ambiguity sets, performance analysis and tractability. *Optimization Online 22567*, 2023.

[83] E. Zhong, W. Fan, Q. Yang, O. Verscheure, and J. Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *Machine Learning and Knowledge Discovery in Databases*, volume 6323, 2010.

A Notation

Throughout the paper, bold lower case letters denote vectors, while standard lower case letters are reserved for scalars. A generic data instance is modeled as $(\boldsymbol{x}, y) \in \Xi := \mathbb{R}^n \times \{-1, +1\}$. For any $p > 0$, $\|\boldsymbol{x}\|_p$ denotes the rational norm $\left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ and $\|\boldsymbol{x}\|_{p^\star}$ is its dual norm where $\frac{1}{p} + \frac{1}{p^\star} = 1$ with the convention of $1/1 + 1/\infty = 1$. The set of probability distributions supported on Ξ is denoted by $\mathcal{P}(\Xi)$. The Dirac measure supported on $\boldsymbol{\xi}$ is denoted by $\delta_{\boldsymbol{\xi}}$. The logloss is defined as $\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y) = \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x}))$ and its associated univariate loss is $L(z) = \log(1 + \exp(-z))$ so that $L(y \cdot \boldsymbol{\beta}^\top \boldsymbol{x}) = \ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$. The exponential cone is denoted by $\mathcal{K}_{\exp} = \mathrm{cl}(\{\boldsymbol{\omega} \in \mathbb{R}^3 : \omega_1 \geq \omega_2 \cdot \exp(\omega_3/\omega_2),\ \omega_1 > 0,\ \omega_2 > 0\})$ where cl is the closure operator. The Lipschitz modulus of a univariate function f is defined as $\mathrm{Lip}(f) := \sup_{z,z' \in \mathbb{R}} \left\{ \frac{|f(z) - f(z')|}{|z - z'|} : z \neq z \right\}$ whereas its effective domain is $\mathrm{dom}(f) = \{z : f(z) < +\infty\}$. For a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, its convex conjugate is $f^*(\boldsymbol{z}) = \sup_{\boldsymbol{x} \in \mathbb{R}^n} \boldsymbol{z}^\top \boldsymbol{x} - f(\boldsymbol{x})$. We reserve $\alpha \geq 0$ for the radii of the norms of adversarial attacks on the features and $\varepsilon \geq 0$ for the radii of distributional ambiguity sets.

B Proofs

B.1 Proof of Proposition 4.1

For any $\boldsymbol{\beta} \in \mathbb{R}^n$, with standard robust optimization arguments [4, 8], we can show that

$$
\begin{aligned}
& \sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p \leq \alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x} + \boldsymbol{z}, y)\} \\
\iff & \sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p \leq \alpha} \{\log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top (\boldsymbol{x} + \boldsymbol{z})))\} \\
\iff & \log\left(1 + \exp\left(\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p \leq \alpha} \{-y \cdot \boldsymbol{\beta}^\top (\boldsymbol{x} + \boldsymbol{z})\}\right)\right) \\
\iff & \log\left(1 + \exp\left(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p \leq 1} \{-y \cdot \boldsymbol{\beta}^\top \boldsymbol{z}\}\right)\right) \\
\iff & \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|-y \cdot \boldsymbol{\beta}\|_{p^\star})) \\
\iff & \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})),
\end{aligned}
$$

where the first step follows from the definition of logloss, the second step follows from the fact that $\log$ and $\exp$ are increasing functions, the third step takes the constant terms out of the $\sup$ problem and exploits the fact that the optimal solution of maximizing a linear function will be at an extreme point of the ℓ_p ball, the fourth step uses the definition of dual norm, and finally the redundant $-y \in \{-1, +1\}$ is omitted from the dual norm. We can therefore define the adversarial loss $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y) := \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$ where α models the strength of the adversary, $\boldsymbol{\beta}$ is the decision vector, and $(\boldsymbol{x}, y)$ is an instance. Replacing $\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p \leq \alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x} + \boldsymbol{z}, y)\}$ in ARO with $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$ concludes the equivalence.

Furthermore, to see $\mathrm{Lip}(L^\alpha) = 1$, firstly note that since $L^\alpha(z) = \log(1 + \exp(-z + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$ is differentiable everywhere in z and its gradient $L^{\alpha\prime}$ is bounded everywhere, we have that $\mathrm{Lip}(L^\alpha)$ is equal to $\sup_{z \in \mathbb{R}} \{|L^{\alpha\prime}(z)|\}$. We thus have

$$
L^{\alpha\prime}(z) = \frac{-\exp(-z + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})}{1 + \exp(-z + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})} = \frac{-1}{1 + \exp(z - \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})} \in (-1, 0)
$$

and $|L^{\alpha\prime}(z)| = [1 + \exp(z - \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})]^{-1} \longrightarrow 1$ as $z \longrightarrow -\infty$. □

B.2 Proof of Corollary 4.3

Proposition 4.1 lets us represent DR-ARO as the DR counterpart of empirical minimization of $\ell^\alpha_{\boldsymbol{\beta}}$:

$$
\begin{array}{ll}
\underset{\boldsymbol{\beta}}{\text{minimize}} & \sup\limits_{\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)} \ \mathbb{E}_{\mathbb{Q}}\left[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)\right] \\
\text{subject to} & \boldsymbol{\beta} \in \mathbb{R}^n.
\end{array}
\tag{3}
$$

Since the univariate loss $L^\alpha(z) := \log(1 + \exp(-z + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$ satisfying the identity $L^\alpha(\langle y \cdot \boldsymbol{x}, \boldsymbol{\beta}\rangle) = \ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$ is Lipschitz continuous (*cf.* Proposition 4.1), Theorem 14 *(ii)* of [59] is immediately applicable. We can therefore rewrite (3) as:

$$\begin{array}{lll}
\underset{\boldsymbol{\beta},\, \lambda,\, \boldsymbol{s}}{\text{minimize}} & \lambda \cdot \varepsilon + \dfrac{1}{N} \displaystyle\sum_{i \in [N]} s_i & \\
\text{subject to} & L^\alpha(\langle y^i \cdot \boldsymbol{x}, \boldsymbol{\beta}\rangle) \leq s_i & \forall i \in [N] \\
& L^\alpha(\langle -y^i \cdot \boldsymbol{x}, \boldsymbol{\beta}\rangle) - \lambda \cdot \kappa \leq s_i & \forall i \in [N] \\
& \mathrm{Lip}(L^\alpha) \cdot \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda & \\
& \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N. &
\end{array}$$

Replacing $\mathrm{Lip}(L^\alpha) = 1$ and substituting the definition of L^α concludes the proof. □

B.3 Proof of Proposition 5.2

We prove Proposition 5.2 by constructing the optimization problem in its statement. We will thus dualize the inner sup problem of Inter-ARO for fixed $\boldsymbol{\beta}$. To this end, we present a sequence of reformulations to the inner problem and then exploit strong semi-infinite duality.

By interchanging $\boldsymbol{\xi} = (\boldsymbol{x}, y)$, we first rewrite the inner problem as

$$\begin{array}{lll}
\underset{\mathbb{Q}, \Pi, \widehat{\Pi}}{\text{maximize}} & \displaystyle\int_{\boldsymbol{\xi} \in \Xi} \ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{\xi}) \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) & \\
\text{subject to} & \displaystyle\int_{\boldsymbol{\xi}, \boldsymbol{\xi'} \in \Xi^2} d(\boldsymbol{\xi}, \boldsymbol{\xi'}) \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \leq \varepsilon & \\
& \displaystyle\int_{\boldsymbol{\xi} \in \Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) = \mathbb{P}_N(\mathrm{d}\boldsymbol{\xi'}) & \forall \boldsymbol{\xi'} \in \Xi \\
& \displaystyle\int_{\boldsymbol{\xi'} \in \Xi} \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) = \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) & \forall \boldsymbol{\xi} \in \Xi \\
& \displaystyle\int_{\boldsymbol{\xi}, \boldsymbol{\xi'} \in \Xi^2} d(\boldsymbol{\xi}, \boldsymbol{\xi'}) \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) \leq \widehat{\varepsilon} & \\
& \displaystyle\int_{\boldsymbol{\xi} \in \Xi} \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) = \widehat{\mathbb{P}}_{\widehat{N}}(\mathrm{d}\boldsymbol{\xi'}) & \forall \boldsymbol{\xi'} \in \Xi \\
& \displaystyle\int_{\boldsymbol{\xi'} \in \Xi} \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi'}) = \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}) & \forall \boldsymbol{\xi} \in \Xi \\
& \mathbb{Q} \in \mathcal{P}(\Xi),\ \Pi \in \mathcal{P}(\Xi^2),\ \widehat{\Pi} \in \mathcal{P}(\Xi^2). &
\end{array}$$

Here, the first three constraints specify that $\mathbb{Q}$ and $\mathbb{P}_N$ have a Wasserstein distance bounded by ε from each other, modeled via their coupling Π. The latter three constraints similarly specify that $\mathbb{Q}$ and $\widehat{\mathbb{P}}_{\widehat{N}}$ are at most $\widehat{\varepsilon}$ away from each other, modeled via their coupling $\widehat{\Pi}$. As $\mathbb{Q}$ lies in the intersection of two Wasserstein balls in Inter-ARO, the marginal $\mathbb{Q}$ is shared between Π and $\widehat{\Pi}$. We can now

substitute the third constraint into the objective and the last constraint and obtain:

$$
\begin{array}{ll}
\underset{\Pi,\widehat{\Pi}}{\text{maximize}} & \displaystyle\int_{\boldsymbol{\xi}\in\Xi} \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{\xi}) \int_{\boldsymbol{\xi}'\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') \\
\text{subject to} & \displaystyle\int_{\boldsymbol{\xi},\boldsymbol{\xi}'\in\Xi^2} d(\boldsymbol{\xi},\boldsymbol{\xi}')\Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') \leq \varepsilon \\
& \displaystyle\int_{\boldsymbol{\xi}\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') = \mathbb{P}_N(\mathrm{d}\boldsymbol{\xi}') \qquad\qquad \forall \boldsymbol{\xi}'\in\Xi \\
& \displaystyle\int_{\boldsymbol{\xi},\boldsymbol{\xi}'\in\Xi^2} d(\boldsymbol{\xi},\boldsymbol{\xi}')\widehat{\Pi}(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') \leq \widehat{\varepsilon} \\
& \displaystyle\int_{\boldsymbol{\xi}\in\Xi} \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') = \widehat{\mathbb{P}}_{\widehat{N}}(\mathrm{d}\boldsymbol{\xi}') \qquad\qquad \forall \boldsymbol{\xi}'\in\Xi \\
& \displaystyle\int_{\boldsymbol{\xi}'\in\Xi} \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') = \int_{\boldsymbol{\xi}'\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') \quad \forall \boldsymbol{\xi}\in\Xi \\
& \Pi \in \mathcal{P}(\Xi^2),\ \widehat{\Pi} \in \mathcal{P}(\Xi^2).
\end{array}
$$

Denoting by $\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) := \Pi(\mathrm{d}\boldsymbol{\xi} \mid \boldsymbol{\xi}^i)$ the conditional distribution of Π upon the realization of $\boldsymbol{\xi}' = \boldsymbol{\xi}^i$ and exploiting the fact that $\mathbb{P}_N$ is a discrete distribution supported on the N data points $\{\boldsymbol{\xi}^i\}_{i\in[N]}$, we can use the marginalized representation $\Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{\xi}^i}(\mathrm{d}\boldsymbol{\xi}')\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi})$. Similarly, we can introduce $\widehat{\mathbb{Q}}^i(\mathrm{d}\boldsymbol{\xi}) := \widehat{\Pi}(\mathrm{d}\boldsymbol{\xi} \mid \widehat{\boldsymbol{\xi}}^i)$ for $\{\widehat{\boldsymbol{\xi}}^i\}_{i\in[\widehat{N}]}$ to exploit the marginalized representation $\widehat{\Pi}(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') = \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \delta_{\widehat{\boldsymbol{\xi}}^j}(\mathrm{d}\boldsymbol{\xi}')\widehat{\mathbb{Q}}^j(\mathrm{d}\boldsymbol{\xi})$. By using this marginalization representation, we can use the following simplification for the objective function:

$$
\int_{\boldsymbol{\xi}\in\Xi} \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{\xi}) \int_{\boldsymbol{\xi}'\in\Xi} \Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi}') \;=\; \frac{1}{N}\sum_{i=1}^N \int_{\boldsymbol{\xi}\in\Xi} \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{\xi}) \int_{\boldsymbol{\xi}'\in\Xi} \delta_{\boldsymbol{\xi}^i}(\mathrm{d}\boldsymbol{\xi}')\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) \;=\; \frac{1}{N}\sum_{i=1}^N \int_{\boldsymbol{\xi}\in\Xi} \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{\xi})\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}).
$$

Applying analogous reformulations to the constraints leads to the following reformulation of the inner sup problem of Inter-ARO:

$$
\begin{array}{ll}
\underset{\mathbb{Q},\widehat{\mathbb{Q}}}{\text{maximize}} & \displaystyle\frac{1}{N}\sum_{i=1}^N \int_{\boldsymbol{\xi}\in\Xi} \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{\xi})\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) \\
\text{subject to} & \displaystyle\frac{1}{N}\sum_{i=1}^N \int_{\boldsymbol{\xi}\in\Xi} d(\boldsymbol{\xi},\boldsymbol{\xi}^i)\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) \leq \varepsilon \\
& \displaystyle\frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \int_{\boldsymbol{\xi}\in\Xi} d(\boldsymbol{\xi},\widehat{\boldsymbol{\xi}}^j)\widehat{\mathbb{Q}}^j(\mathrm{d}\boldsymbol{\xi}) \leq \widehat{\varepsilon} \\
& \displaystyle\frac{1}{N}\sum_{i=1}^N \mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) = \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \widehat{\mathbb{Q}}^j(\mathrm{d}\boldsymbol{\xi}) \quad \forall \boldsymbol{\xi}\in\Xi \\
& \mathbb{Q}^i \in \mathcal{P}(\Xi),\ \widehat{\mathbb{Q}}^j \in \mathcal{P}(\Xi) \qquad\qquad \forall i\in[N],\ \forall j\in[\widehat{N}].
\end{array}
$$

We now decompose each $\mathbb{Q}^i$ into two measures corresponding to $y = \pm 1$, so that $\mathbb{Q}^i(\mathrm{d}(\boldsymbol{x},y)) = \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x})$ for $y = +1$ and $\mathbb{Q}^i(\mathrm{d}(\boldsymbol{x},y)) = \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})$ for $y = -1$. We similarly represent each $\widehat{\mathbb{Q}}^j$ via $\widehat{\mathbb{Q}}^j_{+1}$ and $\widehat{\mathbb{Q}}^j_{-1}$ depending on y. Note that these new measures are not probability measures as they do

not integrate to 1, but non-negative measures supported on $\mathbb{R}^n$ (denoted $\in \mathcal{P}_+(\mathbb{R}^n)$). We get:

$$
\begin{aligned}
\underset{\mathbb{Q}_{\pm 1}, \widehat{\mathbb{Q}}_{\pm 1}}{\text{maximize}} \quad & \frac{1}{N} \sum_{i=1}^{N} \int_{\boldsymbol{x} \in \mathbb{R}^n} [\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, +1) \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, -1) \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})] \\
\text{subject to} \quad & \frac{1}{N} \sum_{i=1}^{N} \int_{\boldsymbol{x} \in \mathbb{R}^n} [d((\boldsymbol{x}, +1), \boldsymbol{\xi}^i) \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + d((\boldsymbol{x}, -1), \boldsymbol{\xi}^i) \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})] \leq \varepsilon \\
& \frac{1}{\widehat{N}} \sum_{j=1}^{\widehat{N}} \int_{\boldsymbol{x} \in \mathbb{R}^n} [d((\boldsymbol{x}, +1), \widehat{\boldsymbol{\xi}}^j) \widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + d((\boldsymbol{x}, -1), \widehat{\boldsymbol{\xi}}^j) \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x})] \leq \widehat{\varepsilon} \\
& \int_{\boldsymbol{x} \in \mathbb{R}^n} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) = 1 && \forall i \in [N] \\
& \int_{\boldsymbol{x} \in \mathbb{R}^n} \widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x}) = 1 && \forall j \in [\widehat{N}] \\
& \frac{1}{N} \sum_{i=1}^{N} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) = \frac{1}{\widehat{N}} \sum_{j=1}^{\widehat{N}} \widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) && \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \frac{1}{N} \sum_{i=1}^{N} \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) = \frac{1}{\widehat{N}} \sum_{j=1}^{\widehat{N}} \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x}) && \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \mathbb{Q}^i_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n),\ \widehat{\mathbb{Q}}^j_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n) && \forall i \in [N],\ j \in [\widehat{N}].
\end{aligned}
$$

Next, we explicitly write the definition of the metric $d(\cdot, \cdot)$ in the first two constraints as well as use auxiliary measures $\mathbb{A}_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n)$ to break down the last two equality constraints:

$$
\begin{aligned}
\underset{\mathbb{A}_{\pm 1},\mathbb{Q}_{\pm 1},\widehat{\mathbb{Q}}_{\pm 1}}{\text{maximize}} \quad & \frac{1}{N}\sum_{i=1}^{N}\int_{\boldsymbol{x}\in\mathbb{R}^n}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},+1)\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x})+\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},-1)\mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})] \\
\text{subject to} \quad & \frac{1}{N}\int_{\boldsymbol{x}\in\mathbb{R}^n}\Big[\kappa\cdot\sum_{i\in[N]:y^i=-1}\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x})+\kappa\cdot\sum_{i\in[N]:y^i=+1}\mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})+ \\
& \sum_{i=1}^{N}\|\boldsymbol{x}-\boldsymbol{x}^i\|_q\cdot[\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x})+\mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})]\Big]\leq\varepsilon \\
& \frac{1}{\widehat{N}}\int_{\boldsymbol{x}\in\mathbb{R}^n}\Big[\kappa\cdot\sum_{j\in[N]:\widehat{y}^j=-1}\widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x})+\kappa\cdot\sum_{j\in[N]:\widehat{y}^j=+1}\widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x})+ \\
& \sum_{j=1}^{\widehat{N}}\|\boldsymbol{x}-\widehat{\boldsymbol{x}}^j\|_q\cdot[\widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x})+\widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x})]\Big]\leq\widehat{\varepsilon} \\
& \int_{\boldsymbol{x}\in\mathbb{R}^n}\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x})+\mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})=1 && \forall i\in[N] \\
& \int_{\boldsymbol{x}\in\mathbb{R}^n}\widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x})+\widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x})=1 && \forall j\in[\widehat{N}] \\
& \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x})=\mathbb{A}_{+1}(\mathrm{d}\boldsymbol{x}) && \forall\boldsymbol{x}\in\mathbb{R}^n \\
& \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}}\widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x})=\mathbb{A}_{+1}(\mathrm{d}\boldsymbol{x}) && \forall\boldsymbol{x}\in\mathbb{R}^n \\
& \frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})=\mathbb{A}_{-1}(\mathrm{d}\boldsymbol{x}) && \forall\boldsymbol{x}\in\mathbb{R}^n \\
& \frac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}}\widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x})=\mathbb{A}_{-1}(\mathrm{d}\boldsymbol{x}) && \forall\boldsymbol{x}\in\mathbb{R}^n \\
& \mathbb{A}_{\pm 1}\in\mathcal{P}_+(\mathbb{R}^n),\ \mathbb{Q}^i_{\pm 1}\in\mathcal{P}_+(\mathbb{R}^n),\ \widehat{\mathbb{Q}}^j_{\pm 1}\in\mathcal{P}_+(\mathbb{R}^n) && \forall i\in[N],\ j\in[\widehat{N}].
\end{aligned}
$$

The following semi-infinite optimization problem, obtained by standard algebraic duality, is a strong dual to the above problem since $\varepsilon,\widehat{\varepsilon}>0$ [61].

$$
\begin{aligned}
\underset{\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}},p_{\pm1},\widehat{p}_{\pm1}}{\text{minimize}} \quad & \frac{1}{N}\left[N\varepsilon\lambda + \widehat{N}\widehat{\varepsilon}\widehat{\lambda} + \sum_{i=1}^{N} s_i + \sum_{j=1}^{\widehat{N}} \widehat{s}_j\right] \\
\text{subject to} \quad & \kappa\frac{1-y^i}{2}\lambda + \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q + s_i + \frac{p_{+1}(x)}{N} \geq \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, +1) \quad \forall i \in [N],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \kappa\frac{1-\widehat{y}^j}{2}\widehat{\lambda} + \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q + \widehat{s}_j + \frac{\widehat{p}_{+1}(x)}{\widehat{N}} \geq 0 \qquad \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \kappa\frac{1+y^i}{2}\lambda + \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q + s_i + \frac{p_{-1}(x)}{N} \geq \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, -1) \quad \forall i \in [N],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \kappa\frac{1+\widehat{y}^j}{2}\widehat{\lambda} + \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q + \widehat{s}_j + \frac{\widehat{p}_{-1}(x)}{\widehat{N}} \geq 0 \qquad \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
& p_{+1}(\boldsymbol{x}) + \widehat{p}_{+1}(\boldsymbol{x}) \leq 0 \\
& p_{-1}(\boldsymbol{x}) + \widehat{p}_{-1}(\boldsymbol{x}) \leq 0 \\
& \lambda \in \mathbb{R}_+,\ \widehat{\lambda} \in \mathbb{R}+,\ \boldsymbol{s} \in \mathbb{R}^N,\ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}} \\
& p_{\pm1} : \mathbb{R}^n \mapsto \mathbb{R},\ \widehat{p}_{\pm1} : \mathbb{R}^n \mapsto \mathbb{R}.
\end{aligned}
$$

To eliminate the (function) variables p_{+1} and $\widehat{p}_{+1}$, we first summarize the constraints they appear

$$
\begin{cases}
p_{+1}(\boldsymbol{x}) \geq N \cdot \left[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, +1) - s_i - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \kappa\dfrac{1-y^i}{2}\lambda\right] & \forall i \in [N],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
\widehat{p}_{+1}(\boldsymbol{x}) \geq \widehat{N} \cdot \left[-\widehat{s}_j - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q - \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda}\right] & \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
p_{+1}(\boldsymbol{x}) + \widehat{p}_{+1}(\boldsymbol{x}) \leq 0 & \forall \boldsymbol{x} \in \mathbb{R}^n,
\end{cases}
$$

and notice that this system is equivalent to the epigraph-based reformulation of the following constraint

$$
\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, +1) - s_i - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \kappa\frac{1-y^i}{2}\lambda + \frac{\widehat{N}}{N} \cdot \left[-\widehat{s}_j - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q - \kappa\frac{1-\widehat{y}^j}{2}\widehat{\lambda}\right] \leq 0 \qquad \forall i \in [N],\ \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n.
$$

We can therefore eliminate p_{+1} and $\widehat{p}_{+1}$. We can also eliminate p_{-1} and $\widehat{p}_{-1}$ since we similarly have:

$$
\begin{aligned}
&\begin{cases}
p_{-1}(\boldsymbol{x}) \geq N \cdot \left[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, -1) - s_i - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \kappa\dfrac{1+y^i}{2}\lambda\right] & \forall i \in [N],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
\widehat{p}_{-1}(\boldsymbol{x}) \geq \widehat{N} \cdot \left[-\widehat{s}_j - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q - \kappa\dfrac{1+\widehat{y}^j}{2}\widehat{\lambda}\right] & \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n \\
p_{-1}(\boldsymbol{x}) + \widehat{p}_{-1}(\boldsymbol{x}) \leq 0 & \forall \boldsymbol{x} \in \mathbb{R}^n
\end{cases} \\
\iff & \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, -1) - s_i - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \kappa\frac{1+y^i}{2}\lambda + \frac{\widehat{N}}{N} \cdot \left[-\widehat{s}_j - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q - \kappa\frac{1+\widehat{y}^j}{2}\widehat{\lambda}\right] \leq 0 \\
& \forall i \in [N],\ \forall j \in [\widehat{N}],\ \forall \boldsymbol{x} \in \mathbb{R}^n.
\end{aligned}
$$

This trick of eliminating $p_{\pm1},\ \widehat{p}_{\pm1}$ is due to the auxiliary distributions $\mathbb{A}_{\pm1}$ that we introduced; without them, the dual problem is substantially harder to work with. We therefore obtain the

following reformulation of the dual problem

$$
\begin{aligned}
\underset{\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} \quad & \frac{1}{N}\left[N\varepsilon\lambda + \widehat{N}\widehat{\varepsilon}\widehat{\lambda} + \sum_{i=1}^{N} s_i + \sum_{j=1}^{\widehat{N}} \widehat{s}_j\right] \\
\text{subject to} \quad & \sup_{\boldsymbol{x}\in\mathbb{R}^n}\{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},+1) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \frac{\widehat{N}}{N}\widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq \\
& \qquad s_i + \kappa\frac{1-y^i}{2}\lambda + \frac{\widehat{N}}{N}\cdot\left[\widehat{s}_j + \kappa\frac{1-\widehat{y}^j}{2}\widehat{\lambda}\right] \qquad \forall i \in [N],\ \forall j \in [\widehat{N}] \\
& \sup_{\boldsymbol{x}\in\mathbb{R}^n}\{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},-1) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \frac{\widehat{N}}{N}\widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq \\
& \qquad s_i + \kappa\frac{1+y^i}{2}\lambda + \frac{\widehat{N}}{N}\cdot\left[\widehat{s}_j + \kappa\frac{1+\widehat{y}^j}{2}\widehat{\lambda}\right] \qquad \forall i \in [N],\ \forall j \in [\widehat{N}] \\
& \lambda \geq 0,\ \widehat{\lambda} \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N_+,\ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+
\end{aligned}
$$

where we replaced the $\forall \boldsymbol{x} \in \mathbb{R}^n$ with the worst case realizations by taking the suprema of the constraints over $\boldsymbol{x}$. We also added non-negativity on the definition of $\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$ which is without loss of generality since this is implied by the first two constraints, which is due to the fact that in the primal reformulation the "integrates to 1" constraints (whose associated dual variables are $\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$) can be written as

$$
\begin{aligned}
\int_{\boldsymbol{x}\in\mathbb{R}^n} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) \leq 1 \quad \forall i \in [N] \\
\int_{\boldsymbol{x}\in\mathbb{R}^n} \widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x}) \leq 1 \quad \forall j \in [\widehat{N}]
\end{aligned}
$$

due to the objective pressure. Relabeling $\frac{\widehat{N}}{N}\widehat{\lambda}$ as $\widehat{\lambda}$ and $\frac{\widehat{N}}{N}\widehat{s}_j$ as $\widehat{s}_j$ simplifies the problem to:

$$
\begin{aligned}
\underset{\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} \quad & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \frac{1}{N}\sum_{i=1}^{N} s_i + \frac{1}{\widehat{N}}\sum_{i=1}^{\widehat{N}} \widehat{s}_j \\
\text{subject to} \quad & \sup_{\boldsymbol{x}\in\mathbb{R}^n}\{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},+1) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq \\
& \qquad s_i + \kappa\frac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa\frac{1-\widehat{y}^j}{2}\widehat{\lambda} \qquad \forall i \in [N],\ \forall j \in [\widehat{N}] \\
& \sup_{\boldsymbol{x}\in\mathbb{R}^n}\{\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},-1) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \leq \\
& \qquad s_i + \kappa\frac{1+y^i}{2}\lambda + \widehat{s}_j + \kappa\frac{1+\widehat{y}^j}{2}\widehat{\lambda} \qquad \forall i \in [N],\ \forall j \in [\widehat{N}] \\
& \lambda \geq 0,\ \widehat{\lambda} \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N_+,\ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+.
\end{aligned}
$$

Combining all the sup constraints with the help of an an auxiliary parameter $l \in \{-1, 1\}$ and replacing this problem with the inner problem of Inter-ARO concludes the proof. □

B.4 Proof of Proposition 5.3

We first present a technical lemma that will allow us to rewrite a specific type of difference of convex functions (DC) maximization problem that appears in the constraints of Inter-ARO. Rewriting such DC maximization problems is one of the key steps in reformulating Wasserstein DRO problems, and our lemma is inspired from [59, Lemma 47], [58, Theorem 3.8], and [3, Lemma 1] who reformulate maximizing the difference of a convex function and a norm. Our DRO problem Inter-ARO, however, comprises two ambiguity sets, hence the DC term that we investigate will be the difference between a convex function and the sum of *two norms*. This requires a new analysis and we will see that Inter-ARO is NP-hard due to this additional difficulty.

Lemma B.1. *Suppose that $L : \mathbb{R} \mapsto \mathbb{R}$ is a closed convex function, and $\|\cdot\|_q$ is a norm. For vectors $\boldsymbol{\omega}, \boldsymbol{a}, \widehat{\boldsymbol{a}} \in \mathbb{R}^n$ and scalars $\lambda, \widehat{\lambda} > 0$, we have:*

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda\|\boldsymbol{a} - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{a}} - \boldsymbol{x}\|_q\} \\
&= \sup_{\theta\in\mathrm{dom}(L^*)} \; - L^*(\theta) + \theta \cdot \boldsymbol{\omega}^\top \boldsymbol{a} + \theta \cdot \inf_{\boldsymbol{z}\in\mathbb{R}^n} \{\boldsymbol{z}^\top(\widehat{\boldsymbol{a}} - \boldsymbol{a}) \;:\; |\theta| \cdot \|\boldsymbol{\omega} - \boldsymbol{z}\|_{q^\star} \le \lambda,\; |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \le \widehat{\lambda}\}
\end{aligned}$$

Proof. We denote by $f_{\boldsymbol{\omega}}(\boldsymbol{x}) = \boldsymbol{\omega}^\top \boldsymbol{x}$ and by g the convex function $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) + g_2(\boldsymbol{x})$ where $g_1(\boldsymbol{x}) := \lambda\|\boldsymbol{a} - \boldsymbol{x}\|_q$ and $g_2(\boldsymbol{x}) := \widehat{\lambda}\|\widehat{\boldsymbol{a}} - \boldsymbol{x}\|_q$, and reformulate the sup problem as

$$\sup_{\boldsymbol{x}\in\mathbb{R}^n} L(\boldsymbol{\omega}^\top \boldsymbol{x}) - g(\boldsymbol{x}) \;=\; \sup_{\boldsymbol{x}\in\mathbb{R}^n} (L \circ f_{\boldsymbol{\omega}})(\boldsymbol{x}) - g(\boldsymbol{x}) \;=\; \sup_{\boldsymbol{z}\in\mathbb{R}^n} g^*(\boldsymbol{z}) - (L \circ f_{\boldsymbol{\omega}})^*(\boldsymbol{z}),$$

where the first identity follows from the definition of composition and the second identity employs Toland's duality [69] to rewrite difference of convex functions optimization.

By using infimal convolutions [52, Theorem 16.4], we can reformulate g^*:

$$\begin{aligned}
g^*(\boldsymbol{z}) &= \inf_{\boldsymbol{z}_1,\boldsymbol{z}_2} \{g_1^*(\boldsymbol{z}_1) + g_2^*(\boldsymbol{z}_2) \;:\; \boldsymbol{z}_1 + \boldsymbol{z}_2 = \boldsymbol{z}\} \\
&= \inf_{\boldsymbol{z}_1,\boldsymbol{z}_2} \{\boldsymbol{z}_1^\top \boldsymbol{a} + \boldsymbol{z}_2^\top \widehat{\boldsymbol{a}} \;:\; \boldsymbol{z}_1 + \boldsymbol{z}_2 = \boldsymbol{z},\; \|\boldsymbol{z}_1\|_{q^\star} \le \lambda,\; \|\boldsymbol{z}_2\|_{q^\star} \le \widehat{\lambda}\},
\end{aligned}$$

where the second step uses the definitions of $g_1^*(\boldsymbol{z}_1)$ and $g_2^*(\boldsymbol{z}_2)$. Moreover, we show

$$\begin{aligned}
(L \circ f_{\boldsymbol{\omega}})^*(\boldsymbol{z}) &= \sup_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{z}^\top \boldsymbol{x} - L(\boldsymbol{\omega}^\top \boldsymbol{x}) \\
&= \sup_{t\in\mathbb{R},\; \boldsymbol{x}\in\mathbb{R}^n} \{\boldsymbol{z}^\top \boldsymbol{x} - L(t) \;:\; t = \boldsymbol{\omega}^\top \boldsymbol{x}\} \\
&= \inf_{\theta\in\mathbb{R}} \sup_{t\in\mathbb{R},\; \boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{z}^\top \boldsymbol{x} - L(t) - \theta \cdot (\boldsymbol{\omega}^\top \boldsymbol{x} - t) \\
&= \inf_{\theta\in\mathbb{R}} \sup_{t\in\mathbb{R}} \sup_{\boldsymbol{x}\in\mathbb{R}^n} (\boldsymbol{z} - \theta \cdot \boldsymbol{\omega})^\top \boldsymbol{x} - L(t) + \theta \cdot t \\
&= \inf_{\theta\in\mathbb{R}} \sup_{t\in\mathbb{R}} \begin{cases} -L(t) + \theta \cdot t & \text{if } \theta \cdot \boldsymbol{\omega} = \boldsymbol{z} \\ +\infty & \text{otherwise.} \end{cases} \\
&= \inf_{\theta\in\mathbb{R}} \begin{cases} L^*(\theta) & \text{if } \theta \cdot \boldsymbol{\omega} = \boldsymbol{z} \\ +\infty & \text{otherwise.} \end{cases} \\
&= \inf_{\theta\in\mathrm{dom}(L^*)} \{L^*(\theta) \;:\; \theta \cdot \boldsymbol{\omega} = \boldsymbol{z}\},
\end{aligned}$$

where the first identity follows from the definition of the convex conjugate, the second identity introduces an additional variable to make this an equality-constrained optimization problem, the third identity takes the Lagrange dual (which is a strong dual since the problem maximizes a concave objective with a single equality constraint), the fourth identity rearranges the expressions, the fifth identity exploits unboundedness of $\boldsymbol{x}$, the sixth identity uses the definition of convex conjugates and the final identity replaces the feasible set $\theta \in \mathbb{R}$ with the domain of $L^\star$ without loss of generality as this is an inf problem.

Replacing the conjugates allows us to conclude that the maximization problem equals

$$
\begin{aligned}
& \sup_{\boldsymbol{z}\in\mathbb{R}^n} g^*(\boldsymbol{z}) + \sup_{\theta\in\mathrm{dom}(L^*)} \{-L^*(\theta) \,:\, \theta\cdot\boldsymbol{\omega} = \boldsymbol{z}\} \\
= & \sup_{\boldsymbol{z}\in\mathbb{R}^n,\, \theta\in\mathrm{dom}(L^*)} \{g^*(\boldsymbol{z}) - L^*(\theta) \,:\, \theta\cdot\boldsymbol{\omega} = \boldsymbol{z}\} \\
= & \sup_{\theta\in\mathrm{dom}(L^*)} g^*(\theta\cdot\boldsymbol{\omega}) - L^*(\theta) \\
= & \sup_{\theta\in\mathrm{dom}(L^*)} -L^*(\theta) + \inf_{\boldsymbol{z}_1,\boldsymbol{z}_2\in\mathbb{R}^n} \{\boldsymbol{z}_1^\top\boldsymbol{a} + \boldsymbol{z}_2^\top\widehat{\boldsymbol{a}} \,:\, \boldsymbol{z}_1+\boldsymbol{z}_2 = \theta\cdot\boldsymbol{\omega},\, \|\boldsymbol{z}_1\|_{q^\star}\le\lambda,\, \|\boldsymbol{z}_2\|_{q^\star}\le\widehat{\lambda}\} \\
= & \sup_{\theta\in\mathrm{dom}(L^*)} -L^*(\theta) + \theta\cdot\inf_{\boldsymbol{z}_1,\boldsymbol{z}_2\in\mathbb{R}^n} \{\boldsymbol{z}_1^\top\boldsymbol{a} + \boldsymbol{z}_2^\top\widehat{\boldsymbol{a}} \,:\, \boldsymbol{z}_1+\boldsymbol{z}_2 = \boldsymbol{\omega},\, |\theta|\cdot\|\boldsymbol{z}_1\|_{q^\star}\le\lambda,\, |\theta|\cdot\|\boldsymbol{z}_2\|_{q^\star}\le\widehat{\lambda}\} \\
= & \sup_{\theta\in\mathrm{dom}(L^*)} -L^*(\theta) + \theta\cdot\boldsymbol{\omega}^\top\boldsymbol{a} + \theta\cdot\inf_{\boldsymbol{z}\in\mathbb{R}^n} \{\boldsymbol{z}^\top(\widehat{\boldsymbol{a}}-\boldsymbol{a}) \,:\, |\theta|\cdot\|\boldsymbol{\omega}-\boldsymbol{z}\|_{q^\star}\le\lambda,\, |\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\le\widehat{\lambda}\}.
\end{aligned}
$$

Here, the first identity follows from writing the problem as a single maximization problem, the second identity follows from the equality constraint, the third identity follows from the definition of the conjugate g^*, the fourth identity is due to relabeling $\boldsymbol{z}_1 = \theta\cdot\boldsymbol{z}_1$ and $\boldsymbol{z}_2 = \theta\cdot\boldsymbol{z}_2$, and the fifth identity is due to a variable change ($\boldsymbol{z}_1 = \boldsymbol{\omega} - \boldsymbol{z}_2$ relabeled as $\boldsymbol{z}$). □

DC maximization terms similar to the one dealt by Lemma B.1 appear on the left-hand side of the constraints of Inter-ARO (*cf.* formulation in Proposition 5.2). These constraints would admit a tractable reformulation for the case without auxiliary data because the $\inf$ term in the reformulation presented in Lemma B.1 does not appear in such cases. To see this, eliminate the second norm (the one associated with auxiliary data) by taking $\widehat{\lambda} = 0$, which will cause the constraint $|\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\le\widehat{\lambda}$ to force $\boldsymbol{z} = \boldsymbol{0}$, and the alternative formulation will thus be:

$$
\begin{aligned}
& \begin{cases} \sup_{\theta\in\mathrm{dom}(L^*)} \{-L^*(\theta) + \theta\cdot\boldsymbol{\omega}^\top\boldsymbol{a}\} & \text{if } \sup_{\theta\in\mathrm{dom}(L^*)}\{|\theta|\}\cdot\|\boldsymbol{z}\|_{q^\star}\le\lambda \\ +\infty & \text{otherwise} \end{cases} \\
= & \begin{cases} L(\boldsymbol{\omega}^\top\boldsymbol{a}) & \text{if } \mathrm{Lip}(L)\cdot\|\boldsymbol{z}\|_{q^\star}\le\lambda \\ +\infty & \text{otherwise} \end{cases}
\end{aligned}
$$

where we used the fact that $L = L^{**}$ and $\sup_{\theta\in\mathrm{dom}(L)}|\theta| = \mathrm{Lip}(L)$ since L is closed convex [52, Corollary 13.3.3]. Hence, the DC maximization can be represented with a convex function with an additional convex inequality, making the constraints tractable for the case without auxiliary data. For the case with auxiliary data, however, the $\sup_\theta\inf_{\boldsymbol{z}}$ structure makes these constraints equivalent to two-stage robust constraints (with uncertain parameter θ and adjustable variable $\boldsymbol{z}$), bringing an adjustable robust optimization ([5, 77]) perspective to Inter-ARO. By using the univariate representation $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y) = L^\alpha(y\cdot\boldsymbol{\beta}^\top\boldsymbol{x})$, Inter-ARO can be written as

$$
\begin{aligned}
\underset{\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} \quad & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \frac{1}{N}\sum_{j=1}^{N} s_j + \frac{1}{\widehat{N}}\sum_{i=1}^{\widehat{N}} \widehat{s}_i \\
\text{subject to} \quad & \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L^\alpha(\boldsymbol{\beta}^\top\boldsymbol{x}) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \le \\
& \qquad s_i + \kappa\frac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa\frac{1-\widehat{y}^j}{2}\widehat{\lambda} \qquad\qquad \forall i\in[N],\, \forall j\in[\widehat{N}] \\
& \sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L^\alpha(-\boldsymbol{\beta}^\top\boldsymbol{x}) - \lambda\|\boldsymbol{x}^i - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^j - \boldsymbol{x}\|_q\} \le \\
& \qquad s_i + \kappa\frac{1+y^i}{2}\lambda + \widehat{s}_j + \kappa\frac{1+\widehat{y}^j}{2}\widehat{\lambda} \qquad\qquad \forall i\in[N],\, \forall j\in[\widehat{N}] \\
& \boldsymbol{\beta}\in\mathbb{R}^n,\, \lambda\ge 0,\, \widehat{\lambda}\ge 0,\, \boldsymbol{s}\in\mathbb{R}^N_+,\, \widehat{\boldsymbol{s}}\in\mathbb{R}^{\widehat{N}}_+,
\end{aligned}
$$

and applying Lemma B.1 to the left-hand side of the constraints gives:

$$
\begin{array}{ll}
\underset{\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} s_j + \dfrac{1}{\widehat{N}}\sum_{i=1}^{\widehat{N}} \widehat{s}_i \\
\text{subject to} & \displaystyle\sup_{\theta\in\mathrm{dom}(L^*)} -L^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \theta\cdot \inf_{\boldsymbol{z}\in\mathbb{R}^n}\{\boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \;:\; |\theta|\cdot\|\boldsymbol{\beta}-\boldsymbol{z}\|_{q^\star}\le\lambda,\; |\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\le\widehat{\lambda}\} \le \\
& \qquad s_i + \kappa\dfrac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda} \quad \forall i\in[N],\; \forall j\in[\widehat{N}] \\
& \displaystyle\sup_{\theta\in\mathrm{dom}(L^*)} -L^{\alpha*}(\theta) - \theta\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \theta\cdot \inf_{\boldsymbol{z}\in\mathbb{R}^n}\{\boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \;:\; |\theta|\cdot\|-\boldsymbol{\beta}-\boldsymbol{z}\|_{q^\star}\le\lambda,\; |\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\le\widehat{\lambda}\} \le \\
& \qquad s_i + \kappa\dfrac{1+y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1+\widehat{y}^j}{2}\widehat{\lambda} \quad \forall i\in[N],\; \forall j\in[\widehat{N}] \\
& \boldsymbol{\beta}\in\mathbb{R}^n,\; \lambda\ge 0,\; \widehat{\lambda}\ge 0,\; \boldsymbol{s}\in\mathbb{R}^N_+,\; \widehat{\boldsymbol{s}}\in\mathbb{R}^{\widehat{N}}_+.
\end{array}
\tag{4}
$$

Which, equivalently, can be written as the following problem with $2N\cdot\widehat{N}$ two-stage robust constraints:

$$
\begin{array}{ll}
\underset{\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} s_j + \dfrac{1}{\widehat{N}}\sum_{i=1}^{\widehat{N}} \widehat{s}_i \\
\text{subject to} & \left[\forall\theta\in\mathrm{dom}(L^*),\; \exists\boldsymbol{z}\in\mathbb{R}^n \;:\; \begin{cases} -L^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \le s_i + \kappa\dfrac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda} \\ |\theta|\cdot\|\boldsymbol{\beta}-\boldsymbol{z}\|_{q^\star}\le\lambda \\ |\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\le\widehat{\lambda} \end{cases}\right] \\
& \qquad\qquad \forall i\in[N],\; \forall j\in[\widehat{N}] \\
& \left[\forall\theta\in\mathrm{dom}(L^*),\; \exists\boldsymbol{z}\in\mathbb{R}^n \;:\; \begin{cases} -L^{\alpha*}(\theta) - \theta\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \le s_i + \kappa\dfrac{1+y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1+\widehat{y}^j}{2}\widehat{\lambda} \\ |\theta|\cdot\|-\boldsymbol{\beta}-\boldsymbol{z}\|_{q^\star}\le\lambda \\ |\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\le\widehat{\lambda} \end{cases}\right] \\
& \qquad\qquad \forall i\in[N],\; \forall j\in[\widehat{N}] \\
& \boldsymbol{\beta}\in\mathbb{R}^n,\; \lambda\ge 0,\; \widehat{\lambda}\ge 0,\; \boldsymbol{s}\in\mathbb{R}^N_+,\; \widehat{\boldsymbol{s}}\in\mathbb{R}^{\widehat{N}}_+.
\end{array}
\tag{Inter-adjustable}
$$

By using adjustable robust optimization theory, we show that this problem is NP-hard even in the simplest setting. To this end, take $N = \widehat{N} = 1$ as well as $\kappa = 0$; the formulation presented in Proposition 5.2 reduces to:

$$
\begin{array}{ll}
\underset{\boldsymbol{\beta},\lambda,\widehat{\lambda},s,\widehat{s}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + s + \widehat{s} \\
\text{subject to} & \displaystyle\sup_{\boldsymbol{x}\in\mathbb{R}^n}\{\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x},l) - \lambda\|\boldsymbol{x}^1-\boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^1-\boldsymbol{x}\|_q\} \le s_1 + \widehat{s}_1 \quad \forall l\in\{-1,1\} \\
& \boldsymbol{\beta}\in\mathbb{R}^n,\; \lambda\ge 0,\; \widehat{\lambda}\ge 0,\; s\ge 0,\; \widehat{s}\ge 0.
\end{array}
$$

The worst case realization of $l\in\{-1,1\}$ will always make $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x},l) = \log(1+\exp(-l\cdot\boldsymbol{\beta}^\top\boldsymbol{x} + \alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$ equal to $\varsigma^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}) = \log(1+\exp(|l\cdot\boldsymbol{\beta}^\top\boldsymbol{x}| + \alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$, where ς inherits similar properties from ℓ: it is convex in $\boldsymbol{\beta}$ and its univariate representation S^α has the same Lipschitz constant with L^α. We can thus represent the above problem as

$$
\begin{array}{ll}
\underset{\boldsymbol{\beta},\lambda,\widehat{\lambda},s,\widehat{s}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + s + \widehat{s} \\
\text{subject to} & \displaystyle\sup_{\boldsymbol{x}\in\mathbb{R}^n}\{S^\alpha(\boldsymbol{\beta}^\top\boldsymbol{x}) - \lambda\|\boldsymbol{x}^1-\boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^1-\boldsymbol{x}\|_q\} \le s + \widehat{s} \\
& \boldsymbol{\beta}\in\mathbb{R}^n,\; \lambda\ge 0,\; \widehat{\lambda}\ge 0,\; s\ge 0,\; \widehat{s}\ge 0.
\end{array}
$$

Substituting $s + \widehat{s}$ into the objective (due to the objective pressure) allows us to reformulate the above problem as

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\lambda,\widehat{\lambda}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \sup\limits_{\boldsymbol{x}\in\mathbb{R}^n} \{S^\alpha(\boldsymbol{\beta}^\top \boldsymbol{x}) - \lambda\|\boldsymbol{x}^1 - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^1 - \boldsymbol{x}\|_q\} \\ \text{subject to} & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda} \geq 0, \end{array} \tag{5}$$

and an application of Lemma B.1 leads us to the following reformulation:

$$\inf_{\substack{\boldsymbol{\beta}\in\mathbb{R}^n \\ \lambda\geq 0,\widehat{\lambda}\geq 0}} \sup_{\theta\in\mathrm{dom}(S^*)} \inf_{\boldsymbol{z}\in\mathbb{R}^n} \left\{ \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} - S^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^1 + \underbrace{\theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^1 - \boldsymbol{x}^1)}_{(1)} \ :\ \underbrace{|\theta|\cdot\|\boldsymbol{\beta}-\boldsymbol{z}\|_{q^\star}\leq\lambda}_{(2)},\ |\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\leq\widehat{\lambda} \right\}.$$

The objective term (1) has a product of the uncertain parameter θ and the adjustable variable $\boldsymbol{z}$, and even when (2) is linear such as in the case of $q = 1$ the product of the uncertain parameter with both the decision variable $\boldsymbol{\beta}$ and the adjustable variable $\boldsymbol{z}$ still appear since:

$$|\theta|\cdot\|\boldsymbol{\beta}-\boldsymbol{z}\|_\infty \leq \lambda \iff -\lambda \leq \theta\boldsymbol{\beta} - \theta\boldsymbol{z} \leq \lambda.$$

This reduces problem (5) to a generic two-stage robust optimization problem with random recourse [66, Problem 1] which is proven to be NP-hard even if $S^{\alpha*}$ was constant [31]. □

B.5 Proof of Theorem 5.4

Consider the reformulation Inter-adjustable of Inter-ARO that we introduced in the proof of Proposition 5.3. For any $i \in [N]$ and $j \in [\widehat{N}]$, the corresponding constraint in the first group of 'adjustable robust' ($\forall,\ \exists$) constraints will be:

$$\forall\theta\in\mathrm{dom}(L^*), \exists\boldsymbol{z}\in\mathbb{R}^n : \begin{cases} -L^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \leq s_i + \kappa\dfrac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda} \\ |\theta|\cdot\|\boldsymbol{\beta}-\boldsymbol{z}\|_{q^\star}\leq\lambda \\ |\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\leq\widehat{\lambda}. \end{cases}$$

By changing the order of $\forall$ and $\exists$, we obtain:

$$\exists\boldsymbol{z}\in\mathbb{R}^n, \forall\theta\in\mathrm{dom}(L^*) : \begin{cases} -L^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \leq s_i + \kappa\dfrac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda} \\ |\theta|\cdot\|\boldsymbol{\beta}-\boldsymbol{z}\|_{q^\star}\leq\lambda \\ |\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\leq\widehat{\lambda}. \end{cases}$$

Notice that this is a safe approximation, since any fixed $\boldsymbol{z}$ satisfying the latter system is a feasible static solution in the former system, meaning that for every realization of θ in the first system, the inner $\exists\boldsymbol{z}$ can always 'play' the same $\boldsymbol{z}$ that is feasible in the latter system (hence the latter is named the *static* relaxation, [10]). In the relaxed system, we can drop $\forall\theta$ and keep its worst-case realization instead:

$$\exists\boldsymbol{z}\in\mathbb{R}^n : \begin{cases} \sup_{\theta\in\mathrm{dom}(L^*)}\{-L^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i)\} \leq s_i + \kappa\dfrac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda} \\ \sup_{\theta\in\mathrm{dom}(L^*)}\{|\theta|\}\cdot\|\boldsymbol{\beta}-\boldsymbol{z}\|_{q^\star}\leq\lambda \\ \sup_{\theta\in\mathrm{dom}(L^*)}\{|\theta|\}\cdot\|\boldsymbol{z}\|_{q^\star}\leq\widehat{\lambda}. \end{cases}$$

The term $\sup_{\theta\in\mathrm{dom}(L^*)}\{-L^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\beta}^\top\boldsymbol{x}^i + \theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i)\}$ is the definition of the biconjugate $L^{\alpha**}(\boldsymbol{\beta}^\top\boldsymbol{x}^i + \boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i))$. Since L^α is a closed convex function, we have $L^{\alpha**} = L^\alpha$ [52, Corollary 12.2.1]. Moreover, $\sup_{\theta\in\mathrm{dom}(L^*)}\{|\theta|\}$ is an alternative representation of the Lipschitz constant of the function L^α [52, Corollary 13.3.3], which is equal to 1 as we showed earlier. The adjustable robust constraint thus reduces to:

$$\exists\boldsymbol{z}\in\mathbb{R}^n : \begin{cases} L^\alpha(\boldsymbol{\beta}^\top\boldsymbol{x}^i + \boldsymbol{z}^\top(\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i)) \leq s_i + \kappa\dfrac{1-y^i}{2}\lambda + \widehat{s}_j + \kappa\dfrac{1-\widehat{y}^j}{2}\widehat{\lambda} \\ \|\boldsymbol{\beta}-\boldsymbol{z}\|_{q^\star}\leq\lambda \\ \|\boldsymbol{z}\|_{q^\star}\leq\widehat{\lambda} \end{cases}$$

as a result of the static relaxation. This relaxed reformulation applies to all $i \in [N]$ and $j \in [\widehat{N}]$ as well as to the second group of adjustable robust constraints analogously. Replacing each constraint of Inter-adjustable with this system concludes the proof.

B.6 Proof of Corollary 5.6

To prove the first statement, take $\widehat{\lambda} = 0$ and observe the constraint $\|\boldsymbol{z}_{ij}^l\|_{q^\star} \leq \widehat{\lambda}$ implies $\boldsymbol{z}_{ij}^l = \mathbf{0}$ for all $l \in \{-1, 1\},\ i \in [N],\ j \in [\widehat{N}]$. The optimization problem can thus be written without those variables:

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} & \varepsilon\lambda + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i + \dfrac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}} \widehat{s}_j \\ \text{subject to} & L^\alpha(l\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\dfrac{1-ly^i}{2}\lambda + \widehat{s}_j \quad \forall l \in \{-1,1\},\ \forall i \in [N],\ \forall j \in [\widehat{N}] \\ & \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda \\ & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \boldsymbol{s} \in \mathbb{R}_+^N,\ \widehat{\boldsymbol{s}} \in \mathbb{R}_+^{\widehat{N}}. \end{array}$$

Notice that optimal solutions should satisfy $\widehat{s}_j = \widehat{s}_{j'}$ for all $j, j' \in [N]$. To see this, assume for contradiction that $\exists j, j' \in [N]$ such that $\widehat{s}_j < \widehat{s}_{j'}$. If a constraint indexed with (l, i, j) for arbitrary $l \in \{-1, 1\}$ and $i \in [N]$ is feasible, it means the consraint indexed with (l, i, j') cannot be tight given that these constraints are identical except for the $\widehat{s}_j$ or $\widehat{s}_{j'}$ appearing on the right hand side. Hence, such a solution cannot be optimal as this is a minimization problem, and updating $\widehat{s}_{j'}$ as $\widehat{s}_j$ preserves the feasibility of the problem while decreasing the objective value. We can thus use a single variable $\tau \in \mathbb{R}_+$ and rewrite the problem as

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s},\widehat{\boldsymbol{s}}}{\text{minimize}} & \varepsilon\lambda + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(s_i + \tau) \\ \text{subject to} & L^\alpha(\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\dfrac{1-y^i}{2}\lambda + \tau \qquad \forall i \in [N] \\ & L^\alpha(-\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\dfrac{1+y^i}{2}\lambda + \tau \quad \forall i \in [N] \\ & \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda \\ & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \boldsymbol{s} \in \mathbb{R}_+^N,\ \widehat{\boldsymbol{s}} \in \mathbb{R}_+^{\widehat{N}}, \end{array}$$

where we also eliminated the index $l \in \{-1, 1\}$ by writing the constraints explicitly. Since s_i and τ both appear as $s_i + \tau$ in this problem, we can use a variable change where we relabel $s_i + \tau$ as s_i (or, equivalently set $\tau = 0$ without any optimality loss). Moreover, the constraints with index $i \in [N]$ are

$$\begin{Bmatrix} L^\alpha(\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \tau \\ L^\alpha(-\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\lambda + \tau \end{Bmatrix} = \begin{Bmatrix} L^\alpha(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \tau \\ L^\alpha(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\lambda + \tau \end{Bmatrix}$$

if $y^i = 1$, and similarly they are

$$\begin{Bmatrix} L^\alpha(\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\lambda + \tau \\ L^\alpha(-\boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \tau \end{Bmatrix} = \begin{Bmatrix} L^\alpha(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \kappa\lambda + \tau \\ L^\alpha(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) \leq s_i + \tau \end{Bmatrix}$$

if $y^i = -1$. Since these are identical, the problem can finally be written as

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\lambda,\boldsymbol{s}}{\text{minimize}} & \varepsilon\lambda + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{subject to} & \log(1 + \exp(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) \leq s_i \qquad \forall i \in [N] \\ & \log(1 + \exp(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) - \lambda\kappa \leq s_i \quad \forall i \in [N] \\ & \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda \\ & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \boldsymbol{s} \in \mathbb{R}_+^N, \end{array}$$

where we also used the definition of L^α. This problem is identical to DR-ARO, which means that feasible solutions of DR-ARO are feasible for Inter-ARO* if the additional variables $(\widehat{\lambda}, \widehat{\boldsymbol{s}}, \boldsymbol{z}_{ij}^l)$ are set to zero, concluding the first statement of the corollary.

The second statement is immediate since $\widehat{\varepsilon} \to \infty$ forces $\widehat{\lambda} = 0$ due to the term $\widehat{\varepsilon}\widehat{\lambda}$ in the objective of Inter-ARO*, and this proof shows in such a case Inter-ARO* reduces to DR-ARO (which is identical to Inter-ARO when $\varepsilon \to \infty$ by definition).

B.7 Proof of Proposition 5.7

By standard linearity arguments and from the definition of $\mathbb{Q}_{\text{mix}}$, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{Q}_{\text{mix}}}\left[\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\right] \\
\iff &\int_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\{-1,+1\}}\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\,\mathrm{d}\mathbb{Q}_{\text{mix}}((\boldsymbol{x},y)) \\
\iff &\frac{N}{N+w\widehat{N}}\int_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\{-1,+1\}}\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\,\mathrm{d}\mathbb{P}_N((\boldsymbol{x},y))+ \\
&\qquad\frac{w\widehat{N}}{N+w\widehat{N}}\int_{(\boldsymbol{x},y)\in\mathbb{R}^n\times\{-1,+1\}}\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\,\mathrm{d}\widehat{\mathbb{P}}_{\widehat{N}}((\boldsymbol{x},y)) \\
\iff &\frac{N}{N+w\widehat{N}}\cdot\frac{1}{N}\sum_{i\in[N]}\sup_{\boldsymbol{z}^i\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i+\boldsymbol{z}^i,y^i)\}+\frac{w\widehat{N}}{N+w\widehat{N}}\cdot\frac{1}{\widehat{N}}\sum_{j\in[\widehat{N}]}\sup_{\boldsymbol{z}^j\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\widehat{\boldsymbol{x}}^j+\boldsymbol{z}^j,\widehat{y}^j)\} \\
\iff &\frac{1}{N+w\widehat{N}}\left[\sum_{i\in[N]}\sup_{\boldsymbol{z}^i\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i+\boldsymbol{z}^i,y^i)\}+w\cdot\sum_{j\in[\widehat{N}]}\sup_{\boldsymbol{z}^j\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\widehat{\boldsymbol{x}}^j+\boldsymbol{z}^j,\widehat{y}^j)\}\right],
\end{aligned}
$$

which coincides with the objective function of (1). The proof of Proposition 4.1 shows

$$
\mathbb{E}_{\mathbb{Q}_{\text{mix}}}\left[\sup_{\boldsymbol{z}\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z},y)\}\right]=\mathbb{E}_{\mathbb{Q}_{\text{mix}}}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x},y)]
$$

which concludes the proof.

B.8 Proof of Proposition 5.8

We first prove auxiliary results on mixture distributions. To this end, denote by $\mathcal{C}(\mathbb{Q},\mathbb{P})\subseteq\mathcal{P}(\Xi\times\Xi)$ the set of couplings of the distributions $\mathbb{Q}\in\mathcal{P}(\Xi)$ and $\mathbb{P}\in\mathcal{P}(\Xi)$.

Lemma B.2. *Let $\mathbb{Q},\mathbb{P}^1,\mathbb{P}^2\in\mathcal{P}(\Xi)$ be probability distributions. If $\Pi^1\in\mathcal{C}(\mathbb{Q},\mathbb{P}^1)$ and $\Pi^2\in\mathcal{C}(\mathbb{Q},\mathbb{P}^2)$, then, $\lambda\cdot\Pi^1+(1-\lambda)\cdot\Pi^2\in\mathcal{C}(\mathbb{Q},\lambda\cdot\mathbb{P}^1+(1-\lambda)\cdot\mathbb{P}^2)$ for all $\lambda\in(0,1)$.*

Proof. Let $\Pi=\lambda\cdot\Pi^1+(1-\lambda)\cdot\Pi^2$ and $\mathbb{P}=\lambda\cdot\mathbb{P}^1+(1-\lambda)\cdot\mathbb{P}^2$. To have $\Pi\in\mathcal{C}(\mathbb{Q},\mathbb{P})$ we need $\Pi(\mathrm{d}\boldsymbol{\xi},\Xi)=\mathbb{Q}(\mathrm{d}\boldsymbol{\xi})$ and $\Pi(\Xi,\mathrm{d}\boldsymbol{\xi'})=\mathbb{P}(\mathrm{d}\boldsymbol{\xi'})$. To this end, observe that

$$
\begin{aligned}
\Pi(\mathrm{d}\boldsymbol{\xi},\Xi)&=\lambda\cdot\Pi^1(\mathrm{d}\boldsymbol{\xi},\Xi)+(1-\lambda)\cdot\Pi^2(\mathrm{d}\boldsymbol{\xi},\Xi)\\
&=\lambda\cdot\mathbb{Q}+(1-\lambda)\cdot\mathbb{Q}\;=\;\mathbb{Q}
\end{aligned}
$$

where the second identity uses the fact that $\Pi^1\in\mathcal{C}(\mathbb{Q},\mathbb{P}^1)$. Similarly, we can show:

$$
\begin{aligned}
\Pi(\Xi,\mathrm{d}\boldsymbol{\xi})&=\lambda\cdot\Pi^1(\Xi,\mathrm{d}\boldsymbol{\xi})+(1-\lambda)\cdot\Pi^2(\Xi,\mathrm{d}\boldsymbol{\xi})\\
&=\lambda\cdot\mathbb{P}^1+(1-\lambda)\cdot\mathbb{P}^2\;=\;\mathbb{P},
\end{aligned}
$$

which concludes the proof. □

We further prove the following intermediary result.

Lemma B.3. *Let $\mathbb{Q},\mathbb{P}^1,\mathbb{P}^2\in\mathcal{P}(\Xi)$ and $\mathbb{P}=\lambda\cdot\mathbb{P}^1+(1-\lambda)\cdot\mathbb{P}^2$ for some $\lambda\in(0,1)$. We have:*

$$
\mathrm{W}(\mathbb{Q},\mathbb{P})\leq\lambda\cdot\mathrm{W}(\mathbb{Q},\mathbb{P}^1)+(1-\lambda)\cdot\mathrm{W}(\mathbb{Q},\mathbb{P}^2).
$$

Proof. The Wasserstein distance between $\mathbb{Q},\mathbb{Q}'\in\mathcal{P}(\Xi)$ can be written as:

$$
\mathrm{W}(\mathbb{Q},\mathbb{Q}')=\min_{\Pi\in\mathcal{C}(\mathbb{Q},\mathbb{Q}')}\left\{\int_{\Xi\times\Xi}d(\boldsymbol{\xi},\boldsymbol{\xi'})\Pi(\mathrm{d}\boldsymbol{\xi},\mathrm{d}\boldsymbol{\xi'})\right\},
$$

and since d is a feature-label metric (*cf.* Definition 3.1) the minimum is well-defined [73, Theorem 4.1]. We name the optimal solutions to the above problem the *optimal couplings*. Let Π^1 be an optimal coupling of $\mathrm{W}(\mathbb{Q}, \mathbb{P}^1)$ and let Π^2 be an optimal coupling of $\mathrm{W}(\mathbb{Q}, \mathbb{P}^2)$ and define $\Pi^c = \lambda \cdot \Pi^1 + (1-\lambda) \cdot \Pi^2$. We have

$$\begin{aligned}
\mathrm{W}(\mathbb{Q}, \mathbb{P}) &= \min_{\Pi \in \mathcal{C}(\mathbb{Q}, \mathbb{P})} \left\{ \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi}') \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \right\} \\
&\leq \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi}') \Pi^c(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \\
&= \lambda \cdot \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi}') \Pi^1(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') + (1-\lambda) \cdot \int_{\Xi \times \Xi} d(\boldsymbol{\xi}, \boldsymbol{\xi}') \Pi^2(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \\
&= \lambda \cdot \mathrm{W}(\mathbb{Q}, \mathbb{P}^1) + (1-\lambda) \cdot \mathrm{W}(\mathbb{Q}, \mathbb{P}^2),
\end{aligned}$$

where the first identity uses the definition of the Wasserstein metric, the inequality is due to Lemma B.2 as Π^c is a feasible coupling (not necessarily optimal), the equality that follows uses the definition of Π^c and the linearity of integrals, and the final identity uses the fact that Π^1 and Π^2 were constructed to be the optimal couplings. □

We now prove the proposition (we refer to $\mathbb{Q}_{\mathrm{mix}}$ in the statement of this lemma simply as $\mathbb{Q}$). To prove $\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$, it is sufficient to show that $\mathrm{W}(\mathbb{P}_N, \mathbb{Q}) \leq \varepsilon$ and $\mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \mathbb{Q}) \leq \widehat{\varepsilon}$ jointly hold. By using Lemma B.3, we can derive the following inequalities:

$$\begin{aligned}
\mathrm{W}(\mathbb{P}_N, \mathbb{Q}) &\leq \lambda \cdot \underbrace{\mathrm{W}(\mathbb{P}_N, \mathbb{P}_N)}_{=0} + (1-\lambda) \cdot \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}}) \\
\mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \mathbb{Q}) &\leq \lambda \cdot \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}}) + (1-\lambda) \cdot \underbrace{\mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \widehat{\mathbb{P}}_{\widehat{N}})}_{=0}.
\end{aligned}$$

Therefore, sufficient conditions on $\mathrm{W}(\mathbb{P}_N, \mathbb{Q}) \leq \varepsilon$ and $\mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \mathbb{Q}) \leq \widehat{\varepsilon}$ would be:

$$\begin{cases} (1-\lambda) \cdot \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}}) \leq \varepsilon \\ \lambda \cdot \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}}) \leq \widehat{\varepsilon}. \end{cases}$$

Moreover, given that $\varepsilon + \widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$, the sufficient conditions further simplify to

$$\begin{cases} (1-\lambda) \cdot \widehat{\varepsilon} \leq \lambda \cdot \varepsilon \\ \lambda \cdot \varepsilon \leq (1-\lambda) \cdot \widehat{\varepsilon}. \end{cases} \quad \Longleftrightarrow \; \lambda \cdot \varepsilon = (1-\lambda) \cdot \widehat{\varepsilon},$$

which is implied when $\dfrac{\lambda}{1-\lambda} = \dfrac{\widehat{\varepsilon}}{\varepsilon}$, concluding the proof.

B.9 Proof of Theorem 6.1

Since each result in the statement of this theorem is abridged, we will present these results sequentially as separate results. We review the existing literature to characterize $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$, in a similar fashion with the results presented in [55, Appendix A] for the logistic loss, by revising them to the adversarial loss whenever necessary. The N-fold product distribution of $\mathbb{P}^0$ from which the training set $\mathbb{P}_N$ is constructed is denoted below by $[\mathbb{P}^0]^N$.

Theorem B.4. *Assume there exist $a > 1$ and $A > 0$ such that $\mathbb{E}_{\mathbb{P}^0}[\exp(\|\boldsymbol{\xi}\|^a)] \leq A$ for a norm $\|\cdot\|$ on $\mathbb{R}^n$. Then, there are constants $c_1, c_2 > 0$ that only depend on $\mathbb{P}^0$ through a, A, and n, such that $[\mathbb{P}^0]^N(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)) \geq 1 - \eta$ holds for any confidence level $\eta \in (0,1)$ as long as the Wasserstein ball radius satisfies the following optimal characterization*

$$\varepsilon \geq \begin{cases} \left(\dfrac{\log(c_1/\eta)}{c_2 \cdot N} \right)^{1/\max\{n,2\}} & \text{if } N \geq \dfrac{\log(c_1/\eta)}{c_2} \\ \left(\dfrac{\log(c_1/\eta)}{c_2 \cdot N} \right)^{1/a} & \text{otherwise.} \end{cases}$$

Proof. The statement follows from Theorem 18 of [34]. The presented decay rate $\mathcal{O}(N^{-1/n})$ of ε as N increases is optimal [25]. □

Now that we gave a confidence for the radius ε of $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$, we analyze the underlying optimization problems. Most of the theory is well-established for logistic loss function, and in the following we show that similar results follow for the adversarial loss function. For convenience, we state DR-ARO again by using the adversarial loss function as in the proof of Proposition 4.1:

$$\begin{array}{ll} \underset{\boldsymbol{\beta}}{\text{minimize}} & \sup\limits_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x},y)] \\ \text{subject to} & \boldsymbol{\beta}\in\mathbb{R}^n. \end{array} \tag{DR-ARO}$$

Theorem B.5. *If the assumptions of Theorem B.4 are satisfied and ε is chosen as in the statement of Theorem B.4, then*

$$[\mathbb{P}^0]^N\left(\mathbb{E}_{\mathbb{P}^0}[\ell^\alpha_{\boldsymbol{\beta}^\star}(\boldsymbol{x},y)] \le \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}^\star}(\boldsymbol{x},y)]\right) \ge 1-\eta$$

holds for all $\eta\in(0,1)$ and all optimizers $\boldsymbol{\beta}^\star$ of DR-ARO.

Proof. The statement follows from Theorem 19 of [34] given that $\ell^\alpha_{\boldsymbol{\beta}}$ is a finite-valued continuous loss function. □

Theorem B.5 states that the optimal value of DR-ARO overestimates the true loss with arbitrarily high confidence $1-\eta$. Despite the desired overestimation of the true loss, we show that DR-ARO is still asymptotically consistent if we restrict the set of admissible $\boldsymbol{\beta}$ to a bounded set[3].

Theorem B.6. *If we restrict the hypotheses $\boldsymbol{\beta}$ to a bounded set $\mathcal{H}\subseteq\mathbb{R}^n$, and parameterize ε as ε_N to show its dependency to the sample size, then, under the assumptions of Theorem B.4, we have*

$$\sup_{\mathbb{Q}\in\mathfrak{B}_{\varepsilon_N}(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}^\star}(\boldsymbol{x},y)] \underset{N\to\infty}{\longrightarrow} \mathbb{E}_{\mathbb{P}^0}[\ell^\alpha_{\boldsymbol{\beta}^\star}(\boldsymbol{x},y)] \quad \mathbb{P}^0\text{-almost surely},$$

whenever ε_N is set as specified in Theorem B.4 along with its finite-sample confidence η_N, and they satisfy $\sum_{N\in\mathbb{N}}\eta_N<\infty$ and $\lim_{N\to\infty}\varepsilon_N=0$.

Proof. If we show that there exists $\boldsymbol{\xi}^0\in\Xi$ and $C>0$ such that $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x},y)\le C(1+d(\boldsymbol{\xi},\boldsymbol{\xi}^0))$ holds for all $\boldsymbol{\beta}\in\mathcal{H}$ and $\boldsymbol{\xi}\in\Xi$ (that is, the adversarial loss satisfies a growth condition), the statement will follow immediately from Theorem 20 of [34].

To see that the growth condition is satisfied, we first substitute the definition of $\ell^\alpha_{\boldsymbol{\beta}}$ and d explicitly, and note that we would like to show there exists $\boldsymbol{\xi}^0\in\Xi$ and $C>0$ such that

$$\log(1+\exp(-y\cdot\boldsymbol{\beta}^\top\boldsymbol{x}+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})) \le C(1+\|\boldsymbol{x}-\boldsymbol{x}^0\|_q+\kappa\cdot\mathbb{1}[y\neq y^0])$$

holds for all $\boldsymbol{\beta}\in\mathcal{H}$ and $\boldsymbol{\xi}\in\Xi$. We take $\boldsymbol{\xi}^0=(\mathbf{0},y^0)$ and show that the right-hand side of the inequality can be lower bounded as:

$$\begin{aligned} C(1+\|\boldsymbol{x}-\boldsymbol{x}^0\|_q+\kappa\cdot\mathbb{1}[y\neq y^0]) &= C(1+\|\boldsymbol{x}\|_q+\kappa\cdot\mathbb{1}[y\neq y^0]) \\ &\ge C(1+\|\boldsymbol{x}\|_q). \end{aligned}$$

Moreover, the left-hand side of the inequality can be upper bounded for any $\boldsymbol{\beta}\in\mathcal{H}\subseteq[-M,M]^n$ (for some $M>0$) and $\boldsymbol{\xi}=(\boldsymbol{x},y)\in\Xi$ as:

$$\begin{aligned} \log(1+\exp(-y\cdot\boldsymbol{\beta}^\top\boldsymbol{x}+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})) &\le \log(1+\exp(|\boldsymbol{\beta}^\top\boldsymbol{x}|+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})) \\ &\le \log(2\cdot\exp(|\boldsymbol{\beta}^\top\boldsymbol{x}|+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})) \\ &= \log(2)+|\boldsymbol{\beta}^\top\boldsymbol{x}|+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star} \\ &\le \log(2)+\sup_{\boldsymbol{\beta}\in[-M,M]^n}\{|\boldsymbol{\beta}^\top\boldsymbol{x}|\}+\alpha\cdot\sup_{\boldsymbol{\beta}\in[-M,M]^n}\{\|\boldsymbol{\beta}\|_{p^\star}\} \\ &= \log(2)+M\cdot\|\boldsymbol{x}\|_1+M\cdot\alpha \\ &\le \log(2)+M\cdot n^{(q-1)/q}\cdot\|\boldsymbol{x}\|_1+M\cdot\alpha \end{aligned}$$

[3]Note that, this is without loss of generality given that we can normalize the decision boundary of linear classifiers.

where the final inequality uses Hölder's inequality to bound the 1-norm with the q-norm. Thus, it suffices to show that we have

$$\log(2) + M \cdot n^{(q-1)/q} \cdot \|\boldsymbol{x}\|_1 + M \cdot \alpha \leq C(1 + \|\boldsymbol{x}\|_q) \quad \forall \boldsymbol{\xi} \in \Xi,$$

which is satisfied for any $C \geq \max\{\log(2) + M \cdot \alpha,\ M \cdot n^{(q-1)/q}\}$. This completes the proof by showing the growth condition is satisfied. □

So far, we reviewed tight characterizations for ε so that the ball $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$ includes the true distribution $\mathbb{P}^0$ with arbitrarily high confidence, proved that the DRO problem DR-ARO overestimates the true loss, while converging to the true problem asymptotically as the confidence $1 - \eta$ increases and the radius ε decreases simultaneously. Finally, we discuss that for optimal solutions $\boldsymbol{\beta}^\star$ to DR-ARO, there are worst case distributions $\mathbb{Q}^\star \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$ of nature's problem that are supported on at most $N + 1$ atoms.

Theorem B.7. *If we restrict the hypotheses $\boldsymbol{\beta}$ to a bounded set $\mathcal{H} \subseteq \mathbb{R}^n$, then there are distributions $\mathbb{Q}^\star \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$ that are supported on at most $N + 1$ atoms and satisfy:*

$$\mathbb{E}_{\mathbb{Q}^\star}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)] \ = \sup_{\mathbb{Q} \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)].$$

Proof. The proof follows from [80]. □

See the proof of [55, Theorem 8] and the discussion that follows for insights and further analysis on these results presented.

B.10 Proof of Theorem 6.2

Firstly, since $\widehat{\mathbb{P}}_{\widehat{N}}$ is constructed from i.i.d. samples of $\widehat{\mathbb{P}}$, we can overestimate the distance $\widehat{\varepsilon}_1 = \mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \widehat{\mathbb{P}})$ analogously by applying Theorem B.4, *mutatis mutandis*. This leads us to the following result where the joint (independent) N-fold product distribution of $\mathbb{P}^0$ and the $\widehat{N}$-fold product distribution of $\widehat{\mathbb{P}}$ is denoted below by $[\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N \times \widehat{N}}$.

Theorem B.8. *Assume that there exist $a > 1$ and $A > 0$ such that $\mathbb{E}_{\mathbb{P}^0}[\exp(\|\boldsymbol{\xi}\|^a)] \leq A$, and there exist $\widehat{a} > 1$ and $\widehat{A} > 0$ such that $\mathbb{E}_{\widehat{\mathbb{P}}}[\exp(\|\boldsymbol{\xi}\|^{\widehat{a}})] \leq \widehat{A}$ for a norm $\|\cdot\|$ on $\mathbb{R}^n$. Then, there are constants $c_1, c_2 > 0$ that only depends on $\mathbb{P}^0$ through a, A, and n, and constants $\widehat{c}_1, \widehat{c}_2 > 0$ that only depends on $\widehat{\mathbb{P}}$ through $\widehat{a}$, $\widehat{A}$, and n such that $[\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N \times \widehat{N}}(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) \geq 1 - \eta$ holds for any confidence level $\eta \in (0, 1)$ as long as the Wasserstein ball radii satisfy the following characterization*

$$\varepsilon \geq \begin{cases} \left(\dfrac{\log(c_1/\eta_1)}{c_2 \cdot N}\right)^{1/\max\{n,2\}} & \text{if } N \geq \dfrac{\log(c_1/\eta_1)}{c_2} \\ \left(\dfrac{\log(c_1/\eta_1)}{c_2 \cdot N}\right)^{1/a} & \text{otherwise} \end{cases}$$

$$\widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}}) + \begin{cases} \left(\dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2 \cdot \widehat{N}}\right)^{1/\max\{n,2\}} & \text{if } \widehat{N} \geq \dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2} \\ \left(\dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2 \cdot \widehat{N}}\right)^{1/\widehat{a}} & \text{otherwise} \end{cases}$$

for some $\eta_1, \eta_2 > 0$ satisfying $\eta_1 + \eta_2 = \eta$.

Proof. It immediately follows from Theorem B.4 that $[\mathbb{P}^0]^N(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)) \geq 1 - \eta_1$ holds. If we take $\widehat{\varepsilon}_1 > 0$ as

$$\widehat{\varepsilon}_1 \geq \begin{cases} \left(\dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2 \cdot \widehat{N}}\right)^{1/\max\{n,2\}} & \text{if } \widehat{N} \geq \dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2} \\ \left(\dfrac{\log(\widehat{c}_1/\eta_2)}{\widehat{c}_2 \cdot \widehat{N}}\right)^{1/\widehat{a}} & \text{otherwise} \end{cases}$$

then, we similarly have $[\widehat{\mathbb{P}}]^{\widehat{N}}(\widehat{\mathbb{P}} \in \mathfrak{B}_{\widehat{\varepsilon}_1}(\widehat{\mathbb{P}}_{\widehat{N}})) \geq 1-\eta_2$. Since the following implication follows from the triangle inequality:

$$\widehat{\mathbb{P}} \in \mathfrak{B}_{\widehat{\varepsilon}_1}(\widehat{\mathbb{P}}_{\widehat{N}}) \implies \mathbb{P}^0 \in \mathfrak{B}_{\widehat{\varepsilon}_1+\mathrm{W}(\mathbb{P}^0,\widehat{\mathbb{P}})}(\widehat{\mathbb{P}}_{\widehat{N}}),$$

we have that $[\widehat{\mathbb{P}}]^{\widehat{N}}(\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\widehat{\mathbb{P}}_{\widehat{N}})) \geq 1-\eta_2$. These results, along with the facts that $\widehat{\mathbb{P}}_{\widehat{N}}$ and $\mathbb{P}_N$ are independently sampled from their true distributions, imply:

$$\begin{aligned}
&[\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N\times\widehat{N}}(\mathbb{P}^0 \notin \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \vee \mathbb{P}^0 \notin \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) \\
\leq &[\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N\times\widehat{N}}(\mathbb{P}^0 \notin \mathfrak{B}_{\varepsilon}(\mathbb{P}_N)) + [\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N\times\widehat{N}}(\mathbb{P}^0 \notin \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) \\
= &[\mathbb{P}^0]^{N}(\mathbb{P}^0 \notin \mathfrak{B}_{\varepsilon}(\mathbb{P}_N)) + [\widehat{\mathbb{P}}]^{\widehat{N}}(\mathbb{P}^0 \notin \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) < \eta_1+\eta_2
\end{aligned}$$

implying the desired result $[\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N\times\widehat{N}}(\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) \geq 1-\eta$. □

The second statement immediately follows under the assumptions of Theorem B.8: Inter-ARO overestimates the true loss analogously as Theorem B.5 with an identical proof.

C Exponential cone representation of DR-ARO

For any $i \in [N]$, the constraints of DR-ARO are

$$\begin{cases} \log(1+\exp(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) \leq s_i \\ \log(1+\exp(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) - \lambda \cdot \kappa \leq s_i, \end{cases}$$

which, by using an auxiliary variable u, can be written as

$$\begin{cases} \log(1+\exp(-y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + u)) \leq s_i \\ \log(1+\exp(y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + u)) - \lambda \cdot \kappa \leq s_i \\ \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star} \leq u. \end{cases}$$

Following the conic modeling guidelines of [41], for new variables $v_i^+, w_i^+ \in \mathbb{R}$, the first constraint can be written as

$$\left\{v_i^+ + w_i^+ \leq 1,\ (v_i^+, 1, [-u + y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i) - s_i] \in \mathcal{K}_{\exp},\ (w_i^+, 1, -s_i) \in \mathcal{K}_{\exp},\right.$$

by using the definition of the exponential cone $\mathcal{K}_{\exp}$. Similarly, for new variables $v_i^-, w_i^- \in \mathbb{R}$, the second constraint can be written as

$$\left\{v_i^- + w_i^- \leq 1,\ (v_i^-, 1, [-u - y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i] - s_i - \lambda \cdot \kappa) \in \mathcal{K}_{\exp},\ (w_i^-, 1, -s_i - \lambda \cdot \kappa) \in \mathcal{K}_{\exp}.\right.$$

Applying this for all $i \in [N]$ concludes that the following is the conic formulation of DR-ARO:

$$\begin{array}{lll}
\underset{\substack{\boldsymbol{\beta},\ \lambda,\ \boldsymbol{s},\ u \\ \boldsymbol{v}^+,\boldsymbol{w}^+,\boldsymbol{v}^-,\boldsymbol{w}^-}}{\text{minimize}} & \lambda \cdot \varepsilon + \dfrac{1}{N}\displaystyle\sum_{i\in[N]} s_i & \\
\text{subject to} & v_i^+ + w_i^+ \leq 1 & \forall i \in [N] \\
& (v_i^+, 1, [-u + y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i] - s_i) \in \mathcal{K}_{\exp},\ (w_i^+, 1, -s_i) \in \mathcal{K}_{\exp} & \forall i \in [N] \\
& v_i^- + w_i^- \leq 1 & \forall i \in [N] \\
& (v_i^-, 1, [-u - y^i \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i] - s_i - \lambda \cdot \kappa) \in \mathcal{K}_{\exp},\ (w_i^-, 1, -s_i - \lambda \cdot \kappa) \in \mathcal{K}_{\exp} & \forall i \in [N] \\
& \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star} \leq u & \\
& \|\boldsymbol{\beta}\|_{q^\star} \leq \lambda & \\
& \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N,\ u \in \mathbb{R},\ \boldsymbol{v}^+, \boldsymbol{w}^+, \boldsymbol{v}^-, \boldsymbol{w}^- \in \mathbb{R}^N. &
\end{array}$$

D Further details for numerical experiments

All experiments are implemented in Julia [11] (MIT license) and executed on Intel Xeon 2.66GHz processors with 8GB memory in single-core mode. We use MOSEK 10.1 [42] to solve all exponential conic programs through JuMP [23]. The UCI datasets [22, 39] we use (see Table 4) are subject to CC BY 4.0 license. MNIST is subject to CC BY-SA 3.0 and EMNIST to CC0 1.0 license.

D.1 UCI experiments

Preprocessing UCI datasets Although we reported the first 5 datasets in the main paper, we experiment on 10 UCI datasets [22, 39] (*cf.* Table 4). We use Python3 for preprocessing these datasets. Classification problems with more than two classes are converted to binary classification problems (most frequent class/others). For all datasets, numerical features are standardized, the ordinal categorical features are left as they are, and the nominal categorical features are processed via one-hot encoding. As mentioned in the main paper, we obtain auxiliary (synthetic) datasets via SDV, which is also implemented in Python 3.

Table 4: Size of the UCI datasets.

DataSet	N	$\widehat{N}$	N_{te}	n
absent	111	333	296	74
annealing	134	404	360	41
audiology	33	102	91	102
breast-cancer	102	307	274	90
contraceptive	220	663	590	23
dermatology	53	161	144	99
ecoli	50	151	135	9
spambase	690	2,070	1,841	58
spect	24	72	64	23
prim-tumor	50	153	136	32

Detailed misclassification results on the UCI datasets Table 5 contains detailed results on the out-of-sample error rates of each method on 10 UCI datasets for classification. All parameters are 5-fold cross-validated: Wasserstein radii from the grid $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 1, 2, 5, 10\}$ ($10^{-6}, 10^{-5}, 2, 5, 10$ are rarely selected, but we did not change our grid in order not to introduce a bias), κ from the grid $\{1, \sqrt{n}, n\}$ the weight parameter of ARO+Aux from grid $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0, 1\}$. We fix the norm defining the feature-label metric to the ℓ_1-norm, and test ℓ_2-attacks, but other choices with analogous results are also implemented.

Finally, we demonstrate that our theory, especially DRO+ARO+Aux, contributes to the DRO literature even without adversarial attacks. In this case of $\alpha = 0$, ERM and ARO would be equivalent, and DRO+ARO would reduce to the traditional DR LR model [60]. ARO+Aux would be interpreted as revising the empirical distribution of ERM to a mixture (mixture weight cross-validated) of the empirical and auxiliary distributions. DRO+ARO+Aux, on the other hand, can be interpreted as DRO over a carefully reduced ambiguity set (intersection of the empirical and auxiliary Wasserstein balls). The results are in Table 6. Analogous results follow as before (that is, DRO+ARO+Aux is the 'winning' approach, DRO+ARO and ARO+Aux alternate for the 'second' approach), with the exception of the dataset *contraceptive*, where ARO+Aux outperforms others.

D.2 MNIST/EMNIST experiments

Our setting is analogous to the UCI experiments. However, for auxiliary data, we use the EMNIST dataset. We used the MLDatasets package of Julia to prepare such auxiliary data.

D.3 Artificial experiments

Data generation We sample a 'true' β from a unit ℓ_2-ball, and generate data as summarized in Algorithm 9. Such a dataset generation gives N instances from the same true data-generating distribution. In order to obtain $\hat{N}$ auxiliary dataset instances, we perturb the probabilities p^i with

Table 5: Mean ($\pm$ std) out-of-sample errors of UCI datasets, each with 10 simulations. Results for adversarial (ℓ_2-)attack strengths $\alpha = 0.05$ and $\alpha = 0.2$ are shared.

Data	α	ERM	ARO	ARO+Aux	DRO+ARO	DRO+ARO+Aux
absent	0.05	44.02% ($\pm$ 2.89)	38.82% ($\pm$ 2.86)	35.95% ($\pm$ 3.78)	34.22% ($\pm$ 2.70)	**32.64%** ($\pm$ 2.54)
	0.20	73.65% ($\pm$ 4.14)	51.49% ($\pm$ 3.39)	49.56% ($\pm$ 3.80)	45.61% ($\pm$ 2.32)	**44.90%** ($\pm$ 2.30)
annealing	0.05	18.08% ($\pm$ 1.89)	16.61% ($\pm$ 2.16)	14.97% ($\pm$ 1.39)	13.50% ($\pm$ 2.98)	**12.78%** ($\pm$ 2.78)
	0.20	37.31% ($\pm$ 3.92)	23.08% ($\pm$ 2.82)	21.30% ($\pm$ 1.93)	20.70% ($\pm$ 1.32)	**19.53%** ($\pm$ 1.42)
audiology	0.05	21.43% ($\pm$ 3.64)	21.54% ($\pm$ 3.92)	17.03% ($\pm$ 2.90)	11.76% ($\pm$ 3.28)	**9.01%** ($\pm$ 3.54)
	0.20	37.91% ($\pm$ 6.78)	29.34% ($\pm$ 5.89)	20.44% ($\pm$ 2.75)	20.00% ($\pm$ 3.01)	**17.91%** ($\pm$ 3.28)
breast-cancer	0.05	4.74% ($\pm$ 1.26)	4.93% ($\pm$ 1.75)	3.87% ($\pm$ 1.17)	3.06% ($\pm$ 0.79)	**2.52%** ($\pm$ 0.50)
	0.20	9.93% ($\pm$ 1.73)	8.14% ($\pm$ 2.01)	6.09% ($\pm$ 1.79)	5.04% ($\pm$ 1.11)	**4.67%** ($\pm$ 0.99)
contraceptive	0.05	44.14% ($\pm$ 2.80)	42.86% ($\pm$ 2.59)	40.98% ($\pm$ 0.95)	40.00% ($\pm$ 1.33)	**39.65%** ($\pm$ 1.15)
	0.20	66.19% ($\pm$ 5.97)	43.49% ($\pm$ 2.24)	**42.71%** ($\pm$ 1.47)	**42.71%** ($\pm$ 1.47)	**42.71%** ($\pm$ 1.47)
dermatology	0.05	15.97% ($\pm$ 2.64)	16.46% ($\pm$ 1.67)	13.47% ($\pm$ 1.97)	12.78% ($\pm$ 1.61)	**10.84%** ($\pm$ 1.24)
	0.20	30.07% ($\pm$ 4.24)	28.54% ($\pm$ 3.25)	21.53% ($\pm$ 2.17)	22.64% ($\pm$ 2.15)	**20.21%** ($\pm$ 1.58)
ecoli	0.05	16.30% ($\pm$ 4.42)	14.67% ($\pm$ 5.13)	13.26% ($\pm$ 3.07)	11.11% ($\pm$ 5.52)	**9.78%** ($\pm$ 2.61)
	0.20	51.41% ($\pm$ 3.37)	42.67% ($\pm$ 2.91)	41.85% ($\pm$ 2.95)	39.70% ($\pm$ 2.68)	**38.89%** ($\pm$ 2.57)
spambase	0.05	11.35% ($\pm$ 0.77)	10.23% ($\pm$ 0.54)	10.16% ($\pm$ 0.56)	9.83% ($\pm$ 0.37)	**9.81%** ($\pm$ 0.38)
	0.20	27.32% ($\pm$ 2.11)	15.83% ($\pm$ 0.77)	15.70% ($\pm$ 0.76)	15.67% ($\pm$ 0.72)	**15.50%** ($\pm$ 0.68)
spect	0.05	33.75% ($\pm$ 5.17)	29.69% ($\pm$ 5.46)	25.78% ($\pm$ 3.06)	25.47% ($\pm$ 3.38)	**21.56%** ($\pm$ 2.74)
	0.20	54.22% ($\pm$ 9.88)	37.5% ($\pm$ 3.53)	35.16% ($\pm$ 2.47)	33.75% ($\pm$ 2.68)	**30.16%** ($\pm$ 3.61)
prim-tumor	0.05	21.84% ($\pm$ 4.55)	20.81% ($\pm$ 3.97)	17.35% ($\pm$ 3.59)	16.18% ($\pm$ 3.83)	**14.78%** ($\pm$ 2.89)
	0.20	34.19% ($\pm$ 6.17)	25.37% ($\pm$ 4.58)	21.62% ($\pm$ 3.45)	21.84% ($\pm$ 3.34)	**19.63%** ($\pm$ 2.71)

Table 6: Mean out-of-sample errors of UCI experiments without adversarial attacks.

Data	ERM	ARO	ARO+Aux	DRO+ARO	DRO+ARO+Aux
absent	36.28%	36.28%	31.86%	28.31%	**27.74%**
annealing	10.61%	10.61%	7.64%	**7.14%**	**7.14%**
audiology	14.94%	14.94%	12.97%	10.11%	**7.69%**
breast-cancer	6.64%	6.64%	5.22%	2.55%	**2.15%**
contraceptive	35.00%	35.00%	**33.75%**	34.56%	33.85%
dermatology	16.04%	16.04%	11.60%	9.93%	**8.06%**
ecoli	6.74%	6.74%	4.96%	5.19%	**4.37%**
spambase	8.95%	8.95%	8.52%	8.34%	**8.16%**
spect	30.74%	30.74%	24.69%	22.35%	**18.75%**
prim-tumor	22.79%	22.79%	17.28%	15.07%	**13.97%**

standard random normal noise which is equivalent to sampling i.i.d. from a *perturbed* distribution. Testing is always done on true data, that is, the test set is sampled according to Algorithm 9.

Algorithm 1 Data from a ground truth logistic classifier

Input: set of feature vectors $\boldsymbol{x}^i,\ i \in [N]$; vector $\boldsymbol{\beta}$
  for $i \in \{1, \ldots, N\}$ **do**
    Find the probability $p^i = \left[1 + \exp(-\boldsymbol{\beta}^\top \boldsymbol{x})\right]^{-1}$.
    Sample $u = \mathcal{U}(0, 1)$
    if $p^i \geq u$ **then**
      $y^i = +1$
    else
      $y^i = -1$
    end if
  end for
Output: $(\boldsymbol{x}^i, y^i),\ i \in [N]$.

Strength of the attack and importance of auxiliary data In the main paper we discussed how the strength of an attack determines whether using auxiliary data in ARO (ARO+Aux) or considering distributional ambiguity (DRO+ARO) is more effective, and observed that unifying them to obtain DRO+ARO+Aux yields the best results in all attack regimes. Now we focus on the methods that rely

Table 7: Mean w in problem (1) and $\varepsilon/\widehat{\varepsilon}$ in problem Inter-ARO across 25 simulations of cross-validating ω, ε, and $\widehat{\varepsilon}$.

Attack	ARO+Aux (cross-validated w)	DRO+ARO+Aux (cross-validated $\varepsilon/\widehat{\varepsilon}$)
$\alpha = 0$	0.002	0.0120
$\alpha = 0.1$	0.046	0.172
$\alpha = 0.25$	0.086	0.232
$\alpha = 0.5$	0.290	0.241

on auxiliary data, namely ARO+Aux and DRO+ARO+Aux and explore the importance of auxiliary data $\widehat{\mathbb{P}}_{\widehat{N}}$ in comparison to its empirical counterpart $\mathbb{P}_N$. Table 7 shows the average values of w for problem (1) obtained via cross-validation. We see that the greater the attack strength is the more we should use the auxiliary data in ARO+Aux. The same relationship holds for the average of $\varepsilon/\widehat{\varepsilon}$ obtained via cross-validation in Inter-ARO, which means that the relative size of the Wasserstein ball built around the empirical distribution gets larger compared to the same ball around the auxiliary data, that is, ambiguity around the auxiliary data is smaller than the ambiguity around the empirical data. We highlight as a possible future research direction exploring when a larger attack *per se* implies the intersection will move towards the auxiliary data distribution.

More results on scalability We further simulate 25 cases with an ℓ_2-attack strength of $\alpha = 0.2$, $N = 200$ instances in the training dataset, $\widehat{N} = 200$ instances in the auxiliary dataset, and we vary the number of features n. We report the median ($50\% \pm 15\%$ quantiles shaded) runtimes of each method in Figure 3. The fastest methods are ERM and ARO among which the faster one depends on n (as the adversarial loss includes a regularizer of $\boldsymbol{\beta}$), followed by ARO+Aux, DRO+ARO, and DRO+ARO+Aux, respectively. DRO+ARO+Aux is the slowest, which is expected given that DRO+ARO is its special for large $\widehat{\varepsilon}$. The runtime however scales graciously.

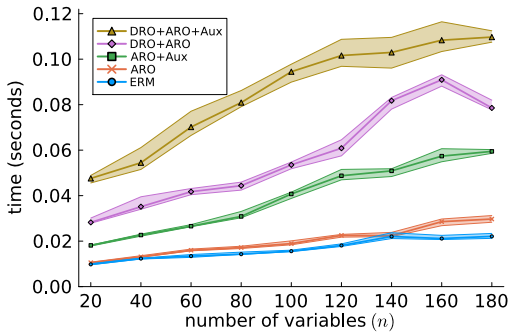

Figure 3: Runtimes under a varying number of features in the artificially generated empirical and auxiliary datasets.

Finally, we focus further on DRO+ARO+Aux which solves problem Inter-ARO with $\mathcal{O}(n \cdot N \cdot \widehat{N})$ variables and exponential cone constraints. For $n = 1,000$ and $N = \widehat{N} = 10,000$, we observe that the runtimes vary between 134 to 232 seconds across 25 simulations.