Distributionally and Adversarially Robust Logistic Regression via Intersecting Wasserstein Balls

Aras Selvi*
Imperial College Business School

Eleonora Kreačić*
JP Morgan AI Research

Mohsen Ghassemi
JP Morgan AI Research

Vamsi K. Potluru
JP Morgan AI Research

Tucker Balch
JP Morgan AI Research

Manuela Veloso
JP Morgan AI Research

Abstract

Adversarially robust optimization (ARO) has emerged as the *de facto* standard for training models that hedge against adversarial attacks in the test stage. While these models are robust against adversarial attacks, they tend to suffer severely from overfitting. To address this issue, some successful methods replace the empirical distribution in the training stage with alternatives including *(i)* a worst-case distribution residing in an ambiguity set, resulting in a distributionally robust (DR) counterpart of ARO; *(ii)* a mixture of the empirical distribution with a distribution induced by an auxiliary (*e.g.*, synthetic, external, out-of-domain) dataset. Inspired by the former, we study the Wasserstein DR counterpart of ARO for logistic regression and show it admits a tractable convex optimization reformulation. Adopting the latter setting, we revise the DR approach by intersecting its ambiguity set with another ambiguity set built using the auxiliary dataset, which offers a significant improvement whenever the Wasserstein distance between the data generating and auxiliary distributions can be estimated. We study the underlying optimization problem, develop efficient solution algorithms, and demonstrate that the proposed method outperforms benchmark approaches on standard datasets.

Preprint. AS completed this work during an internship at JPMorganChase.
* corresponding authors a.selvi19@imperial.ac.uk eleonora.kreacic@jpmorgan.com

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 (Bishop, 2006; Hastie et al., 2009). 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 (Vapnik, 1999, Chapter 2). In real-world applications, however, various challenges, such as data scarcity and the existence of adversarial attacks, lead to deteriorated out-of-sample performance for models trained via ERM.

One of the key limitations of ERM, particularly as it is designed to minimize an expected loss for the empirical distribution, emerges from the finite nature of data in practice. This leads ERM to suffer from the 'optimism bias', also known as overfitting (Murphy, 2022), or the optimizer's curse (DeMiguel and Nogales, 2009; Smith and Winkler, 2006), causing deteriorated out-of-sample performance. A popular approach to prevent this phenomenon, *distributionally robust optimization* (DRO; Delage and Ye 2010), optimizes the expected loss for the worst-case distribution residing within a pre-specified ambiguity set.

Another key challenge faced by ERM in practice is adversarial attacks, where an adversary perturbs the observed features during the testing or deployment phase (Szegedy et al., 2014; Goodfellow et al., 2015). For neural networks, the paradigm of *adversarial training* (AT; Madry et al. 2018) is thus designed to provide adversarial robustness by simulating the attacks in the training stage. Several successful variants of AT, specialized to different losses and attacks, have been proposed in the literature to achieve adversarial robust-

ness without significantly reducing performance on training sets (Shafahi et al., 2019; Zhang et al., 2019; Pang et al., 2022; Gao et al., 2019). While some works (*e.g.*, Carlini et al. 2019; Uesato et al. 2018) examine adversarial robustness guarantees of various training algorithms, a recent stream of research (*e.g.*, Bennouna et al. 2023; Xing et al. 2022b) focuses on *adversarially robust optimization* (ARO), constraining ERM to guarantee an exact pre-specified level of adversarial robustness while maximizing training accuracy.

Recently, it has been observed that the two aforementioned notions of robustness can conflict, as adversarially robust (AR) models suffer from severe overfitting (*robust overfitting*; Raghunathan et al. 2019; Yu et al. 2022; Li and Spratling 2023). Indeed, it is observed that robust overfitting is even more severe than traditional overfitting (Rice et al., 2020). To this end, some works address robust overfitting by revisiting AT algorithms and adding adjustments for better generalization (Chen et al., 2020; Li and Li, 2023). In a recent work, Bennouna et al. (2023, Thm 3.2) decompose the error gap of robust overfitting into the statistical error of estimating the true data-generating distribution via the empirical distribution and an adversarial error resulting from the adversarial attacks, hence proposing the simultaneous adoption of DRO and ARO.

In this work, we study logistic regression (LR) for binary classification that is adversarially robust against ℓ_p-attacks (Croce et al., 2020). To address robust overfitting faced by the adversarially robust LR model, we employ a DRO approach where distributional ambiguity is modeled with the type-1 Wasserstein metric. We base our work on an observation that the worst-case logistic loss under adversarial attacks can be represented as a Lipschitz continuous and convex loss function. This allows us to use existing Wasserstein DRO machinery for Lipschitz losses, and derive an *exact* reformulation of the Wasserstein DR counterpart of adversarially robust LR as a tractable convex problem.

Our main contribution lies in reducing the size of the Wasserstein ambiguity set in the DRO problem mentioned above, in order to create a less conservative problem while preserving the same robustness guarantees. To accomplish this, we draw inspiration from recent work on ARO that leverages auxiliary datasets (*e.g.*, Gowal et al. 2021; Xing et al. 2022b) and revise our DRO problem by intersecting its ambiguity set with another ambiguity set constructed using an auxiliary dataset. Examples of auxiliary data include synthetic data generated from a generative model (*e.g.*, privacy-preserving data release), data in the presence of distributional shifts (*e.g.*, different time periods/regions), noisy data (*e.g.*, measurement errors), or out-of-domain data (*e.g.*, different source);

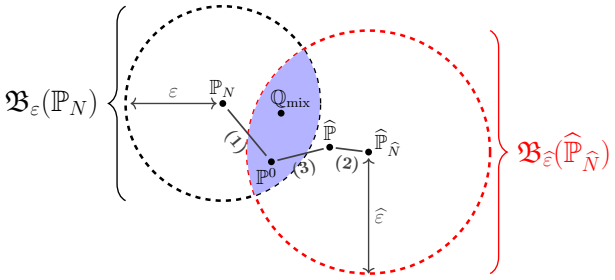

Figure 1: *Traditional ARO optimizes the expected adversarial loss over the empirical distribution $\mathbb{P}_N$ constructed from N i.i.d. samples of the (unknown) true data-generating distribution $\mathbb{P}^0$. Replacing $\mathbb{P}_N$ with a worst-case distribution in the ball $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$ gives us its DR counterpart. To reduce the size of this ball, we intersect it with another ball $\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{P}_{\widehat{N}})$ while ensuring $\mathbb{P}^0$ is still included with high confidence. The latter ball is centered at an empirical distribution $\widehat{\mathbb{P}}_N$ constructed from $\widehat{N}$ i.i.d. samples of some auxiliary distribution $\widehat{\mathbb{P}}$. 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 resides in $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ under some conditions.*

any auxiliary dataset is viable as long as its instances are sampled independently from an underlying data-generating distribution whose Wasserstein distance to the true data-generating distribution is known or can be estimated. Figure 1 illustrates our framework.

The paper unfolds as follows. In Section 2, we review related literature on DRO and ARO, with a focus on their interactions. We examine the use of auxiliary data in ARO and the intersection of Wasserstein balls in DRO. We discuss open questions for LR to motivate our loss function choice in this work. Section 3 gives preliminaries on ERM, ARO, and type-1 Wasserstein DRO. In Section 4, we discuss that the adversarial logistic loss can be reformulated as a Lipschitz convex function, enabling the use of Wasserstein DRO machinery for Lipschitz losses. Our main contribution (*cf.* Figure 1) is in Section 5, where we provide an explicit reformulation of the distributionally and adversarially robust LR problem over the intersection of two Wasserstein balls, prove the NP-hardness of this problem, and derive a convex relaxation of it. Our work is mainly on *optimization* where we focus on how to solve the underlying problems upon cross-validating Wasserstein ball radii, however, in Section 6 we discuss some preliminary statistical approaches to set such radii. We close the paper with numerical experiments on standard benchmark datasets in Section 7. We borrow the standard notation in DR machine learning, which is elaborated on in our Appendices.

2 RELATED WORK

Auxiliary data in ARO. The use of auxiliary data appears in the ARO literature. In particular, it is shown that additional unlabeled data sampled from the same (Carmon et al., 2019; Xing et al., 2022a) or different (Deng et al., 2021) data-generating distributions could provide adversarial robustness. Sehwag et al. (2022) show that adversarial robustness can be certified even when it is provided for a synthetic dataset as long as the distance between its generator and the true data-generating distribution can be quantified. Gowal et al. (2021); Xing et al. (2022b) propose optimizing a weighted combination of ARO over empirical and synthetic datasets. We show that the latter approach can be recovered by our model.

DRO-ARO interactions. In our work, we optimize ARO against worst-case data-generating distributions residing in an ambiguity set, where the type-1 Wasserstein metric is used for distances since it is arguably the most common choice in machine learning (ML) with Lipschitz losses (Shafieezadeh-Abadeh et al., 2019; Gao, 2023). In the literature, it is shown that standard ARO is equivalent to the DRO of the original loss function with a type-∞ Wasserstein metric (Staib and Jegelka, 2017; Khim and Loh, 2018; Pydi and Jog, 2021; Regniez et al., 2022; Frank and Niles-Weed, 2024), or a Lévy-Prokhorov metric (Bennouna and Van Parys, 2022). In other words, in the absence of adversarial attacks, training models adversarially with artificial attacks provide some distributional robustness. 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 residing in an ∞-Wasserstein ball around the empirical distribution. Conversely, Sinha et al. (2018) discuss that while DRO over Wasserstein balls is intractable for generic losses (*e.g.*, neural networks), its Lagrange relaxation resembles ARO and thus ARO yields a certain degree of (relaxed) distributional robustness (Wu et al., 2020; Bui et al., 2022; Phan et al., 2023). However, to the best of our knowledge, there have not been works optimizing a pre-specified level of type-1 Wasserstein distributional robustness (that hedges against overfitting, Kuhn et al. 2019) and adversarial robustness (that hedges against adversarial attacks, Goodfellow et al. 2015) *simultaneously*. To our knowledge, the only approach that considers the exact DR counterpart of ARO is proposed by Bennouna et al. (2023) who model distributional ambiguity with φ-divergences for neural networks.

Intersecting ambiguity sets in DRO. Recent work started to explore the intersection of ambiguity sets for different contexts (Awasthi et al., 2022; Wang et al., 2024) or different metrics (Zhang et al., 2023). Our idea of intersecting Wasserstein balls is originated from the "Surround, then Intersect" strategy (Taskesen et al., 2021, §5.2) to train linear regression under sequential domain adaptation in a non-adversarial setting (see Shafahi et al. 2020 and Song et al. 2019 for robustness in domain adaptation/transfer learning). The aforementioned work focuses on the squared loss function with an ambiguity set using the Wasserstein metric developed for the first and second distributional moments. In a recent study, Rychener et al. (2024) generalize most of the previous results and prove that DRO problems over the intersection of two Wasserstein balls admit tractable convex reformulations whenever the loss function is the maximum of concave functions

Logistic loss in DRO and ARO. Our choice of LR aligns with the current directions and open questions in the related literature. In the DRO literature, even in the absence of adversarial attacks, the aforementioned work of Taskesen et al. (2021) 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 their proof relies on the properties of the squared loss. Similarly, Rychener et al. (2024) discuss that the logistic loss fails to satisfy the piece-wise concavity assumption and is inherently difficult to optimize over the intersection of Wasserstein balls. 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 even without adversarial attacks, and develop specialized approximation techniques. Our problem recovers DR LR (Shafieezadeh-Abadeh et al., 2015; Selvi et al., 2022) as a special case in the absence of adversarial attacks and auxiliary data. Answering theoretical challenges posed by logistic regression has been useful in answering more general questions in the DRO literature, such as DR LR (Shafieezadeh-Abadeh et al., 2015) leading to DR ML (Shafieezadeh-Abadeh et al., 2019) and mixed-feature DR LR (Selvi et al., 2022) leading to mixed-feature DR Lipschitz ML (Belbasi et al., 2023). Finally, in the (non-DR) ARO literature, there are recent theory developments on understanding the effect of auxiliary data (*e.g.*, Xing et al. 2022b) specifically for squared and logistic loss functions.

3 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 via $\text{Prob}[y \mid \boldsymbol{x}] = [1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x})]^{-1}$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$; its associated loss is the *logloss* $\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y) := \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x}))$.

Empirical risk minimization. Let $\mathcal{P}(\Xi)$ denote the set of distributions supported on Ξ and $\mathbb{P}^0 \in \mathcal{P}(\Xi)$ denote the true data-generating distribution. One wants to minimize the expected logloss over $\mathbb{P}^0$, that is

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

In practice, $\mathbb{P}^0$ is hardly ever known, and one 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

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

Distributionally robust optimization. To be able to define a distance between distributions, we first define the following feature-label metric on Ξ.

Definition 1. *The distance between instances* $\boldsymbol{\xi} = (\boldsymbol{x}, y) \in \Xi$ *and* $\boldsymbol{\xi'} = (\boldsymbol{x'}, y') \in \Xi$ *for* $\kappa > 0$ *and* $q \geq 1$ *is*

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

Using this metric, we define the Wasserstein distance.

Definition 2. *The type-1 Wasserstein distance between distributions* $\mathbb{Q}, \mathbb{Q}' \in \mathcal{P}(\Xi)$ *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 $\mathcal{C}(\mathbb{Q}, \mathbb{Q}')$ *is the set of couplings of* $\mathbb{Q}$ *and* $\mathbb{Q}'$.

In finite-data settings, the distance between the true data-generating distribution and the empirical distribution is upper-bounded by some $\epsilon > 0$. The Wasserstein DRO problem is thus defined as

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

where $\mathfrak{B}_\varepsilon(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}(\Xi) : \mathrm{W}(\mathbb{Q}, \mathbb{P}) \leq \varepsilon\}$ denotes the Wasserstein ball centered at $\mathbb{P} \in \mathcal{P}(\Xi)$ with radius ε. We refer to Mohajerin Esfahani and Kuhn (2018) and Kuhn et al. (2019) for the properties of DRO.

Adversarially robust optimization. The goal of adversarial robustness is to provide robustness against adversarial attacks (Goodfellow et al., 2015). An adversarial attack, in the widely studied ℓ_p-noise setting (Croce et al., 2020), 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:

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

ARO reduces to ERM when $\alpha = 0$. Note that ARO is identical to feature robust training (Bertsimas et al., 2019) which is not motivated by adversarial attacks, but the presence of noisy observations in the training set (Ben-Tal et al., 2009; Gorissen et al., 2015).

DRO-ARO connection. A connection between ARO and DRO is noted in the literature (Staib and Jegelka 2017, Proposition 3.1, Khim and Loh 2018, Lemma 22, Pydi and Jog 2021, Lemma 5.1, Regniez et al. 2022, Proposition 2.1, Frank and Niles-Weed 2024, Lemma 3, and Bennouna et al. 2023, §3). Namely, problem ARO is equivalent to a DRO problem

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathfrak{B}^\infty_\alpha(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)], \tag{1}$$

where the ambiguity set $\mathfrak{B}^\infty_\alpha(\mathbb{P}_N)$ is a type-∞ Wasserstein ball (Givens and Shortt, 1984) with radius α. Hence, in non-adversarial settings, ARO provides robustness with respect to the type-∞ Wasserstein distance. In the case of adversarial attacks, it suffers from robust overfitting as discussed earlier. To address this issue, one straightforward approach is to revisit (1) and replace α with some $\alpha' > \alpha$. This approach, however, does not provide improvements for the out-of-sample performance since *(i)* the type-∞ Wasserstein distance employed in problem (1) uses a metric on the feature space, ignoring labels; *(ii)* type-∞ Wasserstein distances do not provide strong out-of-sample performances in ML (unlike, *e.g.*, the type-1 Wasserstein distance) since the required radii to provide meaningful robustness guarantees are typically too large (Bennouna and Van Parys, 2022, §1.2.2, and references therein). We thus study the type-1 Wasserstein counterpart of ARO, which we initiate in the next section.

4 DISTRIBUTIONALLY AND ADVERSARIALLY ROBUST LR

Here we derive the Wasserstein DR counterpart of ARO that will set the ground for our main result in the next section. We impose the following assumption.

Assumption 1. *We are given a finite* $\varepsilon > 0$ *value satisfying* $\mathrm{W}(\mathbb{P}^0, \mathbb{P}_N) \leq \varepsilon$.

The assumption implies that we know an $\varepsilon > 0$ value satisfying $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$. Typically, however, ε is either estimated through cross-validation or finite sample statistics, with the assumption then regarded as holding with high confidence (see §6 for a review of

related results we can borrow). The distributionally and adversarially robust LR problem is thus:

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

By employing a simple duality trick for the inner sup-problem, as commonly applied in robust optimization (Ben-Tal et al., 2009; Bertsimas and Den Hertog, 2022), we can represent DR-ARO as a standard non-adversarial DRO problem with an updated loss function, which we name the *adversarial loss*.

Observation 1. *Problem* DR-ARO *is equivalent to*

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

where the adversarial loss $\ell^\alpha_{\boldsymbol{\beta}}$ *is defined as*

$$\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x},y) := \log(1+\exp(-y\cdot\boldsymbol{\beta}^\top\boldsymbol{x}+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star})),$$

for $p^\star$ *satisfying* $1/p+1/p^\star=1$. *The univariate representation* $L^\alpha(z) := \log(1+\exp(-z+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$ *of* $\ell^\alpha_{\boldsymbol{\beta}}$ *is convex and has a Lipschitz modulus of* 1.

As a corollary of Observation 1, we can directly employ the techniques proposed by Shafieezadeh-Abadeh et al. (2019) to dualize the inner sup-problem of DR-ARO and obtain a tractable reformulation.

Corollary 1. *Problem* DR-ARO *admits the following tractable convex optimization reformulation:*

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

for $q^\star$ *satisfying* $1/q+1/q^\star=1$.

The constraints of this problem are exponential cone representable (derivation is in the appendices) and for $q\in\{1,2,\infty\}$, the yielding problem can be solved with the exponential cone solver of MOSEK (MOSEK ApS, 2023) in polynomial time (Nesterov, 2018).

5 MAIN RESULT

In §4 we discussed the traditional DRO setting where we have access to an empirical distribution $\mathbb{P}_N$ constructed from N i.i.d. samples of the true data-generating distribution $\mathbb{P}^0$, and we are given (or we estimate) some ε so that $\mathbb{P}^0\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)$. Recently in DRO literature, it became a key focus to study the case where we have access to an additional auxiliary empirical distribution $\widehat{\mathbb{P}}_{\widehat{N}}$ constructed from $\widehat{N}$ i.i.d. samples $\widehat{\boldsymbol{\xi}}^j=(\widehat{\boldsymbol{x}}^j,\widehat{y}^j),\ j\in[\widehat{N}]$, of some other distribution $\widehat{\mathbb{P}}$; given the increasing availability of useful auxiliary data in the ARO domain, we explore this direction here. We start with the following assumption.

Assumption 2. *We are given finite* $\varepsilon,\widehat{\varepsilon}>0$ *values satisfying* $\mathrm{W}(\mathbb{P}^0,\mathbb{P}_N)\leq\varepsilon$ *and* $\mathrm{W}(\mathbb{P}^0,\widehat{\mathbb{P}}_{\widehat{N}})\leq\widehat{\varepsilon}$.

The assumption implies that we know $\varepsilon,\widehat{\varepsilon}>0$ values satisfying $\mathbb{P}^0\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)\cap\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$. In practice, this assumption is ensured to hold with high confidence by estimating the ε and $\widehat{\varepsilon}$ values; methods across various domains which we can adopt are reviewed in §6. Under Assumption 2, we want to optimize the adversarial loss over the intersection $\mathfrak{B}_\varepsilon(\mathbb{P}_N)\cap\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$:

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

This formulation is expected to outperform DR-ARO as the ambiguity set is smaller while still including $\mathbb{P}^0$. However, problem Inter-ARO is challenging to solve even in the absence of adversarial attacks ($\alpha=0$) as we reviewed in §2. To address this challenge, we first reformulate Inter-ARO as a semi-infinite optimization problem with finitely many variables.

Proposition 1. Inter-ARO *is equivalent to:*

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

Even though this problem recovers the tractable problem DR-ARO as $\widehat{\varepsilon}\to\infty$, it is NP-hard in the finite radius settings. We reformulate Inter-ARO as an adjustable robust optimization problem (Ben-Tal et al., 2004; Yanıkoğlu et al., 2019), and borrow tools from this literature to obtain the following result.

Proposition 2. Inter-ARO *is equivalent to an adjustable RO problem with* $\mathcal{O}(N\cdot\widehat{N})$ *two-stage robust constraints, which is NP-hard even when* $N=\widehat{N}=1$.

The adjustable RO literature has developed a rich arsenal of relaxations that can be leveraged for Inter-ARO. We adopt the 'static relaxation technique' (Bertsimas et al., 2015) to restrict the feasible region of Inter-ARO and obtain a tractable approximation.

Theorem 1 (main)**.** *The following convex optimiza-*

tion problem is a feasible relaxation of Inter-ARO*:*

$$\begin{array}{ll}\inf\limits_{\substack{\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}\\ \boldsymbol{z}_{ij}^{+},\boldsymbol{z}_{ij}^{-}}} & \varepsilon\lambda+\widehat{\varepsilon}\widehat{\lambda}+\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}s_i+\dfrac{1}{\widehat{N}}\sum_{j=1}^{\widehat{N}}\widehat{s}_j\\ \text{s.t.} & \begin{bmatrix} L^{\alpha}(l\cdot\boldsymbol{\beta}^{\top}\boldsymbol{x}^i+\boldsymbol{z}_{ij}^{l\top}(\widehat{\boldsymbol{x}}^j-\boldsymbol{x}^i))\\ \quad\le s_i+\dfrac{\kappa(1-ly^i)}{2}\lambda+\widehat{s}_j+\dfrac{\kappa(1-l\widehat{y}^j)}{2}\widehat{\lambda},\\ \|l\boldsymbol{\beta}-\boldsymbol{z}_{ij}^{l}\|_{q^\star}\le\lambda,\ \|\boldsymbol{z}_{ij}^{l}\|_{q^\star}\le\widehat{\lambda}\end{bmatrix}\\ & \qquad\forall i\in[N],\ j\in[\widehat{N}],\ l\in\{-1,1\}\\ & \boldsymbol{\beta}\in\mathbb{R}^n,\ \lambda\ge 0,\ \widehat{\lambda}\ge 0,\ \boldsymbol{s}\in\mathbb{R}_{+}^{N},\ \widehat{\boldsymbol{s}}\in\mathbb{R}_{+}^{\widehat{N}}\\ & \boldsymbol{z}_{ij}^{l}\in\mathbb{R}^n,\ (i,j,l)\in[N]\times[\widehat{N}]\times\{-1,1\}.\end{array}\tag{Inter-ARO*}$$

Similarly to DR-ARO, the constraints of Inter-ARO* are exponential cone representable (*cf.* appendices).

Recall that for $\widehat{\varepsilon}$ large enough, Inter-ARO reduces to DR-ARO. The following corollary shows that, despite Inter-ARO* being a relaxation of Inter-ARO, a similar property holds. That is, "not learning anything from auxiliary data" remains feasible: the static relaxation does not force learning from $\widehat{\mathbb{P}}_{\widehat{N}}$, and it learns from auxiliary data only if the objective improves.

Corollary 2. Feasibility of ignoring auxiliary data: *Any feasible solution $(\boldsymbol{\beta},\lambda,\boldsymbol{s})$ of* DR-ARO *can be used to recover a feasible solution $(\boldsymbol{\beta},\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}},\boldsymbol{z}_{ij}^{+},\boldsymbol{z}_{ij}^{-})$ for* Inter-ARO* *with $\widehat{\lambda}=0$, $\widehat{\boldsymbol{s}}=\boldsymbol{0}$, and $\boldsymbol{z}_{ij}^{+}=\boldsymbol{z}_{ij}^{-}=\boldsymbol{0}$.* Convergence to Inter-ARO*: The optimal value of* Inter-ARO* *converges to the optimal value of* Inter-ARO*, with the same set of $\boldsymbol{\beta}$ solutions, as $\widehat{\varepsilon}\to\infty$.*

We close the section by discussing how Inter-ARO can recover some problems in the DRO and ARO literature. Firstly, recall that Inter-ARO can 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 Shafieezadeh-Abadeh et al. (2015). We next relate Inter-ARO to the problems in the ARO literature that use auxiliary data. The works in this literature (Gowal et al., 2021; Xing et al., 2022b) solve the following for some $w>0$:

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n}\ \frac{1}{N+w\widehat{N}}\Big[\sum_{i\in[N]}\sup_{\boldsymbol{z}^i\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i+\boldsymbol{z}^i,y^i)\}+ w\sum_{j\in[\widehat{N}]}\sup_{\boldsymbol{z}^j\in\mathcal{B}_p(\alpha)}\{\ell_{\boldsymbol{\beta}}(\widehat{\boldsymbol{x}}^j+\boldsymbol{z}^j,\widehat{y}^j)\}\Big],\tag{2}$$

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 $\widehat{\mathbb{P}}_{\widehat{N}}$:

Observation 2. *Problem (2) is equivalent to*

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n}\ \mathbb{E}_{\mathbb{Q}_{\text{mix}}}[\ell_{\boldsymbol{\beta}}^{\alpha}(\boldsymbol{x},y)]\tag{3}$$

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

We give a condition on ε and $\widehat{\varepsilon}$ to guarantee that the mixture distribution introduced in Proposition 2 lives in $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)\cap\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$, that is, the distribution $\mathbb{Q}_{\text{mix}}$ will be feasible in the sup problem of Inter-ARO.

Proposition 3. *For any $\lambda\in(0,1)$ and $\mathbb{Q}_{\text{mix}}=\lambda\cdot\mathbb{P}_N+(1-\lambda)\cdot\widehat{\mathbb{P}}_{\widehat{N}}$, we have $\mathbb{Q}_{\text{mix}}\in\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)\cap\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ whenever $\varepsilon+\widehat{\varepsilon}\ge\mathrm{W}(\mathbb{P}_N,\widehat{\mathbb{P}}_{\widehat{N}})$ and $\frac{\widehat{\varepsilon}}{\varepsilon}=\frac{\lambda}{1-\lambda}$.*

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 3 implies that a sufficient condition for this intersection to include $\mathbb{Q}_{\text{mix}}$ is $\widehat{\varepsilon}/\varepsilon=N/\widehat{N}$, which is intuitive since the radii of Wasserstein ambiguity sets are chosen inversely proportional to the number of samples (Kuhn et al., 2019, Theorem 18).

6 SETTING WASSERSTEIN RADII

Thus far, we have assumed knowledge of DRO ball radii ε and $\widehat{\varepsilon}$ that satisfy Assumptions 1 and 2. In this section, we employ Wasserstein finite-sample statistics techniques to estimate these values.

Setting ϵ for DR-ARO. In the following theorem, we present 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. We show that for an ε chosen in such a manner, DR-ARO is well-defined. The full description of this result is available in our appendices.

Theorem 2 (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 the expected loss for $\mathbb{P}^0$ 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.*

We next derive an analogous result for Inter-ARO.

Choosing ϵ and ϵ' in Inter-ARO. Recall that Inter-ARO revises DR-ARO by intersecting $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$ with $\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$. We need a nonempty intersection for Inter-ARO to be well-defined. A sufficient condition follows from the triangle inequality: $\varepsilon+\widehat{\varepsilon}\ge\mathrm{W}(\mathbb{P}_N,\widehat{\mathbb{P}}_{\widehat{N}})$. We also want this intersection to include $\mathbb{P}^0$ with high confidence, in order to satisfy Assumption 2. We next provide a tight characterization for such $\varepsilon,\widehat{\varepsilon}$. The full description of this result is available in our appendices.

Theorem 3 (abridged)**.** *For light-tailed $\mathbb{P}^0$ and $\widehat{\mathbb{P}}$, if $\varepsilon\ge\mathcal{O}(\frac{\log(\eta_1^{-1})}{N})^{1/n}$ and $\widehat{\varepsilon}\ge\mathrm{W}(\mathbb{P}^0,\widehat{\mathbb{P}})+\mathcal{O}(\frac{\log(\eta_2^{-1})}{\widehat{N}})^{1/n}$ 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 1. 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 constant 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.*

Knowledge of $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$**.** In Theorem 3, we use $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ explicitly. This distance, however, is typically unknown, and a common approach is to cross-validate it[1]. This would be applicable in our setting thanks to Corollary 2, because the relaxation problem Inter-ARO* is not forced to learn from the auxiliary data unless it is useful, that is, one can find evidence for the usefulness of the auxiliary data via cross-validation. Moreover, there are several domains where $\mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ is known exactly. For some special cases, we can use direct domain knowledge (*e.g.*, the "Uber vs Lyft" example of Taskesen et al. 2021). A very recent example comes from learning from multi-source data, where $\mathbb{P}^0$ is named the target distribution and $\widehat{\mathbb{P}}$ is the source distribution (Rychener et al., 2024, §1). Another domain is private data release, where 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. The (privately generated) synthetic distribution has a known nonzero Wasserstein distance from the true data-generating distribution (Dwork et al., 2014; Ullman and Vadhan, 2020). Alternatively, one can directly rely 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 (Arjovsky et al., 2017).

7 EXPERIMENTS

We conduct a series of experiments, each having a different source of auxiliary data, to test the proposed DR ARO models using empirical and auxiliary datasets. We use the following abbreviations, where 'solution' refers to the optimal $\boldsymbol{\beta}$ to make decisions:

- `ERM`: Solution of problem ERM (*i.e.*, naïve LR);
- `ARO`: Solution of problem ARO (*i.e.*, adversarially robust LR);
- `ARO+Aux`: Solution of problem (2) (*i.e.*, replacing the empirical distribution of ARO with its mixture with auxiliary data);
- `DRO+ARO`: Solution of DR-ARO (*i.e.*, the Wasserstein DR counterpart of ARO);
- `DRO+ARO+Aux`: Solution of Inter-ARO* (*i.e.*, relaxation of Inter-ARO that revises DR-ARO and intersects its ambiguity set with a Wasserstein ball built using auxiliary data);

Recall that `DRO+ARO` and `DRO+ARO+Aux` are the DR models that we propose. Note also that, `ERM`, `ARO`, and `DRO+ARO` are oblivious to auxiliary data. All Wasserstein radii of DR models, and the weight parameters of `ARO+Aux` are cross-validated from the same grids. All experiments are conducted in Julia (Bezanson et al., 2014) and executed on Intel Xeon 2.66GHz processors with 16GB memory in single-core mode. We use MOSEK's exponential cone optimizer to solve all problems. Implementation details are in the appendices.

7.1 UCI Datasets (Auxiliary Data is Synthetically Generated)

We compare the out-of-sample error rates of each method on 10 UCI datasets for binary classification (Kelly et al., 2023). For each dataset, we run 10 simulations as follows: *(i)* Select 40% of the data as a test set ($N_{\text{te}} \propto 0.4$); *(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 (Patki et al., 2016), which is then used to generate auxiliary data. The mean errors on the test set are reported in Table 1 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 our appendices, we report similar results for attack strengths $\alpha \in \{0, 0.05, 0.2\}$, and share data preprocessing details and standard deviations of out-of-sample errors.

7.2 MNIST/EMNIST Datasets (Auxiliary Data is Out-of-Domain)

We use the MNIST digits dataset (LeCun et al., 1998) to classify whether a digit is 1 or 7. For an auxiliary dataset, we use the larger EMNIST digits dataset (Cohen et al., 2017), whose authors summarize that this dataset has additional samples collected from a different group of individuals (high school students). Since EMNIST digits include MNIST digits, we remove the latter from the EMNIST dataset. We simulate the following 25 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 2 reports the mean out-of-sample errors in various adversarial attack regimes. The results are analogous to the UCI experiments.

[1] Note that, in practice, the distance between the unknown true and auxiliary data-generating distributions is also cross-validated in the transfer learning and domain adaptation literature (Zhong et al., 2010).

Table 1: 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%**
derma	15.97%	16.46%	13.47%	12.78%	**10.84%**
ecoli	16.30%	14.67%	13.26%	11.11%	**9.78%**
spam	11.35%	10.23%	10.16%	9.83%	**9.81%**
spect	33.75%	29.69%	25.78%	25.47%	**21.56%**
p-tumor	21.84%	20.81%	17.35%	16.18%	**14.78%**

Table 2: 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 (more details in the appendices). We simulate 25 cases, each with $N = 100$ training, $\widehat{N} = 200$ auxiliary, and $N_{\text{te}} = 10,000$ test instances and $n = 100$ features. The performance of benchmark models with varying ℓ_2-attacks is available in Figure 2 (left). ERM provides the worst performance, followed by ARO. 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, right). The fastest methods is ARO, followed by ERM, ARO+Aux, DRO+ARO, and DRO+ARO+Aux. The slowest is DRO+ARO+Aux, but the runtime scales graciously.

8 CONCLUSIONS

We formulate the distributionally robust counterpart of adversarially robust logistic regression. Additionally, we demonstrate how to effectively utilize appropriately curated auxiliary data by intersecting Wasserstein balls. 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, recent breakthroughs in the area of foundation models naturally pose the question of whether our ideas in this work apply to these models. For example, Ye et al. (2022) use pre-trained language model (PLM) to generate synthetic pairs of text sequences and labels which are then used to train downstream models. It would be interesting to adapt our ideas to the text domain to explore robustness in the presence of two PLMs.

References

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

Awasthi, P., Jung, C., and Morgenstern, J. (2022). Distributionally robust data join. *arXiv:2202.05797*.

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

Ben-Tal, A., Ghaoui, L. E., and Nemirovski, A. (2009). *Robust Optimization*. Princeton University Press.

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

Bennouna, A., Lucas, R., and Van Parys, B. (2023).

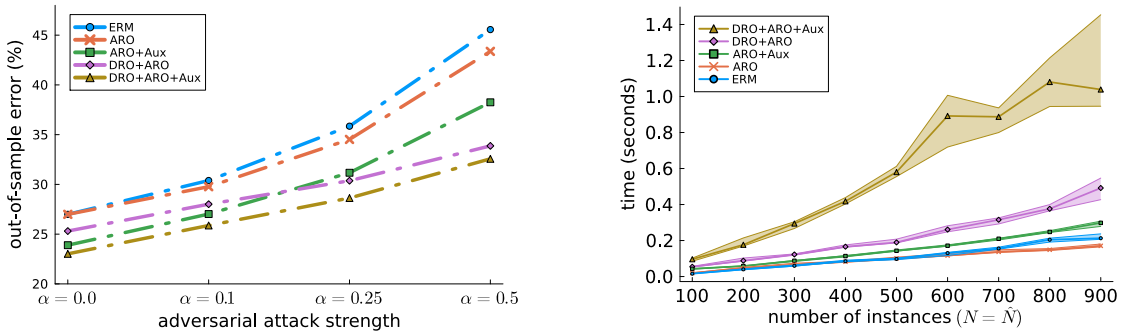

Figure 2: *Out-of-sample errors under varying attack strengths (left) and runtimes under varying numbers of empirical and auxiliary instances (right) of artificial experiments.*

Certified robust neural networks: Generalization and corruption resistance. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202.

Bennouna, A. and Van Parys, B. (2022). Holistic robust data-driven decisions. *arXiv:2207.09560*.

Bertsimas, D. and Den Hertog, D. (2022). *Robust and Adaptive Optimization*. Dynamic Ideas.

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

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

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

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

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

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

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

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

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

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

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

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

Deng, Z., Zhang, L., Ghorbani, A., and Zou, J. (2021). 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.

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

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

Frank, N. S. and Niles-Weed, J. (2024). Existence and minimax theorems for adversarial surrogate risks in binary classification. *Journal of Machine Learning Research*, 25(58):1–41.

Gao, R. (2023). Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking

the curse of dimensionality. *Operations Research*, 71(6):2291–2306.

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

Givens, C. R. and Shortt, R. M. (1984). A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240.

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

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

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

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

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

Kelly, M., Longjohn, R., and Nottingham, K. (2023). The uci machine learning repository. *URL https://archive. ics. uci. edu*.

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

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

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

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

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

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

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

MOSEK ApS (2023). Modeling cookbook. https://docs.mosek.com/MOSEKModelingCookbook-letter.pdf.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

Nesterov, Y. (2018). *Lectures on Convex Optimization*. Springer, 2nd edition.

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

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

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

Pydi, M. S. and Jog, V. (2021). The many faces of adversarial risk. *Advances in Neural Information Processing Systems*, 34:10000–10012.

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

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

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

Rockafellar, R. T. (1997). *Convex Analysis*. Princeton University Press.

Rychener, Y., Esteban-Pérez, A., Morales, J. M., and Kuhn, D. (2024). Wasserstein distributionally robust optimization with heterogeneous data sources. *arXiv:2407.13582*.

Sehwag, V., Mahloujifar, S., Handina, T., Dai, S., Xiang, C., Chiang, M., and Mittal, P. (2022). Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In

Proceedings of the 10th International Conference on Learning Representations.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wang, T., Chen, N., and Wang, C. (2024). Contextual optimization under covariate shift: A robust approach by intersecting wasserstein balls. *arXiv:2406.02426*.

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

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

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

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

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

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

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

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

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

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

Distributionally and Adversarially Robust Logistic Regression via Intersecting Wasserstein Balls
Supplementary Materials

9 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 $\boldsymbol{\xi} = (\boldsymbol{x}, y) \in \Xi := \mathbb{R}^n \times \{-1, +1\}$. For any $p > 0$, $\|\boldsymbol{x}\|_p$ denotes the rational norm $(\sum_{i=1}^n |x_i|^p)^{1/p}$ and $\|\boldsymbol{x}\|_{p^\star}$ is its dual norm where $1/p + 1/p^\star = 1$ 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 $\mathcal{K}_{\exp} = \text{cl}(\{\boldsymbol{\omega} \in \mathbb{R}^3 : \omega_1 \geq \omega_2 \cdot \exp(\omega_3/\omega_2),\ \omega_1 > 0,\ \omega_2 > 0\})$ where cl is the closure operator. The Lipschitz modulus of a univariate function f is defined as $\text{Lip}(f) := \sup_{z,z' \in \mathbb{R}} \{|f(z) - f(z')|/|z - z'| : z \neq z\}$ whereas its effective domain is $\text{dom}(f) = \{z : f(z) < +\infty\}$. 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.

10 MISSING PROOFS

10.1 Proof of Observation 1

For any $\boldsymbol{\beta} \in \mathbb{R}^n$, with standard robust optimization arguments (Ben-Tal et al., 2009; Bertsimas and Den Hertog, 2022), 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 DR-ARO with $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$ concludes the equivalence of the optimization problem.

Furthermore, to see $\text{Lip}(L^\alpha) = 1$, 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 $\text{Lip}(L^\alpha)$ 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 $z \longrightarrow -\infty$. □

10.2 Proof of Corollary 1

Observation 1 lets us represent DR-ARO as the DR counterpart of empirical minimization of $\ell^\alpha_{\boldsymbol{\beta}}$:

$$\begin{array}{ll} \underset{\boldsymbol{\beta}}{\text{minimize}} & \sup\limits_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)} \mathbb{E}_{\mathbb{Q}}\left[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x},y)\right] \\ \text{subject to} & \boldsymbol{\beta}\in\mathbb{R}^n. \end{array} \tag{4}$$

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, Theorem 14 *(ii)* of Shafieezadeh-Abadeh et al. (2019) is immediately applicable. We can therefore rewrite (4) as:

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

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

10.3 Proof of Proposition 1

We prove Proposition 1 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

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

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:

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

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:

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

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}_{\pm1},\mathbb{Q}_{\pm1},\widehat{\mathbb{Q}}_{\pm1}}{\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]\le\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]\le\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}_{\pm1}\in\mathcal{P}_+(\mathbb{R}^n),\ \mathbb{Q}^i_{\pm1}\in\mathcal{P}_+(\mathbb{R}^n),\ \widehat{\mathbb{Q}}^j_{\pm1}\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 $\varepsilon,\widehat{\varepsilon}>0$ (Shapiro, 2001).

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

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:

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

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

This trick of eliminating $p_{\pm 1},\ \widehat{p}_{\pm 1}$ is due to the auxiliary distributions $\mathbb{A}_{\pm 1}$ 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

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

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

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. □

10.4 Proof of Proposition 2

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 Shafieezadeh-Abadeh et al. (2019, Lemma 47), Shafieezadeh-Abadeh et al. (2023, Theorem 3.8), and Belbasi et al. (2023, 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 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 (Toland, 1978) to rewrite difference of convex functions optimization.

By using infimal convolutions (Rockafellar, 1997, 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 1 appear on the left-hand side of the constraints of Inter-ARO (*cf.* formulation in Proposition 1). These constraints would admit a tractable reformulation for the case without auxiliary data because the inf term in the reformulation presented in Lemma 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 $|\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}$ to force $\boldsymbol{z} = \boldsymbol{0}$, and the alternative formulation will thus be:

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

where we used the fact that $L = L^{**}$ and $\sup_{\theta \in \mathrm{dom}(L)} |\theta| = \mathrm{Lip}(L)$ since L is closed convex (Rockafellar, 1997, 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 (Ben-Tal et al., 2004; Yanıkoğlu et al., 2019) 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

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

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

$$\begin{aligned}
\underset{\boldsymbol{\beta}, \lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}}{\text{minimize}} \quad & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \frac{1}{N}\sum_{j=1}^{N} s_j + \frac{1}{\widehat{N}}\sum_{i=1}^{\widehat{N}} \widehat{s}_i \\
\text{subject to} \quad & \sup_{\theta \in \mathrm{dom}(L^*)} - L^{\alpha *}(\theta) + \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \inf_{\boldsymbol{z} \in \mathbb{R}^n} \{\boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \ :\ |\theta| \cdot \|\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda,\ |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}\} \leq \\
& \qquad s_i + \kappa \frac{1 - y^i}{2} \lambda + \widehat{s}_j + \kappa \frac{1 - \widehat{y}^j}{2} \widehat{\lambda} \quad \forall i \in [N],\ \forall j \in [\widehat{N}] \\
& \sup_{\theta \in \mathrm{dom}(L^*)} - L^{\alpha *}(\theta) - \theta \cdot \boldsymbol{\beta}^\top \boldsymbol{x}^i + \theta \cdot \inf_{\boldsymbol{z} \in \mathbb{R}^n} \{\boldsymbol{z}^\top (\widehat{\boldsymbol{x}}^j - \boldsymbol{x}^i) \ :\ |\theta| \cdot \|-\boldsymbol{\beta} - \boldsymbol{z}\|_{q^\star} \leq \lambda,\ |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}\} \leq \\
& \qquad s_i + \kappa \frac{1 + y^i}{2} \lambda + \widehat{s}_j + \kappa \frac{1 + \widehat{y}^j}{2} \widehat{\lambda} \quad \forall i \in [N],\ \forall j \in [\widehat{N}] \\
& \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda} \geq 0,\ \boldsymbol{s} \in \mathbb{R}^N_+,\ \widehat{\boldsymbol{s}} \in \mathbb{R}^{\widehat{N}}_+.
\end{aligned} \tag{5}$$

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

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

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 1 reduces to:

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

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

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

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

$$\begin{array}{ll}\underset{\boldsymbol{\beta},\lambda,\widehat{\lambda}}{\text{minimize}} & \varepsilon\lambda + \widehat{\varepsilon}\widehat{\lambda} + \sup\limits_{\boldsymbol{x}\in\mathbb{R}^n} \{S^\alpha(\boldsymbol{\beta}^\top \boldsymbol{x}) - \lambda\|\boldsymbol{x}^1 - \boldsymbol{x}\|_q - \widehat{\lambda}\|\widehat{\boldsymbol{x}}^1 - \boldsymbol{x}\|_q\} \\ \text{subject to} & \boldsymbol{\beta} \in \mathbb{R}^n,\ \lambda \geq 0,\ \widehat{\lambda} \geq 0, \end{array} \tag{6}$$

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

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

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 (6) to a generic two-stage robust optimization problem with random recourse (Subramanyam et al., 2020, Problem 1) which is proven to be NP-hard even if $S^{\alpha *}$ was constant (Guslitser, 2002). □

10.5 Proof of Theorem 1

Consider the reformulation Inter-adjustable of Inter-ARO that we introduced in the proof of Proposition 2. 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, Bertsimas et al. 2015). 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 $L^{\alpha **} = L^\alpha$ (Rockafellar, 1997, 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 (Rockafellar, 1997, 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. □

10.6 Proof of Corollary 2

To prove the first statement, take $\widehat{\lambda} = 0$ and observe the constraint $\|\boldsymbol{z}_{ij}^l\|_{q^\star} \leq \widehat{\lambda}$ implies $\boldsymbol{z}_{ij}^l = \boldsymbol{0}$ 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

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

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

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

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

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

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

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

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 $(\widehat{\lambda}, \widehat{\boldsymbol{s}}, \boldsymbol{z}^l_{ij})$ 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). $\square$

10.7 Proof of Observation 2

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 (2). Since we have

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

we can conclude the proof. □

10.8 Proof of Proposition 3

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 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 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 1) the minimum is well-defined (Villani et al., 2009, 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 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 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

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

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

10.9 Proof of Theorem 2

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 (Selvi et al., 2022, 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 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 Kuhn et al. (2019). The presented decay rate $\mathcal{O}(N^{-1/n})$ of ε as N increases is optimal (Fournier and Guillin, 2015). □

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 defined in Observation 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 5. *If the assumptions of Theorem 4 are satisfied and ε is chosen as in the statement of Theorem 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 Kuhn et al. (2019) given that $\ell^\alpha_{\boldsymbol{\beta}}$ is a finite-valued continuous loss function. □

Theorem 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[2].

Theorem 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 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 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 (Kuhn et al., 2019).

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

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

$M > 0$) and $\boldsymbol{\xi} = (\boldsymbol{x}, y) \in \Xi$ as:

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

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 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 (Yue et al., 2022). □

See the proof of Selvi et al. (2022, Theorem 8) and the discussion that follows for insights and further analysis on these results presented.

10.10 Proof of Theorem 3

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 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 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 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 8: Inter-ARO overestimates the true loss analogously as Theorem 5 with an identical proof.

11 EXPONENTIAL CONIC REFORMULATION 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 MOSEK ApS (2023), 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:

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

12 FURTHER DETAILS ON NUMERICAL EXPERIMENTS

12.1 UCI Experiments

Preprocessing UCI datasets We experiment on 10 UCI datasets (Kelly et al., 2023) (*cf.* Table 3). We use Python 3 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 3: 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 4 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 (Shafieezadeh-Abadeh et al., 2015). `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 5. 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.

Table 4: Mean (± 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% (± 2.89)	38.82% (± 2.86)	35.95% (± 3.78)	34.22% (± 2.70)	**32.64%** (± 2.54)
	0.20	73.65% (± 4.14)	51.49% (± 3.39)	49.56% (± 3.80)	45.61% (± 2.32)	**44.90%** (± 2.30)
annealing	0.05	18.08% (± 1.89)	16.61% (± 2.16)	14.97% (± 1.39)	13.50% (± 2.98)	**12.78%** (± 2.78)
	0.20	37.31% (± 3.92)	23.08% (± 2.82)	21.30% (± 1.93)	20.70% (± 1.32)	**19.53%** (± 1.42)
audiology	0.05	21.43% (± 3.64)	21.54% (± 3.92)	17.03% (± 2.90)	11.76% (± 3.28)	**9.01%** (± 3.54)
	0.20	37.91% (± 6.78)	29.34% (± 5.89)	20.44% (± 2.75)	20.00% (± 3.01)	**17.91%** (± 3.28)
breast-cancer	0.05	4.74% (± 1.26)	4.93% (± 1.75)	3.87% (± 1.17)	3.06% (± 0.79)	**2.52%** (± 0.50)
	0.20	9.93% (± 1.73)	8.14% (± 2.01)	6.09% (± 1.79)	5.04% (± 1.11)	**4.67%** (± 0.99)
contraceptive	0.05	44.14% (± 2.80)	42.86% (± 2.59)	40.98% (± 0.95)	40.00% (± 1.33)	**39.65%** (± 1.15)
	0.20	66.19% (± 5.97)	43.49% (± 2.24)	**42.71%** (± 1.47)	**42.71%** (± 1.47)	**42.71%** (± 1.47)
dermatology	0.05	15.97% (± 2.64)	16.46% (± 1.67)	13.47% (± 1.97)	12.78% (± 1.61)	**10.84%** (± 1.24)
	0.20	30.07% (± 4.24)	28.54% (± 3.25)	21.53% (± 2.17)	22.64% (± 2.15)	**20.21%** (± 1.58)
ecoli	0.05	16.30% (± 4.42)	14.67% (± 5.13)	13.26% (± 3.07)	11.11% (± 5.52)	**9.78%** (± 2.61)
	0.20	51.41% (± 3.37)	42.67% (± 2.91)	41.85% (± 2.95)	39.70% (± 2.68)	**38.89%** (± 2.57)
spambase	0.05	11.35% (± 0.77)	10.23% (± 0.54)	10.16% (± 0.56)	9.83% (± 0.37)	**9.81%** (± 0.38)
	0.20	27.32% (± 2.11)	15.83% (± 0.77)	15.70% (± 0.76)	15.67% (± 0.72)	**15.50%** (± 0.68)
spect	0.05	33.75% (± 5.17)	29.69% (± 5.46)	25.78% (± 3.06)	25.47% (± 3.38)	**21.56%** (± 2.74)
	0.20	54.22% (± 9.88)	37.5% (± 3.53)	35.16% (± 2.47)	33.75% (± 2.68)	**30.16%** (± 3.61)
prim-tumor	0.05	21.84% (± 4.55)	20.81% (± 3.97)	17.35% (± 3.59)	16.18% (± 3.83)	**14.78%** (± 2.89)
	0.20	34.19% (± 6.17)	25.37% (± 4.58)	21.62% (± 3.45)	21.84% (± 3.34)	**19.63%** (± 2.71)

Table 5: 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%**

12.2 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.

12.3 Artificial Experiments

Data generation We sample a 'true' $\boldsymbol{\beta}$ 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 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.

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

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: $(\boldsymbol{x}^i, y^i),\ i \in [N]$.

Table 6: Mean w in problem (2) 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

results in all attack regimes. Now we focus on the methods that rely 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 6 shows the average values of w for problem (2) 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%±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.

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.

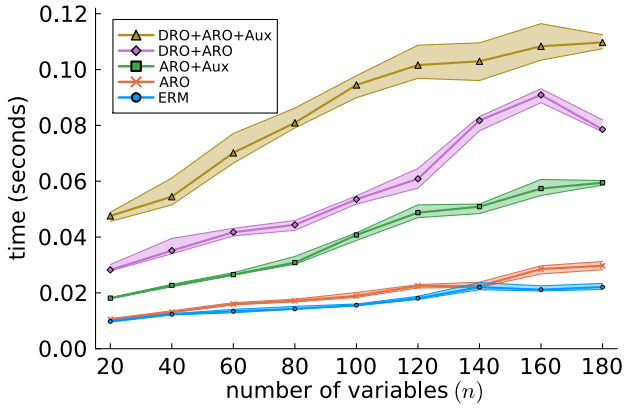

Figure 3: *Runtimes under a varying number of features in the artificially generated empirical and auxiliary datasets.*