CAN SYNTHETIC DATA REDUCE CONSERVATISM OF DISTRIBUTIONALLY ROBUST ADVERSARIAL TRAINING?

Anonymous authors
Paper under double-blind review

ABSTRACT

When the inputs of a machine learning model are subject to adversarial attacks, standard stationarity assumptions on the training and test sets are violated, typically making empirical risk minimization (ERM) ineffective. Adversarial training, which imitates the adversary during the training stage, has thus emerged as the *de facto* standard for hedging against adversarial attacks. Although adversarial training provides some robustness over ERM, it can still be subject to overfitting, which explains why recent work mixing the training set with synthetic data obtains improved out-of-sample performances. Inspired by these observations, we develop a Wasserstein distributionally robust (DR) counterpart of adversarial training for improved generalization and provide a recipe for further reducing the conservatism of this approach by adjusting its ambiguity set with respect to synthetic data. The underlying optimization problem, DR adversarial training with synthetic data, is nonconvex and comprises infinitely many constraints. To this end, by using results from robust optimization and convex analysis, we develop tractable relaxations. We focus our analyses on the logistic loss function and provide discussions for adapting this framework to several other loss functions. We demonstrate the superiority of this approach on artificial as well as standard benchmark problems.

1 INTRODUCTION

In recent years, there has been a surge of interest in utilizing synthetic data in order to improve adversarial robustness of ML models. Carmon et al. (2019) demonstrate that training a classifier with additional unlabelled data from the same distribution helps adversarial robustness. Deng et al. (2021) show how unlabelled data from a different domain/distribution improves adversarial robustness in the original domain. Sehwag et al. (2022) investigate how adversarial robustness of a classifier trained on synthetic data from a proxy distribution translates to the robustness on the real data, and highlight the importance of quantifying the distance between real and proxy data distribution. In comparison to data arising from a related domain/proxy distribution, the advantage of relying on a synthetic data generator trained on real data is that control over the distance between real and synthetic distribution often comes for free as a consequence of theoretical guarantees on the fidelity of the chosen generator (Goodfellow et al., 2014a; Arjovsky et al., 2017; Li et al., 2017). In particular, Wasserstein GAN (Arjovsky et al., 2017) for example achieves the closeness between the training data and the generator in terms of Wasserstein-1 distance and so the generator is guaranteed to live within a small ball around the training distribution.

The concept of adversarial robustness is designed to protect against adversarial attacks, however, it is typically still prone to overfitting Wong et al. (2020). On the other hand, for non-adversarial settings, distributionally robust optimization hedges against overfitting by learning over the worst-case data distribution realization from an ambiguity set (ball) built around the empirical (real data) distribution. In this paper, we explore how synthetic data helps us achieve both adversarial and distributional robustness. We find that synthetic data provides us a ‘direction’ along which to travel from the center of the ball in our search for the true distribution. That is, we rely on synthetic data in order to identify appropriate *fraction* of the ball and so *reduce the conservatism* of distributionally robust optimization. Thus, we aim to utilize synthetic data for model generalization.

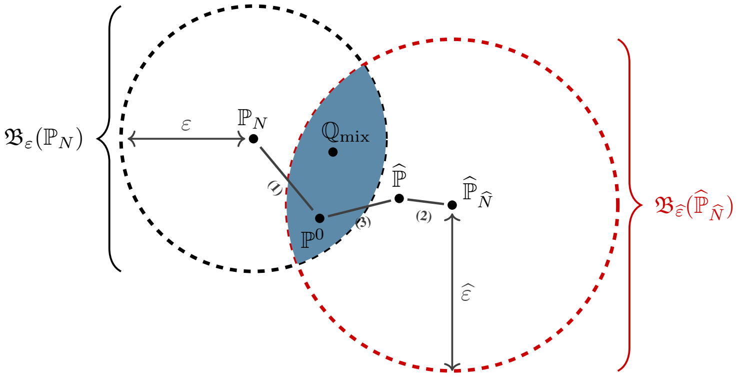

$\mathfrak{B}_\varepsilon(\mathbb{P}_N)$
ε
$\mathbb{P}_N$
$\mathbb{Q}_{\text{mix}}$
(1)
$\widehat{\mathbb{P}}$
$\widehat{\mathbb{P}}_{\widehat{N}}$
(2)
(3)
$\mathbb{P}^0$
$\widehat{\varepsilon}$
$\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$

Figure 1: *Adversarial training Adv optimizes the expected adversarial loss over the empirical distribution $\mathbb{P}_N$. Replacing $\mathbb{P}_N$ with the worst-case distribution in an ambiguity set $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$ built around it gives us the DR adversarial training problem AdvDRO. To reduce conservatism of $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$, we intersect it with another ball around synthetic empirical distribution $\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{P}_{\widehat{N}})$ (Section 4). This intersection includes $\mathbb{P}^0$ if $\widehat{\varepsilon}$ overestimates $(2) + (3)$ for which confidence bounds can be provided (*e.g.*, fidelity guarantees on the synthetic generator along with finite-sample statistics; see Appendix D.2). The recent literature on synthetic data for adversarial training (Gowal et al., 2021; Xing et al., 2022), instead of solving an ERM over $\mathbb{P}_N$, solves an ERM over $\mathbb{Q}_{\text{mix}}$ which is also included in the intersection $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ under some conditions on ε and $\widehat{\varepsilon}$ (*cf.* Lemma 3).*

The overarching question we address in this paper is: *What is the value of synthetic data in the context of model robustness?* See Figure 1 for an illustration of our problem setting and Appendix A for notation. Our contributions are:

- We show that logistic loss adversarial training is equivalent to empirical risk minimization (for a new loss function) and is consequently prone to overfitting.
- We provide a tractable convex optimization reformulation for both adversarially *and* Wasserstein distributionally robust optimization with logistic loss (*cf.* Section 3).
- To reduce the conservatism of adversarially and distributionally robust optimization, in Section 4, we derive a variant of it where the distributional ambiguity is replaced with the intersection of Wasserstein balls built around the real and synthetic data (*cf.* Figure 1). We provide discussions on when this approach is meaningful and provide a rigorous theoretical analysis to derive tractable approximation schemes by unifying the robust optimization, adversarial training, and synthetic data fields.
- We provide extensive experimentation to showcase the benefits of our approach and in particular on all the UCI datasets we considered, we beat all the competing state-of-the-art results.

2 Problem setting and preliminaries

In this work, we consider a binary classification problem where an instance is modeled as $(\boldsymbol{x}, y) \in \Xi := \mathbb{R}^n \times \{-1, +1\}$. We focus specifically on logistic regression as it provides a particularly favourable ground to explore the relationship between adversarial robustness and ERM. More precisely, the labels depend on the features probabilistically with

$$\text{Prob}[y \mid \boldsymbol{x}] \;=\; [1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x})]^{-1},$$

for some $\boldsymbol{\beta} \in \mathbb{R}^n$; its associated loss is the *logloss* function $\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y) := \log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x}))$. We review and discuss several training paradigms (*cf.* Table 1 for a summary).

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

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

Distributional ambiguity measures and sets Let $\mathcal{P}(\Xi)$ denote the set of probability distributions supported on Ξ. In this work, we employ the *Wasserstein distance* to model distributional ambiguity. Before defining the Wasserstein distance on $\mathcal{P}(\Xi)$, we introduce the following feature-label metric on Ξ.

Definition 1. *The distance $d(\boldsymbol{\xi}, \boldsymbol{\xi}')$ between $\boldsymbol{\xi} = (\boldsymbol{x}, y) \in \Xi$ and $\boldsymbol{\xi}' = (\boldsymbol{x}', y') \in \Xi$ is the standard feature-label metric*

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

for $\kappa > 0$ controlling the label weight and $q > 0$ specifying a rational norm on $\mathbb{R}^n$.

Using Definition 1, we next define the Wasserstein distance.

Definition 2. *The type-1 Wasserstein distance between distributions $\mathbb{Q} \in \mathcal{P}(\Xi)$ and $\mathbb{Q}' \in \mathcal{P}(\Xi)$, for the feature-label metric $d(\boldsymbol{\xi}, \boldsymbol{\xi}')$ on Ξ, is defined as*

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

For a fixed $\varepsilon > 0$, we define the Wasserstein ambiguity set of $\mathbb{P} \in \mathcal{P}(\Xi)$ as $\mathfrak{B}_\varepsilon(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}(\Xi) : \mathrm{W}(\mathbb{Q}, \mathbb{P}) \leq \varepsilon\}$.

Empirical Risk Minimization Let us denote by $\mathbb{P}^0$ the true data generating distribution, then one ideally wants to minimize the expected loss over $\mathbb{P}^0$, represented as the following problem

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

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

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\text{minimize}} \quad \mathbb{E}_{\mathbb{P}_N}[\ell_{\boldsymbol{\beta}}(\boldsymbol{x}, y)] = \frac{1}{N} \sum_{i \in [N]} \ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i, y^i). \tag{ERM}$$

Distributionally Robust Optimization It is well understood that in non-asymptotic settings, ERM may suffer from *overfitting* or *optimism bias* (Bishop, 2006; Hastie et al., 2009; Murphy, 2022; DeMiguel & Nogales, 2009; Michaud, 1989; Smith & Winkler, 2006). *Distributionally robust optimization* (DRO, Delage & Ye 2010) is an optimization approach attempting at addressing this issue. DRO is is motivated by the fact that in the finite-data setting, the ambiguity between the true and empirical distributions is positive but upper-bounded by some $\epsilon > 0$. When the ambiguity is measured by Wasserstein distance 2, this means that $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$. The goal in DRO is to optimize the expected loss over the worst possible realization of the true distribution in $\mathfrak{B}_\varepsilon(\mathbb{P}_N)$. That is,

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

We refer the readers to Mohajerin Esfahani & Kuhn (2018) and Kuhn et al. (2019a) for the generalization guarantees and finite sample statistics for DRO.

Adversarial Robustness Another popular approach to improve over the generalization performance of ERM is *adversarial robustness* where the goal is to provide robustness against *adversarial attacks* (Goodfellow et al., 2014b). An adversarial attack, in the widely studied l_p-noise setting (see e.g. 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, in adversarial training one can solve the following optimization problem in the training stage to hedge against adversarial perturbations at test time

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

Note that Adv reduces to ERM when $\alpha = 0$. It is worth noting that Adv is identical to feature robust training (Bertsimas et al., 2019) which does not have adversarial attacks but the training set comprises noisy observations of the features hence one employs robust optimization (Ben-Tal et al., 2009; Gorissen et al., 2015).

3 Distributionally Robust Adversarial Training

Problems Adv and ERM of logistic regression differ from one another in their objective functions. However, the following lemma shows that Problem Adv is nothing but an ERM problem for a different loss function.

Lemma 1. *Problem Adv is identical to*

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \quad \mathbb{E}_{\mathbb{P}_N}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y)],$$

where $\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y) := \log(1+\exp(-y\cdot\boldsymbol{\beta}^\top\boldsymbol{x} + \alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$ *is the* adversarial loss *associated with the logloss. Moreover, the corresponding univariate loss* $L^{\alpha}(z) := \log(1+\exp(-z+\alpha\cdot\|\boldsymbol{\beta}\|_{p^\star}))$ *satisfies* $\operatorname{Lip}(L^{\alpha}) = 1$ *for any* $\alpha > 0$.

Reducing the Adv to problem ERM with loss function $\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$ highlights that adversarial training is prone to overfitting. In order to circumvent this obstacle, we devote the remainder of this section to derive a distributionally robust counterpart of adversarial training. We introduce the following *distributionally and adversarially robust* optimization problem:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q}\in\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)} \quad \mathbb{E}_{\mathbb{Q}}\left[\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p\leq\alpha}\{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z}, y)\}\right]. \tag{AdvDRO}$$

Solving Problem AdvDRO requires the following assumption.

Assumption 1. *There exists* $\varepsilon > 0$ *known to the decision maker so that* $\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$.

We discuss relaxing Assumption 1 in Appendix§D.1. As a corollary of Lemma 1, AdvDRO has a tractable reformulation.

Corollary 1. *Problem AdvDRO admits the following tractable convex optimization reformulation:*

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

In Appendix C.3, we show how one can further rewrite AdvDRO as a conic optimization problem with the exponential cone and the cone associated with the $q^\star$-norm. For $q, q^\star \in \{1, 2, \infty\}$, the yielding problem can be solved with the exponential cone solver of MOSEK (MOSEK ApS, 2023a), in polynomial time (with respect to their input size, Nesterov 2018). We derive the regularized counterparts of these problems, and discuss that their complexity remain the same.

4 Reducing Conservatism with Synthetic Data

So far we have discussed the setting where we have access to the empirical distribution $\mathbb{P}_N$ that is constructed from N i.i.d. samples of $\mathbb{P}^0$. Suppose that we have an additional empirical distribution $\widehat{\mathbb{P}}_{\widehat{N}}$ which is constructed from $\widehat{N}$ i.i.d. samples $\{\widehat{\boldsymbol{\xi}}^j = (\widehat{\boldsymbol{x}}^j, \widehat{y}^j)\}_{j\in[\widehat{N}]}$ of another related but non-identical distribution $\widehat{\mathbb{P}}$. We first start with a strong and unrealistic assumption, that additional data is close enough to $\mathbb{P}^0$:

Assumption 2. *There exists $\widehat{\varepsilon} > 0$ known to the decision maker so that $W(\mathbb{P}^0, \widehat{\mathbb{P}}_{\widehat{N}}) \leq \widehat{\varepsilon}$.*

As mentioned earlier, DRO assumes $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)$. Under Assumption 2, it also follows that $\mathbb{P}^0 \in \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$, meaning that $\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$. We thus want to solve the following variant of AdvDRO where the ambiguity set is modeled as the intersection of two balls, hence providing us a weakly less conservative ambiguity set than of AdvDRO:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\text{minimize}} \quad \sup_{\mathbb{Q}\in\mathfrak{B}_\varepsilon(\mathbb{P}_N)\cap\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})} \quad \mathbb{E}_{\mathbb{Q}}[\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y)], \tag{Synth}$$

which reduces to AdvDRO when $\widehat{\varepsilon}$ is sufficiently large so that $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}}) = \mathfrak{B}_\varepsilon(\mathbb{P}_N)$. Note that for Synth to be well-defined, $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ needs to be non-empty (to be characterized in Section §4).

The inner problem in Synth involves optimizing the worst-case realization $\mathbb{Q}$ from an unusual constraint set in the distribution space. To avoid the difficulty of solving such a problem, we borrow techniques from the *adjustable robust optimization* literature to find a tractable convex relaxation of Problem Synth. Discussion on the difficulty of directly solving Synth and the need for our relaxation is in Appendix §B. The next theorem is our main theoretical contribution, which presents a conservative relaxation to Problem Synth.

Theorem 1. *Problem Synth admits the following tractable convex conservative relaxation:*

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

This formulation admits an exponential cone reformulation, with the same techniques applied to AdvDRO which is summarized in Appendix §C.3. Moreover, in Appendix §C.9, we review tailored methods for norms with $q \in \{1, 2, \infty\}$ as well as present alternative relaxations using techniques from robust optimization.

The case of uninformative synthetic data When the synthetic data does not provide any useful information, we can select $\widehat{\epsilon}$ large enough to have $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}}) = \mathfrak{B}_\varepsilon(\mathbb{P}_N)$. In this case, problem Synth simply reduces to Problem

AdvDRO. The next lemma discusses the same behavior for the relaxed problem ProxSynth, meaning that the presented relaxation will not force learning from the synthetic data, and recovering AdvDRO remains a feasible solution.

Lemma 2. *As $\widehat{\varepsilon} \to \infty$, the optimal value of ProxSynth converges to the optimal value of AdvDRO.*

The case of unknown ϵ and $\widehat{\epsilon}$ Let us discuss how to tune the parameters ε and $\widehat{\varepsilon}$ when they are unknown. This discussion requires understanding the statistical properties of AdvDRO and Synth, which we provide in Appendix §D. First, recall that we need $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}}) \neq \emptyset$ for Synth to be well-defined. To ensure this, a sufficient condition follows from the triangle inequality: $\varepsilon + \widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}})$.

Let us now discuss tuning ε. We aim to find a tight ε value so that $\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$ with high confidence. To this end, we use the rich arsenal of finite-sample statistics from the Wasserstein DRO literature. In Appendix D.1, we review existing results applicable to AdvDRO, including the optimal value of ε given a confidence level for including $\mathbb{P}^0$ in $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N)$, asymptotic consistency of AdvDRO, and the existence of sparse worst-case distributions in the nature's problem.

On the other hand, deciding $\widehat{\varepsilon}$ is a much more challenging task since we want to have $\mathbb{P}^0 \in \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$, but $\widehat{\mathbb{P}}_{\widehat{N}}$ is constructed of i.i.d. samples of another distribution $\widehat{\mathbb{P}}$. Thus, one needs to estimate both $\widehat{\varepsilon}_1 := \mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \widehat{\mathbb{P}})$ and $\widehat{\varepsilon}_2 := \mathrm{W}(\widehat{\mathbb{P}}, \mathbb{P}^0)$. We then would choose $\widehat{\varepsilon} \geq \widehat{\varepsilon}_1 + \widehat{\varepsilon}_2$ to include $\mathbb{P}^0$ in $\mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$. In Appendix D.2, we first assume that $\widehat{\varepsilon}_2$ is known, and show that Synth enjoys an optimal characterization of ε and $\widehat{\varepsilon}$ values that guarantee $\mathbb{P}^0 \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ with arbitrarily high confidence. Note that when no knowledge of $\widehat{\varepsilon}_2$ exists, we use cross-validation to tune this parameter.

Connection to literature on synthetic data for adversarial training Here, we investigate the literature on using synthetic data $\{(\widehat{\boldsymbol{x}}^j, \widehat{y}^j)\}_{j \in [\widehat{N}]}$ for adversarial training and relate it to the problem Synth we propose. The works in this literature (Gowal et al., 2021; Xing et al., 2022) propose solving the following problem, for some $w > 0$.

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{N + w \cdot \widehat{N}} \left[\sum_{i \in [N]} \sup_{\boldsymbol{z}^i : \|\boldsymbol{z}^i\|_p \leq \alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}^i + \boldsymbol{z}^i, y^i)\} + w \cdot \sum_{j \in [\widehat{N}]} \sup_{\boldsymbol{z}^j : \|\boldsymbol{z}^j\|_p \leq \alpha} \{\ell_{\boldsymbol{\beta}}(\widehat{\boldsymbol{x}}^j + \boldsymbol{z}^j, \widehat{y}^j)\} \right], \qquad (1)$$

Proposition 1. *Problem* (1) *is equivalent to* $\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \mathbb{E}_{\mathbb{Q}_{\text{mix}}}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$ *where* $\mathbb{Q}_{\text{mix}} := \lambda \mathbb{P}_N + (1 - \lambda)\widehat{\mathbb{P}}_{\widehat{N}}$ *for* $\lambda = \frac{N}{N + w \cdot \widehat{N}}$.

The following lemma shows that under reasonable conditions for ε and $\widehat{\varepsilon}$, the mixture distribution introduced in Proposition 1 is included in $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$. This means that the distribution $\mathbb{Q}_{\text{mix}}$ used in the literature on synthetic data for adversarial training, belongs to the set of distributions considered in Problem Synth.

Lemma 3. *For any $\lambda \in (0, 1)$ and distribution $\mathbb{Q}_{\text{mix}} := \lambda \cdot \mathbb{P}_N + (1 - \lambda) \cdot \widehat{\mathbb{P}}_{\widehat{N}}$, we have:*

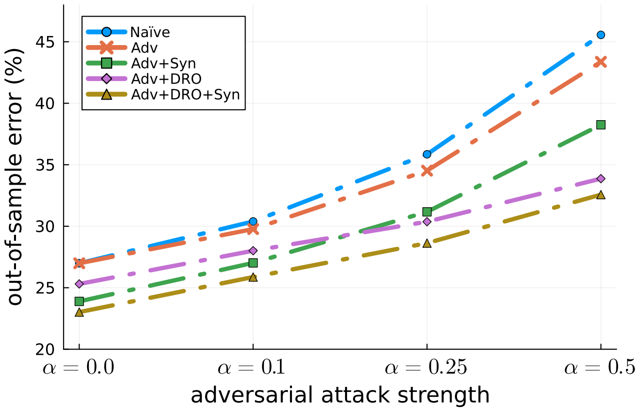
$$\varepsilon + \widehat{\varepsilon} \geq \mathrm{W}(\mathbb{P}_N, \widehat{\mathbb{P}}_{\widehat{N}}) \ \textit{ and } \ \frac{\lambda}{1 - \lambda} = \frac{\widehat{\varepsilon}}{\varepsilon} \implies \mathbb{Q}_{\text{mix}} \in \mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}}).$$

For $\lambda = \frac{N}{N + \widehat{N}}$, provided that $\mathfrak{B}_{\varepsilon}(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ is nonempty, Lemma 3 shows that a sufficient condition for this intersection to include the mixture $\mathbb{Q}_{\text{mix}}$ is $\widehat{\varepsilon}/\varepsilon = N/\widehat{N}$, which is intuitive since the radii of the Wasserstein ambiguity sets are typically chosen inversely proportional to the number of samples (Kuhn et al., 2019b, Theorem 18).

5 Experiments

We compare the proposed DRO problem over the intersection of empirical and synthetic Wasserstein balls with several benchmark methods. Code and more experiments are on an anonymous repo (we will share the URL with reviewers when discussion opens). The following abbreviations will be used throughout the experiments:

- Naïve: Solving ERM with the logloss function;



Case	CV'ed w	CV'ed $\varepsilon/\widehat{\varepsilon}$
$\alpha = 0$	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

Figure 2: *Left: mean out-of-sample error rates of each method over different attack strengths. Right: average importance of synthetic data for empirical and robust risk minimization.*

- Adv: Adversarial training Adv on the empirical distribution (*e.g.*, Madry et al. 2017);
- Adv+Syn: Solving (1) for adversarial training on mixtures of real and synthetic data (*e.g.*, Gowal et al. 2021);
- Adv+DRO: Distributionally robust (DR) counterpart AdvDRO of Adv with the ambiguity set centered at $\mathbb{P}_N$;
- Adv+DRO+Syn: DR counterpart Synth of Adv over intersection of ambiguity sets built around $\mathbb{P}_N$ and $\widehat{\mathbb{P}}_{\widehat{N}}$.

Artificial Data We sample instances i.i.d. from a generated artificial ground truth distribution to form training and test sets. Similarly, we sample instances i.i.d. from another related but different ground truth distribution to form synthetic sets. The details are in Appendix §E. We simulate 25 cases, each with $N = 100$ training instances, $\widehat{N} = 200$ synthetic instances, and $N_{\text{te}} = 10,000$ test instances. For each case, we further simulate different adversarial attack strengths $\alpha \in \{0, 0.1, 0.25, 0.5\}$ using ℓ_2 attacks (ℓ_1 and ℓ_∞ are also available in our repository with similar implementation and results). In each problem, we apply 5-fold cross-validation to select the best parameter combinations for both our and the benchmark methods (the specific parameters and grids for validation are shared in our repository). The performance in terms of the out-of-sample misclassification (error) rates is available on the left-hand side of Figure 2. As expected, Naïve training provides the worst out-of-sample errors. While Adv improves these errors, Adv+Syn and Adv+DRO enhance its performance further by mixing the empirical distribution with synthetic data and by applying DRO, respectively. Notice also that in low attack regimes using synthetic data is more important than adding distributional robustness, where a higher attack regime makes distributional robustness more important. Finally, Adv+DRO+Syn always provides the lowest errors by applying both these techniques simultaneously (that is, by solving Synth). On the right-hand side of Figure 2, we display the average CV'ed w values for problem (1) which shows that the greater the attack strength is the more we should use the synthetic data; same relationship holds for $\varepsilon/\widehat{\varepsilon}$ in Synth, which means that the relative size of the Wasserstein ball built around the empirical distribution gets larger compared to the same ball around the synthetic data. We remark this as a possible future research direction exploring whether a larger attack *per se* implies the intersection of the Wasserstein ball will move towards the synthetic data distribution.

Since the relaxation problems we solve for Synth comprise $\mathcal{O}(n \cdot N \cdot \widehat{N})$ variables and exponential cone constraints, it is natural to ask whether the method scales to practical settings. Indeed, we find that in practice our methods scale better than that. To this end, we simulate problems with the same procedure as explained in the performance experiments above and report the mean runtimes in Figure 3. On the left-hand side of this figure, we simulate 25 cases that fix $\alpha = 0.2$, $N = 200$, $\widehat{N} = 200$ and vary $n \in \{20, 40, \ldots, 180\}$. On the right-hand side, we similarly simulate 25 cases that instead fix $n = 100$ and vary $N = \widehat{N} \in \{100, 200, \ldots, 1000\}$. We observe that the fastest methods are Naïve and Adv among which the faster one depends on n (as the adversarial loss includes a regularizer of $\boldsymbol{\beta}$), followed by Adv+Syn, Adv+DRO, and Adv+DRO+Syn, respectively. The slowest method is Adv+DRO+Syn, which is expected

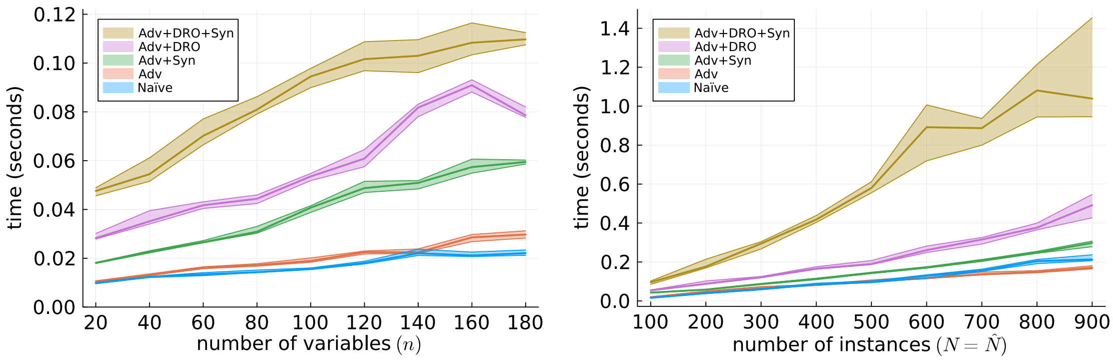


Figure 3: *Runtimes over varying problem dimensions. Left:* $N = \widehat{N} = 200$ *is fixed. Right:* $n = 100$ *is fixed.*

Data Set	N	$\widehat{N}$	N_{te}	n	Attack	Naïve	Adv	Adv+Syn	Adv+DRO	Adv+DRO+Syn
absent	111	333	296	74	$\alpha = 0.05$	44.02% (± 2.89)	38.82% (± 2.86)	35.95% (± 3.78)	34.22% (± 2.70)	**32.64%** (± 2.54)
					$\alpha = 0.20$	73.65% (± 4.14)	51.49% (± 3.39)	49.56% (± 3.80)	45.61% (± 2.32)	**44.90%** (± 2.30)
annealing	134	404	360	41	$\alpha = 0.05$	18.08% (± 1.89)	16.61% (± 2.16)	14.97% (± 1.39)	13.50% (± 2.98)	**12.78%** (± 2.78)
					$\alpha = 0.20$	37.31% (± 3.92)	23.08% (± 2.82)	21.30% (± 1.93)	20.70% (± 1.32)	**19.53%** (± 1.42)
audiology	33	102	91	102	$\alpha = 0.05$	21.43% (± 3.64)	21.54% (± 3.92)	17.03% (± 2.90)	11.76% (± 3.28)	**9.01%** (± 3.54)
					$\alpha = 0.20$	37.91% (± 6.78)	29.34% (± 5.89)	20.44% (± 2.75)	20.00% (± 3.01)	**17.91%** (± 3.28)
breast-cancer	102	307	274	90	$\alpha = 0.05$	4.74% (± 1.26)	4.93% (± 1.75)	3.87% (± 1.17)	3.06% (± 0.79)	**2.52%** (± 0.50)
					$\alpha = 0.20$	9.93% (± 1.73)	8.14% (± 2.01)	6.09% (± 1.79)	5.04% (± 1.11)	**4.67%** (± 0.99)
contraceptive	220	663	590	23	$\alpha = 0.05$	44.14% (± 2.80)	42.86% (± 2.59)	40.98% (± 0.95)	40.00% (± 1.33)	**39.65%** (± 1.15)
					$\alpha = 0.20$	66.19% (± 5.97)	43.49% (± 2.24)	**42.71%** (± 1.47)	**42.71%** (± 1.47)	**42.71%** (± 1.47)
dermatology	53	161	144	99	$\alpha = 0.05$	15.97% (± 2.64)	16.46% (± 1.67)	13.47% (± 1.97)	12.78% (± 1.61)	**10.84%** (± 1.24)
					$\alpha = 0.20$	30.07% (± 4.24)	28.54% (± 3.25)	21.53% (± 2.17)	22.64% (± 2.15)	**20.21%** (± 1.58)
ecoli	50	151	135	9	$\alpha = 0.05$	16.30% (± 4.42)	14.67% (± 5.13)	13.26% (± 3.07)	11.11% (± 5.52)	**9.78%** (± 2.61)
					$\alpha = 0.20$	51.41% (± 3.37)	42.67% (± 2.91)	41.85% (± 2.95)	39.70% (± 2.68)	**38.89%** (± 2.57)
spambase	690	2,070	1,841	58	$\alpha = 0.05$	11.35% (± 0.77)	10.23% (± 0.54)	10.16% (± 0.56)	9.83% (± 0.37)	**9.81%** (± 0.38)
					$\alpha = 0.20$	27.32% (± 2.11)	15.83% (± 0.77)	15.70% (± 0.76)	15.67% (± 0.72)	**15.50%** (± 0.68)
spect	24	72	64	23	$\alpha = 0.05$	33.75% (± 5.17)	29.69% (± 5.46)	25.78% (± 3.06)	25.47% (± 3.38)	**21.56%** (± 2.74)
					$\alpha = 0.20$	54.22% (± 9.88)	37.5% (± 3.53)	35.16% (± 2.47)	33.75% (± 2.68)	**30.16%** (± 3.61)
primacy-tumor	50	153	136	32	$\alpha = 0.05$	21.84% (± 4.55)	20.81% (± 3.97)	17.35% (± 3.59)	16.18% (± 3.83)	**14.78%** (± 2.89)
					$\alpha = 0.20$	34.19% (± 6.17)	25.37% (± 4.58)	21.62% (± 3.45)	21.84% (± 3.34)	**19.63%** (± 2.71)

Table 2: *Mean out-of-sample errors (± standard deviation) over* 10 *UCI datasets.*

given that Adv+DRO is a special case with large $\widehat{\varepsilon}$; however, the runtime is observed to scale graciously. To further concentrate on the runtime of Adv+DRO, we simulated cases with $n = 1,000$ and $N = \widehat{N} = 10,000$, and recorded that the runtimes vary between 134 to 232 seconds.

Real Data We compare the out-of-sample error rates of each method on 10 of the most popular UCI datasets for classification (Dua & Graff, 1998). Classification problems with more than two classes are converted to binary classification problems (most frequent class/others). The numerical features are standardized, the ordinal categorical features are left as they are, and the nominal categorical features are processed via one-hot encoding. Further details about preprocessing are documented in our repository. For each dataset, we applied 10 simulations of the following procedure: *(i)* Select 40% of the dataset as a test set ($N_{\text{te}} \propto 0.4$); *(ii)* Sample 25% of the remaining instances 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 distribution by using the Gaussian Copula from the SDV Patki et al. (2016). We sample $\widehat{N}$ synthetic instances and then 5-fold cross-validate the parameters of each method (to have equal comparison, we used the same grids with the benchmark methods) similarly to the previous experiments. The mean errors on the test set are reported in Table 2 for attacks $\alpha = 0.05$ and $\alpha = 0.20$. The best error is always achieved by Adv+DRO+Syn, followed by Adv+DRO, Adv+Syn, Adv, Naïve, respectively. Adv+DRO achieves better errors than Adv+Syn (as observed from the mean errors) but there are cases where this is reversed. On the other hand, we verify that there is no case where any method is better than Adv+DRO+Syn.

6 Related work

Interactions between adversarial and distributional robustness are not new. Sinha et al. (2017), for example, show that distributional robust optimization over Wasserstein balls is intractable for generic functions (*e.g.*, neural networks) but its Lagrange relaxation resembles adversarial training hence applying adversarial training still gives some distributional robustness guarantees. Similarly, there have been works showing how one can obtain adversarial robustness via distributionally robust optimization (Regniez et al., 2022). However, to our knowledge, there has not been any work addressing Wasserstein distributional robustness (that hedges against overfitting, Kuhn et al. 2019b) and adversarial robustness (that hedges against adversarial attacks, Goodfellow et al. 2014b) simultaneously. To our knowledge, the only existing work that considers generalization and adversarial robustness together is the work of Bennouna et al. (2023) where the distributional ambiguity is modeled with φ-divergences and the prediction model is a neural network.

The idea of intersecting Wasserstein balls is inspired by the "Surround, then Intersect" (SI) strategy (Taskesen et al., 2021, §5.2) to learn linear regression coefficients under sequential domain adaptation (see Shafahi et al. 2020 and Song et al. 2018 for a deeper understanding of robustness in domain adaptation/transfer learning). The authors use a Wasserstein metric constructed around the first and second moments and show that problem Synth admits a tractable convex optimization reformulation under this metric. Their proof uses minimax equality which exploits the squared loss function. Indeed, we showed that in the logistic classification setting, the distributionally robust adversarial training (classification) problem becomes substantially harder, and we thus developed different duality techniques (*e.g.*, Toland's duality) and relaxations (*e.g.*, adjustable robust optimization). Recent work started to explore the intersection of ambiguity sets in different contexts (Awasthi et al., 2022) or different setups (Zhang et al., 2023). Given the increasing interest in this stream of research, we hope our work and analysis will be complementary to the theory. For a deeper understanding of domain adaptation (and transfer learning in general), we refer the interested reader to some works from transfer learning and domain adaptation literature.

7 Conclusion

One of the two motivations of this study was that adversarial training is subject to overfitting, which led us to formulate and experiment with the adversarially and distributionally robust training models. In the numerical results, we observed that adversarial training is improved by taking its Wasserstein DR counterpart. This is due to the same reason behind the success of recent studies that mix empirical data with synthetic data, as using synthetic data, to some extent, regularizes decisions. Motivated by these observations, our main contribution was to unify the practices of adversarial training, DRO, and using synthetic data: instead of achieving adversarial robustness with synthetic data, we achieve it by distributional robustness and use the synthetic data to manipulate the underlying ambiguity sets. The experiments show a drastic improvement over the benchmark methods.

Despite the appealing performance, one must also be mindful that the proposed model needs to tune an additional parameter ($\widehat{\varepsilon}$) that can recover as special cases the DRO problem ($\widehat{\varepsilon} \gg \varepsilon$) and the adversarial training problem ($\varepsilon = 0$ with large $\widehat{\varepsilon}$), and can even resemble non-robust adversarial training with synthetic data when the boundaries of the Wasserstein balls intersect. Regardless, we observe that in most cases this model achieves much better error rates than any of these special cases. Moreover, we focused on the logloss function to show the theoretical challenges and present tractable reformulation/approximation techniques; it could be fruitful to extend these results to neural network classifiers as is typical for theoretical studies stemming from logistic regression (Dreiseitl & Ohno-Machado, 2002). The extension of our findings to different loss functions is sketched in the appendices.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/arjovsky17a.html`.

Pranjal Awasthi, Christopher Jung, and Jamie Morgenstern. Distributionally robust data join. *arXiv preprint arXiv:2202.05797*, 2022.

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

Aharon Ben-Tal, Alexander Goryashko, Elana Guslitzer, and Arkadi Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical programming*, 99(2):351–376, 2004.

Amine Bennouna, Ryan Lucas, and Bart Van Parys. Certified robust neural networks: Generalization and corruption resistance. *arXiv preprint arXiv:2303.02251*, 2023.

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

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

Dimitris Bertsimas, Jack Dunn, Colin Pawlowski, and Ying Daisy Zhuo. Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34, 2019.

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

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

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/32e0bd1497aa43e02a42f47d9d6515ad-Paper.pdf`.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

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

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

Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. Improving adversarial robustness via unlabeled out-of-domain data. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2845–2853. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/deng21b.html`.

Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359, 2002.

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

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

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014a. URL `https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf`.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Bram L Gorissen, İhsan Yanıkoğlu, and Dick Den Hertog. A practical guide to robust optimization. *Omega*, 53:124–137, 2015.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

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

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

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

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pp. 130–166. Informs, 2019b.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: towards deeper understanding of moment matching network. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2203–2213, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/dfd7468ac613286cdbb40872c8ef3b06-Abstract.html`.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

R. O. Michaud. The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42, 1989.

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

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

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

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

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

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410, 2016. doi: 10.1109/DSAA.2016.49.

Chiara Regniez, Gauthier Gidel, and Hugo Berard. A distributional robustness perspective on adversarial training with the ∞-wasserstein distance, 2022. URL `https://openreview.net/forum?id=z7DAilcTx7`.

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

Ernst Roos, Dick den Hertog, Aharon Ben-Tal, FJCT de Ruiter, and Jianzhe Zhen. Tractable approximation of hard uncertain optimization problems. *Available on Optimization Online*, 2018.

Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=WVX0NNVBBkV`.

Aras Selvi, Mohammad Reza Belbasi, Martin Haugh, and Wolfram Wiesemann. Wasserstein logistic regression with mixed features. *Advances in Neural Information Processing Systems*, 35:16691–16704, 2022.

Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning, 2020.

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

Soroosh Shafieezadeh-Abadeh, Liviu Aolaritei, Florian Dörfler, and Daniel Kuhn. New perspectives on regularization and computation in optimal transport-based distributionally robust optimization. *arXiv preprint arXiv:2303.03900*, 2023.

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

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

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

Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Improving the generalization of adversarial training with domain adaptation. *arXiv preprint arXiv:1810.00740*, 2018.

Bahar Taskesen, Man-Chung Yue, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. Sequential domain adaptation by synthesizing distributionally robust experts. In *International Conference on Machine Learning*, pp. 10162–10172. PMLR, 2021.

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

Ioannis Tsaknakis, Mingyi Hong, and Shuzhong Zhang. Minimax problems with coupled linear constraints: computational complexity, duality and solution methods. *arXiv preprint arXiv:2110.11210*, 2021.

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

Alex L Wang and Fatma Kilinc-Karzan. A geometric view of sdp exactness in qcqps and its applications. *arXiv preprint arXiv:2011.07155*, 2020.

Eric Wong, Leslie Rice, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, 2020. URL `https://api.semanticscholar.org/CorpusID:211505864`.

Yue Xing, Qifan Song, and Guang Cheng. Why do artificially generated data help adversarial robustness. *Advances in Neural Information Processing Systems*, 35:954–966, 2022.

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

Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. On linear optimization over wasserstein balls. *Mathematical Programming*, 195(1-2):1107–1122, 2022.

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

Erheng Zhong, Wei Fan, Qiang Yang, Olivier Verscheure, and Jiangtao Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pp. 547–562. Springer, 2010.

Can Synthetic Data Reduce Conservatism of Distributionally Robust Adversarial Training?

Supplementary Materials

A Notation

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

B Discussion on solving Problem Synth

The following lemma transforms problem Synth into a minimization problem with vector variables instead of $\mathbb{Q}$. To this end, assume that Synth is well-defined, that is, $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})$ is non-empty (to be characterized in §D).

Lemma 4. *For $\varepsilon,\ \widehat{\varepsilon} > 0$ satisfying $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}}) \neq \emptyset$, the inner* sup *problem (nature's problem) of Synth admits the following reformulation:*

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

Although the equivalent problem presented in Lemma 4 optimizes a linear objective function over $\mathcal{O}(N \cdot \widehat{N})$ many constraints, the sup terms on the left-hand side of these constraints are maximization of difference of convex (DC) functions. We first reformulate these constraints by using techniques from adjustable robust optimization (Ben-Tal et al., 2004; Yanıkoğlu et al., 2019). To this end, by using the univariate-loss formulation $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y) = L^\alpha(y \cdot \boldsymbol{\beta}^\top \boldsymbol{x})$, and by representing $t(\lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}})$ as a linear function, we can model any of the above constraints as

$$\sup_{\boldsymbol{x} \in \mathbb{R}^n} \{L^\alpha(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \cdot \|\boldsymbol{a} - \boldsymbol{x}\|_q - \widehat{\lambda} \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{x}\|_q\} \leq t(\lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}) \tag{2}$$

for $\boldsymbol{\omega} = \pm\boldsymbol{\beta}$, $\boldsymbol{a} = \boldsymbol{x}^i$ and $\widehat{\boldsymbol{a}} = \widehat{\boldsymbol{x}}^j$ for some $i \in [N]$ and $j \in [\widehat{N}]$. The next lemma shows how to write this constraint as an adjustable robust optimization constraint.

Lemma 5. *The constraint* (2) *and the following system define the identical feasible set for the variables* $\boldsymbol{\omega}, \lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}$ *where* $L^{\alpha*}$ *denotes the conjugate of* L^α*:*

$$\forall \theta \in \operatorname{dom}(L^{\alpha*}),\ \exists \boldsymbol{z} \in \mathbb{R}^n : \begin{cases} -L^{\alpha*}(\theta) + \theta \cdot \boldsymbol{\omega}^\top \boldsymbol{a} + \theta \cdot \boldsymbol{z}^\top (\widehat{\boldsymbol{a}} - \boldsymbol{a}) \leq t(\lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}) \\ |\theta| \cdot \|\boldsymbol{\omega} - \boldsymbol{z}\|_{q^\star} \leq \lambda \\ |\theta| \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}. \end{cases}$$

This is a generalization of the existing reformulations for DC constraints with a single norm (*e.g.*, Lemma 47 of Shafieezadeh-Abadeh et al. 2019 and Theorem 3.8 of Shafieezadeh-Abadeh et al. 2023) to multiple norms. Indeed, consider the case of a single norm, that is, $\widehat{\lambda} = 0$. As this will enforce $\boldsymbol{z} = \boldsymbol{0}$, the system presented in Lemma 5 will reduce to:

$$\forall \theta \in \operatorname{dom}(L^{\alpha*}) : \begin{cases} -L^{\alpha*}(\theta) + \theta \cdot \boldsymbol{\omega}^\top \boldsymbol{a} \leq t(\lambda, 0, \boldsymbol{s}, \boldsymbol{0}) \\ |\theta| \cdot \|\boldsymbol{\omega}\|_{q^\star} \leq \lambda \end{cases}$$

which will not be feasible if $\sup_{\theta \in \operatorname{dom}(L^{\alpha*})}\{|\theta|\} \cdot \|\boldsymbol{\omega}\|_{q^\star} > \lambda$. This system is only feasible when $\sup_{\theta \in \operatorname{dom}(L^{\alpha*})}\{|\theta|\} \cdot \|\boldsymbol{\omega}\|_{q^\star} = \operatorname{Lip}(L^\alpha) \cdot \|\boldsymbol{\omega}\|_{q^\star} \leq \lambda$ where the identity is due to (Rockafellar, 1997, Corollary 13.3.3). Thus, provided that $\operatorname{Lip}(L^\alpha) \cdot \|\boldsymbol{\omega}\|_{q^\star} \leq \lambda$ holds, it reduces to

$$\underbrace{\sup_{\theta \in \operatorname{dom}(L^{\alpha*})} \{-L^{\alpha*}(\theta) + \theta \cdot \boldsymbol{\omega}^\top \boldsymbol{a}\}}_{= L^\alpha(\boldsymbol{\omega}^\top \boldsymbol{a})} \leq t(\lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}).$$

This discussion implies that, for $\widehat{\lambda} = 0$, an equivalent set of constraints to represent (2) is:

$$\begin{cases} L^\alpha(\boldsymbol{\omega}^\top \boldsymbol{a}) \leq t(\lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}) \\ \operatorname{Lip}(L^\alpha) \cdot \|\boldsymbol{\omega}\|_{q^\star} \leq \lambda \end{cases}$$

which coincides with the aforementioned tractable reformulations. For the general case, however, such a tractable reformulation is not possible due to the fact that the uncertain parameter θ and the adjustable variable $\boldsymbol{z}$ share constraints, and traditional methods including minimax theorem do not hold (Tsaknakis et al., 2021). For such problems, a common practice is to seek for tight relaxations, and we thus propose the following safe approximation to constraint (2) by using the formulation presented in Lemma 5 by employing the static relaxation of the adjustable variables (Bertsimas et al., 2015).

Corollary 2. *The following system, upon the inclusion of the new variable $\boldsymbol{z} \in \mathbb{R}^n$, gives a safe approximation*[1] *of constraint* (2)*:*

$$\begin{cases} L^{\alpha}(\boldsymbol{\omega}^\top \boldsymbol{a} + \boldsymbol{z}^\top(\widehat{\boldsymbol{a}} - \boldsymbol{a})) \leq t(\lambda, \widehat{\lambda}, \boldsymbol{s}, \widehat{\boldsymbol{s}}) \\ \underbrace{\mathrm{Lip}(L^{\alpha})}_{=1} \cdot \|\boldsymbol{w} - \boldsymbol{z}\|_{q^\star} \leq \lambda \\ \underbrace{\mathrm{Lip}(L^{\alpha})}_{=1} \cdot \|\boldsymbol{z}\|_{q^\star} \leq \widehat{\lambda}. \end{cases}$$

We now present a tractable convex reformulation to have a safe approximation for Synth.

C PROOFS

C.1 PROOF OF LEMMA 1

Proof. For any $\boldsymbol{\beta} \in \mathbb{R}^n$, using standard robust optimization arguments (Ben-Tal et al., 2009; Bertsimas & 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 ellipsoid, the fourth step uses the definition of the 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 Adv with $\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$ concludes the equivalence.

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

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

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

[1] the terminology of 'safe approximation' is borrowed from the robust optimization literature, meaning that the feasibility of this approximation will imply the feasibility of the original system

C.2 Proof of Corollary 1

Lemma 1 lets us rewrite AdvDRO as the DRO counterpart of empirical minimization of ℓ^{α}:

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

Since the univariate loss $L^{\alpha}(z) := \log(1 + \exp(-z + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$ satisfying the identity $L^{\alpha}(\langle y \cdot \boldsymbol{x}, \boldsymbol{\beta}\rangle) = \ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y)$ is Lipschitz continuous (*cf.* Lemma 1), Theorem 14 *(ii)* of Shafieezadeh-Abadeh et al. (2019) is immediately applicable. We can therefore rewrite (ADV) as:

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

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

C.3 Exponential Cone Representation of AdvDRO

Lemma 6. *Problem AdvDRO is equivalent to*

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

Proof. For any $i \in [N]$, the first two constraints of AdvDRO 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 (2023b), 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 $\mathcal{K}_{\text{exp}}$. Similarly, for new variables $v_i^-, w_i^- \in \mathbb{R}$, the second constraint can be written as

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

Applying this for all $i \in [N]$ concludes the derivation. □

C.4 Distributionally and Adversarially Robust Regularized Model

Remark 1 (Regularization). *Our discussions on robustness extend to include regularization:*

(i) The problems ERM, Adv, DRO, and AdvDRO are analogously extended to their regularized variants. We add $\lambda_r \cdot \|\boldsymbol{\beta}\|_r$ to the objective functions where $r \in \{1, 2\}$ specifies the LASSO- or Ridge-regularizer, respectively, and $\lambda_r > 0$ is the regularization penalty.

(ii) The problem AdvDRO reduces to a variant of Adv with an additional regularizer $\varepsilon \cdot \|\boldsymbol{\beta}\|_{q^\star}$ in the objective when $\kappa \to \infty$, but it is different in nature for $\kappa < \infty$ (Shafieezadeh-Abadeh et al., 2019, Remark 18).

(iii) If there are additional categorical features $\boldsymbol{z}$ in the training set, one can convert them via dummy encoding. Alternatively, to get better results, one can treat them separately by adding the number of disagreeing categorical variables to the feature-label metric (cf. Definition 1). Although the latter case brings exponentially many constraints to AdvDRO, one can adopt the column-and-constraint algorithm of Selvi et al. (2022).

For the completeness of Remark 1 *(i)*, we include the adversarially and distributionally robust regularized model next. Adding the regularization term to AdvDRO gives us:

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

As noted in the remark, despite the repetitive appearance of the norm of $\boldsymbol{\beta}$, regularization and distributional robustness do not reduce to each other as the last constraint is not necessarily tight unless $\kappa \to \infty$. Moreover, adversarial robustness similarly does not resemble either of them, since the norm is used to perturb the input of the logloss.

C.5 Proof of Lemma 4

We prove Lemma 4 by constructing the presented optimization problem. To this end, we first dualize the inner $\sup$ problem of Synth for fixed $\boldsymbol{\beta}$. By interchanging $\boldsymbol{\xi} = (\boldsymbol{x}, y)$, we can write the inner problem as

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

Here, the first three constraints specify that $\mathbb{Q}$ and $\mathbb{P}_N$ have a Wasserstein distance of at most 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 also at most $\widehat{\varepsilon}$ away from each other, modeled via their coupling $\widehat{\Pi}$. The fact that $\mathbb{Q}$ is constrained to be in the intersection of two balls as specified in (Synth) causes Π and $\widehat{\Pi}$ share the same marginal $\mathbb{Q}$. We can now substitute the third constraint in the objective as well as the last constraint to get:

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

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 these representations, 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 similar reformulation in the constraints, we obtain:

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

We now decompose each $\mathbb{Q}^i$ into two sub-measures corresponding to $y = \pm 1$, so that $\mathbb{Q}^i(\mathrm{d}(\boldsymbol{x}, y)) = \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x})$ if $y = +1$ and $\mathbb{Q}^i(\mathrm{d}(\boldsymbol{x}, y)) = \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})$ if $y = -1$. We similarly divide each $\widehat{\mathbb{Q}}^j$ into $\widehat{\mathbb{Q}}^j_{+1}$ and $\widehat{\mathbb{Q}}^j_{-1}$. Note that the sub measures are not probability measures as they do not integrate to 1, but they are non-negative measures supported on $\mathbb{R}^n$ (denoted $\in \mathcal{P}_+(\mathbb{R}^n)$). We obtain:

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

We substitute 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 constraints.

$$\begin{array}{ll}
\underset{\mathbb{A}_{\pm 1}, \mathbb{Q}_{\pm 1}, \widehat{\mathbb{Q}}_{\pm 1}}{\text{maximize}} & \displaystyle \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} & \displaystyle \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}) + \\
& \displaystyle \qquad \qquad \sum_{i=1}^{N} \|\boldsymbol{x} - \boldsymbol{x}^i\|_q \cdot [\mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x})] \Big] \leq \varepsilon \\
& \displaystyle \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}) + \\
& \displaystyle \qquad \qquad \sum_{j=1}^{\widehat{N}} \|\boldsymbol{x} - \widehat{\boldsymbol{x}}^j\|_q \cdot [\widehat{\mathbb{Q}}^j_{+1}(\mathrm{d}\boldsymbol{x}) + \widehat{\mathbb{Q}}^j_{-1}(\mathrm{d}\boldsymbol{x})] \Big] \leq \widehat{\varepsilon} \\
& \displaystyle \int_{\boldsymbol{x} \in \mathbb{R}^n} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) + \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) = 1 \qquad \qquad \forall i \in [N] \\
& \displaystyle \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 \qquad \qquad \forall j \in [\widehat{N}] \\
& \displaystyle \frac{1}{N} \sum_{i=1}^{N} \mathbb{Q}^i_{+1}(\mathrm{d}\boldsymbol{x}) = \mathbb{A}_{+1}(\mathrm{d}\boldsymbol{x}) \qquad \qquad \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \displaystyle \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}) \qquad \qquad \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \displaystyle \frac{1}{N} \sum_{i=1}^{N} \mathbb{Q}^i_{-1}(\mathrm{d}\boldsymbol{x}) = \mathbb{A}_{-1}(\mathrm{d}\boldsymbol{x}) \qquad \qquad \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \displaystyle \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}) \qquad \qquad \forall \boldsymbol{x} \in \mathbb{R}^n \\
& \mathbb{A}_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n), \ \mathbb{Q}^i_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n), \ \widehat{\mathbb{Q}}^j_{\pm 1} \in \mathcal{P}_+(\mathbb{R}^n) \qquad \qquad \forall i \in [N], \ j \in [\widehat{N}].
\end{array}$$

We derive the following semi-infinite dual problem (strong duality holds whenever $\varepsilon, \widehat{\varepsilon} > 0$ Shapiro 2001)

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

To eliminate the variables p_{+1} and $\widehat{p}_{+1}$, notice that their constraints

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

are the epigraph-based reformulation of the following constraint

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

We can thus eliminate p_{+1} and $\widehat{p}_{+1}$. We can similarly eliminate p_{-1} and $\widehat{p}_{-1}$ since

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

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 dual problem

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

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

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

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

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

and concludes the proof. □

C.6 Proof of Lemma 5

We prove a more general case for any closed convex function $L : \mathbb{R} \mapsto \mathbb{R}_+$, vectors $\boldsymbol{\omega}, \boldsymbol{a}, \widehat{\boldsymbol{a}} \in \mathbb{R}^n$, scalars $\lambda, \widehat{\lambda} > 0$ and a norm $\|\cdot\|_q$. Consider the following DC maximization form:

$$\sup_{\boldsymbol{x}\in\mathbb{R}^n} \{L(\boldsymbol{\omega}^\top \boldsymbol{x}) - \lambda \cdot \|\boldsymbol{a} - \boldsymbol{x}\|_q - \widehat{\lambda} \cdot \|\widehat{\boldsymbol{a}} - \boldsymbol{x}\|_q\}.$$

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

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

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 derive 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, inspired from (Shafieezadeh-Abadeh et al., 2019, Lemma 47), we show

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

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 reminds that $\boldsymbol{x}$ is unbounded unless its coefficients are zero, the sixth identity uses the definition of convex conjugates and the final identity uses 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

$$\sup_{\boldsymbol{z}\in\mathbb{R}^n} g^*(\boldsymbol{z}) + \sup_{\theta\in\mathrm{dom}(L^*)} \{-L^*(\theta) \;:\; \theta \cdot \boldsymbol{\omega} = \boldsymbol{z}\}$$

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

Here, the first identity follows by writing the problem as a single maximization problem, the second identity follows by a variable change, 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 substituting the equality constraint.

Since Lemma 4 has this sup problem on the left-hand side of an inequality constraint, we can replace $\sup\theta\ \inf\boldsymbol{z}$ with $\forall\theta,\ \exists\boldsymbol{z}$, which concludes this proof. □

C.7 Proof of Corollary 2

By using Lemma 4 we expressed constraint (2) as

$$
\forall\theta\in\mathrm{dom}(L^{\alpha*}),\ \exists\boldsymbol{z}\in\mathbb{R}^n : \begin{cases} -L^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\omega}^\top\boldsymbol{a} + \theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{a}} - \boldsymbol{a}) \le t(\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}) \\ |\theta|\cdot\|\boldsymbol{\omega} - \boldsymbol{z}\|_{q^\star} \le \lambda \\ |\theta|\cdot\|\boldsymbol{z}\|_{q^\star} \le \widehat{\lambda}. \end{cases}
$$

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

$$
\exists\boldsymbol{z}\in\mathbb{R}^n,\ \forall\theta\in\mathrm{dom}(L^{\alpha*}) : \begin{cases} -L^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\omega}^\top\boldsymbol{a} + \theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{a}} - \boldsymbol{a}) \le t(\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}) \\ |\theta|\cdot\|\boldsymbol{\omega} - \boldsymbol{z}\|_{q^\star} \le \lambda \\ |\theta|\cdot\|\boldsymbol{z}\|_{q^\star} \le \widehat{\lambda}. \end{cases}
$$

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).

The relaxation can equivalently be written as

$$
\exists\boldsymbol{z}\in\mathbb{R}^n : \begin{cases} \sup_{\theta\in\mathrm{dom}(L^{\alpha*})}\{-L^{\alpha*}(\theta) + \theta\cdot\boldsymbol{\omega}^\top\boldsymbol{a} + \theta\cdot\boldsymbol{z}^\top(\widehat{\boldsymbol{a}} - \boldsymbol{a})\} \le t(\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}) \\ \sup_{\theta\in\mathrm{dom}(L^{\alpha*})}\{|\theta|\}\cdot\|\boldsymbol{\omega} - \boldsymbol{z}\|_{q^\star} \le \lambda \\ \sup_{\theta\in\mathrm{dom}(L^{\alpha*})}\{|\theta|\}\cdot\|\boldsymbol{z}\|_{q^\star} \le \widehat{\lambda}. \end{cases}
$$

Since $L^{\alpha*}$ is a closed convex function, we have $\sup_{\theta\in\mathrm{dom}(L^{\alpha*})}\{|\theta|\} = \mathrm{Lip}(L^\alpha)$ as well as $L^{\alpha**} = L^\alpha$, hence this system can be written as

$$
\exists\boldsymbol{z}\in\mathbb{R}^n : \begin{cases} L^\alpha(\boldsymbol{\omega}^\top\boldsymbol{a} + \boldsymbol{z}^\top(\widehat{\boldsymbol{a}} - \boldsymbol{a})) \le t(\lambda,\widehat{\lambda},\boldsymbol{s},\widehat{\boldsymbol{s}}) \\ \mathrm{Lip}(L^\alpha)\cdot\|\boldsymbol{\omega} - \boldsymbol{z}\|_{q^\star} \le \lambda \\ \mathrm{Lip}(L^\alpha)\cdot\|\boldsymbol{z}\|_{q^\star} \le \widehat{\lambda} \end{cases}
$$

and keeping $\boldsymbol{z}$ as a new variable concludes the proof. □

C.8 Proof of Theorem 1

Lemma 4, which dualizes the nature's $\sup$ problem of Synth, allows us to represent Synth as:

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

Using Lemma 5 to reformulate each of the constraints with the adjustable robust optimization reformulation, and then employing the relaxation presented in Corollary 2 concludes the proof. □

C.9 Further Relaxation Techniques

From proof of Lemma 5 we can see that the left-hand side of the DC maximization constraint can be written as

$$\sup_{\theta\in\mathrm{dom}(L^*)} -L^*(\theta)+\theta\cdot\boldsymbol{\omega}^\top\boldsymbol{a}+\theta\cdot\inf_{\boldsymbol{z}\in\mathbb{R}^n}\{\boldsymbol{z}^\top(\widehat{\boldsymbol{a}}-\boldsymbol{a})\ :\ |\theta|\cdot\|\boldsymbol{\omega}-\boldsymbol{z}\|_{q^\star}\leq\lambda,\ |\theta|\cdot\|\boldsymbol{z}\|_{q^\star}\leq\widehat{\lambda}\}.$$

If $q\in\{1,\infty\}$, then the $\inf$-problem is a linear optimization problem (LP), hence we can use LP duality to eliminate the inner problem. Similarly, if $q=2$, we can rewrite the inner problem exactly as a semidefinite problem (SDP) since it is linear optimization over two quadratic constrains (Wang & Kilinc-Karzan, 2020) and take its semidefinite (strong) dual problem. After we replace the inner $\inf$ problem, we can write the constraint with a single $\sup$ problem, where the main nonconvexity will be due to $-L^\star(\theta)$. This, however, can be dealt with in the 'convex uncertainty' literature by lifting $-L^\star(\theta)$ to the uncertainty set and using adjustable robust optimization techniques (*e.g.*, use linear decision rules to relax the problem to a convex optimization problem). See (Roos et al., 2018) for more details.

Alternatively, instead of solving Synth that first intersects the Wasserstein balls and optimizes over the intersection, we can take a mixture between the empirical and synthetic datasets, and build a Wasserstein ball around this distribution. Although this problem will not have the desired finite-sample unlike Synth, it is a special case of AdvDRO, hence admits a tractable exact reformulation.

C.10 Proof of Lemma 2

Notice that in problem ProxSynth, the term $\widehat{\varepsilon}$ appears in the objective as $\widehat{\varepsilon}\cdot\widehat{\lambda}$. Hence, as $\widehat{\varepsilon}\to\infty$, optimal solutions satisfy $\widehat{\lambda}=0$. Thus, the constraints $\|\boldsymbol{z}^-_{ij}\|_{q^\star}\leq\widehat{\lambda}$ and $\|\boldsymbol{z}^+_{ij}\|_{q^\star}\leq\widehat{\lambda}$ specify $\boldsymbol{z}^+_{ij}=\boldsymbol{z}^-_{ij}=0$ for all $i\in[N], j\in[\widehat{N}]$. The problem, therefore, can be written as

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

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 any constraint indexed with j is feasible, it means the same constraint indexed with 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

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

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. Moreover, we can substitute the definition of L^α and obtain:

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

For any $i \in [N]$ if $y^i = 1$, the first two constraints can be written as

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

On the other hand, if $y^i = -1$, the first two constraints reduce to the exactly same system with the reverse order. Hence, the above problem can be represented as

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

which is identical to the reformulation of AdvDRO presented in Corollary 1.

Note that, although this proof used $\widehat{\varepsilon} \to \infty$, in an extended version we will provide a proof that works for any $\widehat{\varepsilon}$ as long as $\mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}}) = \mathfrak{B}_\varepsilon(\mathbb{P}_N)$.

C.11 Proof of Proposition 1

By standard linearity arguments and from the definition of $\mathbb{Q}_{\text{mix}}$ (which will simply be denoted by $\mathbb{Q}$ in this proof), we have

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

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

$$\mathbb{E}_{\mathbb{Q}}\left[\sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_p \leq \alpha} \{\ell_{\boldsymbol{\beta}}(\boldsymbol{x}+\boldsymbol{z}, y)\}\right] = \mathbb{E}_{\mathbb{Q}}[\ell^{\alpha}_{\boldsymbol{\beta}}(\boldsymbol{x}, y)]$$

which concludes the proof.

C.12 Proof of Lemma 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 $\mathbb{P} \in \mathcal{P}(\Xi)$.

Observation 1. *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. □

As a corollary of this observation, we can prove the following result.

Corollary 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 Observation 1 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 lemma (we refer to $\mathbb{Q}_{\text{mix}}$ 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 Corollary 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

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

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

C.13 Different Loss Functions

Notice that although we used the logloss function in our explicit representations, we kept our analyses as general as possible. The theory can be revised to several different loss functions, with the following notes:

- The original loss function should admit a closed-form convex adversarial loss representation in order to eliminate the adversary's problem (*cf.*, Lemma 1). This holds for the Hinge loss function, for example (Bertsimas et al., 2019).
- The adversarial loss function should be closed, convex, and Lipschitz continuous with known Lipschitz constant (*cf.* Corollary 2.
- We use exponential cone optimization to solve the underlying optimization problems, as the logloss constraints are exponential cone representable. New algorithms/solvers should be used for different loss functions (*e.g.*, second-order conic optimization for smooth SVM constraints Hsu et al. 2003).

D Statistical Properties

D.1 Properties of AdvDRO

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 2. *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 depends 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., 2019b). The presented decay rate $\mathcal{O}(N^{-1/n})$ of ε as N increases is optimal (Fournier & 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. In this subsection, we start our analysis with AdvDRO, which is a distributionally robust optimization problem over a

Wasserstein ambiguity set built around $\mathbb{P}_N$ to optimize the worst-case expected adversarial loss. 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 AdvDRO again by using the adversarial loss function as in the proof of Lemma 1:

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

Theorem 3. *If the assumptions of Theorem 2 are satisfied and ε is chosen accordingly as in the statement of Theorem 2, then*

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

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

Proof. The statement follows from Theorem 19 of (Kuhn et al., 2019b) given that $\ell^\alpha_{\boldsymbol{\beta}}$ is a finite-valued continuous loss function. □

Theorem 3 states that the optimal value of AdvDRO overestimates the true loss with arbitrarily high confidence $1 - \eta$. Despite the desired overestimation of the true loss, we show that AdvDRO is still asymptotically consistent if we restrict the set of admissible $\boldsymbol{\beta}$ to a bounded set[2].

Theorem 4. *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 2, we have*

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

whenever ε_N is set as specified in Theorem 2 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 $\ell^\alpha_{\boldsymbol{\beta}}(\boldsymbol{x}, y) \leq C(1 + d(\boldsymbol{\xi}, \boldsymbol{\xi}^0))$ 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., 2019b).

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

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

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

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

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

$$\log(1 + \exp(-y \cdot \boldsymbol{\beta}^\top \boldsymbol{x} + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star})) \leq \log(1 + \exp(|\boldsymbol{\beta}^\top \boldsymbol{x}| + \alpha \cdot \|\boldsymbol{\beta}\|_{p^\star}))$$

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

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

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 AdvDRO 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 AdvDRO, 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 5. *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 this result.

D.2 Properties of Synth

The previous subsection derived statistical properties of AdvDRO to tune ε. This problem can be interpreted as Synth with $\widehat{\varepsilon} = \infty$. Keeping in mind that not learning from the synthetic data is always a feasible solution (*cf.* Lemma 2), we next analyze how to tune $\widehat{\varepsilon}$ for cases where synthetic data is useful.

Firstly, since $\widehat{\mathbb{P}}_{\widehat{N}}$ is constructed of i.i.d. samples of $\widehat{\mathbb{P}}$, we can overestimate the first distance $\widehat{\varepsilon}_1 = \mathrm{W}(\widehat{\mathbb{P}}_{\widehat{N}}, \widehat{\mathbb{P}})$ analogously by applying Theorem 2, *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 6. *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*

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

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

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

Proof. It immediately follows from Theorem 2 that $[\mathbb{P}^0]^N(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N)) \geq 1 - \eta_1$ holds. Moreover, 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 $[\mathbb{P}^0 \times \widehat{\mathbb{P}}]^{N \times \widehat{N}}(\mathbb{P}^0 \in \mathfrak{B}_\varepsilon(\mathbb{P}_N) \cap \mathfrak{B}_{\widehat{\varepsilon}}(\widehat{\mathbb{P}}_{\widehat{N}})) \geq 1 - \eta$. □

One can also show that, under the assumptions of Theorem 6, Synth overestimates the true loss analogously as Theorem 3. However, Synth does not satisfy any asymptotic consistencies, given that $\widehat{N} \to \infty$ will let $\widehat{\varepsilon}_1 \to 0$, but $\widehat{\varepsilon}_2$ is always a constant that arises from the distance between the true distribution $\mathbb{P}^0$ and the synthetic distribution $\widehat{\mathbb{P}}$. Hence, Synth cannot be useful in asymptotic data regimes, which is expected given that the motivation of this problem is to reduce the conservatism of AdvDRO that is intended to prevent overfitting in non-asymptotic settings. See also (Taskesen et al., 2021) for a relevant discussion.

In the above results, we assumed that $\widehat{\varepsilon}_2 = \mathrm{W}(\mathbb{P}^0, \widehat{\mathbb{P}})$ is known. However, as discussed in the main paper, this is not possible in most real-life settings. We can either cross-validate this parameter (as in the transfer learning and domain adaptation literature Zhong et al. 2010) or use domain knowledge in some special settings. This is for example the case when $\widehat{\mathbb{P}}$ corresponds to a synthetic data generator trained on the population $\mathbb{P}^0$ which is guaranteed to reach within $\widehat{\varepsilon}_2$ distance from $\mathbb{P}^0$ (e.g. Wasserstein GAN Arjovsky et al. (2017)). Note that in this setting $\mathbb{P}^0$ represents overall population that is available for the training of synthetic data generator, however only a sample from it $\mathbb{P}_N$ is available for training of the ML model for a downstream task together with a corresponding synthetic sample $\widehat{\mathbb{P}}_{\widehat{N}}$ from $\mathbb{P}$. At

Algorithm 1 Construction of artificial data.

```
Sample the components of $\boldsymbol{\beta}$ i.i.d. from a standard normal distribution, normalize for $\|\boldsymbol{\beta}\|_2 = 1$
for $i \in \{1, \ldots, N\}$ do
    Find the true probability of $y^i = +1$ assuming $\boldsymbol{\beta}$ defines the true logistic classifier:
        $p^i = \left[1 + \exp(-\boldsymbol{\beta}^\top \boldsymbol{x})\right]^{-1}.$
    If $p^i \geq \mathcal{U}(0,1)$, then $y^i = +1$. Otherwise, $y^i = -1$.
end for
The artificial dataset is the collection of all $(\boldsymbol{x}^i, y^i),\ i \in [N]$, constructed above
```

first sight, it seems somewhat unrealistic to assume that the overall population is available for training of a synthetic data generator but not for training of a downstream ML model, however, this is a plausible setup in highly regulated industries (e.g. healthcare, finance) where data usage is guarded with a strict set of rules and subject to explicit consent provided by users/costumers.

Data sharing in a financial institution Data usage in highly regulated industries (e.g. finance, healthcare) requires adherence to a broad set of regulations (FCRA, GDPR, HIPPA etc.). These often restrict the usage of customer data for the specific purpose. Appropriately curated synthetic data is a potential way to utilize real datasets in a legally complaint manner, e.g. without compromising privacy inherent to the raw costumer data. Our setting is motivated by real-world applications where only a portion of available real data can be used for building a particular ML model (e.g. due to explicit costumer consent), together with a synthetic data from a privacy preserving synthetic generator trained on the entire dataset. The subtle point here is that although a costumer did not explicitly express consent for a particular purpose, in some settings it might be possible to use its data for training a generator with privacy guarantees.

E Further Details for Numerical Experiments

All experiments are conducted in Julia (Bezanson et al.) (MIT license) and executed on Intel Xeon 2.66GHz processors with 8GB memory in single-core mode. We use MOSEK 10.1 (MOSEK ApS, 2023a) to solve all exponential conic programs through JuMP (Dunning et al., 2017).

We sample instances by Algorithm 1, where $\mathcal{U}(0,1)$ denotes sampling from a continuous $(0,1)$-uniform distribution. Since the true $\boldsymbol{\beta}$ remains unchanged for the sampled instances, this approach corresponds to i.i.d. sampling from a true distribution. On the other hand, to obtain synthetic data, we perturb the probabilities p^i with standard random normal noise; this changes the true distribution while still sampling i.i.d. from it.