Regularization via Mass Transportation

Soroosh Shafieezadeh-Abadeh[†], Daniel Kuhn[†], and Peyman Mohajerin Esfahani[‡]

[†]Risk Analytics and Optimization Chair, EPFL Lausanne
`soroosh.shafiee,daniel.kuhn@epfl.ch`
[‡]Delft Center for Systems and Control, TU Delft
`p.mohajerinesfahani@tudelft.nl`

July 12, 2019

Abstract

The goal of regression and classification methods in supervised learning is to minimize the empirical risk, that is, the expectation of some loss function quantifying the prediction error under the empirical distribution. When facing scarce training data, overfitting is typically mitigated by adding regularization terms to the objective that penalize hypothesis complexity. In this paper we introduce new regularization techniques using ideas from distributionally robust optimization, and we give new probabilistic interpretations to existing techniques. Specifically, we propose to minimize the worst-case expected loss, where the worst case is taken over the ball of all (continuous or discrete) distributions that have a bounded transportation distance from the (discrete) empirical distribution. By choosing the radius of this ball judiciously, we can guarantee that the worst-case expected loss provides an upper confidence bound on the loss on test data, thus offering new generalization bounds. We prove that the resulting regularized learning problems are tractable and can be tractably kernelized for many popular loss functions. The proposed approach to regluarization is also extended to neural networks. We validate our theoretical out-of-sample guarantees through simulated and empirical experiments.

Keywords: Distributionally robust optimization, optimal transport, Wasserstein distance, robust optimization, regularization

1 Introduction

The fields of machine learning and optimization are closely intertwined. On the one hand, optimization algorithms are routinely used for the solution of classical machine learning problems. Conversely, recent advances in optimization under uncertainty have inspired many new machine learning models.

From a conceptual point of view, many statistical learning tasks give naturally rise to stochastic optimization problems. Indeed, they aim to find an estimator from within a prescribed hypothesis space that minimizes the expected value of some loss function. The loss function quantifies the estimator's ability to correctly predict random outputs (*i.e.*, dependent variables or labels) from random inputs (*i.e.*, independent variables or features). Unfortunately, such stochastic optimization problems cannot be solved exactly because the input-output distribution, which is needed to evaluate the expected loss in the objective function, is not accessible and only indirectly observable through finitely many training samples. Approximating the expected loss with the empirical loss, that is, the average loss across all training samples, yields fragile estimators that are sensitive to perturbations in the data and suffer from overfitting.

Regularization is the standard remedy to combat overfitting. Regularized learning models minimize the sum of the empirical loss and a penalty for hypothesis complexity, which is typically chosen as a norm of the hypothesis. There is ample empirical evidence that regularization reduces a model's generalization error. Statistical learning theory reasons that regularization implicitly restricts the hypothesis space, thereby controlling the gap between the training error and the testing error, see, *e.g.*, Bartlett and Mendelson [5]. However, alternative explanations for the practical success of regularization are possible. In particular, ideas from modern robust optimization (Ben-Tal *et al.* [6]) recently led to a fresh perspective on regularization.

Robust regression and classification models seek estimators that are immunized against adversarial perturbations in the training data. They have received considerable attention since the seminal treatise on robust least-squares regression by El Ghaoui and Lebret [28], who seem to be the first authors to discover an intimate connection between robustification and regularization. Specifically, they show that minimizing the worst-case residual error with respect to all perturbations in a Frobenius norm-uncertainty set is equivalent to a Tikhonov regularization procedure. Xu *et al.* [77] disclose a similar equivalence between robust least-squares regression with a feature-wise independent uncertainty set and the celebrated Lasso (least absolute shrinkage and selection operator) algorithm. Leveraging this new robustness interpretation, they extend Lasso to a wider class of regularization schemes tailored to regression problems with disturbances that are coupled across features. In the context of classification, Xu *et al.* [76] provide a linkage between robustification over non-box-typed uncertainty sets and the standard regularization scheme of support vector machines. A comprehensive characterization of the conditions under which robustification and regularization are equivalent has recently been compiled by Bertsimas and Copenhaver [10].

New learning models have also been inspired by recent advances in the emerging field of *distributionally* robust optimization, which bridges the gap between the conservatism of robust optimization and the specificity of stochastic programming. Distributionally robust optimization seeks to minimize a worst-case expected loss, where the worst case is taken with respect to all distributions in an ambiguity set, that is, a family of distributions consistent with the given prior information on the uncertainty, see, *e.g.*, Calafiore and El Ghaoui [18], Delage and Ye [25], Goh and Sim [35], Wiesemann *et al.* [75] and the references therein. Ambiguity sets are often characterized through generalized moment conditions. For instance, Lanckriet *et al.* [43] propose a distributionally robust minimax probability machine for binary classification, where both classes are encoded by the first and second moments of their features, and the goal is to find a linear classifier that minimizes the worst-case misclassification error in view of all possible input distributions consistent with the given moment information. By construction, this approach forces the worst-case accuracies of both classes to be equal. Huang *et al.* [40] propose a generalization of the minimax probability machine that allows for uneven worst-case classification accuracies. Lanckriet *et al.* [44] extend the minimax probability machine to account for estimation errors in the mean vectors and covariance matrices. Strohmann and Grudic [68] and Bhattacharyya [11] develop minimax probability machines for regression and feature selection, respectively. Shivaswamy *et al.* [66] study linear classification problems trained with incomplete and noisy features, where each training sample is modeled by an ambiguous distribution with known first and second-order moments. The authors propose to address such classification problems with a distributionally robust soft margin support vector machine and then prove that it is equivalent to a classical robust support vector machine with a feature-wise uncertainty set. Farnia and Tse [31] investigate distributionally robust learning models with moment ambiguity sets that restrict the marginal of the features to the empirical marginal. The authors highlight similarities and differences to classical regression models.

Ambiguity sets containing all distributions that share certain low-order moments are computationally attractive but fail to converge to a singleton when the number N of training samples tends to infinity. Thus, they preclude any asymptotic consistency results. A possible remedy is to design spherical ambiguity sets with respect to some probability distance functions and to drive their radii to zero as N grows. Examples include the ϕ-divergence ambiguity sets proposed by Ben-Tal *et al.* [7] or the Wasserstein ambiguity sets studied by Mohajerin Esfahani and Kuhn [52] and Zhao and Guan [80]. Blanchet and Murthy [12] and Gao and Kleywegt [34] consider generalized Wasserstein ambiguity sets defined over Polish spaces.

In this paper we investigate distributionally robust learning models with Wasserstein ambiguity sets. The Wasserstein distance between two distributions is defined as the minimum cost of transporting one distribution to the other, where the cost of moving a unit point mass is determined by the ground metric on the space of uncertainty realizations. In computer science the Wasserstein distance is therefore sometimes aptly termed the 'earth mover's distance' (Rubner *et al.* [57]). Following Mohajerin Esfahani and Kuhn [52], we define Wasserstein ambiguity sets as balls with respect to the Wasserstein distance that are centered at the empirical distribution on the training samples. These ambiguity sets contain all (continuous or discrete) distributions that can be converted to the (discrete) empirical distribution at bounded transportation cost.

Wasserstein distances are widely used in machine learning to compare histograms. For example, Rubner *et al.* [57] use the Wasserstein distance as a metric for image retrieval with a focus on applications to color

and texture. Cuturi [21] and Benamou *et al.* [8] propose fast iterative algorithms to compute a regularized Wasserstein distance between two high-dimensional discrete distributions for image classification tasks. Moreover, Cuturi and Doucet [23] develop first-order algorithms to compute the Wasserstein barycenter between several empirical probability distributions, which has applications in clustering. Arjovsky *et al.* [3] utilize the Wasserstein distance to measure the distance between the data distribution and the model distribution in generative adversarial networks. Furthermore, Frogner *et al.* [33] propose a learning algorithm based on the Wasserstein distance to predict multi-label outputs.

Distributionally robust optimization models with Wasserstein ambiguity sets were introduced to the realm of supervised learning by Shafieezadeh-Abadeh *et al.* [61], who show that distributionally robust logistic regression problems admit a tractable reformulation and encapsulate the classical as well as the popular regularized logistic regression problems as special cases. When the Wasserstein ball is restricted to distributions on a compact set, the problem becomes intractable but can still be addressed with an efficient decomposition algorithm due to Luo and Mehrotra [50]. Support vector machine models with distributionally robust chance constraints over Wasserstein ambiguity sets are studied by Lee and Mehrotra [47]. These models are equivalent to hard semi-infinite programs and can be solved approximately with a cutting plane algorithm.

Wasserstein ambiguity sets are popular for their attractive statistical properties. For example, Fournier and Guillin [32] prove that the empirical distribution on N training samples converges in Wasserstein distance to the true distribution at rate $\mathcal{O}(N^{-1/(n+1)})$, where n denotes the feature dimension. This implies that properly scaled Wasserstein balls constitute natural confidence regions for the data-generating distribution. The worst-case expected prediction loss over all distributions in a Wasserstein ball thus provides an upper confidence bound on the expected loss under the unknown true distribution; see Mohajerin Esfahani and Kuhn [52]. Blanchet *et al.* [13] show, however, that radii of the order $\mathcal{O}(N^{-1/2})$ are asymptotically optimal even though the corresponding Wasserstein balls are too small to contain the true distribution with constant confidence. For Wasserstein distances of type two (where the transportation cost equals the squared ground metric) Blanchet *et al.* [14] develop a systematic methodology for selecting the ground metric. Generalization bounds for the *worst-case* prediction loss with respect to a Wasserstein ball centered at the *true* distribution are derived by Lee and Raginsky [48] in order to address emerging challenges in domain adaptation problems, where the distributions of the training and test samples can differ.

This paper extends the results by Shafieezadeh-Abadeh *et al.* [61] on distributionally robust logistic regression along several dimensions. Our main contributions can be summarized as follows:

- **Tractability:** We propose data-driven distributionally robust regression and classification models that hedge against all input-output distributions in a Wasserstein ball. We demonstrate that the emerging semi-infinite optimization problems admit equivalent reformulations as tractable convex programs for many commonly used loss functions and for spaces of linear hypotheses. We also show that lifted variants of these new learning models are kernelizable and thus offer an efficient procedure for optimizing over all nonlinear hypotheses in a reproducible kernel Hilbert space. Finally, we study distributionally robust learning models over families of feed-forward neural networks. We show that these models can be approximated by regularized empirical loss minimization problems with a convex regularization term and can be addressed with a stochastic proximal gradient descent algorithm.

- **Probabilistic Interpretation of Existing Regularization Techniques:** We show that the classical regularized learning models emerge as special cases of our framework when the cost of moving probability mass along the output space tends to infinity. In this case, the regularization function and its regularization weight are determined by the transportation cost on the input space and the radius of the Wasserstein ball underlying the distributionally robust optimization model, respectively.

- **Generalization Bounds:** We demonstrate that the proposed distributionally robust learning models enjoy new generalization bounds that can be obtained under minimal assumptions. In particular, they do not rely on any notions of hypothesis complexity and may therefore even extend to hypothesis spaces with infinite VC-dimensions. A naïve generalization bound is obtained by leveraging modern measure concentration results, which imply that Wasserstein balls constitute confidence sets for the unknown data-generating distribution. Unfortunately, this generalization bound suffers from a curse

of dimensionality and converges slowly for high input dimensions. By imposing bounds on the hypothesis space, however, we can derive an improved generalization bound, which essentially follows a dimension-independent square root law reminiscent of the central limit theorem.

- **Relation to Robust Optimization:** In classical robust regression and classification the training samples are viewed as uncertain variables that range over a joint uncertainty set, and the best hypothesis is found by minimizing the worst-case loss over this set. We prove that the classical robust and new distributionally robust learning models are equivalent if the data satisfies a dispersion condition (for regression) or a separability condition (for classification). While there is no efficient algorithm for solving the robust learning models in the absence of this condition, the distributionally robust models are efficiently solvable irrespective of the underlying training datasets.

- **Confidence Intervals for Error and Risk:** Using distributionally robust optimization techniques based on the Wasserstein ball, we develop two tractable linear programs whose optimal values provide a confidence interval for the absolute prediction error of any fixed regressor or the misclassification risk of any fixed classifier.

- **Worst-Case Distributions:** We formulate tractable convex programs that enable us to efficiently compute a worst-case distribution in the Wasserstein ball for any fixed hypothesis. This worst-case distribution can be useful for stress tests or contamination experiments.

The rest of the paper develops as follows. Section 2 introduces our new distributionally robust learning models. Section 3 provides finite convex reformulations for learning problems over linear and nonlinear hypothesis spaces and describes efficient procedures for constructing worst-case distributions. Moreover, it compares the new distributionally robust method against existing robust optimization and regularization approaches. Section 4 develops new generalization bounds, while Section 5 addresses error and risk estimation. Numerical experiments are reported in Section 6. All proofs are relegated to the appendix.

1.1 Notation

Throughout this paper, we adopt the conventions of extended arithmetics, whereby $\infty \cdot 0 = 0 \cdot \infty = 0/0 = 0$ and $\infty - \infty = -\infty + \infty = 1/0 = \infty$. The inner product of two vectors $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$ is denoted by $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$, and for any norm $\|\cdot\|$ on $\mathbb{R}^n$, we use $\|\cdot\|_*$ to denote its dual norm defined through $\|\boldsymbol{x}\|_* = \sup\ \{\langle \boldsymbol{x}, \boldsymbol{x}' \rangle : \|\boldsymbol{x}'\| \leq 1\}$. The conjugate of an extended real-valued function $f(\boldsymbol{x})$ on $\mathbb{R}^n$ is defined as $f^*(\boldsymbol{x}) = \sup_{\boldsymbol{x}'} \langle \boldsymbol{x}, \boldsymbol{x}' \rangle - f(\boldsymbol{x}')$. The indicator function of a set $\mathbb{X} \subseteq \mathbb{R}^n$ is defined as $\delta_{\mathbb{X}}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \mathbb{X}; = \infty$ otherwise. Its conjugate $S_{\mathbb{X}}(\boldsymbol{x}) = \sup\{\langle \boldsymbol{x}', \boldsymbol{x} \rangle : \boldsymbol{x}' \in \mathbb{X}\}$ is termed the support function of $\mathbb{X}$. The characteristic function of $\mathbb{X}$ is defined through $\mathbb{1}_{\mathbb{X}}(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in \mathbb{X};\ = 0$ otherwise. For a proper cone $\mathcal{C} \subseteq \mathbb{R}^n$ the relation $\boldsymbol{x} \succeq_{\mathcal{C}} \boldsymbol{x}'$ indicates that $\boldsymbol{x} - \boldsymbol{x}' \in \mathcal{C}$. The cone dual to $\mathcal{C}$ is defined as $\mathcal{C}^* = \{\boldsymbol{x}' : \langle \boldsymbol{x}', \boldsymbol{x} \rangle \geq 0 \,\forall \boldsymbol{x} \in \mathcal{C}\}$. The Lipschitz modulus of a function $L : \mathbb{X} \to \mathbb{R}$ is denoted by $\mathrm{lip}(L) = \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}}\{|L(\boldsymbol{x}) - L(\boldsymbol{x}')|/\|\boldsymbol{x} - \boldsymbol{x}'\| : \boldsymbol{x} \neq \boldsymbol{x}'\}$. If $\mathbb{P}$ is a distribution on a set Ξ, then $\mathbb{P}^N$ denotes the N-fold product of $\mathbb{P}$ on the Cartesian product Ξ^N. For $N \in \mathbb{N}$ we define $[N] = \{1, \ldots, N\}$. A list of commonly used symbols is provided in the following table.

$\mathbb{X}$	input space		$\mathbb{Y}$	output space
ℓ	loss function		L	univariate loss function
$\mathbb{H}$	hypothesis space		$\mathcal{K}$	kernel matrix
f^*	conjugate of f		$\mathrm{lip}(f)$	Lipschitz modulus of f
$\mathcal{C}^*$	dual cone of $\mathcal{C}$		$\|\cdot\|_*$	dual norm of $\|\cdot\|$
$S_{\mathbb{X}}$	support function of $\mathbb{X}$		$\delta_{\mathbb{X}}$	indicator function of $\mathbb{X}$
$\mathbb{1}_{\mathbb{X}}$	characteristic function of $\mathbb{X}$		$[N]$	$\{1, \ldots, N\}$

2 Problem Statement

We first introduce the basic terminology and then describe our new perspective on regularization.

2.1 Classical Statistical Learning

The goal of supervised learning is to infer an unknown target function $f : \mathbb{X} \to \mathbb{Y}$ from limited data. The target function maps any input $\boldsymbol{x} \in \mathbb{X}$ (*e.g.*, information on the frequency of certain keywords in an email) to some output $y \in \mathbb{Y}$ (*e.g.*, a label $+1$ (-1) if the email is likely (unlikely) to be a spam message). If the true target function was accessible, it could be used as a means to reliably predict outputs from inputs (*e.g.*, it could be used to recognize spam messages in an automated fashion). In a supervised learning framework, however, one has only access to finitely many input-output examples $(\widehat{\boldsymbol{x}}_i, \widehat{y}_i)$ for $i = 1, \ldots, N$ (*e.g.*, a database of emails which have been classified by a human as legitimate or as spam messages). We will henceforth refer to these examples as the *training data* or the *in-sample data*. It is assumed that the training samples are mutually independent and follow an unknown distribution $\mathbb{P}$ on $\mathbb{X} \times \mathbb{Y}$.

The supervised learning problems are commonly subdivided into *regression problems*, where the output y is continuous and $\mathbb{Y} = \mathbb{R}$, and *classification problems*, where y is categorical and $\mathbb{Y} = \{+1, -1\}$. As the space of all functions from $\mathbb{X}$ to $\mathbb{Y}$ is typically vast, it may be very difficult to learn the target function from finitely many training samples. Thus, it is convenient to restrict the search space to a structured family of candidate functions $\mathbb{H} \subseteq \mathbb{R}^{\mathbb{X}}$ such as the space of all linear functions, some reproducible kernel Hilbert space or the family of all feed-forward neural networks with a fixed number of layers. We henceforth refer to each candidate function $h \in \mathbb{H}$ as a *hypothesis* and to $\mathbb{H}$ as the *hypothesis space*.

A *learning algorithm* is a method for finding a hypothesis $h \in \mathbb{H}$ that faithfully replicates the unknown target function f. Specifically, in regression we seek to approximate f with a hypothesis h, and in classification we seek to approximate f with a *thresholded* hypothesis $\text{sign}(h)$. Many learning algorithms achieve this goal by minimizing the in-sample error, that is, the empirical average of a loss function $\ell : \mathbb{R} \times \mathbb{Y} \to \mathbb{R}_+$ that estimates the mismatch between the output predicted by $h(\boldsymbol{x})$ and the actual output y for a particular input-output pair $(\boldsymbol{x}, y)$. Any such algorithm solves a minimization problem of the form

$$\inf_{h \in \mathbb{H}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(h(\widehat{\boldsymbol{x}}_i), \widehat{y}_i) = \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\ell(h(\boldsymbol{x}), y) \right] \right\}, \tag{1}$$

where $\widehat{\mathbb{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{(\widehat{\boldsymbol{x}}_i, \widehat{y}_i)}$ denotes the *empirical distribution*, that is, the uniform distribution on the training data. For different choices of the the loss function ℓ, the generic supervised learning problem (1) reduces to different popular regression and classification problems from the literature.

Examples of Regression Models

For ease of exposition, we focus here on learning models with $\mathbb{X} \subseteq \mathbb{R}^n$ and $\mathbb{Y} \subseteq \mathbb{R}$, where $\mathbb{H}$ is set to the space of all linear hypotheses $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ with $\boldsymbol{w} \in \mathbb{R}^n$. Thus, there is a one-to-one correspondence between hypotheses and weight vectors $\boldsymbol{w}$. Moreover, we focus on loss functions of the form $\ell(h(\boldsymbol{x}), y) = L(h(\boldsymbol{x}) - y) = L(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) - y)$ that are generated by a univariate loss function L.

1. A rich class of *robust regression* problems is obtained from (1) if ℓ is generated by the *Huber loss function* with robustness parameter $\delta > 0$, which is defined as $L(z) = \frac{1}{2}z^2$ if $|z| \leq \delta$; $= \delta(|z| - \frac{1}{2}\delta)$ otherwise. Note that the Huber loss function is both convex and smooth and reduces to the squared loss $L(z) = \frac{1}{2}z^2$ for $\delta \uparrow \infty$, which is routinely used in ordinary least squares regression. Problem (1) with squared loss seeks a hypothesis $\boldsymbol{w}$ under which $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ approximates the mean of y conditional on $\boldsymbol{x}$. The Huber loss function for finite δ favors similar hypotheses but is less sensitive to outliers.

2. The *support vector regression* problem [67] emerges as a special case of (1) if ℓ is generated by the *ϵ-insensitive loss function* $L(z) = \max\{0, |z| - \epsilon\}$ with $\epsilon \geq 0$. In this setting, a training sample $(\widehat{\boldsymbol{x}}_i, \widehat{y}_i)$ is penalized in (1) only if the output $\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle$ predicted by hypothesis $\boldsymbol{w}$ differs from the true output $\widehat{y}_i$ by more than ϵ. Support vector regression thus seeks hypotheses $\boldsymbol{w}$ under which all training samples reside within a slab of width 2ϵ centered around the hyperplane $\{(\boldsymbol{x}, y) : \langle \boldsymbol{w}, \boldsymbol{x} \rangle = y\}$.

3. The *quantile regression* problem [41] is obtained from (1) if ℓ is generated by the *pinball loss function* $L(z) = \max\{-\tau z, (1 - \tau)z\}$ defined for $\tau \in [0, 1]$. Quantile regression seeks hypotheses that approximate the $\tau \times 100\%$-quantile of the output conditional on the input. More precisely, it seeks hypotheses $\boldsymbol{w}$ for which $\tau \times 100\%$ of all training samples lie in the halfspace $\{(\boldsymbol{x}, y) : \langle \boldsymbol{w}, \boldsymbol{x} \rangle \geq y\}$.

Examples of Classification Models

We focus here on linear learning models with $\mathbb{X} \subseteq \mathbb{R}^n$ and $\mathbb{Y} = \{+1, -1\}$, where $\mathbb{H}$ is again identified with the space of all linear hypotheses $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ with $\boldsymbol{w} \in \mathbb{R}^n$. Moreover, we focus on loss functions of the form $\ell(\boldsymbol{x}, y) = L(yh(\boldsymbol{x})) = L(y\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$ generated by a univariate loss function L.

1. The *support vector machine* problem [20] is obtained from (1) if ℓ is generated by the *hinge loss function* $L(z) = \max\{0, 1 - z\}$, which is large if z is small. Thus, a training sample $(\widehat{\boldsymbol{x}}_i, \widehat{y}_i)$ is penalized in (1) if the output $\text{sign}(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle)$ predicted by hypothesis $\boldsymbol{w}$ and the true output $\widehat{y}_i$ have opposite signs. More precisely, support vector machines seek hypotheses $\boldsymbol{w}$ under which the inputs of all training samples with output $+1$ reside in the halfspace $\{\boldsymbol{x} : \langle \boldsymbol{w}, \boldsymbol{x} \rangle \geq 1\}$, while the inputs of training samples with output -1 are confined to $\{\boldsymbol{x} : \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leq -1\}$.

2. An alternative support vector machine problem is obtained from (1) if ℓ is generated by the *smooth hinge loss function*, which is defined as $L(z) = \frac{1}{2} - z$ if $z \leq 0$; $= \frac{1}{2}(1 - z)^2$ if $0 < z < 1$; $= 0$ otherwise. The smooth hinge loss inherits many properties of the ordinary hinge loss but has a continuous derivative. Thus, it may be amenable to faster optimization algorithms.

3. The *logistic regression* problem [39] emerges as a special case of (1) if ℓ is generated by the *logloss function* $L(z) = \log(1 + e^{-z})$, which is large if z is small—similar to the hinge loss function. In this case the objective function of (1) can be viewed as the log-likelihood function corresponding to the logistic model $\mathbb{P}(y = 1|\boldsymbol{x}) = [1 + \exp(-\langle \boldsymbol{w}, \boldsymbol{x} \rangle)]^{-1}$ for the conditional probability of $y = 1$ given $\boldsymbol{x}$. Thus, logistic regression allows us to learn the conditional distribution of y given $\boldsymbol{x}$.

Remark 2.1 (Convex approximation). Note that the hinge loss and the logloss functions represent convex approximations for the (discontinuous) *one-zero loss* defined through $L(z) = 1$ if $z \leq 0$; $= 0$ otherwise.

In practice there may be many hypotheses that are compatible with the given training data and thus achieve a small empirical loss in (1). Any such hypothesis would accurately predict outputs from inputs *within the training dataset* [24]. However, due to overfitting, these hypotheses might constitute poor predictors *beyond the training dataset*, that is, on inputs that have not yet been recorded in the database. Mathematically, even if the in-sample error $\mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)]$ of a given hypothesis $\boldsymbol{w}$ is small, the out-of-sample error $\mathbb{E}^{\mathbb{P}}[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)]$ with respect to the unknown true input-output distribution $\mathbb{P}$ may be large.

Regularization is the standard remedy to combat overfitting. Instead of naïvely minimizing the in-sample error as is done in (1), it may thus be advisable to solve the regularized learning problem

$$\inf_{\boldsymbol{w}} \ \mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)] + c\,\Omega(\boldsymbol{w}), \tag{2}$$

which minimizes the sum of the emiprial average loss and a penalty for hpothesis complexity, which consists of a regularization function $\Omega(\boldsymbol{w})$ and its associated regularization weight c. Tikhonov regularization [71], for example, corresponds to the choice $\Omega(\boldsymbol{w}) = \|\boldsymbol{\Gamma}\boldsymbol{w}\|_2^2$ for some Tikhonov matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times n}$. Setting $\boldsymbol{\Gamma}$ to the identity matrix gives rise to standard L_2-regularization. Similarly, Lasso (least absolute shrinkage and selection operator) regularization or L_1-regularization [70] is obtained by setting $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$. Lasso regularization has gained popularity because it favors parsimonious interpretable hypotheses.

Most popular regularization methods admit probabilistic interpretations. However, these interpretations typically rely on prior distributional assumptions that remain to some extent arbitrary (*e.g.*, L_2- and L_1-regularization can be justified if $\boldsymbol{w}$ is governed by a Gaussian or Laplacian prior distribution, respectively [70]). Thus, in spite of their many desirable theoretical properties, there is a consensus that "*most of the (regularization) methods used successfully in practice are heuristic methods*" [1].

2.2 A New Perspective on Regularization

When *linear* hypotheses are used, problem (1) minimizes the in-sample error $\mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)]$. However, a hypothesis $\boldsymbol{w}$ enjoying a low in-sample error may still suffer from a high out-of-sample error $\mathbb{E}^{\mathbb{P}}[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)]$ due to overfitting. This is unfortunate as we seek hypotheses that offer high prediction accuracy on *future* data, meaning that the out-of-sample error is the actual quantity of interest. An ideal learning model would therefore minimize the out-of-sample error. This is impossible, however, for the following reasons:

- The true input-output distribution $\mathbb{P}$ is unknown and only indirectly observable through the N training samples. Thus, we lack essential information to compute the out-of-sample error.

- Even if the distribution $\mathbb{P}$ was known, computing the out-of-sample error would typically be hard due to the intractability of high-dimensional integration; see, *e.g.*, [38, Corollary 1].

The regularized loss $\mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\langle \boldsymbol{w}, \boldsymbol{x}\rangle, y)] + c\,\Omega(\boldsymbol{w})$ used in (2), which consists of the in-sample error and an overfitting penalty, can be viewed as an in-sample estimate of the out-of-sample error. Yet, problem (2) remains difficult to justify rigorously. Therefore, we advocate here a more principled approach to regularization. Specifically, we propose to take into account the expected loss of hypothesis $\boldsymbol{w}$ under *every* distribution $\mathbb{Q}$ that is close to the empirical distribution $\widehat{\mathbb{P}}_N$, that is, every $\mathbb{Q}$ that could have generated the training data with high confidence. To this end, we first introduce a distance measure for distributions. For ease of notation, we henceforth denote the input-output pair $(\boldsymbol{x}, y)$ by $\boldsymbol{\xi}$, and we set $\Xi = \mathbb{X} \times \mathbb{Y}$.

Definition 2.2 (Wasserstein metric). The Wasserstein distance between two distributions $\mathbb{Q}$ and $\mathbb{Q}'$ supported on Ξ is defined as

$$W(\mathbb{Q}, \mathbb{Q}') := \inf_{\Pi} \left\{ \int_{\Xi^2} d(\boldsymbol{\xi}, \boldsymbol{\xi}')\, \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \ : \ \begin{array}{l} \Pi \text{ is a joint distribution of } \boldsymbol{\xi} \text{ and } \boldsymbol{\xi}' \\ \text{with marginals } \mathbb{Q} \text{ and } \mathbb{Q}', \text{ respectively} \end{array} \right\},$$

where d is a metric on Ξ.

By definition, $W(\mathbb{Q}, \mathbb{Q}')$ represents the solution of an infinite-dimensional transportation problem, that is, it corresponds to the minimal cost for moving the distribution $\mathbb{Q}$ to $\mathbb{Q}'$, where the cost for moving a unit probability mass from $\boldsymbol{\xi}$ to $\boldsymbol{\xi}'$ is given by the transportation distance $d(\boldsymbol{\xi}, \boldsymbol{\xi}')$. Due to this interpretation, the metric d is often referred to as the transportation cost [72] or ground metric [22], while the Wasserstein metric is sometimes termed the mass transportation distance or earth mover's distance [57].

Consider now the Wasserstein ball of radius $\rho \geq 0$ around the empirical distribution $\widehat{\mathbb{P}}_N$,

$$\mathbb{B}_\rho(\widehat{\mathbb{P}}_N) = \left\{ \mathbb{Q} \ : \ \mathbb{Q}(\Xi) = 1,\ W(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \rho \right\}, \tag{3}$$

which contains all input-output distributions $\mathbb{Q}$ supported on Ξ whose Wasserstein distance from $\widehat{\mathbb{P}}_N$ does not exceed ρ. This means that $\mathbb{Q}$ can be transported to $\widehat{\mathbb{P}}_N$ (or vice versa) at a cost of at most ρ. The hope is that a large enough Wasserstein ball will contain distributions that are representative of the unknown true input-output distribution $\mathbb{P}$, such that the worst-case expectation $\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\ell(\langle \boldsymbol{w}, \boldsymbol{x}\rangle, y)]$ can serve as an upper confidence bound on the out-of-sample error $\mathbb{E}^{\mathbb{P}}[\ell(\langle \boldsymbol{w}, \boldsymbol{x}\rangle, y)]$. This motivates us to introduce a new regularized learning model, which minimizes precisely this worst-case expectation.

$$\inf_{\boldsymbol{w}} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x}\rangle, y)\right] \tag{4}$$

Problem (4) represents a distributionally robust convex program of the type considered in [52]. Note that if $\ell(\langle \boldsymbol{w}, \boldsymbol{x}\rangle, y)$ is convex in $\boldsymbol{w}$ for every fixed $(\boldsymbol{x}, y)$, *i.e.*, if ℓ is convex in its first argument, then the objective function of (4) is convex because convexity is preserved under integration and maximization. Note also that if ρ is set to zero, then (4) collapses to the unregularized in-sample error minimization problem (1).

Remark 2.3 (Support information). The uncertainty set Ξ captures prior information on the range of the inputs and outputs. In image processing, for example, pixel intensities range over a known interval. Similarly, in diagnostic medicine, physiological parameters such as blood glucose or cholesterol concentrations are restricted to be non-negative. Sometimes it is also useful to construct Ξ as a confidence set that covers the support of $\mathbb{P}$ with a prescribed probability. Such confidence sets are often constructed as ellipsoids, as intersections of different norm balls [6, 25] or as sublevel sets of kernel expansions [60].

In the remainder we establish that the distributionally robust learning problem (4) has several desirable properties. (i) Problem (4) is computationally tractable under standard assumptions about the loss function ℓ, the input-output space Ξ and the transportation metric d. For specific choices of d it even reduces to a regularized learning problem of the form (2). (ii) For all univariate loss functions reviewed in

Section 2.1, a tight conservative approximation of (4) is kernelizable, that is, it can be solved implicitly over high-dimensional spaces of nonlinear hypotheses at the same computational cost required for linear hypothesis spaces. (iii) Leveraging modern measure concentration results, the optimal value of (4) can be shown to provide an upper confidence bound on the out-of-sample error. This obviates the need to mobilize the full machinery of VC theory and, in particular, to estimate the VC dimension of the hypothesis space in order to establish generalization bounds. (iv) If the number of training samples tends to infinity while the Wasserstein ball shrinks at an appropriate rate, then problem (4) asymptotically recovers the *ex post* optimal hypothesis that attains the minimal out-of-sample error.

3 Tractable Reformulations

In this section we demonstrate that the distributionally robust learning problem (4) over linear hypotheses is amenable to efficient computational solution procedures. We also discuss generalizations to nonlinear hypothesis classes such as reproducing kernel Hilbert spaces and families of feed-forward neural networks.

3.1 Distributionally Robust Linear Regression

Throughout this section we focus on linear regression problems, where $\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) = L(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - y)$ for some convex univariate loss function L. We also assume that $\mathbb{X}$ and $\mathbb{Y}$ are both convex and closed and that the transportation metric d is induced by a norm $\| \cdot \|$ on the input-output space $\mathbb{R}^{n+1}$. In this setting, the distributionally robust regression problem (4) admits an equivalent reformulation as a finite convex optimization problem if either (i) the univariate loss function L is piecewise affine or (ii) $\Xi = \mathbb{R}^{n+1}$ and L is Lipschitz continuous (but not necessarily piecewise affine).

Theorem 3.1 (Distributionally robust linear regression)**.** The following statements hold.

(i) If $L(z) = \max_{j \leq J}\{a_j z + b_j\}$, then (4) is equivalent to

$$
\begin{cases}
\displaystyle\inf_{\substack{\boldsymbol{w}, \lambda, s_i \\ \boldsymbol{p}_{ij}, u_{ij}}} & \lambda \rho + \dfrac{1}{N} \sum_{i=1}^{N} s_i \\
\text{s.t.} & S_{\Xi}(a_j \boldsymbol{w} - \boldsymbol{p}_{ij}, -a_j - u_{ij}) + b_j + \langle \boldsymbol{p}_{ij}, \widehat{\boldsymbol{x}}_i \rangle + u_{ij} \widehat{y}_i \leq s_i & i \in [N], j \in [J] \\
& \|(\boldsymbol{p}_{ij}, u_{ij})\|_* \leq \lambda & i \in [N], j \in [J],
\end{cases}
\tag{5}
$$

where S_{Ξ} denotes the support function of Ξ.

(ii) If $\Xi = \mathbb{R}^{n+1}$ and $L(z)$ is Lipschitz continuous, then (4) is equivalent to

$$
\inf_{\boldsymbol{w}} \ \frac{1}{N} \sum_{i=1}^{N} L(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle - \widehat{y}_i) + \rho \operatorname{lip}(L) \, \|(\boldsymbol{w}, -1)\|_*.
\tag{6}
$$

In the following, we exemplify Theorem 3.1 for the Huber, ϵ-insensitive and pinball loss functions under the assumption that the uncertainty set Ξ admits the conic representation

$$
\Xi = \{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} : \boldsymbol{C}_1 \boldsymbol{x} + \boldsymbol{c}_2 y \preceq_{\mathcal{C}} \boldsymbol{d}\}
\tag{7}
$$

for some matrix $\boldsymbol{C}_1$, vectors $\boldsymbol{c}_2$ and $\boldsymbol{d}$ and proper convex cone $\mathcal{C}$ of appropriate dimensions. We also assume that Ξ admits a Slater point $(\boldsymbol{x}_{\mathrm{S}}, y_{\mathrm{S}}) \in \mathbb{R}^{n+1}$ with $\boldsymbol{C}_1 \boldsymbol{x}_{\mathrm{S}} + \boldsymbol{c}_2 y_{\mathrm{S}} \prec_{\mathcal{C}} \boldsymbol{d}$.

Corollary 3.2 (Robust regression)**.** If L represents the Huber loss function with threshold $\delta \geq 0$ and $\Xi = \mathbb{R}^{n+1}$, then (4) is equivalent to

$$
\inf_{\boldsymbol{w}, z_i} \ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} z_i^2 + \delta |\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle - \widehat{y}_i - z_i| + \rho \, \delta \, \|(\boldsymbol{w}, -1)\|_*.
\tag{8}
$$

Corollary 3.3 (Support vector regression). If L represents the ϵ-insensitive loss function for some $\epsilon \geq 0$ and Ξ is of the form (7), then (4) is equivalent to

$$
\begin{cases}
\displaystyle\inf_{\substack{\boldsymbol{w},\lambda,s_i \\ \boldsymbol{p}_i^+,\boldsymbol{p}_i^-}} \quad \lambda\rho + \frac{1}{N}\sum_{i=1}^N s_i \\
\text{s.t.} \quad \widehat{y}_i - \langle\boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle - \epsilon + \langle\boldsymbol{p}_i^+,\boldsymbol{d} - \boldsymbol{C}_1\widehat{\boldsymbol{x}}_i - \boldsymbol{c}_2\widehat{y}_i\rangle \leq s_i & i \in [N] \\
\qquad \langle\boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i - \epsilon + \langle\boldsymbol{p}_i^-,\boldsymbol{d} - \boldsymbol{C}_1\widehat{\boldsymbol{x}}_i - \boldsymbol{c}_2\widehat{y}_i\rangle \leq s_i & i \in [N] \\
\qquad \left\|\left(\boldsymbol{C}_1^\top\boldsymbol{p}_i^+ + \boldsymbol{w}, \boldsymbol{c}_2^\top\boldsymbol{p}_i^+ - 1\right)\right\|_* \leq \lambda & i \in [N] \\
\qquad \left\|\left(\boldsymbol{C}_1^\top\boldsymbol{p}_i^- - \boldsymbol{w}, \boldsymbol{c}_2^\top\boldsymbol{p}_i^- + 1\right)\right\|_* \leq \lambda & i \in [N] \\
\qquad \boldsymbol{p}_i^+, \boldsymbol{p}_i^- \in \mathcal{C}^* & i \in [N] \\
\qquad s_i \geq 0 & i \in [N].
\end{cases}
\tag{9}
$$

Corollary 3.4 (Quantile regression). If L represents the pinball loss function for some $\tau \in [0,1]$ and Ξ is of the form (7), then (4) is equivalent to

$$
\begin{cases}
\displaystyle\inf_{\substack{\boldsymbol{w},\lambda,s_i \\ \boldsymbol{p}_i^+,\boldsymbol{p}_i^-}} \quad \lambda\rho + \frac{1}{N}\sum_{i=1}^N s_i \\
\text{s.t.} \quad \tau\left(\widehat{y}_i - \langle\boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle\right) + \langle\boldsymbol{p}_i^+,\boldsymbol{d} - \boldsymbol{C}_1\widehat{\boldsymbol{x}}_i - \boldsymbol{c}_2\widehat{y}_i\rangle \leq s_i & i \in [N] \\
\qquad (1-\tau)\left(\langle\boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i\right) + \langle\boldsymbol{p}_i^-,\boldsymbol{d} - \boldsymbol{C}_1\widehat{\boldsymbol{x}}_i - \boldsymbol{c}_2\widehat{y}_i\rangle \leq s_i & i \in [N] \\
\qquad \left\|\left(\boldsymbol{C}_1^\top\boldsymbol{p}_i^+ + \tau\boldsymbol{w}, \boldsymbol{c}_2^\top\boldsymbol{p}_i^+ - \tau\right)\right\|_* \leq \lambda & i \in [N] \\
\qquad \left\|\left(\boldsymbol{C}_1^\top\boldsymbol{p}_i^- - (1-\tau)\boldsymbol{w}, \boldsymbol{c}_2^\top\boldsymbol{p}_i^- + 1 - \tau\right)\right\|_* \leq \lambda & i \in [N] \\
\qquad \boldsymbol{p}_i^+, \boldsymbol{p}_i^- \in \mathcal{C}^* & i \in [N] \\
\qquad s_i \geq 0 & i \in [N].
\end{cases}
\tag{10}
$$

Remark 3.5 (Relation to classical regularization). Assume now that the mass transportation costs are additively separable with respect to inputs and outputs, that is,

$$
d((\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2)) = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \kappa|y_1 - y_2|
\tag{11}
$$

for some $\kappa > 0$.[1] Note that κ captures the costs of moving probability mass along the output space. For $\kappa = \infty$ all distributions in the Wasserstein ball $\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ are thus obtained by reshaping $\widehat{\mathbb{P}}_N$ only along the input space. It is easy to verify that for $\kappa = \infty$ and $\Xi = \mathbb{R}^{n+1}$ the learning models portrayed in Corollaries 3.2-3.4 all simplify to

$$
\inf_{\boldsymbol{w}} \quad \frac{1}{N}\sum_{i=1}^N L(\langle\boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i) + c\|\boldsymbol{w}\|_*,
\tag{12}
$$

where $c = \rho\delta$ for robust regression with Huber loss, $c = \rho$ for support vector regression with ϵ-insensitive loss and $c = \max\{\tau, 1-\tau\}\rho$ for quantile regression with pinball loss. Thus, (12) is easily identified as an instance of the classical regularized learning problem (2), where the dual norm term $\|\boldsymbol{w}\|_*$ plays the role of the regularization function, while c represents the usual regularization weight. By definition of the dual norm, the penalty $\|\boldsymbol{w}\|_*$ assigned to a hypothesis $\boldsymbol{w}$ is maximal (minimal) if the cost of moving probability mass along $\boldsymbol{w}$ is minimal (maximal). We emphasize that if $\kappa = \infty$, then the marginal distribution of y corresponding to *every* $\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ coincides with the empirical distribution $\frac{1}{N}\sum_{i=1}^N \delta_{\widehat{y}_i}$. Thus, classical regularization methods, which correspond to $\kappa = \infty$, are explained by a counterintuitive probabilistic model, which pretends that any training sample must have an output that has already been recorded in the training dataset. In other words, classical regularization implicitly assumes that there is no uncertainty in the outputs. More intuitively appealing regularization schemes are obtained for finite values of κ.

---

[1]By slight abuse of notation, the symbol $\|\cdot\|$ now denotes a norm on $\mathbb{R}^n$.

To establish a connection between distributionally robust and classical robust regression as discussed in [28, 77], we further investigate the worst-case expected loss of a fixed linear hypothesis $\boldsymbol{w}$.

$$\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[L(\langle \boldsymbol{w}, \boldsymbol{x}\rangle - y)] \tag{13}$$

Theorem 3.6 (Extremal distributions in linear regression). The following statements hold.

(i) If $L(z) = \max_{j\leq J}\{a_j z + b_j\}$, then the worst-case expectation (13) coincides with

$$\begin{cases} \sup_{\alpha_{ij},\boldsymbol{q}_{ij},v_{ij}} & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\alpha_{ij}\big(a_j(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i) + b_j\big) + a_j(\langle \boldsymbol{w}, \boldsymbol{q}_{ij}\rangle - v_{ij}) \\[2mm] \text{s.t.} & \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\|(\boldsymbol{q}_{ij}, v_{ij})\| \leq \rho \\[2mm] & \sum_{j=1}^{J}\alpha_{ij} = 1 & i\in[N] \\[2mm] & (\widehat{\boldsymbol{x}}_i + \boldsymbol{q}_{ij}/\alpha_{ij}, \widehat{y}_i + v_{ij}/\alpha_{ij}) \in \Xi & i\in[N], j\in[J] \\[1mm] & \alpha_{ij} \geq 0 & i\in[N], j\in[J] \end{cases} \tag{14}$$

for any fixed hypothesis $\boldsymbol{w}$. Moreover, if $(\alpha_{ij}^\star, \boldsymbol{q}_{ij}^\star, v_{ij}^\star)$ maximizes (14), then the discrete distribution

$$\mathbb{Q}^* = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\alpha_{ij}^\star\,\delta_{(\widehat{\boldsymbol{x}}_i + \boldsymbol{q}_{ij}^\star/\alpha_{ij}^\star,\,\widehat{y}_i + v_{ij}^\star/\alpha_{ij}^\star)},$$

represents a maximizer for (13).

(ii) If $\Xi = \mathbb{R}^{n+1}$ and $L(z)$ is Lipschitz continuous, then the discrete distributions

$$\mathbb{Q}_\gamma = \frac{1}{N}\sum_{i=2}^{N}\delta_{(\widehat{\boldsymbol{x}}_i,\widehat{y}_i)} + \frac{1-\gamma}{N}\delta_{(\widehat{\boldsymbol{x}}_1,\widehat{y}_1)} + \frac{\gamma}{N}\delta_{(\widehat{\boldsymbol{x}}_1 + \frac{\rho N}{\gamma}\boldsymbol{x}^\star,\widehat{y}_1 + \frac{\rho N}{\gamma}y^\star)} \quad \text{for} \quad \gamma\in(0,1],$$

where $(\boldsymbol{x}^\star, y^\star)$ solves $\max_{\|(\boldsymbol{x},y)\|\leq 1}\langle \boldsymbol{w}, \boldsymbol{x}\rangle - y$, are feasible and asymptotically optimal in (13) for $\gamma\downarrow 0$.

Recall that $0/0 = 0$ and $1/0 = \infty$ by our conventions of extended arithmetic. Thus, any solution feasible in (14) with $\alpha_{ij} = 0$ must satisfy $\boldsymbol{q}_{ij} = \boldsymbol{0}$ and $v_{ij} = 0$ because otherwise $(\widehat{\boldsymbol{x}}_i + \boldsymbol{q}_{ij}/\alpha_{ij}, \widehat{y}_i + v_{ij}/\alpha_{ij}) \notin \Xi$.

Theorem 3.6 shows how one can use convex optimization to construct a sequence of distributions that are asymptotically optimal in (13). Next, we argue that the worst-case expected cost (13) is equivalent to a (robust) worst-case cost over a suitably defined uncertainty set if the following assumption holds.

Assumption 3.7 (Minimal dispersion). For every $\boldsymbol{w}\in\mathbb{R}^n$ there is a training sample $(\widehat{\boldsymbol{x}}_k, \widehat{y}_k)$ for some $k\leq N$ such that the derivative L' exists at $\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k\rangle - \widehat{y}_k$ and satisfies $|L'(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k\rangle - \widehat{y}_k)| = \mathrm{lip}(L)$.

Remark 3.8 (Minimal dispersion). Assumption 3.7 is reminiscent of the non-separability condition in [76, Theorem 3], which is necessary to prove the equivalence of robust and regularized support vector machines. In the regression context studied here, Assumption 3.7 ensures that, for every $\boldsymbol{w}$, there exists a training sample that activates the largest absolute slope of L.

For instance, in support vector regression, it means that for every $\boldsymbol{w}$ there exists a data point outside of the slab of width $2\epsilon/\|(\boldsymbol{w}, -1)\|_2$ centered around the hyperplane $H_{\boldsymbol{w}} = \{(\boldsymbol{x}, y)\in\mathbb{R}^n\times\mathbb{R} : \langle \boldsymbol{w}, \boldsymbol{x}\rangle - y = 0\}$ (*i.e.*, the empirical ϵ-insensitive loss is not zero). Similarly, in robust regression with the Huber loss function, Assumption 3.7 stipulates that for every $\boldsymbol{w}$ there exists a data point outside of the slab of width $2\delta/\|(\boldsymbol{w}, -1)\|_2$ centered around $H_{\boldsymbol{w}}$. However, quantile regression with $\tau\neq 0.5$ fails to satisfy Assumption 3.7. Indeed, for any training dataset there always exists some $\boldsymbol{w}$ such that all data points reside on the side of $H_{\boldsymbol{w}}$ where the pinball loss function is less steep.

Theorem 3.9 (Robust regression). If $\Xi = \mathbb{R}^{n+1}$ and the loss function $L(z)$ is Lipschitz continuous, then the worst-case expected loss (13) provides an upper bound on the (robust) worst-case loss

$$
\begin{cases}
\displaystyle \sup_{\Delta \boldsymbol{x}_i, \Delta y_i} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} [L(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i + \Delta \boldsymbol{x}_i \rangle - \widehat{y}_i - \Delta y_i)] \\
\text{s.t.} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \|(\Delta \boldsymbol{x}_i, \Delta y_i)\| \le \rho.
\end{cases}
\tag{15}
$$

Moreover, if Assumption 3.7 holds, then (13) and (15) are equal.

Remark 3.10 (Tractability of robust regression). Assume that $\Xi = \mathbb{R}^{n+1}$, while L and $\| \cdot \|$ both admit a tractable conic representation. By Theorem 3.1, the worst-case *expected* loss (13) can then be computed in polynomial time by solving a tractable convex program. Theorem 3.9 thus implies that the worst-case loss (28) can also be computed in polynomial time if Assumption 3.7 holds. To our best knowledge, there exists no generic efficient method for computing (28) if Assumption 3.7 fails to hold and L is *not* piecewise affine. This reinforces our belief that a distributionally robust approach to regression is more natural.

3.2 Distributionally Robust Linear Classification

Throughout this section we focus on linear classification problems, where $\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) = L(y\langle \boldsymbol{w}, \boldsymbol{x} \rangle)$ for some convex univariate loss function L. We also assume that $\mathbb{X}$ is both convex and closed and that $\mathbb{Y} = \{+1, -1\}$. Moreover, we assume that the transportation metric d is defined via

$$
d((\boldsymbol{x}, y), (\boldsymbol{x}', y')) = \|\boldsymbol{x} - \boldsymbol{x}'\| + \kappa \mathbb{1}_{\{y \ne y'\}},
\tag{16}
$$

where $\| \cdot \|$ represents a norm on the input space $\mathbb{R}^n$, and $\kappa > 0$ quantifies the cost of switching a label. In this setting, the distributionally robust classification problem (4) admits an equivalent reformulation as a finite convex optimization problem if either (i) the univariate loss function L is piecewise affine or (ii) $\mathbb{X} = \mathbb{R}^n$ and L is Lipschitz continuous (but not necessarily piecewise affine).

Theorem 3.11 (Distributionally robust linear classification). The following statements hold.

(i) If $L(z) = \max_{j \in J}\{a_j z + b_j\}$, then (4) is equivalent to

$$
\begin{cases}
\displaystyle \inf_{\substack{\boldsymbol{w}, \lambda, s_i \\ \boldsymbol{p}_{ij}^+, \boldsymbol{p}_{ij}^-}} & \displaystyle \lambda \rho + \frac{1}{N} \sum_{i=1}^{N} s_i \\
\text{s.t.} & S_{\mathbb{X}}(a_j \widehat{y}_i \boldsymbol{w} - \boldsymbol{p}_{ij}^+) + b_j + \langle \boldsymbol{p}_{ij}^+, \widehat{\boldsymbol{x}}_i \rangle \le s_i & i \in [N], j \in [J] \\
& S_{\mathbb{X}}(-a_j \widehat{y}_i \boldsymbol{w} - \boldsymbol{p}_{ij}^-) + b_j + \langle \boldsymbol{p}_{ij}^-, \widehat{\boldsymbol{x}}_i \rangle - \kappa \lambda \le s_i & i \in [N], j \in [J] \\
& \|\boldsymbol{p}_{ij}^+\|_* \le \lambda, \|\boldsymbol{p}_{ij}^-\|_* \le \lambda & i \in [N], j \in [J],
\end{cases}
\tag{17}
$$

where $S_{\mathbb{X}}$ denotes the support function of $\mathbb{X}$.

(ii) If $\mathbb{X} = \mathbb{R}^n$ and L is Lipschitz continuous, then (4) is equivalent to

$$
\begin{cases}
\displaystyle \inf_{\boldsymbol{w}, \lambda, s_i} & \displaystyle \lambda \rho + \frac{1}{N} \sum_{i=1}^{N} s_i \\
\text{s.t.} & L(\widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle) \le s_i & i \in [N] \\
& L(-\widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle) - \kappa \lambda \le s_i & i \in [N] \\
& \mathrm{lip}(L)\|\boldsymbol{w}\|_* \le \lambda.
\end{cases}
\tag{18}
$$

In the following, we exemplify Theorem 3.11 for the hinge loss, logloss and smoothed hinge loss functions under the assumption that the input space $\mathbb{X}$ admits the conic representation

$$
\mathbb{X} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{C}\boldsymbol{x} \preceq_{\mathcal{C}} \boldsymbol{d}\}
\tag{19}
$$

for some matrix $\boldsymbol{C}$, vector $\boldsymbol{d}$ and proper convex cone $\mathcal{C}$ of appropriate dimensions. We also assume that $\mathbb{X}$ admits a Slater point $\boldsymbol{x}_{\mathrm{S}} \in \mathbb{R}^n$ with $\boldsymbol{C}\boldsymbol{x}_{\mathrm{S}} \prec_{\mathcal{C}} \boldsymbol{d}$.

Corollary 3.12 (Support vector machine). If L represents the hinge loss function and $\mathbb{X}$ is of the form (19), then (4) is equivalent to

$$
\begin{cases}
\inf_{\substack{\boldsymbol{w},\lambda \\ s_i,\boldsymbol{p}_i^+,\boldsymbol{p}_i^-}} & \lambda\rho + \dfrac{1}{N}\sum_{i=1}^N s_i & \\
\text{s.t.} & 1 - \widehat{y}_i\left(\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle\right) + \langle \boldsymbol{p}_i^+, \boldsymbol{d} - \boldsymbol{C}\widehat{\boldsymbol{x}}_i\rangle \le s_i & i \in [N] \\
& 1 + \widehat{y}_i\left(\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle\right) + \langle \boldsymbol{p}_i^-, \boldsymbol{d} - \boldsymbol{C}\widehat{\boldsymbol{x}}_i\rangle - \kappa\lambda \le s_i & i \in [N] \\
& \|\boldsymbol{C}^\top \boldsymbol{p}_i^+ + \widehat{y}_i\boldsymbol{w}\|_* \le \lambda, \quad \|\boldsymbol{C}^\top \boldsymbol{p}_i^- - \widehat{y}_i\boldsymbol{w}\|_* \le \lambda & i \in [N] \\
& s_i \ge 0, \quad \boldsymbol{p}_i^+, \boldsymbol{p}_i^- \in \mathcal{C}^* & i \in [N].
\end{cases}
\tag{20}
$$

Corollary 3.13 (Support vector machine with smooth hinge loss). If L represents the smooth hinge loss function and $\mathbb{X} = \mathbb{R}^n$, then (4) is equivalent to

$$
\begin{cases}
\min_{\substack{\boldsymbol{w},\lambda,s_i \\ z_i^+,z_i^-,t_i^+,t_i^-}} & \lambda\rho + \dfrac{1}{N}\sum_{i=1}^N s_i & \\
\text{s.t.} & \frac{1}{2}(z_i^+ - \widehat{y}_i\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle)^2 + t_i^+ \le s_i & i \in [N] \\
& \frac{1}{2}(z_i^- + \widehat{y}_i\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle)^2 + t_i^- - \kappa\lambda \le s_i & i \in [N] \\
& 1 - z_i^+ \le t_i^+, \quad 1 - z_i^- \le t_i^- & i \in [N] \\
& t_i^+, t_i^- \ge 0 & i \in [N] \\
& \|\boldsymbol{w}\|_* \le \lambda.
\end{cases}
\tag{21}
$$

Corollary 3.14 (Logistic regression). If L represents the logloss function and $\mathbb{X} = \mathbb{R}^n$, then (4) is equivalent to

$$
\begin{cases}
\min_{\boldsymbol{w},\lambda,s_i} & \lambda\rho + \dfrac{1}{N}\sum_{i=1}^N s_i & \\
\text{s.t.} & \log\left(1 + \exp\left(-\widehat{y}_i\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle\right)\right) \le s_i & i \in [N] \\
& \log\left(1 + \exp\left(\widehat{y}_i\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle\right)\right) - \kappa\lambda \le s_i & i \in [N] \\
& \|\boldsymbol{w}\|_* \le \lambda.
\end{cases}
\tag{22}
$$

Remark 3.15 (Relation to classical regularization). If $\mathbb{X} = \mathbb{R}^n$ and the weight parameter κ in the transportation metric (16) is set to infinity, then the learning problems portrayed in Corollaries 3.12–3.14 all simplify to

$$
\inf_{\boldsymbol{w}} \ \frac{1}{N}\sum_{i=1}^N L(\widehat{y}_i\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle) + \rho\|\boldsymbol{w}\|_*.
\tag{23}
$$

Thus, in analogy to the case of regression, (23) reduces to an instance of the classical regularized learning problem (2), where the dual norm term $\|\boldsymbol{w}\|_*$ plays the role of the regularization function, while the Wasserstein radius ρ represents the usual regularization weight. Note that if $\kappa = \infty$, then mass transportation along the output space is infinitely expensive, that is, any distribution $\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ can smear out the training samples along $\mathbb{X}$, but it cannot flip outputs from $+1$ to -1 or vice versa. Thus, classical regularization schemes, which are recovered for $\kappa = \infty$, implicitly assume that output measurements are exact. As this belief is not tenable in most applications, an approach with $\kappa < \infty$ may be more satisfying. We remark that alternative approaches for learning with noisy labels have previously been studied by Lawrence and Schölkopf [45], Natarajan *et al.* [53], and Yang *et al.* [78].

Remark 3.16 (Relation to Tikhonov regularization). The learning problem

$$
\inf_{\boldsymbol{w}} \ \frac{1}{N}\sum_{i=1}^N L(\widehat{y}_i\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle) + c\|\boldsymbol{w}\|_2^2
\tag{24}
$$

with Tikhonov regularizer enjoys wide popularity. If L represents the hinge loss, for example, then (24) reduces to the celebrated soft margin support vector machine problem. However, the Tikhonov regularizer appearing in (24) is not explained by a distributionally robust learning problem of the form (4). It is known, however, that (23) with $\|\cdot\|_* = \|\cdot\|_2$ and (24) are equivalent in the sense that for every $\rho \geq 0$ there exists $c \geq 0$ such that the solution of (23) also solves (24) and vice versa [76, Corollary 6].

To establish a connection between distributionally robust and classical robust classification as discussed in [76], we further investigate the worst-case expected loss of a fixed linear hypothesis $\boldsymbol{w}$.

$$\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[L(y\langle \boldsymbol{w}, \boldsymbol{x}\rangle)] \tag{25}$$

Theorem 3.17 (Extremal distributions in linear classification). The following statements hold.

(i) If $L(z) = \max_{j \in J}\{a_j z + b_j\}$, then the worst-case expectation (25) coincides with

$$
\begin{cases}
\displaystyle\sup_{\substack{\alpha_{ij}^+, \alpha_{ij}^- \\ \boldsymbol{q}_{ij}^+, \boldsymbol{q}_{ij}^-}} \quad \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^J (\alpha_{ij}^+ - \alpha_{ij}^-)a_j \widehat{y}_i\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i\rangle + a_j \widehat{y}_i\langle \boldsymbol{w}, \boldsymbol{q}_{ij}^+ - \boldsymbol{q}_{ij}^-\rangle + \sum_{j=1}^J b_j \\[2ex]
\text{s.t.} \quad \displaystyle\sum_{i=1}^N \sum_{j=1}^J \|\boldsymbol{q}_{ij}^+\| + \|\boldsymbol{q}_{ij}^-\| + \kappa \alpha_{ij}^- \leq N\rho \\[2ex]
\qquad \displaystyle\sum_{j=1}^J \alpha_{ij}^+ + \alpha_{ij}^- = 1 \qquad\qquad\qquad\qquad\qquad\quad i \in [N] \\[2ex]
\qquad \widehat{\boldsymbol{x}}_i + \boldsymbol{q}_{ij}^+/\alpha_{ij}^+ \in \mathbb{X}, \quad \widehat{\boldsymbol{x}}_i + \boldsymbol{q}_{ij}^-/\alpha_{ij}^- \in \mathbb{X} \qquad i \in [N], j \in [J] \\[1ex]
\qquad \alpha_{ij}^+, \alpha_{ij}^- \geq 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad i \in [N], j \in [J]
\end{cases}
\tag{26}
$$

for any fixed $\boldsymbol{w}$. Also, if $(\alpha_{ij}^{+\star}, \alpha_{ij}^{-\star}, \boldsymbol{q}_{ij}^{+\star}, \boldsymbol{q}_{ij}^{-\star})$ maximizes (26), then the discrete distribution

$$\mathbb{Q} = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^J \alpha_{ij}^{+\star} \delta_{(\widehat{\boldsymbol{x}}_i - \boldsymbol{q}_{ij}^{+\star}/\alpha_{ij}^{+\star},\, \widehat{y}_i)} + \alpha_{ij}^{-\star} \delta_{(\widehat{\boldsymbol{x}}_i - \boldsymbol{q}_{ij}^{-\star}/\alpha_{ij}^{-\star},\, -\widehat{y}_i)}$$

represents a maximizer for (25).

(ii) If $\mathbb{X} = \mathbb{R}^n$, then the worst-case expectation (25) coincides with the optimal value of

$$
\begin{cases}
\displaystyle\sup_{\alpha_i, \theta} \quad \mathrm{lip}(L)\|(\boldsymbol{w})\|_*\theta + \frac{1}{N}\sum_{i=1}^N (1-\alpha_i)L(\widehat{y}_i\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i\rangle) + \alpha_i L(-\widehat{y}_i\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i\rangle) \\[2ex]
\text{s.t.} \quad \theta + \dfrac{\kappa}{N}\displaystyle\sum_{i=1}^N \alpha_i = \rho - \gamma \\[2ex]
\qquad 0 \leq \alpha_i \leq 1 \qquad\qquad i \in [N] \\[1ex]
\qquad \theta \geq 0
\end{cases}
\tag{27}
$$

for $\gamma = 0$. Moreover, if $(\alpha_i^\star(\gamma), \theta^\star(\gamma))$ maximizes (27) for $\gamma > 0$, $\eta(\gamma) = \gamma/(\theta^\star(\gamma) + \kappa - \rho + \gamma + 1)$ and $\boldsymbol{x}^\star$ solves $\max_{\boldsymbol{x}}\{\langle \boldsymbol{w}, \boldsymbol{x}\rangle : \|\boldsymbol{x}\| \leq 1\}$, then the discrete distributions

$$\mathbb{Q}_\gamma = \frac{1}{N}\sum_{i=2}^N (1-\alpha_i^\star(\gamma))\,\delta_{(\widehat{\boldsymbol{x}}_i,\, \widehat{y}_i)} + \alpha_i^\star(\gamma)\,\delta_{(\widehat{\boldsymbol{x}}_i,\, -\widehat{y}_i)} + \frac{1-\eta(\gamma)}{N}\left[(1-\alpha_1^\star(\gamma))\,\delta_{(\widehat{\boldsymbol{x}}_1,\, \widehat{y}_1)} + \alpha_1^\star(\gamma)\,\delta_{(\widehat{\boldsymbol{x}}_1,\, -\widehat{y}_1)}\right]$$
$$+ \frac{\eta(\gamma)}{N}\,\delta_{(\widehat{\boldsymbol{x}}_1 + \frac{\theta^\star(\gamma)N}{\eta(\gamma)}\boldsymbol{x}^\star,\, \widehat{y}_1)}$$

for $\gamma \in [0, \min\{\rho, 1\}]$ are feasible and asymptotically optimal in (25) for $\gamma \downarrow 0$.

13

Theorem 3.17 shows how one can use convex optimization to construct a sequence of distributions that are asymptotically optimal in (25). Next, we show that the worst-case expected cost (25) is equivalent to a (robust) worst-case cost over a suitably defined uncertainty set if the following assumption holds.

Assumption 3.18 (Non-separability). For every $\boldsymbol{w} \in \mathbb{R}^n$ there is a training sample $(\widehat{\boldsymbol{x}}_k, \widehat{y}_k)$ for some $k \leq N$ such that the derivative L' exists at $\widehat{y}_k \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle$ and satisfies $|L'(\widehat{y}_k \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle)| = \mathrm{lip}(L)$.

Remark 3.19 (Non-separability). Assumption 3.18 generalizes the non-separability condition in [76, Theorem 3] for the classical and smooth hinge loss functions to more general Lipschitz continuous losses. Note that, in the case of the hinge loss, Assumption 3.18 effectively stipulates that for any $\boldsymbol{w}$ there exists a training sample $(\widehat{\boldsymbol{x}}_k, \widehat{y}_k)$ with $\widehat{y}_k \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle < 1$, implying that the dataset cannot be perfectly separated by any linear hypothesis $\boldsymbol{w}$. An equivalent requirement is that the empirical hinge loss is nonzero for every $\boldsymbol{w}$. Similarly, in the case of the smooth hinge loss, Assumption 3.18 ensures that for any $\boldsymbol{w}$ there is a training sample with $\widehat{y}_k \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle < 0$, which implies again that the dataset admits no perfect linear separation. Note, however, that the logloss fails to satisfy Assumption 3.18 as its steepest slope is attained at infinity.

Theorem 3.20 (Robust classification). Suppose that $\mathbb{X} = \mathbb{R}^n$, the loss function L is Lipschitz continuous and the cost of flipping a label in the transportation metric (16) is set to $\kappa = \infty$. Then, the worst-case expected loss (25) provides an upper bound on the (robust) worst-case loss

$$
\begin{cases}
\displaystyle \sup_{\Delta \boldsymbol{x}_i} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} L\left(\widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i + \Delta \boldsymbol{x}_i \rangle\right) \\
\text{s.t.} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \|\Delta \boldsymbol{x}_i\| \leq \rho.
\end{cases}
\tag{28}
$$

Moreover, if Assumption 3.18 holds, then (25) and (28) are equal.

Remark 3.21 (Tractability of robust classification). Assume that $\mathbb{X} = \mathbb{R}^n$, while L and $\|\cdot\|$ both admit a tractable conic representation. By Theorem 3.11, the worst-case expected loss (25) can then be computed in polynomial time by solving a tractable convex program. Theorem 3.20 thus implies that the worst-case loss (28) can also be computed in polynomial time if Assumption 3.18 holds. This confirms Proposition 4 in [76]. No efficient method for computing (28) is know if Assumption 3.18 fails to hold.

3.3 Nonlinear Hypotheses: Reproducing Kernel Hilbert Spaces

We now generalize the learning models from Sections 3.1 and 3.2 to nonlinear hypotheses that range over a *reproducing kernel Hilbert space* (RKHS) $\mathbb{H} \subseteq \mathbb{R}^{\mathbb{X}}$ with inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$. By definition, $\mathbb{H}$ thus constitutes a complete metric space with respect to the norm $\|\cdot\|_{\mathbb{H}}$ induced by the inner product, and the point evaluation $h \mapsto h(\boldsymbol{x})$ of the functions $h \in \mathbb{H}$ represents a continuous linear functional on $\mathbb{H}$ for any fixed $\boldsymbol{x} \in \mathbb{X}$. The Riesz representation theorem then implies that for every $\boldsymbol{x} \in \mathbb{X}$ there exists a unique function $\Phi(\boldsymbol{x}) \in \mathbb{H}$ such that $h(\boldsymbol{x}) = \langle h, \Phi(\boldsymbol{x}) \rangle_{\mathbb{H}}$ for all $h \in \mathbb{H}$. We henceforth refer to $\Phi : \mathbb{X} \to \mathbb{H}$ as the *feature map* and to $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$ with $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle_{\mathbb{H}}$ as the *kernel function*. By construction, the kernel function is symmetric and positive definite, that is, the *kernel matrix* $\boldsymbol{\mathcal{K}} \in \mathbb{R}^{N \times N}$ defined through $\boldsymbol{\mathcal{K}}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is positive definite for all $N \in \mathbb{N}$ and $\{\boldsymbol{x}_i\}_{i \leq N} \subseteq \mathbb{X}$.

By the Moore-Aronszajn theorem, any symmetric and positive definite kernel function k on $\mathbb{X}$ induces a unique RKHS $\mathbb{H} \subseteq \mathbb{R}^{\mathbb{X}}$, which can be represented as

$$
\mathbb{H} = \left\{ h \in \mathbb{R}^{\mathbb{X}} : \exists \beta_i \in \mathbb{R}, \ \boldsymbol{x}_i \in \mathbb{X} \ \forall i \in \mathbb{N} \text{ with } h(\boldsymbol{x}) = \sum_{i=1}^{\infty} \beta_i k(\boldsymbol{x}_i, \boldsymbol{x}) \text{ and } \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \beta_i k(\boldsymbol{x}_i, \boldsymbol{x}_j) \beta_j < \infty \right\},
$$

and where the inner product of two arbitrary functions $h_1, h_2 \in \mathbb{H}$ with $h_1(\boldsymbol{x}) = \sum_{i=1}^{\infty} \beta_i k(\boldsymbol{x}_i, \boldsymbol{x})$ and $h_2(\boldsymbol{x}) = \sum_{j=1}^{\infty} \beta'_j k(\boldsymbol{x}'_j, \boldsymbol{x})$ is defined as $\langle h_1, h_2 \rangle_{\mathbb{H}} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \beta_i k(\boldsymbol{x}_i, \boldsymbol{x}'_j) \beta'_j$. One may now use the kernel function to define the feature map Φ through $[\Phi(\boldsymbol{x}')](\boldsymbol{x}) = k(\boldsymbol{x}', \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}$. This choice is admissible because it respects the consistency condition $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle_{\mathbb{H}} = k(\boldsymbol{x}, \boldsymbol{x}')$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{H}$, and

because it implies the desired *reproducing property* $\langle f, \Phi(\boldsymbol{x}')\rangle_{\mathbb{H}} = \sum_{i=1}^{\infty} \beta_i k(\boldsymbol{x}_i, \boldsymbol{x}') = f(\boldsymbol{x}')$ for all $f \in \mathbb{H}$ and $\boldsymbol{x}' \in \mathbb{X}$.

In summary, given a symmetric and positive definite kernel function k, there exists an associated RKHS $\mathbb{H}$ and a feature map Φ with the reproducing property. As we will see below, however, to optimize over nonlinear hypotheses in $\mathbb{H}$, knowledge of k is sufficient, and there is no need to construct $\mathbb{H}$ and Φ explicitly.

Assume now that we are given any symmetric and positive definite kernel function k, and construct a distributionally robust learning problem over all *nonlinear* hypotheses in the corresponding RKHS $\mathbb{H}$ via

$$\widehat{J}(\rho) = \inf_{h \in \mathbb{H}} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\ell(h(\boldsymbol{x}), y)\right], \tag{29}$$

where the transportation metric is given by the Euclidean norm on $\mathbb{X} \times \mathbb{Y}$ (for regression problems) or the separable metric (16) with the Euclidean norm on $\mathbb{X}$ (for classification problems). While problem (29) is hard to solve in general due to the nonlinearity of the hypotheses $h \in \mathbb{H}$, it is easy to solve a lifted learning problem where the inputs $\boldsymbol{x} \in \mathbb{X}$ are replaced with features $\boldsymbol{x}_{\mathbb{H}} \in \mathbb{H}$, while each *nonlinear* hypothesis $h \in \mathbb{H}$ over the input space $\mathbb{X}$ is identified with a *linear* hypothesis $h_{\mathbb{H}} \in \mathbb{H}$ over the feature space $\mathbb{H}$ through the identity $h_{\mathbb{H}}(\boldsymbol{x}_{\mathbb{H}}) = \langle h, \boldsymbol{x}_{\mathbb{H}}\rangle_{\mathbb{H}}$. Thus, the lifted learning problem can be represented as

$$\widehat{J}_{\mathbb{H}}(\rho) = \inf_{h \in \mathbb{H}} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N^{\mathbb{H}})} \mathbb{E}^{\mathbb{Q}}\left[\ell(\langle h, \boldsymbol{x}_{\mathbb{H}}\rangle_{\mathbb{H}}, y)\right], \tag{30}$$

where $\widehat{\mathbb{P}}_N^{\mathbb{H}} = 1/N \sum_{i=1}^{N} \delta_{(\Phi(\widehat{\boldsymbol{x}}_i), \widehat{y}_i)}$ on $\mathbb{H} \times \mathbb{Y}$ denotes the pushforward measure of the emprical distribution $\widehat{\mathbb{P}}_N$ under the feature map Φ induced by k, while $\mathbb{B}_\rho(\widehat{\mathbb{P}}_N^{\mathbb{H}})$ constitutes the Wasserstein ball of radius ρ around $\widehat{\mathbb{P}}_N^{\mathbb{H}}$ corresponding to the transportation metric

$$d_{\mathbb{H}}\left((\boldsymbol{x}_{\mathbb{H}}, y), (\boldsymbol{x}_{\mathbb{H}}', y')\right) = \begin{cases} \sqrt{\|\boldsymbol{x}_{\mathbb{H}} - \boldsymbol{x}_{\mathbb{H}}'\|_{\mathbb{H}}^2 + (y - y')^2} & \text{for regression problems,} \\ \|\boldsymbol{x}_{\mathbb{H}} - \boldsymbol{x}_{\mathbb{H}}'\|_{\mathbb{H}} + \kappa \mathbb{1}_{\{y \neq y'\}} & \text{for classification problems.} \end{cases}$$

Even though $\widehat{\mathbb{P}}_N^{\mathbb{H}}$ constitutes the pushforward measure of $\widehat{\mathbb{P}}_N$ under Φ, not every distribution $\mathbb{Q}^{\mathbb{H}} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N^{\mathbb{H}})$ can be obtained as the pushforward measure of some $\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$. Thus, we should not expect (29) to be equivalent to (30). Instead, one can show that under a judicious transformation of the Wasserstein radius, (30) provides an upper bound on (29) whenever the kernel function satisfies a calmness condition.

Assumption 3.22 (Calmness of the kernel)**.** The kernel function k is calm from above, that is, there exist a concave smooth growth function $g : \mathbb{R}_+ \to \mathbb{R}_+$ with $g(0) = 0$ and $g'(z) \geq 1$ for all $z \in \mathbb{R}_+$ such that

$$\sqrt{k(\boldsymbol{x}_1, \boldsymbol{x}_1) - 2k(\boldsymbol{x}_1, \boldsymbol{x}_2) + k(\boldsymbol{x}_2, \boldsymbol{x}_2)} \leq g(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2) \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{X}.$$

The calmness condition is non-restrictive. In fact, it is satisfied by most commonly used kernels.

Example 3.23 (Growth Functions for Popular Kernels)**.** For most commonly used kernels k on $\mathbb{X} \subseteq \mathbb{R}^n$, we can construct an explicit growth function g that certifies the calmness of k in the sense of Assumption 3.22. This construction typically relies on elementary estimates. Derivations are omitted for brevity.

1. **Linear kernel:** For $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle$, we may set $g(z) = z$.

2. **Gaussian kernel:** For $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = e^{-\gamma \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2}$ with $\gamma > 0$, we may set $g(z) = \max\{\sqrt{2\gamma}, 1\}z$.

3. **Laplacian kernel:** For $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = e^{-\gamma \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_1}$ with $\gamma > 0$, we may set $g(z) = \sqrt{2\gamma z \sqrt{n}}$ if $0 \leq z \leq \gamma\sqrt{n}/2$ and $g(z) = z + \gamma\sqrt{n}/2$ otherwise.

4. **Polynomial kernel:** The kernel $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\gamma\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle + 1)^d$ with $\gamma > 0$ and $d \in \mathbb{N}$ fails to satisfy the calmness condition if $\mathbb{X}$ is unbounded and $d > 1$, in which case $\sqrt{k(\boldsymbol{x}_1, \boldsymbol{x}_1) - 2k(\boldsymbol{x}_1, \boldsymbol{x}_2) + k(\boldsymbol{x}_2, \boldsymbol{x}_2)}$ grows superlinearly. If $\mathbb{X} \subseteq \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \leq R\}$ for some $R > 0$, however, the polynomial kernel is calm with respect to the growth function

$$g(z) = \begin{cases} \max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d}, 1\}z & \text{if } d \text{ is even,} \\ \max\{\frac{1}{2R}\sqrt{2(\gamma R^2 + 1)^d - 2(1 - \gamma R^2)^d}, 1\}z & \text{if } d \text{ is odd.} \end{cases}$$

15

Theorem 3.24 (Lifted learning problems). If Assumption 3.22 holds for some growth function g, then the following statements hold for all Wasserstein radii $\rho \geq 0$.

(i) For regression problems we have $\widehat{J}(\rho) \leq \widehat{J}_{\mathbb{H}}(\sqrt{2}g(\rho))$.

(ii) For classification problems we have $\widehat{J}(\rho) \leq \widehat{J}_{\mathbb{H}}(g(\rho))$.

We now argue that the lifted learning problem (30) can be solved efficiently by leveraging the following representer theorem, which generalizes [59, Theorem 4.2] to non-separable loss functions.

Theorem 3.25 (Representer theorem). Assume that we are given a symmetric positive definite kernel k on $\mathbb{X}$ with corresponding RKHS $\mathbb{H}$, a set of training samples $(\widehat{\boldsymbol{x}}_i, \widehat{y}_i) \in \mathbb{X} \times \mathbb{Y}$, $i \leq N$, and an arbitrary loss function $f : (\mathbb{X} \times \mathbb{Y} \times \mathbb{R})^N \times \mathbb{R}_+ \to \mathbb{R}$ that is non-decreasing in its last argument. Then, there exist $\beta_i \in \mathbb{R}$, $i \leq N$, such that the learning problem

$$\min_{h \in \mathbb{H}} \ f((\widehat{\boldsymbol{x}}_1, \widehat{y}_1, h(\widehat{\boldsymbol{x}}_1)), \dots, (\widehat{\boldsymbol{x}}_N, \widehat{y}_N, h(\widehat{\boldsymbol{x}}_N)), \|h\|_{\mathbb{H}}) \tag{31}$$

is solved by a hypothesis $h^\star \in \mathbb{H}$ representable as $h^\star(\boldsymbol{x}) = \sum_{i=1}^N \beta_i k(\boldsymbol{x}, \widehat{\boldsymbol{x}}_i)$.

The subsequent results involve the Kernel matrix $\boldsymbol{\mathcal{K}} = [\boldsymbol{\mathcal{K}}_{ij}]$ defined through $\boldsymbol{\mathcal{K}}_{ij} = k(\widehat{\boldsymbol{x}}_i, \widehat{\boldsymbol{x}}_j)$, $i, j \leq N$. The following theorems demonstrate that the lifted learning problem (30) admits a kernel representation.

Theorem 3.26 (Kernelized distributionally robust regression). Suppose that $\mathbb{X} = \mathbb{R}^n$, $\mathbb{Y} = \mathbb{R}$ and k is a symmetric positive definite kernel on $\mathbb{X}$ with associated RKHS $\mathbb{H}$. If ℓ is generated by a convex and Lipschitz continuous loss function L, that is, $\ell(h(\boldsymbol{x}), y) = L(h(\boldsymbol{x}) - y)$, then (30) is equivalent to

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^N} \frac{1}{N} \sum_{i=1}^N L\Big(\sum_{j=1}^N \boldsymbol{\mathcal{K}}_{ij} \beta_j - \widehat{y}_i \Big) + \rho \operatorname{lip}(L) \| (\boldsymbol{\mathcal{K}}^{\frac{1}{2}} \boldsymbol{\beta}, 1) \|_2,$$

and for any of its minimizers $\boldsymbol{\beta}^\star$ the hypothesis $h^\star(\boldsymbol{x}) = \sum_{i=1}^N \beta_i^\star k(\boldsymbol{x}, \widehat{\boldsymbol{x}}_i)$ is optimal in (30).

Theorem 3.27 (Kernelized distributionally robust classification). Suppose that $\mathbb{X} = \mathbb{R}^n$, $\mathbb{Y} = \{+1, -1\}$ and k is a symmetric positive definite kernel on $\mathbb{X} = \mathbb{R}^n$ with associated RKHS $\mathbb{H}$. If ℓ is generated by a convex and Lipschitz continuous loss function L, that is, $\ell(h(\boldsymbol{x}), y) = L(yh(\boldsymbol{x}))$, then (30) is equivalent to

$$\begin{cases} \min_{\beta_i, \lambda, s_i} & \lambda \rho + \dfrac{1}{N} \sum_{i=1}^N s_i \\[2mm] \text{s.t.} & L\Big(\sum_{j=1}^N \widehat{y}_i \boldsymbol{\mathcal{K}}_{ij} \beta_j \Big) \leq s_i & i \in [N] \\[2mm] & L\Big(-\sum_{j=1}^N \widehat{y}_i \boldsymbol{\mathcal{K}}_{ij} \beta_j \Big) - \kappa \lambda \leq s_i & i \in [N] \\[2mm] & \operatorname{lip}(L) \| \boldsymbol{\mathcal{K}}^{\frac{1}{2}} \boldsymbol{\beta} \|_2 \leq \lambda, \end{cases} \tag{32}$$

and for any of its minimizers $\boldsymbol{\beta}^\star$ the hypothesis $h^\star(\boldsymbol{x}) = \sum_{i=1}^N \beta_i^\star k(\boldsymbol{x}, \widehat{\boldsymbol{x}}_i)$ is optimal in (30).

Theorems 3.26 and 3.27 show that the lifted learning problem (30) can be solved with similar computational effort as problem (4), that is, optimizing over a possibly infinite-dimensional RKHS of nonlinear hypotheses is not substantially harder than optimizing over the space of linear hypotheses.

Remark 3.28 (Kernelization in robust regression and classification). Recall from Theorem 3.9 that distributionally robust and classical robust linear regression are equivalent if $\Xi = \mathbb{R}^{n+1}$ and the training samples are sufficiently dispersed in the sense of Assumption 3.7. Similarly, Theorem 3.20 implies that distributionally robust and classical robust linear classification are equivalent if $\kappa = \infty$ and the training samples are non-separable in the sense of Assumption 3.18. One can show that Theorems 3.9 and 3.20

naturally extend to nonlinear regression and classification models over an RKHS induced by some symmetric and positive definite kernel. Specifically, one can show that some *lifted* robust learning problem is equivalent to the *lifted* distributionally robust learning problem (30) whenever the lifted training samples $(\Phi(\widehat{\boldsymbol{x}}_1), \widehat{y}_1), \cdots, (\Phi(\widehat{\boldsymbol{x}}_N), \widehat{y}_N)$ satisfy Assumption 3.7 (for regression) or 3.18 (for classification). Theorems 3.26 and 3.27 thus imply that the lifted robust regression and classification problems can be solved efficiently under mild regularity conditions whenever Assumptions 3.7 and 3.18 hold, respectively. Unfortunately, these conditions are often violated for popular kernels. For example, the lifted samples are always linearly separable under the Gaussian kernel [76, p. 1496]. In this case, the lifted robust classification problem can never be reduced to an efficiently solvable lifted distributionally robust classification problem of the form (30). In fact, no efficient method for solving the lifted robust classification problem seems to be known. In contrast, the lifted *distributionally* robust learning problems are always efficiently solvable under standard regularity conditions.

3.4 Nonlinear Hypotheses: Neural Networks[2]

Families of *neural networks* represent particularly expressive classes of nonlinear hypotheses. In the following, we characterize a family $\mathbb{H}$ of neural networks with $M \in \mathbb{N}$ layers through M continuous activation functions $\sigma_m : \mathbb{R}^{n_{m+1}} \to \mathbb{R}^{n_{m+1}}$ and M weight matrices $\boldsymbol{W}_m \in \mathbb{R}^{n_{m+1} \times n_m}$, $m \in [M]$. The weight matrices can encode *fully connected* or *convolutional* layers, for example. If $n_1 = n$ and $n_{M+1} = 1$, then we may set

$$\mathbb{H} = \left\{ h \in \mathbb{R}^{\mathbb{X}} \;:\; \exists \boldsymbol{W}_m \in \mathbb{R}^{n_{m+1} \times n_m}, \; m \in [M], \; h(\boldsymbol{x}) = \sigma_M \Big(\boldsymbol{W}_M \cdots \sigma_2 \big(\boldsymbol{W}_2 \sigma_1 (\boldsymbol{W}_1 \boldsymbol{x}) \big) \cdots \Big) \right\}.$$

Each hypothesis $h \in \mathbb{H}$ constitutes a neural network and is uniquely determined by the collection of all weight matrices $\boldsymbol{W}_{[M]} := (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M)$. In order to emphasize the dependence on $\boldsymbol{W}_{[M]}$, we will sometimes use $h(\boldsymbol{x}; \boldsymbol{W}_{[M]})$ to denote the hypotheses in $\mathbb{H}$. Setting $\boldsymbol{x}_1 = \boldsymbol{x}$, the features of the neural network are defined recursively through $\boldsymbol{x}_{m+1} = \sigma_m(\boldsymbol{z}_m)$, where $\boldsymbol{z}_m = \boldsymbol{W}_m \boldsymbol{x}_m$, $m \in [M]$. The features $\boldsymbol{x}_m$, $m = 2, \ldots, M$, correspond to the *hidden* layers of the neural network, while $\boldsymbol{x}_{M+1}$ determines its output.

Example 3.29 (Activation functions)**.** The following activation functions are most widely used.

1. **Hyperbolic tangent**: $[\sigma_m(\boldsymbol{z}_m)]_i = (\exp([\boldsymbol{z}_m]_i) - \exp(-[\boldsymbol{z}_m]_i))/(\exp([\boldsymbol{z}_m]_i) + \exp(-[\boldsymbol{z}_m]_i))$

2. **Sigmoid:** $[\sigma_m(\boldsymbol{z}_m)]_i = 1/(1 + \exp(-[\boldsymbol{z}_m]_i))$

3. **Softmax:** $[\sigma_m(\boldsymbol{z}_m)]_i = \exp([\boldsymbol{z}_m]_i)/\sum_{j=1}^{n_{m+1}} \exp([\boldsymbol{z}_m]_j)$

4. **Rectified linear unit (ReLU):** $[\sigma_m(\boldsymbol{z}_m)]_i = \max\{0, [\boldsymbol{z}_m]_i\}$

5. **Exponential linear unit (ELU):** $[\sigma_m(\boldsymbol{z}_m)]_i = \max\{0, [\boldsymbol{z}_m]_i\} + \min\{0, \alpha(\exp([\boldsymbol{z}_m]_i) - 1)\}$

The distributionally robust learning model over the hypothesis class $\mathbb{H}$ can now be represented as

$$\inf_{h \in \mathbb{H}} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}} \left[\ell\big(h(\boldsymbol{x}), y\big) \right] = \inf_{\boldsymbol{W}_{[M]}} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}} \left[\ell\big(h(\boldsymbol{x}; \boldsymbol{W}_{[M]}), y\big) \right], \tag{33}$$

where we use the transportation metrics (11) and (16) for regression and classification problems, respectively. Moreover, we adopt the standard convention that $\ell(h(\boldsymbol{x}), y) = L(h(\boldsymbol{x}) - y)$ for regression problems and $\ell(h(\boldsymbol{x}, y)) = L(yh(\boldsymbol{x}))$ for classification problems, where L is a convex and Lipschitz continuous univariate loss function. In the following we equip each feature space $\mathbb{R}^{n_m}$ with a norm $\| \cdot \|$, $m \in [M+1]$. By slight abuse of notation, we use the same symbol for all norms even though the norms on different feature spaces may differ. Using the norm on $\mathbb{R}^{n_{m+1}}$, we define the Lipschitz modulus of σ_m as

$$\text{lip}(\sigma_m) := \sup_{\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^{n_{m+1}}} \left\{ \frac{\|\sigma(\boldsymbol{z}) - \sigma(\boldsymbol{z}')\|}{\|\boldsymbol{z} - \boldsymbol{z}'\|} : \boldsymbol{z} \neq \boldsymbol{z}' \right\}.$$

We are now ready to state the main result of this section, which provides a conservative upper bound on the distributionally robust learning model (33).

---

[2]We are grateful to an anonymous referee for encouraging us to write this section.

Theorem 3.30 (Distributionally robust learning with neural networks)**.** The distributionally robust learning model (33) is bounded above by the regularized empirical loss minimization problem

$$\inf_{\boldsymbol{W}_{[M]}} \frac{1}{N} \sum_{i=1}^{N} \ell(h(\widehat{\boldsymbol{x}}_i; \boldsymbol{W}_{[M]}), \widehat{y}_i) + \rho \operatorname{lip}(L) \max \left\{ \prod_{m=1}^{M} \operatorname{lip}(\sigma_m) \|\boldsymbol{W}_m\|, \frac{c}{\kappa} \right\}, \tag{34}$$

where $c = 1$ for regression problems and $c = \max\{1, 2 \sup_{h \in \mathbb{H}, \boldsymbol{x} \in \mathbb{X}} |h(\boldsymbol{x})|\}$ for classification problems. Moreover, $\|\boldsymbol{W}_m\| = \sup_{\|\boldsymbol{x}_m\|=1} \|\boldsymbol{W}_m \boldsymbol{x}_m\|$ is the operator norm induced by the norms on $\mathbb{R}^{n_m}$ and $\mathbb{R}^{n_{m+1}}$.

Remark 3.31 (Uniform upper bound on all neural networks)**.** For classification problems the constant c in (34) represents a uniform upper bound on all neural networks and may be difficult to evaluate in general. It is easy to estimate c, however, if the last activation function is itself bounded such as the softmax function, which yields a probability distribution over the output space. In this case one may simply set $c = 2$.

The product term $\prod_{m=1}^{M} \operatorname{lip}(\sigma_m) \|\boldsymbol{W}_m\|$ in (34) represents an upper bound on the Lipschitz modulus of $h(\boldsymbol{x}; \boldsymbol{W}_{[M]})$. We emphasize that computing the exact Lipschitz modulus of a neural network is NP-hard even if there are only two layers and all activation functions are of the ReLU type [58, Theorem 2]. In contrast, the upper bound at hand is easy to compute as all activation functions listed in Example 3.29 have Lipschitz modulus 1 with respect to the Euclidean norms on the domain and range spaces [36, 74]. For more details on how to estimate the Lipschitz moduli of neural networks we refer to [36, 51, 54, 69].

Note that even though (34) represents a finite-dimensional optimization problem over the weight matrices of the neural network, both the empirical prediction loss as well as the regularization term are non-convex in $\boldsymbol{W}_{[M]}$, which complicates numerical solution. If $\kappa = \infty$, however, one can derive an alternative upper bound on the distributionally robust learning model (33) with a *convex* regularization term.

Corollary 3.32 (Convex regularization term)**.** If $\kappa = \infty$, then there is $\overline{\rho} \geq 0$ such that the distributionally robust learning model (33) is bounded above by the regularized empirical loss minimization problem

$$\inf_{\boldsymbol{W}_{[M]}} \frac{1}{N} \sum_{i=1}^{N} \ell(h(\widehat{\boldsymbol{x}}_i; \boldsymbol{W}_{[M]}), \widehat{y}_i) + \overline{\rho} \sum_{m=1}^{M} \|\boldsymbol{W}_m\|. \tag{35}$$

As the empirical prediction loss remains non-convex, it is expedient to address problem (35) with local optimization methods such as stochastic gradient descent algorithms. For a comprehensive review of first- and the second-order stochastic gradient algorithms we refer to [2] and the references therein. In the numerical experiments we will use a stochastic proximal gradient descent algorithm that exploits the convexity of the regularization term and generates iterates $\boldsymbol{W}_{[M]}^k$ for $k \in \mathbb{N}$ according to the update rule

$$\boldsymbol{W}_m^{k+1} = \operatorname{prox}_{\eta_k \overline{\rho} \|\boldsymbol{W}_m\|} \left(\boldsymbol{W}_m^k - \eta_k \nabla_{\boldsymbol{W}_m} \ell(h(\widehat{\boldsymbol{x}}_{i_k}; \boldsymbol{W}_{[M]}^k), \widehat{y}_{i_k}) \right) \quad \forall m \in [M],$$

where $\eta_k > 0$ is a given step size and i_k is drawn randomly from the index set $[N]$, see, *e.g.*, [55]. Here, the proximal operator associated with a convex function $\varphi : \mathbb{R}^{n_{m+1} \times n_m} \to \mathbb{R}$ is defined through

$$\operatorname{prox}_{\varphi}(\boldsymbol{W}_m) := \arg \min_{\boldsymbol{W}_m'} \ \varphi(\boldsymbol{W}_m') + \frac{1}{2} \|\boldsymbol{W}_m' - \boldsymbol{W}_m\|_F^2,$$

where $\| \cdot \|_F$ stands for the Frobenius norm. The algorithm is stopped as soon as the improvement of the objective value falls below a prescribed threshold. As the empirical prediction loss is non-convex and potentially non-smooth, the algorithm fails to offer any strong performance guarantees. For the scalability of the algorithm, however, it is essential that the proximal operator can be evaluated efficiently.

Example 3.33 (Proximal operator)**.** Suppose that all feature spaces $\mathbb{R}^{n_m}$ are equipped with the p-norm for some $p \in \{1, 2, \infty\}$, which implies that all parameter spaces $\mathbb{R}^{n_{m+1} \times n_m}$ are equipped with the corresponding matrix p-norm. In this case the proximal operator of $\varphi(\boldsymbol{W}_m) = \eta \|\boldsymbol{W}_m\|_p$ for some fixed $\eta > 0$ can be evaluated highly efficiently.

1. **MACS ($p = 1$):** The matrix 1-norm returns the maximum absolute column sum (MACS). Evaluating the proximal operator of $\varphi(\boldsymbol{W}_m) = \eta\|\boldsymbol{W}_m\|_1$ amounts to solving the minimization problem

$$
\operatorname{prox}_\varphi(\boldsymbol{W}_m) = \begin{cases} \min\limits_{\boldsymbol{W}'_m, u} & \eta u + \sum\limits_{i=1}^{n_m} \|[\boldsymbol{W}'_m]_{:,i} - [\boldsymbol{W}_m]_{:,i}\|_2^2 \\ \text{s.t.} & \|[\boldsymbol{W}'_m]_{:,i}\|_1 \leq u \quad i \in [n_m], \end{cases}
$$

where $[\boldsymbol{W}_m]_{:,i}$ and $[\boldsymbol{W}'_m]_{:,i}$ represent the i-th columns of $\boldsymbol{W}_m$ and $\boldsymbol{W}'_m$, respectively. For any fixed u, the above problem decomposes into n_m projections of the vectors $[\boldsymbol{W}_m]_{:,i}$, $i \in [n_m]$, to the ℓ_1-ball of radius u centered at the origin. Each of these projections can be computed via an efficient sorting algorithm proposed in [27]. Next, we can use any line search method such as the golden-section search algorithm to optimize over u, thereby solving the full proximal problem.

2. **Spectral ($p = 2$):** The matrix 2-norm coincides with the spectral norm, which returns the maximum singular value. In this case, the proximal problem for $\varphi(\boldsymbol{W}_m) = \eta\|\boldsymbol{W}_m\|_2$ can be solved analytically via singular value thresholding [17, Theorem 2.1], that is, given the singular value decomposition $\boldsymbol{W}_m = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ with $\boldsymbol{U} \in \mathbb{R}^{n_{m+1} \times n_{m+1}}$ orthogonal, $\boldsymbol{S} \in \mathbb{R}_+^{n_{m+1} \times n_m}$ diagonal and $\boldsymbol{V} \in \mathbb{R}^{n_m \times n_m}$ orthogonal, the proximal operator satisfies

$$
\operatorname{prox}_\varphi(\boldsymbol{W}_m) = \operatorname{prox}_\varphi(\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top) = \boldsymbol{U}\widetilde{\boldsymbol{S}}\boldsymbol{V}^\top, \qquad \text{where} \quad \widetilde{\boldsymbol{S}}_{ij} = \max\{\boldsymbol{S}_{ij} - \eta, 0\}.
$$

The singular value decomposition can be accelerated using a randomized algorithm proposed in [37].

3. **MARS ($p = \infty$):** The matrix ∞-norm returns the maximum absolute row sum (MARS) and thus satisfies $\|\boldsymbol{W}_m\|_\infty = \|\boldsymbol{W}_m^\top\|_1$. Therefore, one can use the iterative scheme developed for MACS to compute the proximal operator of $\varphi(\boldsymbol{W}_m) = \eta\|\boldsymbol{W}_m\|_\infty$ by simply transposing the weight matrix $\boldsymbol{W}_m$.

The convergence behavior of the stochastic proximal gradient descent algorithm can be further improved by including a momentum term inside the proximal operator, see, *e.g.*, [49].

4 Generalization Bounds

Generalization bounds constitute upper confidence bounds on the out-of-sample error. Traditionally, generalization bounds are derived by controlling the complexity of the hypothesis space, which is typically quantified in terms of its VC-dimension or via covering numbers or Rademacher averages [62]. Strengthened generalization bounds for large margin classifiers can be obtained by improving the estimates of the VC-dimension and the Rademacher average [64, 65]. We will now demonstrate that distributionally robust learning models of the type (4) or (30) enjoy simple new generalization bounds that can be obtained under minimal assumptions. In particular, they do not rely on any notions of hypothesis complexity and may therefore even extend to hypothesis spaces with infinite VC-dimensions. Our approach is reminiscent of the generalization theory for robust support vector machines portrayed in [76], which also replaces measures of hypothesis complexity with robustness properties. However, we derive explicit finite sample guarantees, while [76] establishes asymptotic consistency results. Moreover, we relax some technical conditions used in [76] such as the compactness of the input space $\mathbb{X}$.

The key enabling mechanism of our analysis is a measure concentration property of the Wasserstein metric, which holds whenever the unknown data-generating distribution has exponentially decaying tails.

Assumption 4.1 (Light-tailed distribution). There exist constants $a > 1$ and $A > 0$ and a reference point $\boldsymbol{\xi}' \in \mathbb{R}^{n+1}$ such that $\mathbb{E}^{\mathbb{P}}[\exp(d(\boldsymbol{\xi}, \boldsymbol{\xi}')^a)] \leq A$, where d denotes the usual mass transportation cost.

Theorem 4.2 (Measure concentration [32, Theorem 2]). If Assumption 4.1 holds, then we have

$$
\mathbb{P}^N\left\{W(\mathbb{P}, \widehat{\mathbb{P}}_N) \geq \rho\right\} \leq \begin{cases} c_1 \exp\left(-c_2 N \rho^{\max\{n+1,2\}}\right) & \text{if } \rho \leq 1, \\ c_1 \exp\left(-c_2 N \rho^a\right) & \text{if } \rho > 1, \end{cases} \tag{36}
$$

for all $N \geq 1$, $n \neq 1$, and $\rho > 0$, where the constants $c_1, c_2 > 0$ depend only on a, A, d and n.[3]

---

[3] A similar but slightly more complicated inequality also holds for the special case $n = 1$; see [32, Theorem 2] for details.

Theorem 36 asserts that the empirical distribution $\widehat{\mathbb{P}}_N$ converges exponentially fast to the unknown data-generating distribution $\mathbb{P}$, in probability with respect to the Wasserstein metric, as the sample size N tends to infinity. We can now derive simple generalization bounds by increasing the Wasserstein radius ρ until the violation probability on the of the right hand side of (36) drops below a prescribed significance level $\eta \in (0,1]$. Specifically, Theorem 36 implies that $\mathbb{P}^N\{\mathbb{P} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)\} \geq 1 - \eta$ for any $\rho \geq \rho_N(\eta)$, where

$$
\rho_N(\eta) = \begin{cases} \left(\dfrac{\log(c_1/\eta)}{c_2 N}\right)^{\frac{1}{\max\{n+1,2\}}} & \text{if } N \geq \frac{\log(c_1/\eta)}{c_2}, \\[3mm] \left(\dfrac{\log(c_1/\eta)}{c_2 N}\right)^{\frac{1}{a}} & \text{if } N < \frac{\log(c_1/\eta)}{c_2}. \end{cases}
\tag{37}
$$

Theorem 4.3 (Basic generalization bound [52, Theorem 3.5]). *If Assumption 4.1 holds, then*

$$
\mathbb{P}^N\left\{ \mathbb{E}^{\mathbb{P}}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x}\rangle, y)\right] \leq \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x}\rangle, y)\right] \ \forall \boldsymbol{w} \in \mathbb{R}^n \right\} \geq 1 - \eta
\tag{38}
$$

for any $N \geq 1$, $n \neq 1$, $\eta \in (0,1]$ and $\rho \geq \rho_N(\eta)$.

Remark 4.4 (Discussion of basic generalization bound). The following comments are in order.

I. **Performance guarantees for optimal hypotheses:** If $\widehat{J}(\rho)$ denotes the minimum and $\widehat{\boldsymbol{w}}$ a minimizer of the distributionally robust learning problem (4), then Theorem 4.3 implies that

$$
\mathbb{P}^N\left\{ \mathbb{E}^{\mathbb{P}}\left[\ell(\langle \widehat{\boldsymbol{w}}, \boldsymbol{x}\rangle, y)\right] \leq \widehat{J}(\rho) \right\} \geq 1 - \eta
$$

for any $N \geq 1$, $n > 1$, $\eta \in (0,1]$ and $\rho \geq \rho_N(\eta)$.

II. **Light-tail assumption:** Assumption 4.1 is restrictive but unavoidable for any measure concentration result of the type described in Theorem 4.2. It is automatically satisfied if the input-output pair has bounded support or is known to follow a Gaussian or exponential distribution, for instance.

III. **Asymptotic consistency:** It is clear from (37) that for any fixed $\eta \in (0,1]$, the radius $\rho_N(\eta)$ tends to 0 as N increases. Moreover, Theorem 3.6 in [52] implies that if η_N converges to 0 at a carefully chosen rate (*e.g.*, $\eta_N = \exp(-\sqrt{N})$), then the solution of the distributionally robust learning problem (4) with Wasserstein radius $\rho = \rho_N(\eta_N)$ converges almost surely to the solution of the ideal learning problem that minimizes the out-of-sample error under the unknown true distribution $\mathbb{P}$.

IV. **Curse of dimensionality:** The Wasserstein radius (37) has two decay regimes. For small N, $\rho_N(\eta)$ decays as $N^{-\frac{1}{a}}$, and for large N it is proportional to $N^{-\frac{1}{n+1}}$. We thus face a curse of dimensionality for large sample sizes. In order to half the Wasserstein radius, one has to increase N by a factor of 2^n. This curse of dimensionality is fundamental, *i.e.*, the dependence of the measure concentration result in Theorem 4.2 on the input dimension n cannot be improved for generic distributions $\mathbb{P}$; see [73] or [32, Section 1.3]. Improvements are only possible in special cases, *e.g.*, if $\mathbb{P}$ is finitely supported.

V. **Extension to nonlinear hypotheses:** Theorem 4.3 directly extends to any distributionally robust learning problem over an RKHS $\mathbb{H}$ induced by some symmetric and positive definite kernel function k. Specifically, if k is calm in the sense of Assumption 3.22 with growth function g, then we have

$$
\mathbb{P}^N\left\{ \mathbb{E}^{\mathbb{P}}[\ell(h(\boldsymbol{x}), y)] \leq \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N^{\mathbb{H}})} \mathbb{E}^{\mathbb{Q}}\left[\ell(\langle h, \boldsymbol{x}_{\mathbb{H}}\rangle_{\mathbb{H}}, y)\right] \ \forall h \in \mathbb{H} \right\} \geq 1 - \eta
\tag{39}
$$

for any $N \geq 1$, $n \neq 1$, $\eta \in (0,1]$ and $\rho \geq cg(\rho_N(\eta))$, where $c = \sqrt{2}$ for regression problems and $c = 1$ for classification problems. To see this, note that the inclusion $\mathbb{P} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ implies

$$
\mathbb{E}^{\mathbb{P}}[\ell(h(\boldsymbol{x}), y)] \leq \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\ell(h(\boldsymbol{x}), y)\right] \leq \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N^{\mathbb{H}})} \mathbb{E}^{\mathbb{Q}}\left[\ell(\langle h, \boldsymbol{x}_{\mathbb{H}}\rangle_{\mathbb{H}}, y)\right] \quad \forall h \in \mathbb{H},
\tag{40}
$$

where the second inequality follows from the proof of Theorem 3.24. The generalization bound (39) thus holds because $\mathbb{P}^N\{\mathbb{P} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)\} \geq 1-\eta$ for any $\rho \geq \rho_N(\eta)$. Note that the rightmost term in (40) can be computed for any finitely generated hypothesis $h \in \mathbb{H}$ representable as $h(\boldsymbol{x}) = \sum_{i=1}^N \beta_i k(\boldsymbol{x}, \widehat{\boldsymbol{x}}_i)$, which follows from Theorems 3.26 and 3.27, while the middle term is hard to compute. We emphasize that the generalization bound (39) does not rely on any notion of hypothesis complexity and remains valid even if $\mathbb{H}$ has infinite VC-dimension (*e.g.*, if $\mathbb{H}$ is generated by the Gaussian kernel).

Theorem 4.2 provides a confidence set for the unknown probability distribution $\mathbb{P}$, and Theorem 4.3 uses this confidence set to construct a uniform generalization bound on the prediction error under $\mathbb{P}$. The radius of the confidence set for $\mathbb{P}$ decreases slowly due to a curse of dimensionality, but the decay rate is essentially optimal. This does not imply that the decay rate of the generalization bound (38) is optimal, too. In fact, the worst-case expected error over a Wasserstein ball of radius ρ can be a $(1 - \eta)$-confidence bound on the expected error under $\mathbb{P}$ even if the Wasserstein ball fails to contain $\mathbb{P}$ with confidence $1 - \eta$. Thus, the measure concentration result of Theorem 4.2 is too powerful for our purposes and leads to an over-conservative generalization bound. Below we will show that the curse of dimensionality in the generalization bound (38) can be broken if we impose the following restriction on the hypothesis space.

Assumption 4.5 (Hypothesis space)**.** The space of admissible hypotheses in (4) is restricted to $\mathbb{W} \subseteq \mathbb{R}^n$. There exists $\underline{\Omega} > 0$ with $\inf_{\boldsymbol{w} \in \mathbb{W}} \|(\boldsymbol{w}, -1)\|_* \geq \underline{\Omega}$ if (4) is a regression problem and $\inf_{\boldsymbol{w} \in \mathbb{W}} \|\boldsymbol{w}\|_* \geq \underline{\Omega}$ if (4) is a classification problem. Similarly, there exists $\overline{\Omega} \geq 0$ with $\sup_{\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{W}} \|\boldsymbol{w} - \boldsymbol{w}'\|_\infty \leq \overline{\Omega}$.

Theorem 4.6 (Improved generalization bound)**.** Suppose that Assumptions 4.1 and 4.5 hold, and the function L is Lipschitz continuous. Moreover, assume that $\Xi = \mathbb{R}^{n+1}$ and $M_n = \max_{i \leq n} \|\boldsymbol{e}_i^{n+1}\|_*$ if (4) is a regression problem, while $\Xi = \mathbb{R}^n \times \{-1, 1\}$ and $M_n = \max_{i \leq n} \|\boldsymbol{e}_i^n\|_*$ if (4) is a classification problem, where $\boldsymbol{e}_i^n$ is the i-th standard basis vector in $\mathbb{R}^n$. Then, there exist constants $c_3 \geq 1$, $c_4 > 0$ depending only on the light tail constants a and A such that the generalization bound (38) holds for any $N \geq \max\left\{(16n/c_4)^2, 16\log(c_3/\eta)/c_4\right\}$, $\eta \in (0, 1]$ and $\rho \geq \rho'_N(\eta)$, where

$$\rho'_N(\eta) = \frac{2\overline{\Omega}}{\sqrt{N}\underline{\Omega}}\left[M_n n A + \sqrt{\frac{n\log(\sqrt{N}) + \log(c_3/\eta)}{c_4}}\right].$$

The improved generalization bound from Theorem 4.5 does not suffer from a curse of dimensionality. In fact, in order to half the Wasserstein radius $\rho'_N(\eta)$, it suffices to increase the sample size N by a factor of 4, irrespective of the input dimension n.

Remark 4.7 (Discussion of improved generalization bound)**.** The following comments are in order.

I. **Bounds on hypothesis space:** Assumption 4.5 imposes upper and lower bounds on $\mathbb{W}$. The upper bound enables us to control the difference between the empirical and the true expected loss uniformly across all admissible hypotheses. This bound is less restrictive than the uniform bound on the loss function used to derive Rademacher generalization bounds (see, *e.g.*, [62, Theorem 26.4]), which essentially imposes upper bounds both on the hypotheses and the input-output pairs. The lower bound in Assumption 4.5 is restrictive for classification problems but trivially holds for regression problems because $\|(\boldsymbol{w}, -1)\|_*$ is uniformly bounded away from zero for any (dual) norm on $\mathbb{R}^{n+1}$.

II. **Breaking the curse of dimensionality:** By leveraging Assumption 4.5, Theorem 4.6 reduces the critical Wasserstein radius in the generalization bound (38) from $\rho_N(\eta) \propto \mathcal{O}([\log(\eta^{-1})/N]^{1/(n+1)})$, which suffers from a curse of dimensionality, to $\rho'_N(\eta) \propto \mathcal{O}([\log(\eta^{-1}) + n\log(N))/N]^{1/2})$, which essentially follows a square root law reminiscent of the central limit theorem.

5   Error and Risk Estimation

Once a hypothesis $h(\boldsymbol{x})$ has been chosen, it is instructive to derive pessimistic *and* optimistic estimates of its out-of-sample prediction error (in the case of regression) or its out-of-sample risk (in the case of classification). We will argue below that the distributionally robust optimization techniques developed in

this paper also offer new perspectives on error and risk estimation. For ease of exposition, we ignore any support constraints, that is, we set $\mathbb{X} = \mathbb{R}^n$ and $\mathbb{Y} = \mathbb{R}$ (for regression) or $\mathbb{X} = \mathbb{R}^n$ and $\mathbb{Y} = \{+1, -1\}$ (for classification). Moreover, we focus on linear hypotheses of the form $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$. Note, however, that all results extend directly to conic representable support sets and to nonlinear hypotheses.

In the context of regression, we aim to estimate the prediction error $\mathcal{E}(\boldsymbol{w}) = \mathbb{E}^{\mathbb{P}}[|y - \langle \boldsymbol{w}, \boldsymbol{x} \rangle|]$ or, more precisely, the mean absolute prediction error under the unknown data-generating distribution $\mathbb{P}$. As usual, we assume that the transportation metric d is induced by a norm $\| \cdot \|$ on the input-output space $\mathbb{R}^{n+1}$.

Theorem 5.1 (Error bounds in linear regression). *The prediction error admits the following estimates.*

(i) *The worst-case error* $\mathcal{E}_{\max}(\boldsymbol{w}) = \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[|y - \langle \boldsymbol{w}, \boldsymbol{x} \rangle|]$ *is given by*

$$\mathcal{E}_{\max}(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} |\widehat{y}_i - \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle| + \rho \|(\boldsymbol{w}, -1)\|_*. \tag{41a}$$

(ii) *The best-case error* $\mathcal{E}_{\min}(\boldsymbol{w}) = \inf_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[|y - \langle \boldsymbol{w}, \boldsymbol{x} \rangle|]$ *is given by*

$$\mathcal{E}_{\min}(\boldsymbol{w}) = \max \left\{ \frac{1}{N} \sum_{i=1}^{N} |\widehat{y}_i - \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle| - \rho \|(\boldsymbol{w}, -1)\|_*, 0 \right\} \tag{41b}$$

In the context of classification, we aim to quantify the risk $\mathcal{R}(\boldsymbol{w}) = \mathbb{P}[y \neq \text{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle)]$, that is, the misclassification probability under the unknown true distribution $\mathbb{P}$. Note that the risk can equivalently be defined as the expectation of a characteristic function, that is, $\mathcal{R}(\boldsymbol{w}) = \mathbb{E}^{\mathbb{P}}[\mathbb{1}_{\{y \neq \langle \boldsymbol{w}, \boldsymbol{x} \rangle\}}]$. As usual, we assume that the transportation metric d is of the form (16), where $\kappa \geq 0$ is the cost of flipping a label.

Theorem 5.2 (Risk bounds in linear classification). *The risk admits the following estimates.*

(i) *The worst-case risk* $\mathcal{R}_{\max}(\boldsymbol{w}) = \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\mathbb{1}_{\{y \neq \langle \boldsymbol{w}, \boldsymbol{x} \rangle\}}]$ *is given by*

$$\mathcal{R}_{\max}(\boldsymbol{w}) = \begin{cases} \min_{\substack{\lambda, s_i \\ r_i, t_i}} & \lambda \rho + \frac{1}{N} \sum_{i=1}^{N} s_i \\ \text{s.t.} & 1 - r_i \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle \leq s_i & i \in [N] \\ & 1 + t_i \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle - \lambda \kappa \leq s_i & i \in [N] \\ & r_i \|\boldsymbol{w}\|_* \leq \lambda, \ t_i \|\boldsymbol{w}\|_* \leq \lambda & i \in [N] \\ & r_i, t_i, s_i \geq 0 & i \in [N]. \end{cases} \tag{42a}$$

(ii) *The best-case risk* $\mathcal{R}_{\min}(\boldsymbol{w}) = \inf_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\mathbb{1}_{\{y \neq \langle \boldsymbol{w}, \boldsymbol{x} \rangle\}}]$ *is given by*

$$\mathcal{R}_{\min}(\boldsymbol{w}) = 1 - \begin{cases} \min_{\substack{\lambda, s_i \\ r_i, t_i}} & \lambda \rho + \frac{1}{N} \sum_{i=1}^{N} s_i \\ \text{s.t.} & 1 + r_i \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle \leq s_i & i \in [N] \\ & 1 - t_i \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle - \lambda \kappa \leq s_i & i \in [N] \\ & r_i \|\boldsymbol{w}\|_* \leq \lambda, \ t_i \|\boldsymbol{w}\|_* \leq \lambda & i \in [N] \\ & r_i, t_i, s_i \geq 0 & i \in [N]. \end{cases} \tag{42b}$$

We emphasize that, as the hypothesis $\boldsymbol{w}$ is fixed, the error and risk estimation problems (41) and (42) constitute tractable linear programs that can be solved highly efficiently.

Remark 5.3 (Confidence intervals for error and risk). If the Wasserstein radius is set to $\rho_N(\eta/2)$ defined in (37), where $\eta \in (0, 1]$ is a prescribed significance level, then Theorem 4.3 implies that $\mathcal{E}(\widehat{\boldsymbol{w}}) \in [\mathcal{E}_{\min}(\widehat{\boldsymbol{w}}), \mathcal{E}_{\max}(\widehat{\boldsymbol{w}})]$ and $\mathcal{R}(\widehat{\boldsymbol{w}}) \in [\mathcal{R}_{\min}(\widehat{\boldsymbol{w}}), \mathcal{R}_{\max}(\widehat{\boldsymbol{w}})]$ with confidence $1 - \eta$ for any $\widehat{\boldsymbol{w}} \in \mathbb{R}^n$ that may even depend on the training data. Theorem 5.1 implies that the confidence interval for the true error $\mathcal{E}(\widehat{\boldsymbol{w}})$ can be calculated analytically from (41), while Theorem 5.2 implies that the confidence interval for the true risk $\mathcal{R}(\widehat{\boldsymbol{w}})$ can be computed efficiently by solving the tractable linear programs (42).

Remark 5.4 (Extension to nonlinear hypotheses). By using the tools of Section 3.3, Theorems 5.1 and 5.2 generalize immediately to nonlinear hypotheses that range over a RKHS. Specifically, we can formulate lifted error and risk estimation problems where the inputs $\boldsymbol{x} \in \mathbb{X}$ are replaced with features $\boldsymbol{x}_\mathbb{H} \in \mathbb{H}$, while each *nonlinear* hypothesis $h \in \mathbb{H}$ over the input space $\mathbb{X}$ is identified with a *linear* hypothesis $h_\mathbb{H} \in \mathbb{H}$ over the feature space $\mathbb{H}$ through the identity $h_\mathbb{H}(\boldsymbol{x}_\mathbb{H}) = \langle h, \boldsymbol{x}_\mathbb{H} \rangle_\mathbb{H}$. Tractability is again facilitated by Theorem 3.25, which allows us to focus on finitely parameterized hypotheses of the form $h(\boldsymbol{x}) = \sum_{i=1}^N \beta_i k(\boldsymbol{x}, \widehat{\boldsymbol{x}}_i)$.

6 Numerical Results

We showcase the power of regularization via mass transportation in various applications based on standard datasets from the literature. All optimization problems are implemented in Python and solved with Gurobi 7.5.1 All experiments are run on an Intel XEON CPU (3.40GHz), and the corresponding codes are made publicly available at https://github.com/sorooshafiee/Regularization-via-Transportation.

6.1 Regularization with Pre-selected Parameters

We first assess how the out-of-sample performance of a distributionally robust support vector machine (DRSVM) is impacted by the choice of the Wasserstein radius ρ, the cost κ of flipping a label, and the kernel function k. To this end, we solve three binary classification problems from the MNIST database [46] targeted at distinguishing pairs of similar handwritten digits (1-vs-7, 3-vs-8, 4-vs-9). In the first experiment we optimize over linear hypotheses and use the separable transporation metric (16) involving the ∞-norm on the input space. All results are averaged over 100 independent trials. In each trial, we randomly select 500 images to train the DRSVM model (20) and use the remaining 12,000 images for testing. The correct classification rate (CCR) on the test data, averaged across all 100 trials, is visualized in Figure 1 as a function of the Wasserstein radius ρ for each $\kappa \in \{0.1, 0.25, 0.5, 0.75, \infty\}$. The best out-of-sample CCR is obtained for $\kappa = 0.25$ uniformly across all Wasserstein radii, and performance deteriorates significantly when κ is reduced or increased. Recall from Remark 3.15 that, as κ tends to infinity, the DRSVM reduces to the classical regularized support vector machine (RSVM) with 1-norm regularizer. Thus, the results of Figure 1 indicate that regularization via mass transportation may be preferable to classical regularization in terms of the maximum achievable out-of-sample CCR. More specifically, we observe that the out-of-sample CCR of the best DRSVM ($\kappa = 0.25$) displays a slightly higher and significantly wider plateau around the optimal regularization parameter ρ than the classical RSVM ($\kappa = \infty$). This suggests that the regularization parameter in DRSVMs may be easier to calibrate from data than in RSVMs, a conjecture that will be put to scrutiny in Section 6.2. Finally, Figure 1 reveals that the standard (unregularized) support vector machine (SVM), which can be viewed as a special case of the DRSVM with $\rho = 0$, is dominated by the RSVMs and DRSVMs across a wide range of regularization parameters.[4] Note that the SVM problem (20) with $\rho = 0$ reduces to a linear program and may thus suffer from multiple optimal solutions. This explains why the limiting out-of-sample CCR for $\rho \downarrow 0$ changes with κ.

6.2 Regularization with Learned Parameters

It is easy to read off the best regularization parameters ρ and κ from the charts in Figure 1. As these charts are constructed from more than 12,000 test samples, however, they are not accessible in the training phase. In practice, ρ and κ must be calibrated from the training data alone. This motivates us to revisit the three classification problems from Section 6.1 using a fully data-driven procedure, where all free model parameters are calibrated via 5-fold cross validation; see, *e.g.*, [1, § 4.3.3]. Moreover, to evaluate the benefits of kernelization, we now solve a generalized DRSVM model of the form (32), which implicitly optimizes over all nonlinear hypotheses in some RKHS. As explained in Section 3.3, kernelization necessitates the use of the separable transportation metric (16) with the Euclidean norm on the input space.

All free parameters of the resulting DRSVM model are restricted to finite search grids in order to ease the computational burden of cross validation. Specifically, we select the Wasserstein radius ρ from

---

[4]By slight abuse of notation, we use the acronym 'SVM' to refer to the unregularized empirical hinge loss minimization problem even though the traditional formulations of the support vector machine involve a Tikhonov regularization term.
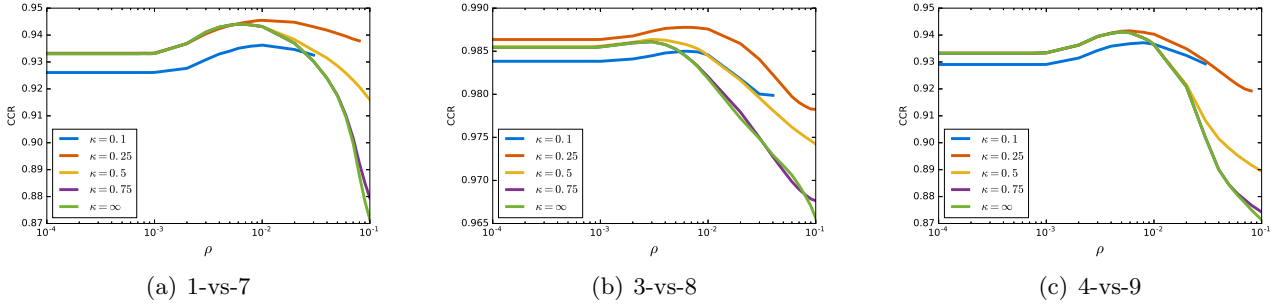
| | (a) 1-vs-7 | (b) 3-vs-8 | (c) 4-vs-9 |

Figure 1: Average out-of-sample CCR scores of the DRSVM with pre-selected parameters.

Table 1: Average out-of-sample CCR scores of the DRSVM with learned parameters.

	Polynomial		Laplacian		Gaussian	
	RSVM	DRSVM	RSVM	DRSVM	RSVM	DRSVM
1-vs-7	98.9 ± 0.2	99.1 ± 0.2	98.3 ± 0.5	98.5 ± 0.4	99.1 ± 0.2	99.2 ± 0.2
3-vs-8	95.2 ± 0.4	97.0 ± 0.4	96.5 ± 0.4	96.8 ± 0.4	97.0 ± 0.3	97.2 ± 0.3
4-vs-9	95.0 ± 0.4	96.5 ± 0.4	95.8 ± 0.6	96.0 ± 0.6	96.8 ± 0.4	96.9 ± 0.4

within $\{b \cdot 10^e : b \in \{1, 5\}, e \in \{1, 2, 3, 4\}\}$ and the label flipping cost κ from within $\{0.1, 0.25, 0.5, 0.75, \infty\}$. Moreover, we select the degree d of the polynomial kernel from within $\{1, 2, 3, 4, 5\}$ and the peakedness parameter γ of the Laplacian and Gaussian kernels from within $\{\frac{1}{100}, \frac{1}{81}, \frac{1}{64}, \frac{1}{49}, \frac{1}{36}, \frac{1}{25}\}$. Otherwise, we use the same experimental setup as in Section 6.1. Table 1 reports the averages and standard deviations of the CCR scores on the test data based on 100 independent trials. We observe that the DRSVM (ρ, κ, d, and γ learned by cross validation) outperforms the RSVM (ρ, d and γ learned by cross validation, $\kappa = \infty$) consistently across all tested kernel functions (Polynomial, Laplacian, Gaussian). Note that the DRSVM with polynomial kernel subsumes the non-kernelized DRSVM (20) as a special case because the polynomial kernel with $d = 1$ coincides with the linear kernel.

In the third experiment, we assess the out-of-sample performance of the DRSVM (20) for different transportation metrics on 10 standard datasets from the UCI repository [4], each containing up to 1,000 samples. Specifically, we use different variants of the separable transportation metric (16), where distances in the input space are measured via a p-norm with $p \in \{1, 2, \infty\}$. We focus exclusively on linear hypotheses because the kernelization techniques described in Section 3.3 are only available for $p = 2$. The DRSVM is compared against the standard (unregularized) SVM and the RSVM with q-norm regularizer ($\frac{1}{p} + \frac{1}{q} = 1$). All results are averaged across 100 independent trials. In each trial, we randomly select 75% of the data for training and the remaining 25% for testing. The inputs are first standardized to zero mean and unit variance along each coordinate axis. The Wasserstein radius ρ and the label flipping cost κ in the DRSVM as well as the regularization weight ρ in the RSVM are estimated via stratified 5-fold cross validation.

Classifier performance is now quantified in terms of the *receiver operating characteristic* (ROC) curve, which plots the true positive rate (percentage of correctly classified test samples with true label $y = 1$) against the false positive rate (percentage of *in*correctly classified test samples with true label $y = -1$) by sweeping the discrimination threshold. Specifically, we use the *area under the ROC curve* (AUC) as a measure of classifier performance. AUC does not bias on the size of the test data and is a more appropriate performance measure than CCR in the presence of an unbalanced label distribution in the training data. We emphasize that most of the considered datasets are indeed imbalanced, and thus a high CCR score would not necessarily provide evidence of superior classifier performance. The averages and standard deviations of the AUC scores based on 100 trials are reported in Table 2. The results suggest that the DRSVM outperforms the RSVM in terms of AUC for all norms by about the same amount by which the RSVM outperforms the classical hinge loss minimization, consistently across all datasets.

Table 2: Average out-of-sample AUC scores of the SVM, RSVM and DRSVM.

		$p = \infty \ / \ q = 1$		$p = 2 \ / \ q = 2$		$p = 1 \ / \ q = \infty$	
	SVM	RSVM	DRSVM	RSVM	DRSVM	RSVM	DRSVM
Australian	91.6 ± 3.0	91.5 ± 3.2	92.0 ± 2.5	92.0 ± 2.2	92.3 ± 2.0	91.9 ± 2.8	92.2 ± 2.4
Blood transfusion	73.7 ± 3.8	73.8 ± 3.8	75.5 ± 3.8	74.9 ± 3.5	75.5 ± 3.7	75.4 ± 3.4	75.4 ± 3.7
Climate model	93.8 ± 3.9	94.4 ± 4.0	94.3 ± 4.0	94.3 ± 3.8	94.0 ± 4.0	93.6 ± 3.9	93.9 ± 4.0
Cylinder	72.0 ± 3.7	71.2 ± 4.0	72.1 ± 4.1	71.3 ± 4.0	71.8 ± 3.8	71.5 ± 3.8	72.2 ± 3.7
Heart	90.4 ± 2.7	90.1 ± 2.7	90.3 ± 2.7	90.6 ± 2.6	90.9 ± 2.5	90.5 ± 2.6	90.7 ± 2.6
Ionosphere	85.0 ± 4.9	89.7 ± 4.5	89.2 ± 4.3	90.4 ± 3.7	89.9 ± 3.9	86.0 ± 4.9	87.2 ± 4.8
Liver disorders	60.5 ± 0.0	61.1 ± 0.7	61.7 ± 0.7	61.2 ± 0.3	61.7 ± 0.5	61.1 ± 0.4	61.8 ± 0.5
QSAR	90.5 ± 1.5	90.5 ± 1.6	91.0 ± 1.6	90.5 ± 1.5	91.2 ± 1.5	90.6 ± 1.5	91.1 ± 1.6
Splice	92.1 ± 0.0	93.0 ± 0.4	93.1 ± 0.1	92.5 ± 0.1	92.6 ± 0.1	92.0 ± 0.1	92.5 ± 0.1
Thoracic surgery	61.7 ± 7.1	61.5 ± 6.5	64.6 ± 6.6	64.4 ± 6.4	64.3 ± 7.0	64.0 ± 6.3	64.6 ± 6.3

6.3 Multi-Label Classification

The aim of object recognition is to discover instances of particular object classes in digital images. We now describe an object recognition experiment based on the PASCAL VOC 2007 dataset [30] consisting of 9,963 images, which are pre-partitioned into 25% for training, 25% for validation and 50% for testing. Each image is annotated with 20 binary labels corresponding to 20 given object categories (the n-th label is set to $+1$ if the image contains the n-th object and to -1 otherwise). A multi-label classifier is a function that predicts all labels of an unlabelled input image. The ability of a classifier to detect objects belonging to any fixed category is measured by the *average precision* (AP), which is defined in [30] as (a proxy for) the area under the classifier's precision-recall curve. The overall performance of a classifier is quantified by the *mean average precision* (mAP), that is, the arithmetic mean of the AP scores across all object categories.

In the first scenario, we train a separate binary RSVM and DRSVM classifier for each of the 20 object categories. This classifier predicts whether an object of the respective category appears in the input image. At the beginning we preprocess the entire dataset by resizing each image to 256×256 pixels and extracting the central patch of 244×244 pixels. As shown in [19, 26, 79], the features generated by the penultimate layer of a deep convolutional neural network trained on a large image dataset provide a powerful image descriptor. Using the ALEXNET neural network trained on the ImageNet dataset [42], we can thus compress each (preprocessed) image of the PASCAL VOC 2007 dataset into 1,000 meaningful features. We normalize these feature vectors to lie on the unit sphere. When training the RSVM and DRSVM classifiers, we can thus work with these feature vectors instead of the corresponding images. Moreover, we restrict attention to linear hypotheses and assume that transportation distances in the input-output space are measured by the separable metric (16) with the Euclidean norm on the input space. We tune the Wasserstein radius $\rho \in \{b \cdot 10^e : b \in \{1, \ldots, 9\}, e \in \{-2, -3, -4\}\}$ and the label flipping cost $\kappa \in \{0.1, 0.2, \ldots, 1, \infty\}$ via the holdout method using the validation data. As usual, we fix $\kappa = \infty$ for RSVM. Table 3 reports the AP scores of the RSVM and DRSVM models for each object category. The ensemble of all 20 binary RSVM or DRSVM classifiers, respectively, can be viewed as a naïve multi-label classifier that predicts all labels of an image. As DRSVM outperforms RSVM on an object-by-object basis, it also wins in terms of mAP.

In the second scenario, we construct a proper multi-label classifier by fine-tuning the last layer of the pre-trained ALEXNET network. To this end, we replace the original M-th layer of the network with a new fully connected layer characterized by a parameter matrix $\boldsymbol{W}_M \in \mathbb{R}^{20 \times 1000}$, and we set σ_M to the Sigmoid activation function. The resulting classifer outputs for each of the 20 object categories a probability that an object from the respective category appears in the input image. The quality of a classifier (which is encoded by $\boldsymbol{W}_M$) is measured by the cross-entropy loss function, which naturally generalizes the logloss to multiple labels. The resulting empirical loss minimization problem is enhanced with a regularization term proportional to $\|\boldsymbol{W}_M\|_{1,1}$ (Lasso), $\|\boldsymbol{W}_M\|_F^2$ (Tikhonov), $\|\boldsymbol{W}_M\|_1$ (MACS), $\|\boldsymbol{W}_M\|_2$ (Spectral) or $\|\boldsymbol{W}_M\|_\infty$ (MARS). By using similar arguments as in Section 3.4, one can show that the empirical cross-entropy with

Table 3: AP scores of different multi-label classifiers.

	Scenario 1		Scenario 2				
	RSVM	DRSVM	Lasso	Tikhonov	MACS	Spectral	MARS
	84.80	84.85	85.50	84.26	84.35	83.89	85.53
	78.54	78.49	76.18	76.55	76.37	75.67	76.22
	82.19	82.19	83.37	83.11	83.52	84.19	83.08
	79.55	79.56	77.65	78.04	77.70	78.92	77.53
	37.52	37.98	38.10	39.73	39.53	38.75	38.10
	72.05	72.03	70.05	69.43	69.17	70.28	69.77
	83.08	83.12	83.10	83.67	83.50	82.92	83.17
	80.60	80.56	79.87	79.86	80.07	79.85	79.82
	54.64	54.64	54.40	54.76	53.94	54.72	54.55
	47.82	53.13	52.06	52.10	51.62	55.54	51.19
	54.26	58.88	63.41	65.23	65.15	66.79	62.95
	75.81	75.81	76.92	77.39	77.26	76.54	76.97
	82.74	82.72	82.17	81.89	81.6	80.81	81.9
	72.69	72.88	74.70	75.41	74.76	76.48	74.31
	90.36	90.36	90.07	90.30	90.22	90.35	90.09
	50.22	51.90	50.20	50.18	50.27	51.39	50.20
	60.75	63.57	71.78	71.3	70.39	71.64	71.40
	56.85	56.98	52.15	54.36	54.65	55.12	51.94
	85.09	85.03	84.89	84.55	84.41	85.43	84.96
	69.02	69.08	64.73	65.85	65.63	64.26	64.48
mAP	69.92	70.69	70.56	70.90	70.71	71.20	70.40

MACS, Spectral or MARS regularization term overestimates the worst-case expected cross-entropy over all distributions of $(\boldsymbol{x}_M, \boldsymbol{x}_{M+1})$ in a Wasserstein ball provided that the transportation cost is given by

$$d((\boldsymbol{x}_M, \boldsymbol{x}_{M+1}), (\boldsymbol{x}'_M, \boldsymbol{x}'_{M+1})) = \|\boldsymbol{x}_M - \boldsymbol{x}'_M\|_p + \kappa \mathbb{1}_{\{\boldsymbol{x}_{M+1} \neq \boldsymbol{x}'_{M+1}\}}$$

for $\kappa = \infty$, whenever $p = 1$, $p = 2$ or $p = \infty$, respectively. Thus, the MACS, Spectral and MARS regularization terms admit a distributionally robust interpretation.

We use the stochastic proximal gradient descent algorithm of Section 3.4 to tune $\boldsymbol{W}_M$, including an additional momentum term with weight 0.9. As in [42], we split the training phase into 100 epochs, each corresponding to a complete pass through the training dataset in a random order. As the ALEXNET requires input images of size 244×244, in each iteration we extract a random patch of 244×244 pixels from the current image and flip it horizontally at random. This procedure effectively augments the training dataset. The initial step size is set to 10^{-3} and then reduced by a factor of 10 after every 7 epochs. The algorithm terminates after 100 epochs. We preprocess the images in the validation and test datasets as in Scenario 1 and tune the regularization weights via the holdout method using the validation data. Table 3 reports the AP and mAP scores of the different classifiers that were tested. These results suggest that fine-tuning the last layer of a pre-trained neural network may improve classifier performance. We observe that the spectral norm regularizer, which has a distributionally robust interpretation, consistently outperforms almost all other methods. For further details on the experimental setup (such as the exact search grids for all hyperparameters) we refer to the code publicized on Github.

6.4 Generalization Bounds

The next experiment estimates the scaling behavior of the smallest Wasserstein radius that verifies the generalization bound (38) for the synthetic `threenorm` classification problem [16]. The experiment in-
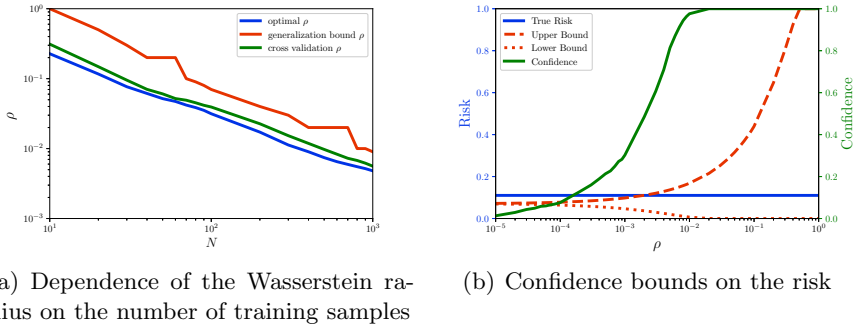
(a) Dependence of the Wasserstein radius on the number of training samples

(b) Confidence bounds on the risk

Figure 2: Results of the `threenorm` classification problem.

volves 1,000 simulation trials. In each trial we generate N training samples for some $N \in \{10, \ldots, 90\} \cup \{100, \ldots, 1,000\}$ as well as 10^5 test samples. Each sample $(\boldsymbol{x}, y) \in \mathbb{R}^{20} \times \{-1, 1\}$ is constructed as follows. The label y is drawn uniformly from $\{-1, 1\}$. If $y = -1$, then $\boldsymbol{x}$ is drawn from a standard multivariate normal distribution shifted by $(c, \ldots, c)$ or $(-c, \ldots, -c)$ with equal probabilities, where $c = 2/\sqrt{20}$. If $\widehat{y} = 1$, on the other hand, then $\boldsymbol{x}$ is drawn from a standard multivariate normal distribution shifted by $(c, -c, +c, \ldots, -c)$.

We now describe three different approaches to choose the Wasserstein radius ρ in the DRSVM (20) with transportation cost (16), where $\kappa = \infty$ and $\|\cdot\|$ represents the ∞-norm on the input space. Throughout the experiment we use $P = \{b \cdot 10^{-e} : b \in \{1, \ldots, 10\}, e \in \{1, \ldots, 5\}\}$ as the search space for ρ. Approach 1 ('cross validation') calibrates the Wasserstein radius as before via 5-fold cross validation based solely on the N training samples. This approach reflects what would typically be done in practice. Approaches 2 and 3 both solve (20) based on the empirical distribution induced by the N training samples and select the Wasserstein radius using the 10^5 test samples. Specifically, approach 2 ('optimal') chooses the Wasserstain radius that leads to the lowest test error, while approach 3 ('generalization bound') selects the smallest Wasserstein radius for which the optimal value of (20) exceeds the expected loss on the test samples in at least 95% of all trials, that is, it approximates the smallest Wasserstein radius that verifies the generalization bound (38) for $\eta = 5\%$. As the test samples are not available in the training phase, the last two approaches are not implementable in practice, and we merely study them to gain insights. Figure 2(a) visualizes all resulting Wasserstein radii as a function of N. Note that the radii obtained with the first two approaches are uncertain as they depend on a particular choice of the training samples. Figure 2(a) thus only shows their averages across all simulation trials. In contrast, the radii obtained with the third approach depend on the training sample sets of all 1,000 trials and are thus essentially deterministic.

We observe that the Wasserstein radii of all three approaches decay approximately as $1/\sqrt{N}$, which is in line with the theoretical generalization bound of Theorem 4.6. We expect this decay rate to be optimal because any faster decay would be in conflict with the central limit theorem. Note also that our results empirically confirm Theorem 4.6 even though we did not impose any restrictions on $\mathbb{W}$ as dictated by Assumption 4.5. This suggests that Theorem 4.6 might remain valid under weaker conditions.

In the experiment underlying Figure 2(b), we first fix $\widehat{\boldsymbol{w}}$ to an optimal solution of (20) for $\rho = 0.1$ and $N = 100$. Figure 2(b) shows the true risk $\mathcal{R}(\widehat{\boldsymbol{w}})$ and its confidence bounds given by Theorem 5.2. As expected, for $\rho = 0$ the upper and lower bounds coincide with the empirical risk on the training data, which is a lower bound for the true risk on the test data due to over-fitting effects. As ρ increases, the confidence interval between the bounds widens and eventually covers the true risk. For instance, at $\rho \approx 0.009$ the confidence interval is given by $[0.008, 0.162]$ and contains the true risk with probability $1 - \eta = 95\%$.

6.5 Worst-Case Distributions

Consider again the 3-vs-8 classification problems from the MNIST database [46] and fix $\boldsymbol{w}^\star$ to an optimal solution of the empirical hinge loss minimization problem. The goal of the last experiment is to evaluate the *worst-case* hinge loss of $\boldsymbol{w}^\star$ for different Wasserstein radii $\rho \in \{0, 0.01, 0.05, 0.1, 0.5, 1\}$ and label flipping costs $\kappa \in \{0, \infty\}$ and to investigate the corresponding worst-case distributions, which are efficiently
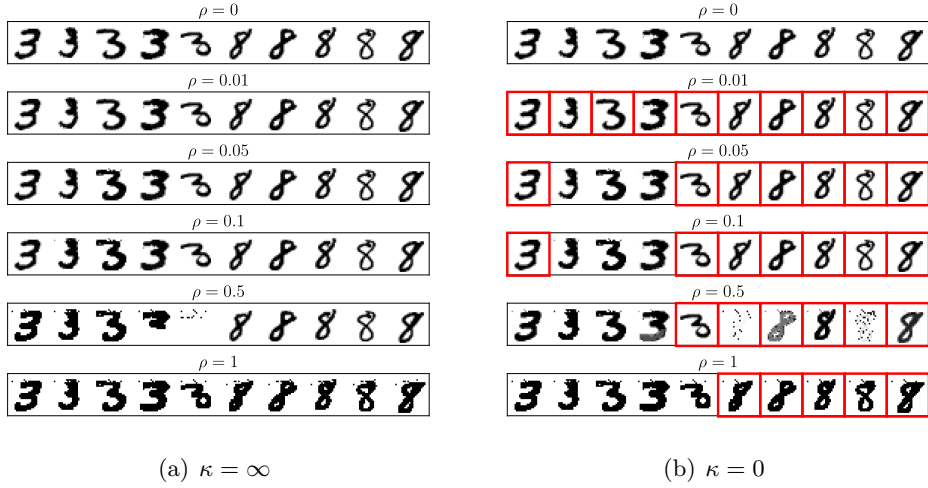
(a) $\kappa = \infty$ (b) $\kappa = 0$

Figure 3: Discretization points (input images) of the worst-case distribution for different ρ and κ. Red frames indicate that the corresponding labels are flipped under the worst-case distribution.

computable by virtue of Theorem 3.17(i). As each input constitutes a vector of pixels intensities between zero and one, we impose support constraints of the form $\boldsymbol{C}\boldsymbol{x} \leq \boldsymbol{d}$ with $\boldsymbol{C} = [\boldsymbol{I}, -\boldsymbol{I}]^{\top}$ and $\boldsymbol{d} = [\boldsymbol{1}^{\top}, \boldsymbol{0}^{\top}]^{\top}$.

For illustrative purposes we only use the $N = 10$ first datapoints in the MNIST dataset as training samples. Each training sample $\widehat{\boldsymbol{x}}_i$ corresponds to four discretization points ($\widehat{\boldsymbol{x}}_i + \boldsymbol{q}^{+\star}_{ij}/\alpha^{+\star}_{ij}$ and $\widehat{\boldsymbol{x}}_i + \boldsymbol{q}^{-\star}_{ij}/\alpha^{-\star}_{ij}$ for $j = 1, 2$) in the worst-case distribution obtained from (26). We observe that for every i exactly one out of these four points has probability $\frac{1}{N}$, while all others have probability 0. Figure 3 depicts only those 10 discretization points that have nonzero probability for a fixed ρ and κ. As expected, the perturbations of the training samples are more severe for larger Wasserstein radii. For $\kappa = \infty$ these scenarios must have the same labels as the corresponding training samples. For $\kappa = 0$, on the other hand, the labels can be flipped at no cost (flipped labels are indicated by red frames). Each scenario group shown in Figure 3 can thus be viewed as a worst-case training dataset for the corresponding Wasserstein radius and label flipping cost.

A   Appendix: Proofs

A.1   Proofs of Section 3

The proof of Theorem 3.1 requires three preparatory Lemmas. The first lemma is adapted from [52] and asserts that the worst-case expectation over a Wasserstein ball can be re-expressed as a classical robust optimization problem.

Lemma A.1 (Robust reformulation). Set $\widehat{\boldsymbol{\xi}}_i = (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)$ for all $i \leq N$. For any measurable integrand $I(\boldsymbol{\xi})$ that is bounded above by a Lipschitz continuous function we have

$$\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[I(\boldsymbol{\xi})] = \inf_{\lambda \geq 0} \; \lambda\rho + \frac{1}{N} \sum_{i=1}^{N} \sup_{\boldsymbol{\xi} \in \Xi} \; I(\boldsymbol{\xi}) - \lambda d(\boldsymbol{\xi}, \widehat{\boldsymbol{\xi}}_i). \tag{43}$$

Proof. By the definition of the Wasserstein ball we have

$$\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[I(\boldsymbol{\xi})] = \begin{cases} \sup_{\Pi} & \int_{\Xi^2} I(\boldsymbol{\xi}) \, \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \\ \text{s.t.} & \Pi \text{ is a joint distribution of } \boldsymbol{\xi} \\ & \text{and } \boldsymbol{\xi}' \text{ with marginals } \mathbb{Q} \text{ and } \widehat{\mathbb{P}}_N \\ & \int_{\Xi^2} d\left(\boldsymbol{\xi}, \boldsymbol{\xi}'\right) \Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') \leq \rho \end{cases}$$

28

$$
= \begin{cases} \sup_{\mathbb{Q}^i} & \dfrac{1}{N} \sum_{i=1}^{N} \int_{\Xi} I(\boldsymbol{\xi})\, \mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) \\[2mm] \text{s.t.} & \displaystyle\int_{\Xi} \mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) = 1 \quad i \in [N] \\[2mm] & \dfrac{1}{N} \sum_{i=1}^{N} \int_{\Xi} d(\boldsymbol{\xi}, \widehat{\boldsymbol{\xi}}_i)\, \mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi}) \le \rho. \end{cases}
$$

Note that the integral of $I(\boldsymbol{\xi})$ exists under every $\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ because $I(\boldsymbol{\xi})$ admits a Lipschitz continuous majorant. The last equality in the above expression holds because the marginal distribution of $\boldsymbol{\xi}'$ is the uniform distribution on the training samples, which implies that Π is completely determined by the conditional distributions $\mathbb{Q}^i$ of $\boldsymbol{\xi}$ given $\boldsymbol{\xi}' = \widehat{\boldsymbol{\xi}}_i$, that is, $\Pi(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\xi}') = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\boldsymbol{\xi}}_i}(\mathrm{d}\boldsymbol{\xi}')\mathbb{Q}^i(\mathrm{d}\boldsymbol{\xi})$. The resulting generalized moment problem over the normalized measures $\mathbb{Q}^i$ admits the semi-infinite dual

$$
\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[I(\boldsymbol{\xi})\right] = \begin{cases} \inf_{\lambda, s_i} & \lambda\rho + \dfrac{1}{N}\sum_{i=1}^{N} s_i \\[2mm] \text{s.t.} & \sup_{\boldsymbol{\xi}\in\Xi} I(\boldsymbol{\xi}) - \lambda d(\boldsymbol{\xi}, \widehat{\boldsymbol{\xi}}_i) \le s_i \qquad i \in [N] \\[2mm] & \lambda \ge 0, \end{cases}
$$

Strong duality holds for any $\rho > 0$ due to [63, Proposition 3.4]. The claim then follows by eliminating s_i. $\qquad\square$

Lemma A.2. For any $a \in \mathbb{R}$, $\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}} \in \mathbb{R}^d$, $\gamma \in \mathbb{R}_+$ and $\boldsymbol{\zeta} \in \mathcal{Z}$, where $\mathcal{Z} \subseteq \mathbb{R}^d$ is a closed convex set, we have

$$
\sup_{\boldsymbol{\zeta}\in\mathcal{Z}} a\langle\boldsymbol{\beta}, \boldsymbol{\zeta}\rangle - \gamma\|\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\| = \begin{cases} \inf_{\boldsymbol{p}} & S_{\mathcal{Z}}(a\boldsymbol{\beta} - \boldsymbol{p}) + \langle\boldsymbol{p}, \widehat{\boldsymbol{\zeta}}\rangle \\[2mm] \text{s.t.} & \|\boldsymbol{p}\|_* \le \gamma. \end{cases}
$$

Proof. We have

$$
\begin{aligned}
\sup_{\boldsymbol{\zeta}\in\mathcal{Z}} a\langle\boldsymbol{\beta}, \boldsymbol{\zeta}\rangle - \gamma\|\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\| &= \sup_{\boldsymbol{\zeta}\in\mathcal{Z}} \inf_{\|\boldsymbol{p}\|_*\le\gamma} a\langle\boldsymbol{\beta}, \boldsymbol{\zeta}\rangle - \langle\boldsymbol{p}, \boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\rangle \\
&= \inf_{\|\boldsymbol{p}\|_*\le\gamma} \sup_{\boldsymbol{\zeta}\in\mathcal{Z}} a\langle\boldsymbol{\beta}, \boldsymbol{\zeta}\rangle - \langle\boldsymbol{p}, \boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\rangle \\
&= \inf_{\|\boldsymbol{p}\|_*\le\gamma} \sup_{\boldsymbol{\zeta}\in\mathbb{R}^d} \langle a\boldsymbol{\beta} - \boldsymbol{p}, \boldsymbol{\zeta}\rangle - \delta_{\mathcal{Z}}(\boldsymbol{\zeta}) + \langle\boldsymbol{p}, \widehat{\boldsymbol{\zeta}}\rangle \\
&= \inf_{\|\boldsymbol{p}\|_*\le\gamma} S_{\mathcal{Z}}(a\boldsymbol{\beta} - \boldsymbol{p}) + \langle\boldsymbol{p}, \widehat{\boldsymbol{\zeta}}\rangle,
\end{aligned}
$$

where the first equality follows from the definition of the dual norm, the second equality holds due to the minimax theorem [9, Proposition 5.5.4], and the last equality holds because the support function $S_{\mathcal{Z}}$ is the conjugate of the indicator function $\delta_{\mathcal{Z}}$. Thus, the claim follows. $\qquad\square$

Lemma A.3. If $L(z)$ is a convex and Lipschitz continuous loss function, $\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}} \in \mathbb{R}^d$ and $\gamma > 0$, then

$$
\sup_{\boldsymbol{\zeta}\in\mathbb{R}^d} L(\langle\boldsymbol{\beta}, \boldsymbol{\zeta}\rangle) - \gamma\|\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\| = \begin{cases} L(\langle\boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}\rangle) & \text{if } \operatorname{lip}(L)\|\boldsymbol{\beta}\|_* \le \gamma \\ +\infty & \text{otherwise.} \end{cases}
$$

Proof. Note that $L(\langle\boldsymbol{\beta}, \boldsymbol{\zeta}\rangle) - \gamma\|\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\|$ constitutes a difference of convex functions and may thus be neither convex nor concave in $\boldsymbol{\zeta}$. In order to maximize this function, we re-write $I(\boldsymbol{\zeta}) = L(\langle\boldsymbol{\beta}, \boldsymbol{\zeta}\rangle)$ as an upper envelope of infinitely many affine functions. To this end, we express the conjugate of $I(\boldsymbol{\zeta})$ as

$$
I^*(\boldsymbol{z}) = \sup_{\boldsymbol{\zeta}} \langle\boldsymbol{z}, \boldsymbol{\zeta}\rangle - L(\langle\boldsymbol{\beta}, \boldsymbol{\zeta}\rangle) = \sup_{t, \boldsymbol{\zeta}} \left\{\langle\boldsymbol{z}, \boldsymbol{\zeta}\rangle - L(t) : t = \langle\boldsymbol{\beta}, \boldsymbol{\zeta}\rangle\right\} = \inf_{\theta} \left\{L^*(\theta) : \theta\boldsymbol{\beta} = \boldsymbol{z}\right\},
$$

where the last equality follows from strong Lagrangian duality, which holds because Slater's constraint qualification is trivially satisfied in the absence of inequality constraints [9, Proposition 5.3.1]. Defining

29

$\Theta = \{\theta \in \mathbb{R} : L^*(\theta) < \infty\}$ as the effective domain of $L^*(\theta)$, we may then replace $\theta \in \mathbb{R}$ with $\theta \in \Theta$ in the last expression. As $I(\boldsymbol{\zeta})$ is convex and continuous, it coincides with its bi-conjugate, that is,

$$I(\boldsymbol{\zeta}) = I^{**}(\boldsymbol{\zeta}) = \sup_{\boldsymbol{z}} \ \langle \boldsymbol{z}, \boldsymbol{\zeta} \rangle - I^*(\boldsymbol{z}) = \sup_{\theta \in \Theta} \ \langle \theta \boldsymbol{\beta}, \boldsymbol{\zeta} \rangle - L^*(\theta).$$

In other words, we have represented $I(\boldsymbol{\zeta})$ as the upper envelope of infinitely many linear functions. Using this representation, we obtain

$$
\begin{aligned}
\sup_{\boldsymbol{\zeta}} \ I(\boldsymbol{\zeta}) - \gamma \|\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\| &= \sup_{\boldsymbol{\zeta}} \ I^{**}(\boldsymbol{\zeta}) - \gamma \|\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\| \\
&= \sup_{\boldsymbol{\zeta}} \ \sup_{\theta \in \Theta} \ \langle \theta \boldsymbol{\beta}, \boldsymbol{\zeta} \rangle - L^*(\theta) - \gamma \|\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\| \\
&= \sup_{\theta \in \Theta} \ \sup_{\boldsymbol{\zeta}} \ \inf_{\|\boldsymbol{p}\|_* \leq \gamma} \ \theta \langle \boldsymbol{\beta}, \boldsymbol{\zeta} \rangle - L^*(\theta) - \langle \boldsymbol{p}, \boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}} \rangle \\
&= \sup_{\theta \in \Theta} \ \inf_{\|\boldsymbol{p}\|_* \leq \gamma} \ \sup_{\boldsymbol{\zeta}} \ \langle \theta \boldsymbol{\beta} - \boldsymbol{p}, \boldsymbol{\zeta} \rangle - L^*(\theta) + \langle \boldsymbol{p}, \widehat{\boldsymbol{\zeta}} \rangle,
\end{aligned}
$$

where the last equality holds due to [9, Proposition 5.5.4]. Evaluating the maximization over $\boldsymbol{\zeta}$ yields

$$
\begin{aligned}
\sup_{\boldsymbol{\zeta}} \ I(\boldsymbol{\zeta}) - \gamma \|\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}}\| &= \sup_{\theta \in \Theta} \ \inf_{\|\boldsymbol{p}\|_* \leq \gamma} \begin{cases} \langle \boldsymbol{p}, \widehat{\boldsymbol{\zeta}} \rangle - L^*(\theta) & \text{if } \boldsymbol{p} = \theta \boldsymbol{\beta} \\ +\infty & \text{otherwise} \end{cases} \\
&= \sup_{\theta \in \Theta} \begin{cases} \langle \theta \boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}} \rangle - L^*(\theta) & \text{if } \|\theta \boldsymbol{\beta}\|_* \leq \gamma \\ +\infty & \text{otherwise} \end{cases} \\
&= \begin{cases} \sup_{\theta \in \Theta} \ \theta \langle \boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}} \rangle - L^*(\theta) & \text{if } \sup_{\theta \in \Theta} \|\theta \boldsymbol{\beta}\|_* \leq \gamma \\ +\infty & \text{otherwise} \end{cases} \\
&= \begin{cases} L(\langle \boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}} \rangle) & \text{if } \sup_{\theta \in \Theta} |\theta| \cdot \|\boldsymbol{\beta}\|_* \leq \gamma \\ +\infty & \text{otherwise.} \end{cases}
\end{aligned}
$$

Thus, the claim follows by noting that $\sup_{\theta}\{|\theta| : L^*(\theta) < \infty\}$ represents the Lipschitz modulus of L. $\qquad \square$

Proof of Theorem 3.1. To prove assertion (i), we apply Lemma A.1 to the integrand $f(\boldsymbol{x}, y) = L(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - y)$ with $L(z) = \max_{j \leq J}\{a_j z + b_j\}$ to obtain

$$
\begin{aligned}
\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)] &= \inf_{\lambda \geq 0} \ \lambda \rho + \frac{1}{N} \sum_{i=1}^{N} \sup_{(\boldsymbol{x},y) \in \Xi} L(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - y) - \lambda \|(\boldsymbol{x}, y) - (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)\| \\
&= \inf_{\lambda \geq 0} \ \lambda \rho + \frac{1}{N} \sum_{i=1}^{N} \max_{j \leq J} \sup_{(\boldsymbol{x},y) \in \Xi} a_j(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - y) + b_j - \lambda \|(\boldsymbol{x}, y) - (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)\| \\
&= \begin{cases} \inf_{\lambda, \boldsymbol{p}_{ij}, u_{ij}} \ \lambda \rho + \sum_{i=1}^{N} \max_{j \leq J} S_\Xi(a_j \boldsymbol{w} - \boldsymbol{p}_{ij}, -a_j - u_{ij}) + \langle \boldsymbol{p}_{ij}, \widehat{\boldsymbol{x}}_i \rangle + u_{ij} \widehat{y}_i + b_j \\ \text{s.t.} \quad \|(\boldsymbol{p}_{ij}, u_{ij})\|_* \leq \lambda \qquad i \in [N], j \in [J], \end{cases}
\end{aligned}
$$

where the last equality follows from Lemma A.2. The claim now follows by introducing auxiliary epigraphical variables s_i for the max-terms in the objective function and by including $\boldsymbol{w}$ as a decision variable.

To prove assertion (ii), we apply Lemma A.1 to the integrand $f(\boldsymbol{x}, y) = L(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - y)$, where L is a Lipschitz continuous convex loss function. Thus we find

$$
\begin{aligned}
\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\ell_{\boldsymbol{w}}(\boldsymbol{x}, y)] &= \inf_{\lambda \geq 0} \ \lambda \rho + \sum_{i=1}^{N} \sup_{\boldsymbol{x}, y} L(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - y) - \lambda \|(\boldsymbol{x}, y) - (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)\| \\
&= \begin{cases} \inf_{\lambda} \ \lambda \rho + \frac{1}{N} \sum_{i=1}^{N} L(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle - \widehat{y}_i) \\ \text{s.t.} \quad \text{lip}(L)|\theta| \cdot \|(\boldsymbol{w}, -1)\|_* \leq \lambda, \end{cases}
\end{aligned}
$$

30

where the last equality uses Lemma A.3. Next, we eliminate λ and include $\boldsymbol{w}$ as a decision variable. $\qquad\square$

Proof of Corollary 3.2. Note that the Huber loss function $L(z)$ coincides with the inf-convolution of $\frac{1}{2}z^2$ and $\delta|z|$ and can thus be expressed as $L(z) = \min_{z_1} \frac{1}{2}z_1^2 + \delta|z - z_1|$. Moreover, the Lipschitz modulus of the Huber loss function is δ. The rest of the proof follows from Theorem 3.1(ii). $\qquad\square$

Proof of Corollary 3.3. Notice that the ϵ-insensitive loss function is a piecewise linear function with $J = 3$ pieces, see Section 2.1. By strong conic duality, the support function of $\Xi = \{(\boldsymbol{x}, y) \in \mathbb{R}^{n+1} : \boldsymbol{C}_1\boldsymbol{x} + \boldsymbol{c}_2 y \preceq_\mathcal{C} \boldsymbol{d}\}$ can be re-expressed as

$$ S_\Xi(\boldsymbol{z}_1, z_2) = \sup_{\boldsymbol{x}, y} \{\langle \boldsymbol{z}_1, \boldsymbol{x}\rangle + z_2 y : \boldsymbol{C}_1\boldsymbol{x} + \boldsymbol{c}_2 y \preceq_\mathcal{C} \boldsymbol{d}\} = \inf_{\boldsymbol{q} \in \mathcal{C}^*} \left\{\langle \boldsymbol{q}, \boldsymbol{d}\rangle : \boldsymbol{C}_1^\top \boldsymbol{q} = \boldsymbol{z}_1, \ \boldsymbol{c}_2^\top \boldsymbol{q} = z_2\right\}. $$

Strong duality holds because Ξ admits a Slater point. The rest of proof follows from Theorem 3.1(i). $\qquad\square$

Proof of Corollary 3.4. The pinball loss function is a piecewise linear function with $J = 2$ pieces, see Section 2.1. The rest of proof follows from the dual representation of the support function $S_\Xi(\boldsymbol{z}_1, z_2)$, which is known from the proof of Corollary 3.3, and from Theorem 3.1(i). $\qquad\square$

The proof of Theorem 3.6 is based on the following preparatory lemma.

Lemma A.4. *If $\mathcal{Z} \subseteq \mathbb{R}^d$ is a non-empty convex closed set, $\widehat{\boldsymbol{\zeta}} \in \mathcal{Z}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\alpha, \gamma \geq 0$, then we have*

$$ \inf_{\boldsymbol{p}} \ \alpha S_\mathcal{Z}(\boldsymbol{\beta} - \boldsymbol{p}) + \alpha\langle \boldsymbol{p}, \widehat{\boldsymbol{\zeta}}\rangle + \gamma\|\boldsymbol{p}\|_* = \left\{ \begin{array}{cl} \sup\limits_{\|\boldsymbol{q}\| \leq \gamma} & \alpha\langle \boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}\rangle + \langle \boldsymbol{\beta}, \boldsymbol{q}\rangle \\ \text{s.t.} & \widehat{\boldsymbol{\zeta}} + \boldsymbol{q}/\alpha \in \mathcal{Z}. \end{array}\right. $$

Proof. If $\alpha = 0$, then the optimal values of both opimization problems vanish due to our conventions of extended arithmetic, and thus the claim trivially holds. If $\alpha > 0$, however, we have

$$ \begin{aligned} \inf_{\boldsymbol{p}} \ \alpha S_\mathcal{Z}(\boldsymbol{\beta} - \boldsymbol{p}) + \alpha\langle \boldsymbol{p}, \widehat{\boldsymbol{\zeta}}\rangle + \gamma\|\boldsymbol{p}\|_* &= \inf_{\boldsymbol{p}} \sup_{\|\boldsymbol{q}\| \leq \gamma} \alpha S_\mathcal{Z}(\boldsymbol{\beta} - \boldsymbol{p}) + \alpha\langle \boldsymbol{p}, \widehat{\boldsymbol{\zeta}}\rangle + \langle \boldsymbol{p}, \boldsymbol{q}\rangle \\ &= \sup_{\|\boldsymbol{q}\| \leq \gamma} \inf_{\boldsymbol{p}} \ \alpha S_\mathcal{Z}(\boldsymbol{\beta} - \boldsymbol{p}) + \langle \boldsymbol{p}, \alpha\widehat{\boldsymbol{\zeta}} + \boldsymbol{q}\rangle \\ &= \sup_{\|\boldsymbol{q}\| \leq \gamma} \inf_{\boldsymbol{z}} \ \alpha S_\mathcal{Z}(\boldsymbol{z}) + \langle \boldsymbol{\beta} - \boldsymbol{z}, \alpha\widehat{\boldsymbol{\zeta}} + \boldsymbol{q}\rangle \\ &= \sup_{\|\boldsymbol{q}\| \leq \gamma} \ \alpha\langle \boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}\rangle + \langle \boldsymbol{\beta}, \boldsymbol{q}\rangle - \alpha\big(\sup_{\boldsymbol{z}}\langle \boldsymbol{z}, \widehat{\boldsymbol{\zeta}} + \boldsymbol{q}/\alpha\rangle - S_\mathcal{Z}(\boldsymbol{z})\big) \\ &= \sup_{\|\boldsymbol{q}\| \leq \gamma} \ \alpha\langle \boldsymbol{\beta}, \widehat{\boldsymbol{\zeta}}\rangle + \langle \boldsymbol{\beta}, \boldsymbol{q}\rangle - \alpha\delta_\mathcal{Z}(\widehat{\boldsymbol{\zeta}} + \boldsymbol{q}/\alpha), \end{aligned} $$

where the first equality follows from the definition of the dual norm, the second equality exploits [9, Proposition 5.5.4], and the last equality holds because, for any convex closed set, the indicator function is the conjugate of the support function. $\qquad\square$

Proof of Theorem 3.6. We first prove assertion (i). By Theorem 3.1(i), the worst-case expectation problem (13) constitutes a restriction of (5) where $\boldsymbol{w}$ is fixed, and thus it coincides with the minimax problem

$$ \begin{aligned} \inf_{\substack{\lambda, s_i \\ \boldsymbol{p}_{ij}, u_{ij}}} \sup_{\alpha_{ij} \geq 0, \gamma_{ij} \geq 0} & \ \lambda\rho + \frac{1}{N}\sum_{i=1}^N s_i + \sum_{i=1}^N\sum_{j=1}^J \alpha_{ij}\big(S_\Xi(-a_j\boldsymbol{w} - \boldsymbol{p}_{ij}, a_j - u_{ij}) + b_j + \langle \boldsymbol{p}_{ij}, \widehat{\boldsymbol{x}}_i\rangle + u_{ij}\widehat{y}_i - s_i\big) \\ & + \sum_{i=1}^N\sum_{j=1}^J \gamma_{ij}\big(\|(\boldsymbol{p}_{ij}, u_{ij})\|_* - \lambda\big). \end{aligned} $$

The minimization and the maximization may be interchanged by strong duality, which holds because the convex program (5) satisfies Slater's constraint qualification for every fixed $\boldsymbol{w}$ [9, Proposition 5.3.1].

Indeed, note that S_Ξ is proper, convex and lower semi-continuous and appears in constraints that are always satisfiable because they involve a free decision variable. Thus, the above minimax problem is equivalent to

$$
\begin{cases}
\displaystyle\sup_{\alpha_{ij},\gamma_{ij}} \quad \inf_{\boldsymbol{p}_{ij},u_{ij}} \sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ij}\big(S_\Xi(a_j\boldsymbol{w} - \boldsymbol{p}_{ij}, a_j - u_{ij}) + b_j + \langle \boldsymbol{p}_{ij}, \widehat{\boldsymbol{x}}_i\rangle + u_{ij}\widehat{y}_i\big) + \sum_{i=1}^{N}\sum_{j=1}^{J}\gamma_{ij}\|\boldsymbol{p}_{ij}\|_* \\[2ex]
\text{s.t.} \quad \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{J}\gamma_{ij} = \rho \\[2ex]
\qquad \displaystyle\sum_{j=1}^{J}\alpha_{ij} = \frac{1}{N} \qquad\qquad i \in [N] \\[2ex]
\qquad \alpha_{ij}, \gamma_{ij} \geq 0 \qquad\qquad i \in [N], j \in [J].
\end{cases}
$$

By Lemma A.4, which applies because $(\widehat{\boldsymbol{x}}_i, \widehat{y}_i) \in \Xi$ for all $i \leq N$, the above dual problem simplifies to

$$
\begin{cases}
\displaystyle\sup_{\substack{\alpha_{ij},\gamma_{ij}\\ \boldsymbol{q}_{ij},v_{ij}}} \quad \sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ij}\big(a_j(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i) + b_j\big) + a_j(\langle \boldsymbol{w}, \boldsymbol{q}_{ij}\rangle - v_{ij}) \\[2ex]
\text{s.t.} \quad \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{J}\gamma_{ij} = \rho \\[2ex]
\qquad \displaystyle\sum_{j=1}^{J}\alpha_{ij} = \frac{1}{N} \qquad\qquad\qquad\qquad i \in [N] \\[2ex]
\qquad \|(\boldsymbol{q}_{ij}, v_{ij})\| \leq \gamma_{ij} \qquad\qquad\quad i \in [N], j \in [J] \\[1ex]
\qquad (\widehat{\boldsymbol{x}}_i - \boldsymbol{q}_{ij}/\alpha_{ij}, \widehat{y}_i - v_{ij}/\alpha_{ij}) \in \Xi \quad i \in [N], j \in [J] \\[1ex]
\qquad \alpha_{ij}, \gamma_{ij} \geq 0 \qquad\qquad\qquad\qquad i \in [N], j \in [J].
\end{cases}
$$

Problem (14) is now obtained by eliminating the variables γ_{ij} and by substituting α_{ij}, $\boldsymbol{q}_{ij}$, and v_{ij} with α_{ij}/N, $\boldsymbol{q}_{ij}/N$, and v_{ij}/N, respectively.

As for assertion (ii), we first show that the discrete distribution $\mathbb{Q}_\gamma$ belongs to the Wasserstein ball $\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ for all $\gamma \in (0,1]$. Indeed, the Wasserstein distance between $\mathbb{Q}_\gamma$ and $\widehat{\mathbb{P}}_N$ amounts to

$$
d(\mathbb{Q}_\gamma, \widehat{\mathbb{P}}_N) \leq \frac{\gamma}{N}\left\|\left(\widehat{\boldsymbol{x}}_1 + \frac{\rho N}{\gamma}\boldsymbol{x}^\star, \widehat{y}_1 + \frac{\rho N}{\gamma}y^\star\right) - (\widehat{\boldsymbol{x}}_1, \widehat{y}_1)\right\| = \rho\|(\boldsymbol{x}^\star, y^\star)\| \leq \rho,
$$

where the first inequality holds because the Wasserstein distance coincides with the optimal mass transportation cost, and the last inequality holds because the norm of $(\boldsymbol{x}^\star, y^\star)$ is at most 1 by construction. Thus, $\mathbb{Q}_\gamma \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ for all $\gamma \in (0,1]$. Denoting the optimal value of (13) by $J^\star(\boldsymbol{w})$ and using $\ell_{\boldsymbol{w}}(\boldsymbol{x}, y)$ as a shorthand for $L(\langle \boldsymbol{w}, \boldsymbol{x}\rangle - y)$, we find

$$
\begin{aligned}
J^\star(\boldsymbol{w}) \ &\geq \ \mathbb{E}^{\mathbb{Q}_\gamma}[\ell_{\boldsymbol{w}}(\boldsymbol{x}, y)] \\
&= \ \frac{1}{N}\sum_{i=1}^{N}\ell_{\boldsymbol{w}}(\widehat{\boldsymbol{x}}_i, \widehat{y}_i) - \frac{\gamma}{N}\ell_{\boldsymbol{w}}(\widehat{\boldsymbol{x}}_1, \widehat{y}_1) + \frac{\gamma}{N}\ell_{\boldsymbol{w}}(\widehat{\boldsymbol{x}}_1 + \frac{\rho N}{\gamma}\boldsymbol{x}^\star, \widehat{y}_1 + \frac{\rho N}{\gamma}y^\star) \\
&\geq \ \frac{1}{N}\sum_{i=1}^{N}\ell_{\boldsymbol{w}}(\widehat{\boldsymbol{x}}_i, \widehat{y}_i) - \frac{\gamma}{N}\ell_{\boldsymbol{w}}(\widehat{\boldsymbol{x}}_1, \widehat{y}_1) + \frac{\gamma}{N}\big(\langle(\boldsymbol{x}, y), (\widehat{\boldsymbol{x}}_1 + \frac{\rho N}{\gamma}\boldsymbol{x}^\star, \widehat{y}_1 + \frac{\rho N}{\gamma}y^\star)\rangle \\
&\qquad -\ell_{\boldsymbol{w}}^*(\boldsymbol{x}, y)\big) \quad \forall(\boldsymbol{x}, y) \in \mathbb{R}^{n+1},
\end{aligned}
$$

where the last estimate follows from Fenchel's inequality. Setting $(\boldsymbol{x}, y) = \mathrm{lip}(L)(\boldsymbol{w}, -1)$ we thus have

$$
\begin{aligned}
J^\star(\boldsymbol{w}) \ &\geq \ \lim_{\gamma\to 0^+} \frac{1}{N}\sum_{i=1}^{N}\ell_{\boldsymbol{w}}(\widehat{\boldsymbol{x}}_i, \widehat{y}_i) - \frac{\gamma}{N}\ell_{\boldsymbol{w}}(\widehat{\boldsymbol{x}}_1, \widehat{y}_1) + \frac{\gamma}{N}\mathrm{lip}(L)(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_1\rangle - \widehat{y}_1) + \rho\,\mathrm{lip}(L)\|(\boldsymbol{w}, -1)\|_* \\
&\qquad -\frac{\gamma}{N}\ell_{\boldsymbol{w}}^*(\mathrm{lip}(L)(\boldsymbol{w}, -1)) \\
&= \ \frac{1}{N}\sum_{i=1}^{N}\ell_{\boldsymbol{w}}(\widehat{\boldsymbol{x}}_i, \widehat{y}_i) + \rho\,\mathrm{lip}(L)\|(\boldsymbol{w}, -1)\|_* = J^\star(\boldsymbol{w}),
\end{aligned}
$$

where the equality follows from Theorem 3.1(ii). The above reasoning implies that $\lim_{\gamma\to 0^+}\mathbb{E}^{\mathbb{Q}_\gamma}[\ell_{\boldsymbol{w}}(\boldsymbol{x}, y)] = J^\star(\boldsymbol{w})$, and thus the claim follows. $\qquad\square$

Proof of Theorem 3.9. Assume first that the loss function L is convex piecewise linear, that is, $L(z) = \max_{j \in J}\{a_j z + b_j\}$. As $\Xi \in \mathbb{R}^{n+1}$, Theorem 3.6(i) implies that the worst-case expectation (13) is given by

$$
\begin{cases}
\displaystyle \sup_{\alpha_{ij}, \boldsymbol{q}_{ij}, v_{ij}} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} \alpha_{ij} \left[a_j(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle - \widehat{y}_i) + b_j \right] + a_j(\langle \boldsymbol{w}, \boldsymbol{q}_{ij} \rangle - v_{ij}) \\[3mm]
\text{s.t.} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} \|(\boldsymbol{q}_{ij}, v_{ij})\| \leq \rho \\[3mm]
& \displaystyle \sum_{j=1}^{J} \alpha_{ij} = 1 \qquad\qquad i \in [N] \\[2mm]
& \alpha_{ij} \geq 0 \qquad\qquad i \in [N], j \in [J]
\end{cases}
$$

$$
\geq
\begin{cases}
\displaystyle \sup_{\alpha_{ij}, \Delta\boldsymbol{x}_{ij}, \Delta y_{ij}} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} \alpha_{ij} \left[a_j(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i + \Delta\boldsymbol{x}_{ij} \rangle - \widehat{y}_i - \Delta y_{ij}) + b_j \right] \\[3mm]
\text{s.t.} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} \alpha_{ij} \|(\Delta\boldsymbol{x}_{ij}, \Delta y_{ij})\| \leq \rho \\[3mm]
& \displaystyle \sum_{j=1}^{J} \alpha_{ij} = 1 \qquad\qquad i \in [N] \\[2mm]
& \alpha_{ij} \geq 0 \qquad\qquad i \in [N], j \in [J]
\end{cases}
$$

$$
\geq
\begin{cases}
\displaystyle \sup_{\alpha_{ij}, \Delta\boldsymbol{x}_i, \Delta y_i} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} \alpha_{ij} \left[a_j(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i + \Delta\boldsymbol{x}_i \rangle - \widehat{y}_i - \Delta y_i) + b_j \right] \\[3mm]
\text{s.t.} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} \alpha_{ij} \|(\Delta\boldsymbol{x}_i, \Delta y_i)\| \leq \rho \\[3mm]
& \displaystyle \sum_{j=1}^{J} \alpha_{ij} = 1 \qquad\qquad i \in [N] \\[2mm]
& \alpha_{ij} \geq 0 \qquad\qquad i \in [N], j \in [J].
\end{cases}
$$

The first inequality holds because for any feasible solution $\{\alpha_{ij}, \Delta\boldsymbol{x}_{ij}, \Delta y_{ij}\}$ to the second optimization problem, the solution $\{\alpha_{ij}, \boldsymbol{q}_{ij}, v_{ij}\}$ with $\boldsymbol{q}_{ij} = \alpha_{ij}\Delta\boldsymbol{x}_{ij}$ and $v_{ij} = \alpha_{ij}\Delta y_{ij}$ is feasible in the first problem and attains the same objective value (conversely, note that the first problem admits feasible solutions with $\alpha_{ij} = 0$ and $\boldsymbol{q}_{ij} \neq \boldsymbol{0}$ that have no counterpart in the second problem). The second inequality follows from the restriction that $\Delta\boldsymbol{x}_{ij}$ and Δy_{ij} must be independent of j. It is easy to verify that the last optimization problem in the above expression is equivalent to (15) because $(\alpha_{i1}, \ldots, \alpha_{iJ})$ ranges over a simplex for every $i \leq N$, and thus (15) provides a lower bound on (13).

Suppose now that Assumption 3.7 holds, and note that the worst-case loss (15) can be expressed as

$$
\begin{cases}
\displaystyle \sup_{\Delta\boldsymbol{x}_i, \Delta y_i} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \max_{j \leq J} \left[a_j(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i + \Delta\boldsymbol{x}_i \rangle - \widehat{y}_i - \Delta y_i) + b_j \right] \\[3mm]
\text{s.t.} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \|(\Delta\boldsymbol{x}_i, \Delta y_i)\| \leq \rho
\end{cases}
$$

$$
\geq
\begin{cases}
\displaystyle \sup_{\Delta\boldsymbol{x}, \Delta y} & \displaystyle \frac{1}{N} \sum_{i \neq k} \max_{j \leq J} \left[a_j(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle - \widehat{y}_i) + b_j \right] + \frac{1}{N} \max_{j \leq J} \left[a_j(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k + \Delta\boldsymbol{x} \rangle - \widehat{y}_k - \Delta y) + b_j \right] \\[3mm]
\text{s.t.} & \displaystyle \frac{1}{N} \|(\Delta\boldsymbol{x}, \Delta y)\| \leq \rho.
\end{cases}
$$

The above inequality follows from setting $\Delta\boldsymbol{x}_i = 0$ and $\Delta y_i = 0$ for all $i \neq k$, where $(\widehat{\boldsymbol{x}}_k, \widehat{y}_k)$ is a training sample satisfying $|L'(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle - \widehat{y}_k)| = \text{lip}(L)$, which exists due to Assumption 3.7. The last expression

equals

$$
\begin{aligned}
=&\frac{1}{N}\sum_{i\neq k}\max_{j\leq J}\left[a_j(\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i)+b_j\right] + \max_{j\leq J}\left\{\begin{array}{ll}\sup\limits_{\Delta\boldsymbol{x},\Delta y} & \frac{1}{N}\left[a_j(\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_k\rangle - \widehat{y}_k)+b_j\right]+\frac{1}{N}\left[a_j(\langle \boldsymbol{w},\Delta\boldsymbol{x}\rangle - \Delta y)\right]\\[2mm] \text{s.t.} & \frac{1}{N}\|(\Delta\boldsymbol{x},\Delta y)\|\leq\rho\end{array}\right.\\[2mm]
=&\frac{1}{N}\sum_{i\neq k}\max_{j\leq J}\left[a_j(\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i)+b_j\right] + \frac{1}{N}\max_{j\leq J}\left[a_j(\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_k\rangle - \widehat{y}_k)+b_j+\rho N\|(\boldsymbol{w},-1)\|_*|a_j|\right]\\[2mm]
=&\frac{1}{N}\sum_{i\neq k}\max_{j\leq J}\left[a_j(\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i)+b_j\right] + \frac{1}{N}\max_{j\leq J}\left[a_j(\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_k\rangle - \widehat{y}_k)+b_j\right]+\max_{j\leq J}\rho\|(\boldsymbol{w},-1)\|_*|a_j|\\[2mm]
=&\frac{1}{N}\sum_{i=1}^{N}L(\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i)+\max_{j\leq J}\rho\|(\boldsymbol{w},-1)\|_*|a_j|,
\end{aligned}
$$

where the penultimate equality holds because $\langle \boldsymbol{w},\widehat{\boldsymbol{x}}_k\rangle - \widehat{y}_k$ resides on the steepest linear piece of the loss function L by virtue of Assumption 3.7. The claim then follows from Theorem 3.1(ii) because $\mathrm{lip}(L)=\max_{j\leq J}|a_j|$. Note that generic convex Lipschitz continuous loss functions can be uniformly approximated as closely as desired with convex piecewise linear functions. Thus, the above argument extends directly to generic convex Lipschitz continuous loss functions. Details are omitted for brevity. $\qquad\square$

Proof of Theorem 3.11. To prove assertion (i), we apply Lemma A.1 to problem (4) with the transportation distance (16), where $\boldsymbol{\xi}=(\boldsymbol{x},y)$ and $\Xi=\mathbb{X}\times\{-1,+1\}$. Thus, we obtain

$$
\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)}\mathbb{E}^{\mathbb{Q}}\left[\ell(\langle \boldsymbol{w},\boldsymbol{x}\rangle,y)\right]=\left\{\begin{array}{ll}\inf\limits_{\lambda,s_i} & \lambda\rho+\dfrac{1}{N}\sum\limits_{i=1}^{N}s_i\\[3mm] \text{s.t.} & \sup\limits_{(\boldsymbol{x},y)\in\Xi}L(y\langle \boldsymbol{w},\boldsymbol{x}\rangle)-\lambda\left(\|\boldsymbol{x}-\widehat{\boldsymbol{x}}_i\|+\tfrac{\kappa}{2}|y-\widehat{y}_i|\right)\leq s_i & i\in[N]\\[3mm] & \lambda\geq 0.\end{array}\right.
$$

$$
=\left\{\begin{array}{ll}\inf\limits_{\lambda,s_i} & \lambda\rho+\dfrac{1}{N}\sum\limits_{i=1}^{N}s_i\\[3mm] \text{s.t.} & \sup\limits_{\boldsymbol{x}\in\mathbb{X}}L(\widehat{y}_i\langle \boldsymbol{w},\boldsymbol{x}\rangle)-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}_i\|\leq s_i & i\in[N]\\[3mm] & \sup\limits_{\boldsymbol{x}\in\mathbb{X}}L(-\widehat{y}_i\langle \boldsymbol{w},\boldsymbol{x}\rangle)-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}_i\|-\kappa\lambda\leq s_i & i\in[N]\\[3mm] & \lambda\geq 0\end{array}\right.
$$

$$
=\left\{\begin{array}{ll}\inf\limits_{\lambda,s_i} & \lambda\rho+\dfrac{1}{N}\sum\limits_{i=1}^{N}s_i\\[3mm] \text{s.t.} & \sup\limits_{\boldsymbol{x}\in\mathbb{X}}a_j\widehat{y}_i\langle \boldsymbol{w},\boldsymbol{x}\rangle+b_j-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}_i\|\leq s_i & i\in[N],j\in[J]\\[3mm] & \sup\limits_{\boldsymbol{x}\in\mathbb{X}}-a_j\widehat{y}_i\langle \boldsymbol{w},\boldsymbol{x}\rangle+b_j-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}_i\|-\kappa\lambda\leq s_i & i\in[N],j\in[J]\\[3mm] & \lambda\geq 0,\end{array}\right.
$$

where the second equality holds because, for every i, y can be either equal to $\widehat{y}_i$ or to $-\widehat{y}_i$. Reformulating the constraints using Lemma A.2 and including $\boldsymbol{w}$ as a decision variable then yields (17).

When $\mathbb{X}=\mathbb{R}^n$ and L is Lipschitz continuous, we can use similar arguments as above to prove that

$$
\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)}\mathbb{E}^{\mathbb{Q}}\left[\ell(\langle \boldsymbol{w},\boldsymbol{x}\rangle,y)\right]=\left\{\begin{array}{ll}\inf\limits_{\lambda,s_i} & \lambda\rho+\dfrac{1}{N}\sum\limits_{i=1}^{N}s_i\\[3mm] \text{s.t.} & \sup\limits_{\boldsymbol{x}\in\mathbb{R}^n}L(\widehat{y}_i\langle \boldsymbol{w},\boldsymbol{x}\rangle)-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}_i\|\leq s_i & i\in[N]\\[3mm] & \sup\limits_{\boldsymbol{x}\in\mathbb{R}^n}L(-\widehat{y}_i\langle \boldsymbol{w},\boldsymbol{x}\rangle)-\lambda\|\boldsymbol{x}-\widehat{\boldsymbol{x}}_i\|-\kappa\lambda\leq s_i & i\in[N]\\[3mm] & \lambda\geq 0.\end{array}\right.
$$

Applying Lemma A.3 to the subordinate maximization problems in the constraints yields

$$
\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)\right] =
\begin{cases}
\inf_{\lambda, s_i} & \lambda\rho + \dfrac{1}{N}\sum_{i=1}^{N} s_i \\
\text{s.t.} & L(\widehat{y}_i\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle) \leq s_i & i \in [N], j \in [J] \\
& L(-\widehat{y}_i\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle) - \kappa\lambda \leq s_i & i \in [N], j \in [J] \\
& \sup_{\theta \in \Theta} |\theta| \cdot \|\boldsymbol{w}\|_* \leq \lambda.
\end{cases}
$$

Thus, assertion (ii) follows by recalling that $\mathrm{lip}(L) = \sup_\theta\{|\theta| : L^*(\theta) < \infty\} = \sup_{\theta \in \Theta}|\theta|$ and by including $\boldsymbol{w}$ as a decision variable. $\qquad\square$

Proof of Corollary 3.12. Notice that the hinge loss function is piecewise linear with $J = 2$ pieces, see Section 2.1. Moreover, by strong conic duality the support function of $\mathbb{X}$ can be re-expressed as

$$
S_{\mathbb{X}}(\boldsymbol{z}) = \sup_{\boldsymbol{x}}\left\{\langle \boldsymbol{z}, \boldsymbol{x}\rangle : \boldsymbol{C}\boldsymbol{x} \preceq_{\mathcal{C}} \boldsymbol{d}\right\} = \inf_{\boldsymbol{q} \in \mathcal{C}^*}\left\{\langle \boldsymbol{q}, \boldsymbol{d}\rangle : \boldsymbol{C}^\top\boldsymbol{q} = \boldsymbol{z}\right\}.
$$

Strong duality holds because $\mathbb{X}$ admits a Slater point. The proof thus follows from Theorem 3.11(i). $\qquad\square$

Proof of Corollary 3.13. The smooth hinge loss $L(z)$ coincides with the inf-convolution of $\frac{1}{2}z^2$ and $\max\{0, 1 - z\}$ and can thus be expressed as $L(z) = \min_{z_1}\frac{1}{2}z_1^2 + \max\{0, 1 - z - z_1\}$. Moreover, the Lipschitz modulus of the smooth hinge loss function is 1. The proof thus follows from Theorem 3.11(ii). $\qquad\square$

Proof of Corollary 3.14. Notice that the logloss function is convex and has Lipschitz modulus 1, see Section 2.1. The rest of proof follows from Theorem 3.11(ii). For more details see [61]. $\qquad\square$

Proof of Theorem 3.17. We first prove assertion (i). By Theorem 3.11(i), the worst-case expectation problem (25) constitutes a restriction of (17) where $\boldsymbol{w}$ is fixed, and thus it is equivalent to the minimax problem

$$
\inf_{\substack{\lambda, s_i \\ \boldsymbol{p}_{ij}^+, \boldsymbol{p}_{ij}^-}} \sup_{\substack{\alpha_{ij}^+ \geq 0, \gamma_{ij}^+ \geq 0 \\ \alpha_{ij}^- \geq 0, \gamma_{ij}^- \geq 0}} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} s_i + \sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ij}^+\left(S_{\mathbb{X}}(a_j\widehat{y}_i\boldsymbol{w} - \boldsymbol{p}_{ij}^+) + b_j + \langle \boldsymbol{p}_{ij}^+, \widehat{\boldsymbol{x}}_i \rangle - s_i\right)
$$

$$
+ \sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ij}^-\left(S_{\mathbb{X}}(-a_j\widehat{y}_i\boldsymbol{w} - \boldsymbol{p}_{ij}^-) + b_j + \langle \boldsymbol{p}_{ij}^-, \widehat{\boldsymbol{x}}_i \rangle - \kappa\lambda - s_i\right)
$$

$$
+ \sum_{i=1}^{N}\sum_{j=1}^{J} \gamma_{ij}^+\left(\|\boldsymbol{p}_{ij}^+\|_* - \lambda\right) + \sum_{i=1}^{N}\sum_{j=1}^{J} \gamma_{ij}^-\left(\|\boldsymbol{p}_{ij}^-\|_* - \lambda\right).
$$

The minimization and the maximization may be interchanged by strong Lagrangian duality, which holds because the convex program (17) satisfies Slater's constraint qualification for any fixed $\boldsymbol{w}$ [9, Proposition 5.3.1]. Indeed, note that $S_{\mathbb{X}}$ is proper, convex and lower semi-continuous and appears in constraints that are always satisfiable because they involve a free decision variable. Thus, problem (25) is equivalent to

$$
\begin{cases}
\sup_{\substack{\alpha_{ij}^+, \gamma_{ij}^+ \\ \alpha_{ij}^-, \gamma_{ij}^-}} \inf_{\boldsymbol{p}_{ij}^+, \boldsymbol{p}_{ij}^-} & \sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ij}^+\left(S_{\mathbb{X}}(a_j\widehat{y}_i\boldsymbol{w} - \boldsymbol{p}_{ij}^+) + b_j + \langle \boldsymbol{p}_{ij}^+, \widehat{\boldsymbol{x}}_i \rangle\right) + \gamma_{ij}^+\|\boldsymbol{p}_{ij}^+\|_* \\
& \quad + \sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ij}^-\left(S_{\mathbb{X}}(-a_j\widehat{y}_i\boldsymbol{w} - \boldsymbol{p}_{ij}^-) + b_j + \langle \boldsymbol{p}_{ij}^-, \widehat{\boldsymbol{x}}_i \rangle\right) + \gamma_{ij}^-\|\boldsymbol{p}_{ij}^-\|_* \\
\text{s.t.} & \sum_{i=1}^{N}\sum_{j=1}^{J} \gamma_{ij}^+ + \gamma_{ij}^- + \kappa\alpha_{ij}^- = \rho \\
& \sum_{j=1}^{J} \alpha_{ij}^+ + \alpha_{ij}^- = \dfrac{1}{N} & i \in [N] \\
& \alpha_{ij}^+, \alpha_{ij}^-, \gamma_{ij}^+, \gamma_{ij}^- \geq 0 & i \in [N], j \in [J].
\end{cases}
$$

By Lemma A.4, the above dual problem simplifies to

$$
\begin{cases}
\displaystyle\sup_{\substack{\alpha_{ij}^+,\gamma_{ij}^+,\boldsymbol{q}_{ij}^+ \\ \alpha_{ij}^-,\gamma_{ij}^-,\boldsymbol{q}_{ij}^-}} & \displaystyle\sum_{i=1}^N\sum_{j=1}^J (\alpha_{ij}^+ - \alpha_{ij}^-)a_j\widehat{y}_i\langle\boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle + a_j\widehat{y}_i\langle\boldsymbol{w},\boldsymbol{q}_{ij}^+ - \boldsymbol{q}_{ij}^-\rangle + \sum_{j=1}^J b_j \\[2ex]
\text{s.t.} & \displaystyle\sum_{i=1}^N\sum_{j=1}^J \gamma_{ij}^+ + \gamma_{ij}^- + \kappa\alpha_{ij}^- = \rho \\[2ex]
& \displaystyle\sum_{j=1}^J \alpha_{ij}^+ + \alpha_{ij}^- = \frac{1}{N} & i \in [N] \\[2ex]
& \|\boldsymbol{q}_{ij}^+\| \le \gamma_{ij}^+, \quad \|\boldsymbol{q}_{ij}^-\| \le \gamma_{ij}^- & i \in [N], j \in [J] \\[1ex]
& \widehat{\boldsymbol{x}}_i + \boldsymbol{q}_{ij}^+/\alpha_{ij}^+ \in \mathbb{X}, \quad \widehat{\boldsymbol{x}}_i + \boldsymbol{q}_{ij}^-/\alpha_{ij}^- \in \mathbb{X} & i \in [N], j \in [J] \\[1ex]
& \alpha_{ij}^+, \alpha_{ij}^-, \gamma_{ij}^+, \gamma_{ij}^- \ge 0 & i \in [N], j \in [J].
\end{cases}
$$

Problem (26) is now obtained by eliminating γ_{ij}^+ and γ_{ij}^- and by substituting α_{ij}^+, α_{ij}^-, $\boldsymbol{q}_{ij}^+$, and $\boldsymbol{q}_{ij}^-$ with α_{ij}^+/N, α_{ij}^-/N, $\boldsymbol{q}_{ij}^+/N$, and $\boldsymbol{q}_{ij}^-/N$, respectively.

As for assertion (ii), we use L_i^+ and L_i^- to abbreviate $L(\widehat{y}_i\langle\boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle)$ and $L(-\widehat{y}_i\langle\boldsymbol{w},\widehat{\boldsymbol{x}}_i\rangle)$, respectively. By Theorem 3.11(ii), we have

$$
\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[L(y\langle\boldsymbol{w},\boldsymbol{x}\rangle)] =
\begin{cases}
\displaystyle\inf_{\boldsymbol{w},\lambda,s_i} & \displaystyle\lambda\rho + \frac{1}{N}\sum_{i=1}^N s_i \\[2ex]
\text{s.t.} & L_i^+ \le s_i & i \in [N] \\[1ex]
& L_i^- - \kappa\lambda \le s_i & i \in [N] \\[1ex]
& \mathrm{lip}(L)\|(\boldsymbol{w})\|_* \le \lambda
\end{cases}
$$

$$
=
\begin{cases}
\displaystyle\sup_{\alpha_i^+,\alpha_i^-,\theta} & \displaystyle\mathrm{lip}(L)\|(\boldsymbol{w})\|_*\theta + \sum_{i=1}^N \alpha_i^+ L_i^+ + \alpha_i^- L_i^- \\[2ex]
\text{s.t.} & \alpha_i^+ + \alpha_i^- = \frac{1}{N} & i \in [N] \\[2ex]
& \displaystyle\theta + \kappa\sum_{i=1}^N \alpha_i^- = \rho \\[2ex]
& \alpha_i^+ \ge 0, \alpha_i^- \ge 0 & i \in [N] \\[1ex]
& \theta \ge 0,
\end{cases}
$$

where the second equality follows from strong linear programming duality, which holds because the primal problem is manifestly feasible. Eliminating the first constraint and replacing α_i^- with α_i/N and α_i^+ with $(1-\alpha_i)/N$ allows us to reformulate the dual linear program as

$$
\begin{cases}
\displaystyle\sup_{\alpha_i,\theta} & \displaystyle\mathrm{lip}(L)(\boldsymbol{w})\|_*\theta + \frac{1}{N}\sum_{i=1}^N (1-\alpha_i)L_i^+ + \alpha_i L_i^- \\[2ex]
\text{s.t.} & \displaystyle\theta + \frac{\kappa}{N}\sum_{i=1}^N \alpha_i = \rho \\[2ex]
& 0 \le \alpha_i \le 1 & i \in [N] \\[1ex]
& \theta \ge 0.
\end{cases}
$$

Thus, the worst-case expectation (25) coincides with the optimal value of (27) for $\gamma = 0$. Next let $(\alpha_i^\star(\gamma), \theta^\star(\gamma))$ be an optimal solution of (27) for $\gamma \ge 0$, and define $\mathbb{Q}_\gamma$ as in the theorem statement. Note that (27) is infeasible for $\gamma > \rho$. Moreover, note that the atoms of $\mathbb{Q}_\gamma$ have non-negative probabilities if $\eta(\gamma) \in [0,1]$, which holds whenever $\gamma \in [0,1]$. We thus focus on parameter values $\gamma \in [0, \min\{\rho,1\}]$. By construction, the Wasserstein distance between $\mathbb{Q}_\gamma$ and the empirical distribution satisfies

$$
\begin{aligned}
d(\mathbb{Q}_\gamma, \widehat{\mathbb{P}}_N) &\le \frac{\kappa}{N}\sum_{i=1}^N \alpha_i^\star(\gamma) - \frac{\eta(\gamma)\kappa}{N}\alpha_1^\star(\gamma) + \frac{\eta(\gamma)}{N}d\big((\widehat{\boldsymbol{x}}_1 + \tfrac{\theta^\star(\gamma)N}{\eta(\gamma)}\boldsymbol{x}^\star, \widehat{y}_1) - (\widehat{\boldsymbol{x}}_1, \widehat{y}_1)\big) \\
&\le \rho - \gamma - \theta^\star(\gamma) + \theta^\star(\gamma)\|\boldsymbol{x}^\star\| \le \rho,
\end{aligned}
$$

where the first inequality holds because the Wasserstein distance is defined as the minimum cost of moving $\mathbb{Q}_\gamma$ to $\widehat{\mathbb{P}}_N$, the second inequality follows from the feasibility of $(\alpha_i^\star(\gamma), \theta^\star(\gamma))$ in (27) and the non-negativity of $\eta(\gamma)$, and the last inequality holds because $\|\boldsymbol{x}^\star\| \leq 1$, $\theta^\star(\gamma) \geq 0$ and $\gamma \geq 0$. Thus, $\mathbb{Q}_\gamma \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ for all $\gamma \in [0, \min\{\rho, 1\}]$. Denoting the optimal value of (25) by $J^\star(\boldsymbol{w})$, we find

$$
\begin{aligned}
J^\star(\boldsymbol{w}) \;\geq\; & \mathbb{E}^{\mathbb{Q}_\gamma}\left[L(y\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\right] \\
=\; & \tfrac{1}{N}\Big(\sum_{i=1}^N (1-\alpha_i^\star(\gamma))L_i^+ + \alpha_i^\star(\gamma)L_i^-\Big) - \tfrac{\eta(\gamma)}{N}\big((1-\alpha_1^\star(\gamma))L_1^+ + \alpha_1^\star(\gamma)L_1^-\big) \\
& + \tfrac{\eta(\gamma)}{N} L\big(\widehat{y}_1\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_1 + \tfrac{\theta^\star(\gamma)N}{\eta(\gamma)}\boldsymbol{x}^\star\rangle\big) \\
\geq\; & \tfrac{1}{N}\Big(\sum_{i=1}^N (1-\alpha_i^\star(\gamma))L_i^+ + \alpha_i^\star(\gamma)L_i^-\Big) - \tfrac{\eta(\gamma)}{N}\big((1-\alpha_1^\star(\gamma))L_1^+ + \alpha_1^\star(\gamma)L_1^-\big) \\
& + \tfrac{\eta(\gamma)}{N}\Big(\langle \boldsymbol{x}, \widehat{\boldsymbol{x}}_1 + \tfrac{\theta^\star(\gamma)N}{\eta(\gamma)}\boldsymbol{x}^\star\rangle - L^*(\widehat{y}_1\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\Big) \qquad \forall \boldsymbol{x} \in \mathbb{R}^n,
\end{aligned}
$$

where the last estimate follows from Fenchel's inequality. Setting $\boldsymbol{x} = \mathrm{lip}(L)\boldsymbol{w}$ and driving γ to zero yields

$$
\begin{aligned}
J^\star(\boldsymbol{w}) \;\geq\; & \lim_{\gamma \to 0^+} \tfrac{1}{N}\Big(\sum_{i=1}^N (1-\alpha_i^\star(\gamma))L_i^+ + \alpha_i^\star(\gamma)L_i^-\Big) - \tfrac{\eta(\gamma)}{N}\big((1-\alpha_1^\star(\gamma))L_1^+ + \alpha_1^\star(\gamma)L_1^-\big) \\
& + \tfrac{\eta(\gamma)}{N}\big(\mathrm{lip}(L)\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_1\rangle - L^*(\widehat{y}_1\,\mathrm{lip}(L)\langle \boldsymbol{w}, \boldsymbol{w}\rangle)\big) + \mathrm{lip}(L)\|\boldsymbol{w}\|_* \cdot \theta^\star(\gamma) \\
=\; & \lim_{\gamma \to 0^+} \tfrac{1}{N}\Big(\sum_{i=1}^N (1-\alpha_i^\star(\gamma))L_i^+ + \alpha_i^\star(\gamma)L_i^-\Big) + \mathrm{lip}(L)\|\boldsymbol{w}\|_* \cdot \theta^\star(\gamma) = J^\star(\boldsymbol{w}),
\end{aligned}
$$

where the first equality follows from the observation that $\eta(\gamma) \in [0, \gamma]$, which implies that $\eta(\gamma)$ converges to zero as γ tends to zero. The second equality holds because the optimal value of (27) is concave and non-increasing and—*a fortiori*—continuous in $\gamma \in [0, \min\{\rho, 1\}]$ and because $J^\star(\boldsymbol{w})$ coincides with the optimal value of (27) when $\gamma = 0$. The above reasoning implies that $\lim_{\gamma \to 0^+} \mathbb{E}^{\mathbb{Q}_\gamma}[L(y\langle \boldsymbol{w}, \boldsymbol{x}\rangle)] = J^\star(\boldsymbol{w})$, and thus the claim follows. □

Proof of Theorem 3.20. Assume first that the loss function L is convex piecewise linear, that is, $L(z) = \max_{j \in J}\{a_j z + b_j\}$. As $\mathbb{X} = \mathbb{R}^n$ and $\kappa = \infty$, Theorem 3.17(i) implies that (25) can be expressed as

$$
\geq
\begin{cases}
\begin{cases}
\displaystyle \sup_{\alpha_{ij}^+, \boldsymbol{q}_{ij}^+} & \displaystyle \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^J \alpha_{ij}^+ a_j \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i\rangle + a_j \widehat{y}_i \langle \boldsymbol{w}, \boldsymbol{q}_{ij}^+\rangle + \sum_{j=1}^J b_j \\
\text{s.t.} & \displaystyle \sum_{i=1}^N \sum_{j=1}^J \|\boldsymbol{q}_{ij}^+\| \leq N\rho \\
& \displaystyle \sum_{j=1}^J \alpha_{ij}^+ = 1 \qquad i \in [N] \\
& \alpha_{ij}^+ \geq 0 \qquad i \in [N], j \in [J]
\end{cases} \\[2em]
\begin{cases}
\displaystyle \sup_{\alpha_{ij}, \Delta\boldsymbol{x}_{ij}} & \displaystyle \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^J \alpha_{ij} a_j \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i + \Delta\boldsymbol{x}_{ij}\rangle + \sum_{j=1}^J b_j \\
\text{s.t.} & \displaystyle \sum_{i=1}^N \sum_{j=1}^J \alpha_{ij}\|\Delta\boldsymbol{x}_{ij}\| \leq N\rho \\
& \displaystyle \sum_{j=1}^J \alpha_{ij} = 1 \qquad i \in [N] \\
& \alpha_{ij} \geq 0 \qquad i \in [N], j \in [J]
\end{cases}
\end{cases}
$$

$$\geq \begin{cases} \sup_{\alpha_{ij}, \Delta \boldsymbol{x}_i} & \dfrac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{J} \alpha_{ij} a_j \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i + \Delta \boldsymbol{x}_i \rangle + \sum_{j=1}^{J} b_j \\ \text{s.t.} & \sum_{i=1}^{N} \sum_{j=1}^{J} \alpha_{ij} \|\Delta \boldsymbol{x}_i\| \leq N\rho \\ & \sum_{j=1}^{J} \alpha_{ij} = 1 \quad i \in [N] \\ & \alpha_{ij} \geq 0 \qquad i \in [N], j \in [J]. \end{cases}$$

The first optimization problem constitutes a special case of (26). Indeed, as $\kappa = \infty$, the first constraint in (26) implies that $\alpha_{ij}^- = 0$, which in turn implies via the fourth constraint and our conventions of extended arithmetics that $\boldsymbol{q}_{ij}^- = \boldsymbol{0}$. The first inequality in the above expression holds because for any feasible solution $\{\alpha_{ij}, \Delta \boldsymbol{x}_{ij}\}$ to the second problem, the solution $\{\alpha_{ij}^+, \boldsymbol{q}_{ij}^+\}$ with $\boldsymbol{q}_{ij}^+ = \alpha_{ij}^+ \Delta \boldsymbol{x}_{ij}$ and $\alpha_{ij} = \alpha_{ij}^+$ for all $i \leq N$ and $j \leq J$ is feasible in the first problem and attains the same objective value. The second inequality in the above expression follows from the restriction that $\Delta \boldsymbol{x}_{ij}$ must be independent of j. It is easy to verify that the last optimization problem is equivalent to (28) because $(\alpha_{i1}, \ldots, \alpha_{iJ})$ ranges over a simplex for every $i \leq N$, and thus (28) provides a lower bound on (25).

Suppose now that Assumption 3.18 holds, and note that (28) can be expressed as

$$\geq \begin{cases} \begin{cases} \sup_{\Delta \boldsymbol{x}_i, \Delta y_i} & \dfrac{1}{N} \sum_{i=1}^{N} \max_{j \leq J} \left[a_j \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i + \Delta \boldsymbol{x}_i \rangle + b_j \right] \\ \text{s.t.} & \dfrac{1}{N} \sum_{i=1}^{N} \|\Delta \boldsymbol{x}_i\| \leq \rho \end{cases} \\ \begin{cases} \sup_{\Delta \boldsymbol{x}, \Delta y} & \dfrac{1}{N} \sum_{i \neq k} \max_{j \leq J} \left[a_j \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle + b_j \right] + \dfrac{1}{N} \max_{j \leq J} \left[a_j \widehat{y}_k \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k + \Delta \boldsymbol{x} \rangle + b_j \right] \\ \text{s.t.} & \dfrac{1}{N} \|\Delta \boldsymbol{x}\| \leq \rho. \end{cases} \end{cases}$$

The above inequality follows from setting $\Delta \boldsymbol{x}_i = 0$ and $\Delta y_i = 0$ for all $i \neq k$, where $(\widehat{\boldsymbol{x}}_k, \widehat{y}_k)$ is a training sample satisfying $|L'(\widehat{y}_k \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle)| = \mathrm{lip}(L)$, which exists due to Assumption 3.18. The last expression equals

$$\frac{1}{N} \sum_{i \neq k} \max_{j \leq J} \left[a_j \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle + b_j \right] + \max_{j \leq J} \begin{cases} \sup_{\Delta \boldsymbol{x}, \Delta y} & \dfrac{1}{N} \left[a_j \widehat{y}_k \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle + b_j \right] + \dfrac{1}{N} \left[a_j \widehat{y}_k \langle \boldsymbol{w}, \Delta \boldsymbol{x} \rangle \right] \\ \text{s.t.} & \dfrac{1}{N} \|(\Delta \boldsymbol{x}, \Delta y)\| \leq \rho \end{cases}$$

$$= \frac{1}{N} \sum_{i \neq k} \max_{j \leq J} \left[a_j \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle + b_j \right] + \frac{1}{N} \max_{j \leq J} \left[a_j \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle + b_j + \rho N \|\boldsymbol{w}\|_* |a_j| \right]$$

$$= \frac{1}{N} \sum_{i \neq k} \max_{j \leq J} \left[a_j \widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle + b_j \right] + \frac{1}{N} \max_{j \leq J} \left[a_j \widehat{y}_k \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle + b_j \right] + \max_{j \leq J} \rho \|\boldsymbol{w}\|_* |a_j|$$

$$= \frac{1}{N} \sum_{i=1}^{N} L(\widehat{y}_i \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle) + \max_{j \leq J} \rho \|\boldsymbol{w}\|_* |a_j|,$$

where the penultimate equality holds because $\widehat{y}_k \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_k \rangle$ resides on the steepest linear piece of the loss function L by virtue of Assumption 3.18. The claim then follows from Theorem 3.11(ii) because $\mathrm{lip}(L) = \max_{j \leq J} |a_j|$. Note that generic convex Lipschitz continuous loss functions can be uniformly approximated as closely as desired with convex piecewise linear functions. Thus, the above arguments extend directly to generic convex Lipschitz continuous loss functions. Details are omitted for brevity. $\qquad \square$

Proof of Theorem 3.24. By the definition of the feature map Φ corresponding to the kernel k, we have

$$\begin{aligned} \|\Phi(\boldsymbol{x}_1) - \Phi(\boldsymbol{x}_2)\|_{\mathbb{H}} &= \sqrt{\langle \Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_1) \rangle_{\mathbb{H}} - 2\langle \Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_2) \rangle_{\mathbb{H}} + \langle \Phi(\boldsymbol{x}_2), \Phi(\boldsymbol{x}_2) \rangle_{\mathbb{H}}} \\ &= \sqrt{k(\boldsymbol{x}_1, \boldsymbol{x}_1) - 2k(\boldsymbol{x}_1, \boldsymbol{x}_2) + k(\boldsymbol{x}_2, \boldsymbol{x}_2)} \leq g(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2) \end{aligned} \tag{45}$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{X}$, where the inequality follows from Assumption 3.22. As for assertion (i), we may use similar argument as in the proof of Lemma A.1 to reformulate the worst-case expectation in (29) as

$$
\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\ell(h(\boldsymbol{x}), y)\right] = \begin{cases} \sup\limits_{\mathbb{Q}^i} & \dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{X} \times \mathbb{Y}} \ell(h(\boldsymbol{x}), y) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) \\ \text{s.t.} & \dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{X} \times \mathbb{Y}} d\big((\boldsymbol{x}, y), (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)\big) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) \leq \rho \\ & \int_{\mathbb{X} \times \mathbb{Y}} \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) = 1 \qquad i \in [N] \end{cases}
$$

$$
= \begin{cases} \sup\limits_{\mathbb{Q}^i} & \dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{X} \times \mathbb{Y}} \ell(h(\boldsymbol{x}), y) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) \\ \text{s.t.} & g\left(\dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{X} \times \mathbb{Y}} \sqrt{2} d\big((\boldsymbol{x}, y), (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)\big) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) \right) \leq g(\sqrt{2}\rho) \\ & \int_{\mathbb{X} \times \mathbb{Y}} \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) = 1 \qquad i \in [N] \end{cases}
$$

$$
\leq \begin{cases} \sup\limits_{\mathbb{Q}^i} & \dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{X} \times \mathbb{Y}} \ell(h(\boldsymbol{x}), y) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) \\ \text{s.t.} & \dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{X} \times \mathbb{Y}} g\left(\sqrt{2} d\big((\boldsymbol{x}, y), (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)\big) \right) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) \leq g(\sqrt{2}\rho) \\ & \int_{\mathbb{X} \times \mathbb{Y}} \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) = 1 \qquad i \in [N], \end{cases} \tag{46}
$$

where the second equality holds because g is strictly monotonically increasing, and the inequality follows from Jensen inequality, which applies because g is concave. By the definition of the transportation metric on $\mathbb{H} \times \mathbb{Y}$ for regression problems, we then find

$$
\begin{aligned} g\left(\sqrt{2} d\big((\boldsymbol{x}, y), (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)\big) \right) = g\left(\sqrt{2\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\|_2^2 + 2(y - \widehat{y}_i)^2} \right) \\ \geq g(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\|_2 + |y - \widehat{y}_i|) \\ \geq g(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\|_2) + |y - \widehat{y}_i| \\ \geq \|\Phi(\boldsymbol{x}) - \Phi(\widehat{\boldsymbol{x}}_i)\|_\mathbb{H} + |y - \widehat{y}_i| \\ \geq \sqrt{\|\Phi(\boldsymbol{x}) - \Phi(\widehat{\boldsymbol{x}}_i)\|_\mathbb{H}^2 + (y - \widehat{y}_i)^2} = d_\mathbb{H}\big((\Phi(\boldsymbol{x}), y), (\Phi(\widehat{\boldsymbol{x}}_i), \widehat{y}_i)\big), \end{aligned} \tag{47}
$$

where the first inequality holds because $2a^2 + 2b^2 \geq (a + b)^2$ for all $a, b \geq 0$ and because g is strictly monotonically increasing, the second inequality exploits the assumption that $g'(z) \geq 1$ for all $z \geq 0$, the third inequality follows form (45), and the last equality holds because $a^2 + b^2 \leq (a + b)^2$ for all $a, b \geq 0$. Substituting the above estimate into (46) and using the reproducing property $h(\boldsymbol{x}) = \langle h, \Phi(\boldsymbol{x}) \rangle_\mathbb{H}$ yields

$$
\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\ell(h(\boldsymbol{x}), y)\right] \leq \begin{cases} \sup\limits_{\mathbb{Q}^i} & \dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{X} \times \mathbb{Y}} \ell(\langle h, \Phi(\boldsymbol{x}) \rangle_\mathbb{H}, y) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) \\ \text{s.t.} & \dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{X} \times \mathbb{Y}} d_\mathbb{H}\big((\Phi(\boldsymbol{x}), y), (\Phi(\widehat{\boldsymbol{x}}_i), \widehat{y}_i)\big) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) \leq g(\sqrt{2}\rho) \\ & \int_{\mathbb{X} \times \mathbb{Y}} \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}, \mathrm{d}y) = 1 \qquad i \in [N] \end{cases}
$$

$$
\leq \begin{cases} \sup\limits_{\mathbb{Q}^i} & \dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{H} \times \mathbb{Y}} \ell(\boldsymbol{x}_\mathbb{H}, y) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}_\mathbb{H}, \mathrm{d}y) \\ \text{s.t.} & \dfrac{1}{N} \sum\limits_{i=1}^{N} \int_{\mathbb{H} \times \mathbb{Y}} d_\mathbb{H}\big((\boldsymbol{x}_\mathbb{H}, y), (\Phi(\widehat{\boldsymbol{x}}_i), \widehat{y}_i)\big) \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}_\mathbb{H}, \mathrm{d}y) \leq g(\sqrt{2}\rho) \\ & \int_{\mathbb{H} \times \mathbb{Y}} \mathbb{Q}^i(\mathrm{d}\boldsymbol{x}_\mathbb{H}, \mathrm{d}y) = 1 \qquad i \in [N] \end{cases}
$$

$$
= \sup_{\mathbb{Q} \in \mathbb{B}_{g(\sqrt{2}\rho)}(\widehat{\mathbb{P}}_N^\mathbb{H})} \mathbb{E}^{\mathbb{Q}}\left[\ell(\boldsymbol{x}_\mathbb{H}, y)\right],
$$

where the second inequality follows from relaxing the implicit condition that the random variable $\boldsymbol{x}_{\mathbb{H}}$ must be supported on $\{\Phi(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{X}\} \subseteq \mathbb{H}$. This completes the proof of assertion (i) for regression problems.

The proof of assertion (ii) parallels that of assertion (i) with obvious modifications. Due to the different transportation metric for classification problems, however, the estimate (47) changes to

$$
\begin{aligned}
g\left(d((\boldsymbol{x}, y), (\widehat{\boldsymbol{x}}_i, \widehat{y}_i))\right) &= g\left(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\|_2 + \kappa|y - \widehat{y}_i|\right) \\
&\geq g\left(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\|_2\right) + \kappa|y - \widehat{y}_i| \\
&\geq \|\Phi(\boldsymbol{x}) - \Phi(\widehat{\boldsymbol{x}}_i)\|_{\mathbb{H}} + \kappa|y - \widehat{y}_i| = d_{\mathbb{H}}\left((\Phi(\boldsymbol{x}), y), (\Phi(\widehat{\boldsymbol{x}}_i), \widehat{y}_i)\right),
\end{aligned}
$$

where the first inequality exploits the assumption that $g'(z) \geq 1$ for all $z \geq 0$, and the second inequality follows form (45). Further details are omitted for brevity. $\qquad\square$

Proof of Theorem 3.25. The proof follows immediately from [59, Theorem 4.2], which applies to loss functions representable as a sum of an empirical loss depending on $(\widehat{\boldsymbol{x}}_i, \widehat{y}_i, h(\widehat{\boldsymbol{x}}_i))$, $i \leq N$, and a regularization term that is strictly monotonically increasing in $\|h\|_{\mathbb{H}}$. However, the additive separability is not needed for the proof. We remark that the optimal solution of (31) is unique if f is striclty increasing in $\|h\|_{\mathbb{H}}$. If f is only non-decreasing in $\|h\|_{\mathbb{H}}$, on the other hand, uniqueness may be lost. Details are omitted for brevity. $\qquad\square$

Proof of Theorem 3.26. Using similar arguments as in the proof of Theorem 3.1(ii) and observing that any Hilbert norm is self-dual, one can show that

$$
\inf_{h \in \mathbb{H}} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N^{\mathbb{H}})} \mathbb{E}^{\mathbb{Q}}\left[L(\langle h, \boldsymbol{x}_{\mathbb{H}} \rangle_{\mathbb{H}} - y)\right] = \min_{h \in \mathbb{H}} \frac{1}{N} \sum_{i=1}^{N} L(h(\widehat{\boldsymbol{x}}_i) - \widehat{y}_i) + \rho \operatorname{lip}(L) \sqrt{\|h\|_{\mathbb{H}}^2 + 1}.
$$

By the representer theorem, which applies because the objective function of the above optimization problem is non-decreasing in $\|h\|_{\mathbb{H}}$, we may restrict the feasible set from $\mathbb{H}$ to the subset of all linearly parametrized hypotheses of the form $h(\boldsymbol{x}) = \sum_{j=1}^{N} \beta_j k(\boldsymbol{x}, \widehat{\boldsymbol{x}}_j)$ for some $\boldsymbol{\beta} \in \mathbb{R}^N$ without sacrificing optimality. The claim then follows by observing that $h(\widehat{\boldsymbol{x}}_i) = \sum_{j=1}^{N} \mathcal{K}_{ij} \beta_j$ and $\|h\|_2^2 = \langle \boldsymbol{\beta}, \mathcal{K}\boldsymbol{\beta} \rangle$. $\qquad\square$

Proof of Theorem 3.27. Using similar arguments as in the proof of Theorem 3.11(ii) and observing that any Hilbert norm is self-dual, one can show that

$$
\inf_{h \in \mathbb{H}} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N^{\mathbb{H}})} \mathbb{E}^{\mathbb{Q}}\left[L(y\langle h, \boldsymbol{x}_{\mathbb{H}} \rangle_{\mathbb{H}})\right] =
\begin{cases}
\min_{h,\lambda,s_i} & \lambda\rho + \dfrac{1}{N} \sum_{i=1}^{N} s_i \\[2mm]
\text{s.t.} & L(\widehat{y}_i h(\widehat{\boldsymbol{x}}_i)) \leq s_i & i \in [N] \\
& L(-\widehat{y}_i h(\widehat{\boldsymbol{x}}_i)) - \kappa\lambda \leq s_i & i \in [N] \\
& \operatorname{lip}(L)\|h\|_{\mathbb{H}} \leq \lambda,
\end{cases}
$$

see [34, Theorem 1] for a full proof. By the representer theorem, which applies because the loss function

$$
f((\widehat{\boldsymbol{x}}_1, \widehat{y}_1, h(\widehat{\boldsymbol{x}}_1)), \ldots, (\widehat{\boldsymbol{x}}_N, \widehat{y}_N, h(\widehat{\boldsymbol{x}}_N)), \|h\|_{\mathbb{H}}) =
\begin{cases}
\min_{\lambda} & \lambda\rho + \frac{1}{N} \sum_{i=1}^{N} \max\{L(\widehat{y}_i h(\widehat{\boldsymbol{x}}_i)), L(-\widehat{y}_i h(\widehat{\boldsymbol{x}}_i)) - \kappa\lambda\} \\
\text{s.t.} & \lambda \geq \operatorname{lip}(L)\|h\|_{\mathbb{H}}
\end{cases}
$$

is non-decreasing in $\|h\|_{\mathbb{H}}$, we may restrict attention to all linearly parametrized hypotheses of the form $h(\boldsymbol{x}) = \sum_{j=1}^{N} \beta_j k(\boldsymbol{x}, \widehat{\boldsymbol{x}}_j)$ for some $\boldsymbol{\beta} \in \mathbb{R}^N$ without sacrificing optimality. Thus, the claim follows. $\qquad\square$

Proof of Theorem 3.30. The worst-case expectation for regression problems satisfies

$$
\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\ell(h(\boldsymbol{x}; \boldsymbol{W}_{[M]}), y)\right]
$$

$$
= \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N} \sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x} \in \mathbb{X} \\ y \in \mathbb{Y}}} L(h(\boldsymbol{x}; \boldsymbol{W}_{[M]}) - y) - \lambda(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\| + \kappa|y - \widehat{y}_i|)
$$

$$\leq \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x}\in\mathbb{X}\\y\in\mathbb{Y}}} L(h(\widehat{\boldsymbol{x}}_i;\boldsymbol{W}_{[M]}) - \widehat{y}_i) + \mathrm{lip}(L)(|h(\boldsymbol{x};\boldsymbol{W}_{[M]}) - y - h(\widehat{\boldsymbol{x}}_i;\boldsymbol{W}_{[M]}) + \widehat{y}_i|)$$
$$- \lambda(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\| + \kappa|y - \widehat{y}_i|)$$
$$\leq \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x}\in\mathbb{X}\\y\in\mathbb{Y}}} L(h(\widehat{\boldsymbol{x}}_i;\boldsymbol{W}_{[M]}) - \widehat{y}_i) + \mathrm{lip}(L)\,\mathrm{lip}(h)\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\| + \mathrm{lip}(L)|y - \widehat{y}_i|$$
$$- \lambda(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\| + \kappa|y - \widehat{y}_i|),$$

where the equality holds by Lemma A.1, and the first and the second inequalities follow from the Lipschitz continuity of the loss function L and the hypothesis h. Note that composition of functions satisfies the inequality $\mathrm{lip}(f \circ g) \leq \mathrm{lip}(f)\,\mathrm{lip}(g)$; see for example [56, Exercise 9.8]. Thus, the Lipschitz modulus of the hypothesis h admits the upper bound

$$\mathrm{lip}(h) \leq \prod_{m=1}^{m} \mathrm{lip}(\sigma_m(\boldsymbol{W}_m\boldsymbol{z}_m)) \leq \prod_{m=1}^{m} \mathrm{lip}(\sigma_m)\,\|\boldsymbol{W}_m\|. \tag{48}$$

Replacing $\mathrm{lip}(h)$ by its upper bound in (48) and setting

$$\lambda = \mathrm{lip}(L)\max\left\{\prod_{k=1}^{M} \mathrm{lip}\big(\phi(\cdot;\boldsymbol{W}_m)\big), \frac{1}{\kappa}\right\},$$

will complete the proof for regression problems. In a similar way, the worst-case expectation for classification problems can be handled as follows:

$$\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\big[\ell(h(\boldsymbol{x};\boldsymbol{W}_{[M]}), y)\big]$$

$$= \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x}\in\mathbb{X}\\y\in\mathbb{Y}}} L(yh(\boldsymbol{x};\boldsymbol{W}_{[M]})) - \lambda(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\| + \kappa\mathbb{1}_{\{y\neq\widehat{y}_i\}})$$

$$\leq \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x}\in\mathbb{X}\\y\in\mathbb{Y}}} L(\widehat{y}_i h(\widehat{\boldsymbol{x}}_i;\boldsymbol{W}_{[M]})) - \lambda(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\| + \kappa\mathbb{1}_{\{y\neq\widehat{y}_i\}}) + \mathrm{lip}(L)(|yh(\boldsymbol{x}) - \widehat{y}_i h(\widehat{\boldsymbol{x}}_i)|)$$

$$\leq \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x}\in\mathbb{X}\\y\in\mathbb{Y}}} L(\widehat{y}_i h(\widehat{\boldsymbol{x}}_i;\boldsymbol{W}_{[M]})) - \lambda(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\| + \kappa\mathbb{1}_{\{y\neq\widehat{y}_i\}}) + \mathrm{lip}(L)\,\mathrm{lip}(h)\,\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\|\mathbb{1}_{\{y=\widehat{y}_i\}}$$

$$+ \mathrm{lip}(L)\,|h(\boldsymbol{x}) + h(\widehat{x}_i)|\,\mathbb{1}_{\{y\neq\widehat{y}_i\}}$$

$$\leq \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x}\in\mathbb{X}\\y\in\mathbb{Y}}} L(\widehat{y}_i h(\widehat{\boldsymbol{x}}_i;\boldsymbol{W}_{[M]})) - \lambda(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\| + \kappa\mathbb{1}_{\{y\neq\widehat{y}_i\}})$$

$$+ \mathrm{lip}(L)\,\max\{\frac{2\sup_{\boldsymbol{x}\in\mathbb{X},h\in\mathbb{H}}|h(\boldsymbol{x})|}{\kappa}, \mathrm{lip}(h)\}\,(\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_i\| + \kappa\mathbb{1}_{\{y\neq\widehat{y}_i\}}).$$

Using the upper bound in (48) and setting

$$\lambda = \mathrm{lip}(L)\max\left\{\prod_{k=1}^{M} \mathrm{lip}\big(\phi(\cdot;\boldsymbol{W}_m)\big), \frac{\max\{1, 2\sup_{\boldsymbol{x}\in\mathbb{X},h\in\mathbb{H}}|h(\boldsymbol{x})|\}}{\kappa}\right\}$$

will completes the proof for classification problems. $\qquad\square$

Proof of Corollary 3.32. Let $\bar{\sigma} = \max_{m\in[M]} \mathrm{lip}(\sigma_m)$. By Theorem 3.30 and setting $\kappa = \infty$, the worst-case expectation for both classification and regression problems can be upper bounded by

$$\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}\left[\ell(h(\boldsymbol{x};\boldsymbol{W}_{1,\dots,M}), y)\right] \leq \frac{1}{N}\sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x}\in\mathbb{X}\\y\in\mathbb{Y}}} \ell(h(\widehat{\boldsymbol{x}}_i;\boldsymbol{W}_{1,\dots,M}), \widehat{y}_i) + \rho\,\bar{\sigma}\,\mathrm{lip}(\ell)\prod_{k=1}^{M}\|\boldsymbol{W}_k\|$$

$$\leq \frac{1}{N} \sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x} \in \mathbb{X} \\ y \in \mathbb{Y}}} \ell(h(\widehat{\boldsymbol{x}}_i; \boldsymbol{W}_{1,\ldots,M}), \widehat{y}_i) + \rho \, \bar{\sigma} \operatorname{lip}(\ell) \left(\sum_{k=1}^{M} \frac{\|\boldsymbol{W}_k\|}{M} \right)^M,$$

where the last inequality follows from the arithmetic-geometric means inequality. By [29, Theorem 1], if $\boldsymbol{W}_{[M]}^\star$ is a minimizer of the optimization problem

$$\min_{\boldsymbol{W}_{[M]}} \; \frac{1}{N} \sum_{i=1}^{N} \sup_{\substack{\boldsymbol{x} \in \mathbb{X} \\ y \in \mathbb{Y}}} \ell(h(\widehat{\boldsymbol{x}}_i; \boldsymbol{W}_{1,\ldots,M}), \widehat{y}_i) + \rho \, \bar{\sigma} \operatorname{lip}(\ell) \left(\sum_{k=1}^{M} \frac{\|\boldsymbol{W}_k\|}{M} \right)^M,$$

then the same $\boldsymbol{W}_{[M]}^\star$ also minimizes the constrained optimization problem

$$\begin{cases} \displaystyle \inf_{\boldsymbol{W}_{[M]}} & \displaystyle \frac{1}{N} \sum_{i=1}^{N} \ell(h(\widehat{\boldsymbol{x}}_i; \boldsymbol{W}_{[M]}), \widehat{y}_i) \\[2mm] \text{s.t.} & \displaystyle \left(\sum_{k=1}^{M} \frac{\|\boldsymbol{W}_k\|}{M} \right)^M \leq \theta^M, \end{cases}$$

for $\theta = \sum_{i=1}^{M} \|\boldsymbol{W}_k^\star\|/M$. Notice that the constraint in the above optimization problem can be simplified to $\sum_{i=1}^{M} \|\boldsymbol{W}_k^\star\| \leq M\theta$. Hence, there exists a Lagrange multiplier $\overline{\rho}$ for the simplified constraint $\sum_{i=1}^{M} \|\boldsymbol{W}_k^\star\| \leq M\theta$ such that $\boldsymbol{W}_{[M]}^\star$ is an optimizer of the penalized problem

$$\inf_{\boldsymbol{W}_{[M]}} \frac{1}{N} \sum_{i=1}^{N} \ell(h(\widehat{\boldsymbol{x}}_i; \boldsymbol{W}_{[M]}), \widehat{y}_i) + \overline{\rho} \sum_{k=1}^{M} \|\boldsymbol{W}_k\|.$$

This completes the proof. $\qquad\square$

A.2 Proofs of Section 4

The proof of Theorem 4.6 relies on the following preparatory lemma, which basically asserts that the sample average of a linearly growing function of $\boldsymbol{\xi}$ has sub-Gaussian tails.

Lemma A.5 (Sub-Gaussian tails). If Assumption 4.1 holds, then there exist constants $c_3 \geq 1$ and $c_4 > 0$ that depend only on the light tail constants a and A of $\mathbb{P}$ such that

$$\mathbb{P}^N \Big\{ \big| \mathbb{E}^{\mathbb{P}}[f(\boldsymbol{\xi})] - \mathbb{E}^{\widehat{\mathbb{P}}_N}[f(\boldsymbol{\xi})] \big| \geq \delta \Big\} \leq c_3 \exp(-c_4 N \delta^2)$$

for any $N \in \mathbb{N}$, $\delta \in [0,1]$ and function $f : \Xi \to \mathbb{R}$ with $|f(\boldsymbol{\xi}) - f(\boldsymbol{\xi}')| \leq d(\boldsymbol{\xi}, \boldsymbol{\xi}')$ for all $\boldsymbol{\xi} \in \Xi$ and some reference point $\boldsymbol{\xi}' \in \Xi$.

Proof. Assume that $f : \Xi \to \mathbb{R}$ is a linear growth function with $|f(\boldsymbol{\xi}) - f(\boldsymbol{\xi}')| \leq d(\boldsymbol{\xi}, \boldsymbol{\xi}')$ for all $\boldsymbol{\xi} \in \Xi$ and some reference point $\boldsymbol{\xi}' \in \Xi$. Set $\xi_f = f(\boldsymbol{\xi})$ and $\xi_f' = f(\boldsymbol{\xi}')$. Thus, the distribution of the scalar random variable ξ_f is given by the pushforward measure $f_*(\mathbb{P})$ of $\mathbb{P}$. By construction, we have

$$\mathbb{E}^{f_*(\mathbb{P})} \big[\exp \big(|\xi_f - \xi_f'|^a \big) \big] = \mathbb{E}^{\mathbb{P}} \big[\exp \big(|f(\boldsymbol{\xi}) - f(\boldsymbol{\xi}')|^a \big) \big] \leq \mathbb{E}^{\mathbb{P}} \big[\exp \big(d(\boldsymbol{\xi}, \boldsymbol{\xi}')^a \big) \big] \leq A,$$

where the first inequality follows form the growth condition of f, while the second inequality holds because $\mathbb{P}$ satisfies Assumption 4.1. Hence, the distribution $f_*(\mathbb{P})$ satisfies Assumption 4.1 for $n = 0$ when distances on $\mathbb{R}$ are measured by the absolute value, and it inherits the light-tail constants a and A from $\mathbb{P}$. By using Theorem 4.2 for $n = 0$, we may thus conclude that there exist constants $c_3, c_4 > 0$ with

$$\mathbb{P}^N \Big\{ W(f_*(\mathbb{P}), f_*(\widehat{\mathbb{P}}_N)) \geq \delta \Big\} \leq c_3 \exp \big(-c_4 N \delta^2 \big) \quad \forall \delta \in [0,1],$$

where $f_*(\widehat{\mathbb{P}}_N)$ represents the empirical distribution of ξ_f, which coincides with the pushforward measure of $\widehat{\mathbb{P}}_N$ under f. By slight abuse of notation, W stands here for the Wasserstein distance between *univariate* distributions, where the absolute value is used as the ground metric. Note that the above univariate measure concentration result holds for any linear growth function f with asymptotic growth rate ≤ 1. We emphasize that $c_3 \geq 1$ because otherwise the above estimate would fail to hold for $\delta = 0$.

By construction of the Wasserstein distance in Definition 2.2, we have $W(f_*(\mathbb{P}), f_*(\widehat{\mathbb{P}}_N)) < \delta$ if and only if the scalar random variables ξ_f and ξ'_f admit a joint distribution Π with $\mathbb{E}^\Pi[|\xi_f - \xi'_f|] < \delta$ under which ξ_f and ξ'_f have marginals $f_*(\mathbb{P})$ and $f_*(\widehat{\mathbb{P}}_N)$, respectively. The inequality $W(f_*(\mathbb{P}), f_*(\widehat{\mathbb{P}}_N)) < \delta$ thus implies

$$\left| \mathbb{E}^\mathbb{P}[f(\boldsymbol{\xi})] - \mathbb{E}^{\widehat{\mathbb{P}}_N}[f(\boldsymbol{\xi}')] \right| = \left| \mathbb{E}^{f_*(\mathbb{P})}[\xi_f] - \mathbb{E}^{f_*(\widehat{\mathbb{P}}_N)}[\xi'_f] \right| \leq \left| \mathbb{E}^\Pi[\xi_f - \xi'_f] \right| = \mathbb{E}^\Pi\left[|\xi_f - \xi'_f| \right] < \delta.$$

By contraposition, we then obtain the implication

$$\left| \mathbb{E}^\mathbb{P}[f(\boldsymbol{\xi})] - \mathbb{E}^{\widehat{\mathbb{P}}_N}[f(\boldsymbol{\xi}')] \right| \geq \delta \qquad \Longrightarrow \qquad W(f_*(\mathbb{P}), f_*(\widehat{\mathbb{P}}_N)) \geq \delta,$$

which leads to the desired inequality

$$\mathbb{P}^N\left\{ |\mathbb{E}^\mathbb{P}[f(\boldsymbol{\xi})] - \mathbb{E}^{\widehat{\mathbb{P}}_N}[f(\boldsymbol{\xi})]| \geq \delta \right\} \leq \mathbb{P}^N\left\{ W(f_*(\mathbb{P}), f_*(\widehat{\mathbb{P}}_N)) \geq \delta \right\} \leq c_3 \exp\left(-c_4 N \delta^2 \right).$$

This inequality holds for all $\delta \in [0, 1]$ and $N \in \mathbb{N}$, irrespective of the linear growth function f. $\qquad\square$

Remark A.6 (Hoeffding's inequality). If it is known that $\mathbb{P}\{\underline{f} \leq f(\boldsymbol{\xi}) \leq \overline{f}\} = 1$, then Lemma A.5 reduces to Hoeffding's inequality [15, Theorem 2.8], in which case we may set $c_3 = 2$ and $c_4 = 2/(\overline{f} - \underline{f})^2$.

Proof of Theorem 4.6. To avoid cumbersome case distinctions, we prove the theorem only in the case when (4) is a classification problem. Thus, we assume that $\Xi = \mathbb{R}^n \times \{-1, 1\}$ and that the transportation cost is of the form (16), where $\|\cdot\|$ denotes a norm on the input space $\mathbb{R}^n$. The proof for regression problems is similar and only requires minor modifications. It will be omitted for brevity.

From Theorem 3.11(ii) we know that

$$\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^\mathbb{Q}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) \right] = \mathbb{E}^{\widehat{\mathbb{P}}_N}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) \right] + \rho \operatorname{lip}(L)\Omega(\boldsymbol{w}) \quad \forall \boldsymbol{w} \in \mathbb{W},$$

where $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|_*$ can be viewed as a regularization function. For every $\rho \geq \rho'_N(\eta)$ we thus have

$$\begin{aligned}
&\mathbb{P}^N\left\{ \mathbb{E}^\mathbb{P}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) \right] \leq \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}^\mathbb{Q}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) \right] \; \forall \boldsymbol{w} \in \mathbb{W} \right\} \\
&\geq \mathbb{P}^N\left\{ 0 \leq \min_{\boldsymbol{w} \in \mathbb{W}} \mathbb{E}^{\widehat{\mathbb{P}}_N}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) \right] + \rho'_N(\eta) \operatorname{lip}(L)\Omega(\boldsymbol{w}) - \mathbb{E}^\mathbb{P}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) \right] \right\} \\
&= 1 - \mathbb{P}^N\left\{ \min_{\boldsymbol{w} \in \mathbb{W}} \mathbb{E}^{\widehat{\mathbb{P}}_N}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) \right] + \rho'_N(\eta) \operatorname{lip}(L)\Omega(\boldsymbol{w}) - \mathbb{E}^\mathbb{P}\left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) \right] < 0 \right\}. \qquad (49)
\end{aligned}$$

Observe that $\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)$ is Lipschitz continuous in $\boldsymbol{w}$ for every fixed $\boldsymbol{x}$ and y because the underlying univariate loss function L is Lipschitz continuous by assumption. Specifically, we have

$$|\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y) - \ell(\langle \boldsymbol{w}', \boldsymbol{x} \rangle, y)| \leq \operatorname{lip}(L)|\langle \boldsymbol{w} - \boldsymbol{w}', \boldsymbol{x} \rangle| \leq \operatorname{lip}(L)\|\boldsymbol{w} - \boldsymbol{w}'\|_\infty \|\boldsymbol{x}\|_1 \qquad \forall \boldsymbol{w}, \boldsymbol{w}' \in \mathbb{W}.$$

For any $\Delta > 0$ there exists a finite set $\mathbb{W}_\Delta \subseteq \mathbb{W}$ with $\Delta = \sup_{\boldsymbol{w} \in \mathbb{W}} \inf_{\boldsymbol{w}' \in \mathbb{W}_\Delta} \|\boldsymbol{w} - \boldsymbol{w}'\|_\infty$ and whose cardinality satisfies $|\mathbb{W}_\Delta| < (\overline{\Omega}/\Delta - 1)^n < (\overline{\Omega}/\Delta)^n - 1$ where the second inequality holds because $\Delta \leq \overline{\Omega}$ by construction. In the following we set $\Delta = \overline{\Omega}/\sqrt{N}$.

As $\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)$ is Lipschitz continuous in $\boldsymbol{w}$, for every $\boldsymbol{w} \in \mathbb{W}$ there is $\boldsymbol{w}' \in \mathbb{W}_\Delta$ with

$$|\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)| \geq |\ell(\langle \boldsymbol{w}', \boldsymbol{x} \rangle, y)| - \operatorname{lip}(L)\Delta\|\boldsymbol{x}\|_1.$$

Applying this estimate twice and recalling the assumption that $\Omega(\boldsymbol{w}) \geq \underline{\Omega}$ for all $\boldsymbol{w} \in \mathbb{W}$, we may thus conclude that the probability in (49) is smaller or equal to

$$\mathbb{P}^N \left\{ \min_{\boldsymbol{w} \in \mathbb{W}} \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)\right] - \mathbb{E}^{\mathbb{P}} \left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)\right] + \rho'_N(\eta) \operatorname{lip}(L) \underline{\Omega} < 0 \right\}$$

$$\leq \mathbb{P}^N \left\{ \min_{\boldsymbol{w} \in \mathbb{W}_\Delta} \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)\right] - \mathbb{E}^{\mathbb{P}} \left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)\right] - \operatorname{lip}(L) \Delta \left(\mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\|\boldsymbol{x}\|_1\right] + \mathbb{E}^{\mathbb{P}} \left[\|\boldsymbol{x}\|_1\right] \right) + \rho'_N(\eta) \operatorname{lip}(L) \underline{\Omega} < 0 \right\}$$

$$\leq \mathbb{P}^N \left\{ \min_{\boldsymbol{w} \in \mathbb{W}_\Delta} \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)\right] - \mathbb{E}^{\mathbb{P}} \left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)\right] - \operatorname{lip}(L) \Delta \left(\mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\|\boldsymbol{x}\|_1\right] - \mathbb{E}^{\mathbb{P}} \left[\|\boldsymbol{x}\|_1\right] \right) \right.$$

$$\left. < 2 \operatorname{lip}(L) \Delta M_n n A - \rho'_N(\eta) \operatorname{lip}(L) \underline{\Omega} \right\}. \tag{50}$$

The second inequality in the above expression follows from the estimate

$$M_n = \max_{i \leq n} \|\boldsymbol{e}_i^n\|_* = \max_{i \leq n} \sup_{\|\boldsymbol{x}\| \leq 1} \langle \boldsymbol{e}_i^n, \boldsymbol{x} \rangle = \max_{i \leq n} \sup_{\|\boldsymbol{x}\| \leq 1} |x_i| = \sup_{\|\boldsymbol{x}\| \leq 1} \|\boldsymbol{x}\|_\infty \geq \sup_{\boldsymbol{x} \in \mathbb{R}^n} \frac{\|\boldsymbol{x}\|_1}{n\|\boldsymbol{x}\|},$$

which can be paraphrased as $\|\boldsymbol{x}\|_1 \leq M_n n \|\boldsymbol{x}\|$ for every $\boldsymbol{x} \in \mathbb{R}^n$ and thus implies

$$\mathbb{E}^{\mathbb{P}}[\|\boldsymbol{x}\|_1] \leq M_n n \mathbb{E}^{\mathbb{P}}[\|\boldsymbol{x}\|] \leq M_n n \mathbb{E}^{\mathbb{P}}[\exp(\|\boldsymbol{x}\|^a)] \leq M_n n \mathbb{E}^{\mathbb{P}}[\exp\left(d(\boldsymbol{\xi}, \boldsymbol{\xi}')\right)^a)] \leq M_n n A.$$

Next, we introduce an auxiliary parameter

$$\delta = \frac{\rho'_N(\eta) \underline{\Omega} - M_n A}{\Delta + \overline{\Omega}} = \frac{2\sqrt{n \log(\sqrt{N})/c_4 + \log(c_3/\eta)/c_4}}{1 + \sqrt{N}},$$

where the second equality follows from the definition of $\rho'_N(\eta)$ in the theorem statement and the convention that $\Delta = \overline{\Omega}/\sqrt{N}$. One can prove that $\delta \in [0, 1]$. Indeed, the nonnegativity of δ is immediate because $c_3 \geq 1$ and $c_4 > 0$. Moreover, we find

$$\delta \leq \frac{2\sqrt{n \log(\sqrt{N})/c_4}}{1 + \sqrt{N}} + \frac{2\sqrt{\log(c_3/\eta)/c_4}}{1 + \sqrt{N}} \leq 1,$$

where the first inequality follows from the observation that $\sqrt{x_1 + x_2} \leq \sqrt{x_1} + \sqrt{x_2}$ for all $x_1, x_2 \geq 0$, while the second inequality holds because $\log(\sqrt{N}) \leq \sqrt{N}$ for all $N \in \mathbb{N}$ and $N \geq \max\{(16n/c_4)^2, 16\log(c_3/\eta)/c_4\}$, which implies that both fractions in the middle of the above expression are smaller ore qual to $\frac{1}{2}$.

Multiplying the definition of δ with $-\operatorname{lip}(L)(\Delta + \overline{\Omega})$ yields the identity

$$-\operatorname{lip}(L) \Delta \delta - \operatorname{lip}(L) \overline{\Omega} \delta = 2 \operatorname{lip}(L) \Delta M_n n A - \operatorname{lip}(L) \rho_N(\eta) \underline{\Omega},$$

and thus the probability (50) can be bounded above by

$$\mathbb{P}^N \left\{ \min_{\boldsymbol{w} \in \mathbb{W}_\Delta} \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)\right] - \mathbb{E}^{\mathbb{P}} \left[\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)\right] \leq -\operatorname{lip}(L) \overline{\Omega} \delta \quad \text{or} \right.$$

$$\left. -\operatorname{lip}(L) \Delta \left(\mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\|\boldsymbol{x}\|\right] - \mathbb{E}^{\mathbb{P}} \left[\|\boldsymbol{x}\|\right] \right) \leq -\operatorname{lip}(L) \Delta \delta \right\}$$

$$\leq \mathbb{P}^N \left\{ \min_{\boldsymbol{w} \in \mathbb{W}_\Delta} \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\frac{\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)}{\operatorname{lip}(L) \overline{\Omega}} \right] - \mathbb{E}^{\mathbb{P}} \left[\frac{\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)}{\operatorname{lip}(L) \overline{\Omega}} \right] \leq -\delta \right\} \tag{51a}$$

$$+ \mathbb{P}^N \left\{ \left(\mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\|\boldsymbol{x}\|\right] - \mathbb{E}^{\mathbb{P}} \left[\|\boldsymbol{x}\|\right] \right) \geq \delta \right\}, \tag{51b}$$

where the inequality follows from the subadditivity of probability measures.

For any fixed $\boldsymbol{w} \in \mathbb{W}$ one can show that the function $f(\boldsymbol{\xi}) = \ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)/(\operatorname{lip}(L) \overline{\Omega})$ with $\boldsymbol{\xi} = (\boldsymbol{x}, y)$ satisfies the linear growth condition $|f(\boldsymbol{\xi}) - f(\boldsymbol{\xi}')| \leq d(\boldsymbol{\xi}, \boldsymbol{\xi}')$ for all $\boldsymbol{\xi} \in \Xi$ if $\boldsymbol{\xi}' = (\boldsymbol{0}, 1)$. Details are omitted for brevity. By the subadditivity of probability measures, the probability (51a) is thus smaller or equal to

$$\sum_{\boldsymbol{w} \in \mathbb{W}_\Delta} \mathbb{P}^N \left\{ \left| \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\frac{\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)}{\operatorname{lip}(L) \overline{\Omega}} \right] - \mathbb{E}^{\mathbb{P}} \left[\frac{\ell(\langle \boldsymbol{w}, \boldsymbol{x} \rangle, y)}{\operatorname{lip}(L) \overline{\Omega}} \right] \right| > \delta \right\} \leq |\mathbb{W}_\Delta| c_3 \exp(-c_4 N \delta^2),$$

where the inequality follows from Lemma A.5, which applies because $\delta \in [0,1]$. Moreover, the function $f(\boldsymbol{\xi}) = \|\boldsymbol{x}\|$ with $\boldsymbol{\xi} = (\boldsymbol{x}, y)$ trivially satisfies the linear growth condition $|f(\boldsymbol{\xi}) - f(\boldsymbol{\xi}')| \le d(\boldsymbol{\xi}, \boldsymbol{\xi}')$ for all $\boldsymbol{\xi} \in \Xi$ if $\boldsymbol{\xi}' = (\boldsymbol{0}, 1)$. By Lemma A.5, the probability (51b) is thus smaller or equal to

$$\mathbb{P}^N \left\{ \left| \mathbb{E}^{\mathbb{P}} \left[\|\boldsymbol{x}\| \right] - \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\|\boldsymbol{x}\| \right] \right| > \delta \right\} \le c_3 \exp(-c_4 N \delta^2).$$

By combining the above estimates, we may conclude that the probability in (49) does not exceed

$$
\begin{aligned}
(|\mathbb{W}_\Delta| + 1)\, c_3 \exp\left(-c_4 N \delta^2\right) &\le (\overline{\Omega}/\Delta)^n\, c_3 \exp\left(-c_4 N \left(\frac{\rho_N'(\eta)\Omega - 2\Delta M_n n A}{\Delta + \overline{\Omega}}\right)^2\right) \\
&= N^{\frac{n}{2}} c_3 \exp\left(-c_4 N \left(\frac{\rho_N'(\eta)\Omega\sqrt{N} - 2\overline{\Omega} M_n n A}{\overline{\Omega}(\sqrt{N} + 1)}\right)^2\right) \\
&\le N^{\frac{n}{2}} c_3 \exp\left(-c_4 N \left(\frac{2\overline{\Omega}\sqrt{n \log(\sqrt{N})/c_4 + \log(c_3/\eta)/c_4}}{2\overline{\Omega}\sqrt{N}}\right)^2\right) \\
&= N^{\frac{n}{2}} c_3 \exp\left(-n \log(\sqrt{N}) - \log(c_3/\eta)\right) = \eta,
\end{aligned}
$$

where the first inequality follows from the definition of δ and the assumption that $|\mathbb{W}_\Delta| < (\overline{\Omega}/\Delta)^n - 1$, the first equality holds because $\overline{\Omega}/\Delta = \sqrt{N}$, and the second inequality holds due to the definition of $\rho_N'(\eta)$. In summary, we have shown that the probability in (49) is at most η, and thus the claim follows. $\qquad\square$

A.3 Proofs of Section 5

Proof of Theorem 5.1. As for assertion (i), note that the absolute value function coincides with the ϵ-insensitive loss for $\epsilon = 0$. Thus, (41a) follows immediately from Corollary 3.3 by fixing $\boldsymbol{w}$ and by setting $\epsilon = 0$ and $\Xi = \mathbb{R}^{n+1}$. As for assertion (ii), similar arguments as in the proof of Lemma A.1 show that

$$\mathcal{E}_{\min}(\boldsymbol{w}) = \sup_{\lambda \ge 0}\; -\lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \inf_{\boldsymbol{x}, y}\; |y - \langle \boldsymbol{w}, \boldsymbol{x}\rangle| + \lambda\|(\boldsymbol{x}, y) - (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)\|. \tag{52}$$

The subordinate minimization problem in the first constraint of (52) is equivalent to

$$
\begin{aligned}
\inf_{\boldsymbol{x}, y}\; |y - \langle \boldsymbol{w}, \boldsymbol{x}\rangle| + \lambda\|(\boldsymbol{x}, y) - (\widehat{\boldsymbol{x}}_i, \widehat{y}_i)\| &= \inf_{\boldsymbol{x}, y}\; \sup_{\|(\boldsymbol{q}_i, v_i)\|_* \le \lambda}\; |y - \langle \boldsymbol{w}, \boldsymbol{x}\rangle| + \langle \boldsymbol{q}_i, \boldsymbol{x} - \widehat{\boldsymbol{x}}_i\rangle + v_i(y_i - \widehat{y}_i) \\
&= \sup_{\|(\boldsymbol{q}_i, v_i)\|_* \le \lambda} \left\{ \begin{aligned} &\inf_{\boldsymbol{x}, y, z}\; z + \langle \boldsymbol{q}_i, \boldsymbol{x} - \widehat{\boldsymbol{x}}_i\rangle + v_i(y_i - \widehat{y}_i) \\ &\text{s.t.}\quad z \ge y - \langle \boldsymbol{w}, \boldsymbol{x}\rangle,\; z \ge \langle \boldsymbol{w}, \boldsymbol{x}\rangle - y \end{aligned} \right. \\
&= \left\{ \begin{aligned} &\sup_{\boldsymbol{q}_i, v_i, r_i, t_i}\; -\langle \boldsymbol{q}_i, \widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i v_i \\ &\text{s.t.}\quad t_i + r_i = 1,\; t_i - r_i = v_i \\ &\qquad\quad (r_i - t_i)\boldsymbol{w} = \boldsymbol{q}_i,\; \|(\boldsymbol{q}_i, v_i)\|_* \le \lambda \\ &\qquad\quad r_i, t_i \ge 0 \end{aligned} \right. \\
&= \left\{ \begin{aligned} &\sup_{v_i}\; v_i(\langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i\rangle - \widehat{y}_i) \\ &\text{s.t.}\quad -1 \le v_i \le 1 \\ &\qquad\quad v_i\|(\boldsymbol{w}, -1)\|_* \le \lambda,\; -v_i\|(\boldsymbol{w}, -1)\|_* \le \lambda, \end{aligned} \right.
\end{aligned}
$$

where the second equality holds due to Proposition 5.5.4 in [9], and the third equality holds due to strong linear programming duality. By substituting the last optimization problem into (52) and replacing v_i with $-v_i$, we have

$$
\mathcal{E}_{\min}(\boldsymbol{w}) = \left\{ \begin{aligned} &\sup_{v_i, \lambda}\; -\lambda\rho + \frac{1}{N}\sum_{i=1}^{N} v_i(\widehat{y}_i - \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i\rangle) \\ &\text{s.t.}\quad -1 \le v_i \le 1 && i \in [N] \\ &\qquad\quad v_i\|(\boldsymbol{w}, -1)\|_* \le \lambda && i \in [N] \\ &\qquad\quad -v_i\|(\boldsymbol{w}, -1)\|_* \le \lambda && i \in [N] \end{aligned} \right.
$$

$$
= \begin{cases}
\sup_{v_i, \lambda} & -\lambda\rho + \dfrac{1}{N}\sum_{i=1}^{N} v_i|\widehat{y}_i - \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle| \\
\text{s.t.} & 0 \le v_i \le 1 \qquad\qquad\qquad\quad i \in [N] \\
& v_i\|(\boldsymbol{w}, -1)\|_* \le \lambda \qquad\qquad\; i \in [N]
\end{cases}
$$

$$
= \begin{cases}
\sup_{v, \lambda} & -\lambda\rho + \dfrac{1}{N}\sum_{i=1}^{N} v|\widehat{y}_i - \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle| \\
\text{s.t.} & 0 \le v \le 1 \\
& v\|(\boldsymbol{w}, -1)\|_* \le \lambda
\end{cases}
$$

$$
= \begin{cases}
\sup_{v} & v\left(\dfrac{1}{N}\sum_{i=1}^{N} |\widehat{y}_i - \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle| - \rho\|(\boldsymbol{w}, -1)\|_* \right) \\
\text{s.t.} & 0 \le v \le 1
\end{cases}
$$

$$
= \max\left\{ \dfrac{1}{N}\sum_{i=1}^{N} |\widehat{y}_i - \langle \boldsymbol{w}, \widehat{\boldsymbol{x}}_i \rangle| - \rho\|(\boldsymbol{w}, -1)\|_*, 0 \right\}
$$

$\square$

Proof of Theorem 5.2. As for assertion (i), similar arguments as in the proof of Lemma A.1 show that

$$
\mathcal{R}_{\max}(\boldsymbol{w}) = \begin{cases}
\inf_{\lambda, s_i} & \lambda\rho + \dfrac{1}{N}\sum_{i=1}^{N} s_i \\
\text{s.t.} & \sup_{\boldsymbol{x}} \mathbb{1}_{\{\widehat{y}_i \langle \boldsymbol{w}, \boldsymbol{x}\rangle \le 0\}} - \lambda\|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}\| \le s_i & \forall i \le N \\
& \sup_{\boldsymbol{x}} \mathbb{1}_{\{-\widehat{y}_i \langle \boldsymbol{w}, \boldsymbol{x}\rangle \le 0\}} - \lambda\|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}\| - \kappa\lambda \le s_i & \forall i \le N \\
& \lambda \ge 0.
\end{cases}
\tag{53a}
$$

Next, observe that the indicator functions in (53a) can be represented as pointwise maxima of extended real-valued concave functions of the form $\mathbb{1}_{\{\widehat{y}_i \langle \boldsymbol{w}, \boldsymbol{x}\rangle \le 0\}} = \max\{I_1(\boldsymbol{x}), 0\}$ and $\mathbb{1}_{\{\widehat{y}_i \langle \boldsymbol{w}, \boldsymbol{x}\rangle \ge 0\}} = \max\{I_2(\boldsymbol{x}), 0\}$, respectively, where

$$
I_1(\boldsymbol{x}) = \begin{cases} 1 & \widehat{y}_i \langle \boldsymbol{w}, \boldsymbol{x}\rangle \le 0, \\ -\infty & \text{otherwise,} \end{cases} \quad \text{and} \quad I_2(\boldsymbol{x}) = \begin{cases} 1 & \widehat{y}_i \langle \boldsymbol{w}, \boldsymbol{x}\rangle \ge 0, \\ -\infty & \text{otherwise.} \end{cases}
$$

This allows us to reformulate (53a) as

$$
\begin{cases}
\inf_{\lambda, s_i} & \lambda\rho + \dfrac{1}{N}\sum_{i=1}^{N} s_i \\
\text{s.t.} & \sup_{\boldsymbol{x}\in\mathbb{R}^n} I_1(\boldsymbol{x}) - \lambda\|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}\| \le s_i & \forall i \le N \\
& \sup_{\boldsymbol{x}} 0 - \lambda\|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}\| \le s_i & \forall i \le N \\
& \sup_{\boldsymbol{x}} I_2(\boldsymbol{x}) - \lambda\|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}\| - \kappa\lambda \le s_i & i \in [N] \\
& \sup_{\boldsymbol{x}} 0 - \lambda\|\widehat{\boldsymbol{x}}_i - \boldsymbol{x}\| - \kappa\lambda \le s_i & \forall i \le N \\
& \lambda \ge 0.
\end{cases}
$$

Using the definition of the dual norm and applying the duality theorem [9, Proposition 5.5.4], we find

$$
\mathcal{R}_{\max}(\boldsymbol{w}) = \begin{cases}
\inf_{\lambda, s_i, \boldsymbol{p}_i, \boldsymbol{q}_i} & \lambda\rho + \dfrac{1}{N}\sum_{i=1}^{N} s_i \\
\text{s.t.} & \sup_{\boldsymbol{x}} I_1(\boldsymbol{x}) + \langle \boldsymbol{p}_i, \boldsymbol{x}\rangle - \langle \boldsymbol{p}_i, \widehat{\boldsymbol{x}}_i \rangle \le s_i & \forall i \le N \\
& \sup_{\boldsymbol{x}} I_2(\boldsymbol{x}) + \langle \boldsymbol{q}_i, \boldsymbol{x}\rangle - \langle \boldsymbol{q}_i, \widehat{\boldsymbol{x}}_i \rangle - \kappa\lambda \le s_i & \forall i \le N \\
& s_i \ge 0, \|\boldsymbol{p}_i\|_* \le \lambda, \|\boldsymbol{q}_i\|_* \le \lambda & i \in [N].
\end{cases}
\tag{53b}
$$

Moreover, by strong linear programming duality we have

$$\sup_{\boldsymbol{x}} \ I_1(\boldsymbol{x}) + \langle \boldsymbol{p}_i, \boldsymbol{x} \rangle = \sup_{\boldsymbol{x}} \{ 1 + \langle \boldsymbol{p}_i, \boldsymbol{x} \rangle : \widehat{y}_i \langle \boldsymbol{w}, \boldsymbol{x} \rangle \leq 0 \} = \inf_{r_i \geq 0} \{ 1 : \widehat{y}_i r_i \boldsymbol{w} = \boldsymbol{p}_i \} \tag{53c}$$

and

$$\sup_{\boldsymbol{x}} \ I_2(\boldsymbol{x}) + \langle \boldsymbol{q}_i, \boldsymbol{x} \rangle = \sup_{\boldsymbol{x}} \{ 1 + \langle \boldsymbol{q}_i, \boldsymbol{x} \rangle : \widehat{y}_i \langle \boldsymbol{w}, \boldsymbol{x} \rangle \geq 0 \} = \inf_{t_i \geq 0} \{ 1 : \widehat{y}_i t_i \boldsymbol{w} = \boldsymbol{q}_i \} . \tag{53d}$$

Substituting (53c) and (53d) into (53b) yields (42a). The expression (42b) for the best-case risk can be proved in a similar fashion. Details are omitted for brevity. $\qquad\square$

References

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning from Data*. AMLBook, 2012.

[2] N. Agarwal, B. Bullins, and E. Hazan. Second-order stochastic optimization for machine learning in linear time. *Journal of Machine Learning Research*, 18:4148–4187, 2017.

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[4] K. Bache and M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml. Available from http://archive.ics.uci.edu/ml.

[5] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

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

[7] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[8] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[9] D. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[10] D. Bertsimas and M. S. Copenhaver. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 2017.

[11] C. Bhattacharyya. Second order cone programming formulations for feature selection. *Journal of Machine Learning Research*, 5:1417–1433, 2004.

[12] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 2019.

[13] J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

[14] J. Blanchet, Y. Kang, F. Zhang, and K. Murthy. Data-driven optimal transport cost selection for distributionally robust optimization. *arXiv preprint arXiv:1705.07152*, 2017.

[15] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[16] L. Breiman. Bias, variance, and arcing classifiers. Technical report, University of California, Berkeley, 1996.

[17] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[18] G. C. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.

[19] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[20] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[21] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[22] M. Cuturi and D. Avis. Ground metric learning. *Journal of Machine Learning Research*, 15:533–564, 2014.

[23] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

[24] B. Defourny. *Machine Learning Solution Methods for Multistage Stochastic Programming*. PhD thesis, Institut Montefiore, Université de Liege, 2010.

[25] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[26] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014.

[27] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *International Conference on Machine Learning*, pages 272–279, 2008.

[28] L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1035–1064, 1997.

[29] H. Everett III. Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Operations Research*, 11(3):399–417, 1963.

[30] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[31] F. Farnia and D. Tse. A minimax approach to supervised learning. In *Advances in Neural Information Processing Systems*, pages 4240–4248, 2016.

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

[33] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[34] R. Gao and A. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[35] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4):902–917, 2010.

[36] H. Gouk, E. Frank, B. Pfahringer, and M. Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.

[37] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[38] G. A. Hanasusanto, D. Kuhn, and W. Wiesemann. A comment on computational complexity of stochastic programming problems. *Mathematical Programming A*, 159(1-2):557–569, 2016.

[39] D. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 2013.

[40] K. Huang, H. Yang, I. King, M. R. Lyu, and L. Chan. The minimum error minimax probability machine. *Journal of Machine Learning Research*, 5:1253–1286, 2004.

[41] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[43] G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan. Minimax probability machine. In *Advances in Neural Information Processing Systems*, pages 801–807, 2002.

[44] G. R. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.

[45] N. Lawrence and B. Schölkopf. Estimating a kernel Fisher discriminant in the presence of label noise. In *International Conference on Machine Learning*, pages 306–306, 2001.

[46] Y. LeCun, C. Cortes, and C. J. Burges. The MNIST database of handwritten digits, 1998. Available from http://yann.lecun.com/exdb/mnist.

[47] C. Lee and S. Mehrotra. A distributionally-robust approach for finding support vector machines. *Available from Optimization Online*, 2015.

[48] J. Lee and M. Raginsky. Minimax statistical learning and domain adaptation with Wasserstein distances. *arXiv preprint arXiv:1705.07815*, 2017.

[49] N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *arXiv preprint arXiv:1712.09677*, 2017.

[50] F. Luo and S. Mehrotra. Decomposition algorithm for distributionally robust optimization using Wasserstein metric. *arXiv preprint arXiv:1704.03920*, 2017.

[51] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[52] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming A*, 171(1-2):115–166, 2018.

[53] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 1196–1204, 2013.

[54] B. Neyshabur, S. Bhojanapalli, and N. Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

[55] A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2014.

[56] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 2009.

[57] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[58] K. Scaman and A. Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, pages 3835–3844, 2018.

[59] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2001.

[60] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[61] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015.

[62] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

[63] A. Shapiro. On duality theory of conic linear problems. In M. Á. Goberna and M. A. López, editors, *Semi-Infinite Programming*, pages 135–165. Kluwer Academic Publishers, 2001.

[64] P. K. Shivaswamy and T. Jebara. Ellipsoidal kernel machines. In *Artificial Intelligence and Statistics*, pages 484–491, 2007.

[65] P. K. Shivaswamy and T. Jebara. Maximum relative margin and data-dependent regularization. *Journal of Machine Learning Research*, 11:747–788, 2010.

[66] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, 2006.

[67] A. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3): 199–222, 2004.

[68] T. Strohmann and G. Z. Grudic. A formulation for minimax probability machine regression. In *Advances in Neural Information Processing Systems*, pages 785–792, 2003.

[69] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[70] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288, 1996.

[71] A. N. Tikhonov, V. I. Arsenin, and F. John. *Solutions of Ill-Posed Problems.* Winston, 1977.

[72] C. Villani. *Optimal Transport: Old and New.* Springer, 2008.

[73] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

[74] T. Wiatowski, M. Tschannen, A. Stanic, P. Grohs, and H. Bölcskei. Discrete deep feature extraction: A theory and new architectures. In *International Conference on Machine Learning*, pages 2149–2158, 2016.

[75] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[76] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009.

[77] H. Xu, C. Caramanis, and S. Mannor. Robust regression and Lasso. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.

[78] T. Yang, M. Mahdavi, R. Jin, L. Zhang, and Y. Zhou. Multiple kernel learning from noisy labels by stochastic programming. In *International Conference on Machine Learning*, pages 123–130, 2012.

[79] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014.

[80] C. Zhao and Y. Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262–267, 2018.