Les Cahiers du GERAD ISSN: 0711–2440

Distributionally robust local non-parametric conditional estimation

V. A. Nguyen, F. Zhang
J. Blanchet, E. Delage, Y. Ye

G–2020–55

October 2020

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Distributionally robust local non-parametric conditional estimation

Viet Anh Nguyen [a]

Fan Zhang [a]

Jose Blanchet [a]

Erick Delage [b]

Yinyu Ye [a]

[a] Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA

[b] GERAD & Department of Decision Sciences, HEC Montréal, Montréal (Québec) Canada, H3C 3A7

erick.delage@hec.ca

October 2020
Les Cahiers du GERAD
G–2020–55

Abstract: Conditional estimation given specific covariate values (i.e., local conditional estimation or functional estimation) is ubiquitously useful with applications in engineering, social and natural sciences. Existing data-driven non-parametric estimators mostly focus on structured homogeneous data (e.g., weakly independent and stationary data), thus they are sensitive to adversarial noise and may perform poorly under a low sample size. To alleviate these issues, we propose a new distributionally robust estimator that generates non-parametric local estimates by minimizing the worst-case conditional expected loss over all adversarial distributions in a Wasserstein ambiguity set. We show that despite being generally intractable, the local estimator can be efficiently found via convex optimization under broadly applicable settings, and it is robust to the corruption and heterogeneity of the data. Experiments with synthetic and MNIST datasets show the competitive performance of this new class of estimators.

Acknowledgments: Material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550–20–1–0397. Additional support is gratefully acknowledged from NSF grants 1915967, 1820942, 1838676, NSERC grant RGPIN–2016–05208, and from the China Merchant Bank. Finally, This research was enabled in part by support provided by Compute Canada.

1 Introduction

We consider the estimation of conditional statistics of a response variable, $Y \in \mathbb{R}^m$, given the value of a predictor or covariate $X \in \mathbb{R}^n$. The single most important instance of these types of problems involves estimating the conditional mean, or also known as the regression function. Under finite variance assumptions, the conditional mean $\mathbb{E}_{\mathbb{P}}[Y|X = x_0]$ is technically defined as $\psi^\star(x_0)$ for some measurable function $\psi^\star$ that solves the minimum mean square error problem

$$\min_{\psi} \ \mathbb{E}_{\mathbb{P}}[\|Y - \psi(X)\|_2^2],$$

where the minimization is taken over the space of all measurable functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. While the optimal solution $\psi^\star$ is unique up to sets of $\mathbb{P}$-measure zero, unfortunately, solving for $\psi^\star$ is challenging because it is an infinite-dimensional optimization problem. The regression function $\psi^\star$ can be efficiently found only under specific settings, for example, if one assumes that (X, Y) follows a jointly Gaussian distribution. However, these specific situations are overly restrictive in practice.

In order to bypass the infinite-dimensional challenge involved in directly computing $\psi^\star$, we may instead consider a family of optimization problems that are parametrized by x_0. More specifically, in the presence of a regular conditional distribution, the conditional mean $\mathbb{E}_{\mathbb{P}}[Y|X = x_0]$ can be estimated pointwise by $\widehat{\beta}$ defined as

$$\widehat{\beta} \in \arg\min_{\beta} \mathbb{E}_{\mathbb{P}}[\|Y - \beta\|_2^2|X = x_0]$$

for any covariate value x_0 of interest. This presents the challenge of effectively accessing the conditional distribution, which is particularly difficult if the event $X = x_0$ has $\mathbb{P}$-probability zero.

Using an analogous argument, if we are interested in the conditional $(\tau \times 100\%)$-quantile of Y given X, then this conditional statistics can be estimated pointwise at any location x_0 of interest by

$$\widehat{\beta} \in \arg\min_{\beta} \mathbb{E}_{\mathbb{P}}[\max\{-\tau(Y - \beta), (1 - \tau)(Y - \beta)\}|X = x_0].$$

The previous examples illustrate that the estimation of a wide range of conditional statistics can be recast into solving a family of finite-dimensional optimization problems parametrically in x_0

$$\min_{\beta} \ \mathbb{E}_{\mathbb{P}}[\ell(Y, \beta)|X = x_0] \tag{1}$$

with an appropriately chosen statistical loss function ℓ.

Problem (1) poses several challenges, some of which were alluded to earlier. First, it requires the integration with respect to a difficult to compute conditional probability distribution. Second, the probability measure $\mathbb{P}$ is generally unknown, hence we lack a fundamental input to solve (1). Finally, in a data-driven setting, there may be few, or even no, observations with value covariate $X = x_0$.

To alleviate these difficulties, our formulation, as we shall explain, involves two features. First, we consider a relaxation of problem (1) in which the event $X = x_0$ is replaced by a neighborhood $\mathcal{N}_\gamma(x_0)$ of a suitable radius $\gamma \geq 0$ around x_0. Second, we introduce a data-driven distributionally robust optimization (DRO) formulation (e.g. [7, 11, 22]) in order to mitigate the problem that $\mathbb{P}$ is unknown. In turn, the DRO formulation involves a novel class of conditional ambiguity set which copes with the underlying *conditional distribution* being unknown.

In particular, we propose the following *distributionally robust local conditional estimation problem*

$$\min_{\beta} \sup_{\mathbb{Q} \in \mathbb{B}_\rho^\infty, \mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) > 0} \mathbb{E}_{\mathbb{Q}}\big[\ell(Y, \beta)|X \in \mathcal{N}_\gamma(x_0)\big], \tag{2}$$

where the maximization is taken over all probability measures $\mathbb{Q}$ that are within ρ distance in the ∞-Wasserstein sense of a benchmark nominal model, which often corresponds to the empirical distribution of available data. The probability measures $\mathbb{Q}$ are constrained so that $\mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) > 0$ to eliminate the complication of conditioning on a set of measure zero.

Contributions. Resting on formulation (2), our main contributions are summarized as follows.

1. We introduce a novel paradigm of non-parametric local *conditional* estimation based on distributionally robust optimization. In contrast to classical non-parametric conditional estimators, our new class of estimators are endowed by design with robustness features. They are structurally built to mitigate the impact of model contamination and therefore they may be reasonably applied to heterogeneous data (e.g., non i.i.d. input).
2. We demonstrate that when the ambiguity set is a type-∞ Wasserstein ball around the empirical measure, the proposed min-max estimation problem can be efficiently solved in many applicable settings, including notably the local conditional mean and quantile estimation.
3. We show that this class of type-∞ Wasserstein local conditional estimators can be considered as a systematic robustification of the k-nearest neighbor estimator. We also provide further insights on the statistical properties of our approach and empirical evidence, with both a synthetic and real data sets, that our approach can provide more accurate estimations in practically relevant settings.

Related work. One can argue that every single prediction task in machine learning ultimately relates to conditional estimation. So, attempting to provide a full literature survey on non-parametric conditional estimation is an impossible task. Since our contribution is primarily on introducing a novel conceptual paradigm powered by DRO, we focus on discussing well-understood estimators that encompass most of the conceptual ideas used to mitigate the challenges exposed earlier.

The challenges of conditioning on zero probability events and the fact that x_0 may not be a part of the sample are addressed based on the idea of averaging around a neighborhood of the point of interest and smoothing. This gives rise to estimators such as k-NN (see, for example, [12]), and kernel density estimators, including, for instance the Nadaraya-Watson estimator ([29, 39]) and the Epanechnikov estimator [13], among others. Additional averaging methods include, for example, random forests [8] and Classification and Regression Trees (CARTs, [9]), see also [19] for other techniques.

These averaging and smoothing ideas are well understood, leading to the optimal selection (in a suitable sense) of the kernel along with the associated tuning parameters such as the bandwidth size. These choices are then used to deal with the ignorance of the true data generating distribution by assuming a certain degree of homogeneity in the data, such as stationarity and weak dependence, in order to guarantee consistency and recovery of the underlying generating model. However, none of these estimators are directly designed to cope with the problem of general (potentially adversarial) data contamination.

The later issue revolving around the evaluation of an unknown conditional probability model is connected with robustness, another classical topic in statistics [20]. Much of the classical literature on robustness focuses on the impact of outliers. The work of [41] studies robust-against-outliers kernel regression which enjoys asymptotic consistency and normality under i.i.d. assumptions in a setting where the data contamination becomes negligible. In contrast to this type of contamination, our estimators are designed to be min-max optimal in the DRO sense by supplying the best response against a large (non-parametric) class of adversarial contamination.

Our results can also be seen as connected to adversarial training, which has received a significant amount of attention in recent years [18, 23, 26, 32, 34, 38]. Much of the work in this area focuses on designing well-crafted attacks and associated robust learning procedures to mitigate the effect of the attacks. This is the spirit precisely of the work in [25], in the context of k-NN estimation. One can interpret our approach as training conditional estimators against adversarial attacks, the difference, in the k-NN estimation setting for example, is that our attacks are optimal in a specific sense. The proposed estimator is thus provably the best for a uniform class of distributional attacks.

DRO-based estimators have generated a great deal of interest because they possess various desirable properties in connection to various forms of regularization (e.g., variance [30]; norm [33]; shrinkage [31]). The tools that we employ are related to those currently being investigated. Our formulation considers

adversarial perturbations based on the Wasserstein distance [7, 15, 22, 27]. In particular, the type-∞ Wasserstein distance [17] is recently applied in DRO formulations [4, 6, 40]. In particular, the work of [5] considers adversarial conditional estimation, taking as input various classical estimators (e.g., k-NNs, kernel methods, etc.) and proposes a robustification approach considering only perturbation in the response variable. Our method whereas allows perturbations both to the covariate and response variables, which is technically more subtle because of the local conditioning problem. Within the k-NN DRO conditional robustification, our numerical experiments in Section 4 show substantial benefits of our local conditioning approach, especially in dealing with non-homogeneity and sharp variations in the underlying density.

Notations. For any integer $M \in \mathbb{N}_+$, we denote by $[M]$ the set $\{1, \ldots, M\}$. For any set $\mathcal{S}$, $\mathcal{M}(\mathcal{S})$ is the space of all probability measures supported on $\mathcal{S}$.

2 Local conditional estimate using type-∞ Wasserstein ambiguity set

We start by delineating the building blocks of our distributionally robust estimation problem (2). The nominal measure is set to the empirical distribution of the available data, $\widehat{\mathbb{P}} = N^{-1} \sum_{i\in[N]} \delta_{(\widehat{x}_i, \widehat{y}_i)}$, where $\delta_{(\widehat{x},\widehat{y})}$ represents the Dirac distribution at $(\widehat{x}, \widehat{y})$. The ambiguity set $\mathbb{B}_\rho^\infty$ is a Wasserstein ball around $\widehat{\mathbb{P}}$ that contains the true distribution $\mathbb{P}$ with high confidence.

Definition 1 (Wasserstein distance) *Let $\mathbb{D}$ be a metric on Ξ. The type-p $(1 \le p < +\infty)$ Wasserstein distance between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ is defined as*

$$\mathbb{W}_p(\mathbb{Q}_1, \mathbb{Q}_2) \triangleq \inf \left\{ \left(\mathbb{E}_\pi[\mathbb{D}(\xi_1, \xi_2)^p]\right)^{\frac{1}{p}} : \pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2) \right\},$$

where $\Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ is the set of all probability measures on $\Xi \times \Xi$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$, respectively. The type-∞ Wasserstein distance is defined as the limit of $\mathbb{W}_p$ as p tends to ∞ and amounts to

$$\mathbb{W}_\infty(\mathbb{Q}_1, \mathbb{Q}_2) \triangleq \inf \left\{ \operatorname*{ess\,sup}_\pi \left\{ \mathbb{D}(\xi_1, \xi_2) : (\xi_1, \xi_2) \in \Xi \times \Xi \right\} : \pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2) \right\}.$$

We assume that (X, Y) admits values in $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^n \times \mathbb{R}^m$, and the distance $\mathbb{D}$ on $\mathcal{X} \times \mathcal{Y}$ is

$$\mathbb{D}\big((x, y), (x', y')\big) = \mathbb{D}_\mathcal{X}(x, x') + \mathbb{D}_\mathcal{Y}(y, y') \qquad \forall (x, y), (x', y') \in \mathcal{X} \times \mathcal{Y},$$

where $\mathbb{D}_\mathcal{X}$ and $\mathbb{D}_\mathcal{Y}$ are continuous metric on $\mathcal{X}$ and $\mathcal{Y}$, respectively. The joint ambiguity set $\mathbb{B}_\rho^\infty$ is now formally defined as a type-∞ Wasserstein ball in the space of joint probability measures

$$\mathbb{B}_\rho^\infty \triangleq \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) : \mathbb{W}_\infty(\mathbb{Q}, \widehat{\mathbb{P}}) \le \rho \right\}.$$

We assume further that the compact neighborhood $\mathcal{N}_\gamma(x_0)$ around x_0 is prescribed using the distance $\mathbb{D}_\mathcal{X}$ as $\mathcal{N}_\gamma(x_0) \triangleq \{x \in \mathcal{X} : \mathbb{D}_\mathcal{X}(x, x_0) \le \gamma\}$, and the loss function ℓ is jointly continuous in y and β.

To solve the estimation problem (2), we study the worst-case conditional expected loss function

$$f(\beta) \triangleq \sup_{\mathbb{Q} \in \mathbb{B}_\rho, \mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) > 0} \mathbb{E}_\mathbb{Q}\big[\ell(Y, \beta) | X \in \mathcal{N}_\gamma(x_0)\big],$$

which corresponds to the inner maximization problem of (2). To ensure that the value $f(\beta)$ is well-defined, we first investigate the conditions under which the above supremum problem has a non-empty feasible set. Towards this end, for any set $\mathcal{N}_\gamma(x_0) \subset \mathcal{X}$, define the quantities $\kappa_{i,\gamma}$ as

$$0 \le \kappa_{i,\gamma} \triangleq \min_{x \in \mathcal{N}_\gamma(x_0)} \mathbb{D}_\mathcal{X}(x, \widehat{x}_i) + \inf_{y \in \mathcal{Y}} \mathbb{D}_\mathcal{Y}(y, \widehat{y}_i) \qquad \forall i \in [N]. \tag{3}$$

The value $\kappa_{i,\gamma}$ signifies the unit cost of moving a point mass from an observation $(\widehat{x}_i, \widehat{y}_i)$ to the fiber set $\mathcal{N}_\gamma(x_0) \times \mathcal{Y}$. We also define $\widehat{x}_i^p$ as the projection of $\widehat{x}_i$ onto the neighborhood $\mathcal{N}_\gamma(x_0)$, which coincides with the optimal solution in the variable x of the minimization problem in (3). The next proposition asserts that $f(\beta)$ is well-defined if the radius ρ is sufficiently large.

Proposition 1 (Minimum radius) *For any $x_0 \in \mathcal{X}$ and $\gamma \in \mathbb{R}_+$, there exists a distribution $\mathbb{Q} \in \mathbb{B}_\rho$ that satisfies $\mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) > 0$ if and only if $\rho \geq \min_{i\in[N]} \kappa_{i,\gamma}$.*

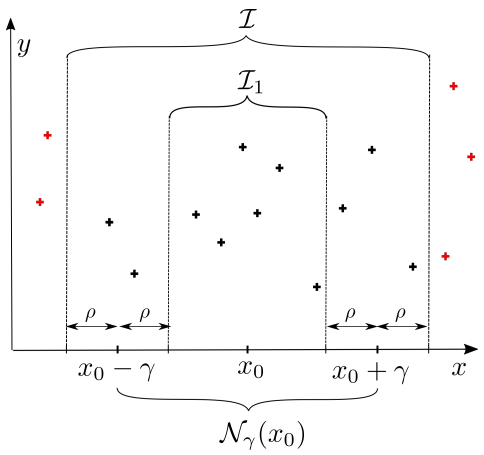


Figure 1: Illustration around the neighborhood of x_0 with $\rho < \gamma$. Black crosses are samples in the set $\mathcal{I}$.

We now proceed to the reformulation of $f(\beta)$. Let $\mathcal{I}$ be the index set defined as

$$\mathcal{I} \triangleq \{i \in [N] : \mathbb{D}_\mathcal{X}(x_0, \widehat{x}_i) \leq \rho + \gamma\}, \tag{4a}$$

and $\mathcal{I}$ is decomposed further into two disjoint subsets

$$\mathcal{I}_1 = \{i \in \mathcal{I} : \mathbb{D}_\mathcal{X}(x_0, \widehat{x}_i) + \rho \leq \gamma\} \quad \text{and} \quad \mathcal{I}_2 = \mathcal{I} \backslash \mathcal{I}_1. \tag{4b}$$

Intuitively speaking, $\mathcal{I}$ contains the indices of data points whose covariate $\widehat{x}_i$ is sufficiently close to x_0 measured by $\mathbb{D}_\mathcal{X}$, and are thus relevant to the local estimation problem. The index set $\mathcal{I}_1$ indicates the data points that lie strictly inside the neighborhood, while the set $\mathcal{I}_2$ contains those points that are on the boundary ring of width ρ around the neighborhood $\mathcal{N}_\gamma(x_0)$. The value $f(\beta)$ can be efficiently computed in a quasi-closed form thanks to the following result.

Theorem 1 (Worst-case conditional expected loss computation) *For any $\gamma \in \mathbb{R}_+$, suppose that $\rho \geq \min_{i\in[N]} \kappa_{i,\gamma}$. For any $\beta \in \mathcal{Y}$, let $v_i^\star(\beta)$ be defined as*

$$v_i^\star(\beta) \triangleq \sup_{y_i} \{\ell(y_i, \beta) : y_i \in \mathcal{Y},\ \mathbb{D}_\mathcal{Y}(y_i, \widehat{y}_i) \leq \rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i)\} \quad \forall i \in \mathcal{I}. \tag{5}$$

The worst-case conditional expected loss is equal to $f(\beta) = \left(\sum_{i\in\mathcal{I}} \alpha_i\right)^{-1} \sum_{i\in\mathcal{I}} \alpha_i v_i^\star(\beta)$, where α admits the value

$$\forall i \in \mathcal{I}: \quad \alpha_i = \begin{cases} 1 & \text{if } i \in \mathcal{I}_1 \text{ or } (\mathcal{I}_1 = \emptyset \text{ and } v_i^\star(\beta) = \max_{j\in\mathcal{I}_2} v_j^\star(\beta)), \\ 1 & \text{if } v_i^\star(\beta) > \dfrac{\sum_{i\in\mathcal{I}_1} v_i^\star(\beta) + \sum_{j\in\mathcal{I}_2 : v_j^\star(\beta) > v_i^\star(\beta)} v_j^\star(\beta)}{|\mathcal{I}_1| + |\{j \in \mathcal{I}_2 : v_j^\star(\beta) > v_i^\star(\beta)\}|}, \\ 0 & \text{otherwise.} \end{cases}$$

If we possess an oracle that evaluates (5) at a complexity $\mathcal{O}$, then by Theorem 1, quantifying $f(\beta)$ is reduced to calculating $|\mathcal{I}|$ values of $v_i^\star(\beta)$ and then sorting these values in order to determine the

value of α. Thus, computing $f(\beta)$ takes an amount of time of order $O\big(|\mathcal{I}|(\log|\mathcal{I}|+\mathcal{O})\big)$. Moreover, $f(\beta)$ depends solely on the observations in the locality of x_0 whose indices belong to the index set $\mathcal{I}$, the cardinality of which can be substantially smaller than the total number of training samples N.

If ℓ is a convex function in β, then a standard result from convex analysis implies that f, being a pointwise supremum of convex functions, is also convex. If $\mathcal{Y}$, and hence β, is unidimensional, a golden section search algorithm can be utilized to identify the local conditional estimate β^* that solves (2) in an amount of time of order $O\big(\log(1/\epsilon)|\mathcal{I}|(\log(|\mathcal{I}|)+\mathcal{O})\big)$, where $\epsilon>0$ is an arbitrary accuracy level. Fortunately, in the case of conditional mean and quantile estimation, we also have access to the closed form expressions of $v_i^\star(\beta)$ as long as $\mathbb{D}_\mathcal{Y}$ is an absolute distance.

Corollary 1 (Value of $v_i^\star(\beta)$) *Suppose that $\mathcal{Y}=[a,b]\subseteq[-\infty,+\infty]$ and $\mathbb{D}_\mathcal{Y}(y_i,\widehat{y}_i)=|y_i-\widehat{y}_i|$.*

(i) Conditional mean estimation: if $\ell(y,\beta)=(y-\beta)^2$, then $\forall i\in\mathcal{I}$

$$v_i^\star(\beta)=\max\big\{(\max\{\widehat{y}_i+\rho-\mathbb{D}_\mathcal{X}(\widehat{x}_i^p,\widehat{x}_i),a\}-\beta)^2,(\min\{\widehat{y}_i+\rho-\mathbb{D}_\mathcal{X}(\widehat{x}_i^p,\widehat{x}_i),b\}-\beta)^2\big\}.$$

(ii) Conditional quantile estimation: if $\ell(y,\beta)=\max\{-\tau(y-\beta),(1-\tau)(y-\beta)\}$, then $\forall i\in\mathcal{I}$

$$v_i^\star(\beta)=\max\big\{-\tau(\max\{\widehat{y}_i+\rho-\mathbb{D}_\mathcal{X}(\widehat{x}_i^p,\widehat{x}_i),a\}-\beta),(1-\tau)(\min\{\widehat{y}_i+\rho-\mathbb{D}_\mathcal{X}(\widehat{x}_i^p,\widehat{x}_i),b\}-\beta)\big\}.$$

If $\mathcal{Y}$ is multidimensional, the structure of $\ell(y,\beta)$ and $\mathbb{D}_\mathcal{Y}$ might be exploited to identify tractable optimization reformulations. The next result focuses on the local conditional mean estimation.

Proposition 2 (Multivariate conditional mean estimation) *Let $\mathcal{Y}=\mathbb{R}^m$ and $\ell(y,\beta)=\|y-\beta\|_2^2$.*

(i) Suppose that $\mathbb{D}_\mathcal{Y}$ is a 2-norm on $\mathcal{Y}$, that is, $\mathbb{D}_\mathcal{Y}(y,\widehat{y})=\|y-\widehat{y}\|_2$. The distributionally robust local conditional estimation problem (2) is equivalent to the second-order cone program

$$\begin{array}{ll}\min & \lambda\\ \text{s.t.} & \beta\in\mathbb{R}^m,\ \lambda\in\mathbb{R},\ u_i\in\mathbb{R}\ \forall i\in\mathcal{I}_1,\ u_i\in\mathbb{R}_+\ \forall i\in\mathcal{I}_2,\ t_i\in\mathbb{R}_+\ \forall i\in\mathcal{I}\\ & \sum_{i\in\mathcal{I}}u_i\le 0,\quad t_i\ge\|y_i-\beta\|_2\quad\forall i\in\mathcal{I}\\ & \|[t_i+\rho-\mathbb{D}_\mathcal{X}(\widehat{x}_i^p,\widehat{x}_i)\ ;\ (1/2)(1-\lambda-u_i)]\|_2\le(1/2)(1+\lambda+u_i)\quad\forall i\in\mathcal{I}.\end{array}$$

(ii) Suppose that $\mathbb{D}_\mathcal{Y}$ is a ∞-norm on $\mathcal{Y}$, that is, $\mathbb{D}_\mathcal{Y}(y,\widehat{y})=\|y-\widehat{y}\|_\infty$. The distributionally robust local conditional estimation problem (2) is equivalent to the second-order cone program

$$\begin{array}{ll}\min & \lambda\\ \text{s.t.} & \beta\in\mathbb{R}^m,\ \lambda\in\mathbb{R},\ T\in\mathbb{R}_+^{|\mathcal{I}|\times m},\ u_i\in\mathbb{R}\ \forall i\in\mathcal{I}_1,\ u_i\in\mathbb{R}_+\ \forall i\in\mathcal{I}_2\\ & \sum_{i\in\mathcal{I}}u_i\le 0,\ ;\ \|[T_{i1}\ ;\ T_{i2}\ ;\ \cdots\ ;\ T_{im}\ ;\ \frac{1}{2}(1-\lambda-u_i)]\|_2\le\frac{1}{2}(1+\lambda+u_i)\ \ \forall i\in\mathcal{I}\\ & \left.\begin{array}{l}T_{ij}\le\widehat{y}_{ij}-\beta_j-\rho+\mathbb{D}_\mathcal{X}(\widehat{x}_i^p,\widehat{x}_i)\le T_{ij}\\ T_{ij}\le\widehat{y}_{ij}-\beta_j+\rho-\mathbb{D}_\mathcal{X}(\widehat{x}_i^p,\widehat{x}_i)\le T_{ij}\end{array}\right\}\forall(i,j)\in\mathcal{I}\times[m],\end{array}$$

where $\widehat{y}_{ij}$ and β_j are the j-th component of $\widehat{y}_i$ and β, respectively.

Both optimization problems presented in Proposition 2 can be solved in large scale by commercial optimization solvers such as MOSEK [28]. For other multivariate conditional estimation problems, there is also a possibility of employing subgradient methods by leveraging on the next proposition.

Proposition 3 (Subgradient of f) *Suppose that $\mathbb{D}_\mathcal{Y}$ is coercive and $\ell(y,\cdot)$ is convex. Under the conditions of Theorem 1, for any $\beta\in\mathbb{R}^m$, a subgradient of the function f at β is given by $\partial f(\beta)=(\sum_{i\in\mathcal{I}}\alpha_i)^{-1}\sum_{i\in\mathcal{I}}\alpha_i\partial_\beta\ell(y_i^\star,\beta)$, where the value of α is as defined in Theorem 1 and $y_i^\star$ satisfies $y_i^\star\in\{y_i\in\mathcal{Y}:\mathbb{D}_\mathcal{Y}(y_i,\widehat{y}_i)\le\rho-\mathbb{D}_\mathcal{X}(\widehat{x}_i^p,\widehat{x}_i),\ \ell(y_i^\star,\beta)=v_i^\star(\beta)\}$ for all $i\in\mathcal{I}$.*

3 Probabilistic theoretical properties

We now study the some statistical properties of our proposed estimator. Under some regularity conditions, the type-∞ Wasserstein ball can be viewed as a confidence set that contains the true

distribution $\mathbb{P}$ with high probability, provided that the radius ρ is chosen judiciously. The value $f(\beta^\star)$ thus constitutes a generalization bound on the out-of-sample performance of the optimal conditional estimate $\beta^\star$. This idea can be formalized as follows.

Proposition 4 (Finite sample guarantee) *Suppose that $\mathcal{X} \times \mathcal{Y}$ is bounded, open, connected with a Lipschitz boundary. Suppose that the true probability measure $\mathbb{P}$ of (X, Y) admits a density function ν satisfying $\bar{\nu}^{-1} \leq \nu(x, y) \leq \bar{\nu}$ for some constant $\bar{\nu} \geq 1$. For any $\gamma > 0$, if*

$$\rho \geq \begin{cases} CN^{-\frac{1}{2}} \log(N)^{\frac{3}{4}} & \text{when } n + m = 2, \\ CN^{-\frac{1}{n+m}} \log(N)^{\frac{1}{n+m}} & \text{otherwise,} \end{cases}$$

where C is a constant dependent on $\mathcal{X} \times \mathcal{Y}$ and $\bar{\nu}$, then for a probability of at least $1 - O(N^{-c})$, where $c > 1$ is a constant dependent on C, we have $\mathbb{E}_{\mathbb{P}}[\ell(Y, \beta^\star) | X \in \mathcal{N}_\gamma(x_0)] \leq f(\beta^\star)$, where $\beta^\star$ is the optimal conditional estimate that solves problem (2).

We now switch gear to study the properties of our estimator in the asymptotic regime, in particular, we focus on the consistency of our estimator. The interplay between the neighborhood radius γ and the ambiguity size ρ often produces tangling effects on the asymptotic convergence of the estimate. We thus showcase two exemplary setups with either γ or ρ is zero, which interestingly produce two opposite outcomes on the consistency of the estimator. This underlines the intricacy of the problem.

Example 1 (Non-consistency when $\gamma = 0$) *Suppose that $\gamma = 0$, $\rho \in \mathbb{R}_{++}$ be a fixed constant, $\mathcal{Y} = \mathbb{R}$, $\ell(y, \beta) = (y - \beta)^2$, and $\mathbb{D}_{\mathcal{Y}}$ is the absolute distance. Let $\beta_N^\star$ be the optimal estimate that solves (2) dependent on $\{(\widehat{x}_i, \widehat{y}_i)\}_{i=1,\ldots,N}$. If under the true distribution $\mathbb{P}$, X is independent of Y, $\mathbb{P}(\mathbb{D}_{\mathcal{X}}(X, x_0) \leq \rho) > 0$, $\mathbb{P}(Y \geq 0) = 1$ and $\mathbb{P}(Y \geq y) > 0$ $\forall y > 0$, then with probability 1, we have $\widehat{\beta}_N \to +\infty$ while $\mathbb{E}_{\mathbb{P}}[Y | X = x_0] < \infty$.*

Example 2 (Consistency when $\rho = 0$) *Suppose that $\rho = 0$, $\mathcal{Y} = \mathbb{R}$, $\ell(y, \beta) = (y - \beta)^2$, $\mathbb{D}_{\mathcal{X}}$ and $\mathbb{D}_{\mathcal{Y}}$ are the Euclidean distance, k_N is a sequence of integer. Let γ be the k_N-th smallest value of $\mathbb{D}_{\mathcal{X}}(x_0, \widehat{x}_i)$, then $\beta_N^\star$ that solves (2) recovers the k_N-nearest neighbor regression estimator. If k_N satisfies $\lim_{N\to\infty} k_N = \infty$ and $\lim_{N\to\infty} k_N/N = 0$, and , then $\lim_{N\to\infty} \beta_N^\star = \mathbb{E}_{\mathbb{P}}[Y | X = x_0]$ by [36, Corollary 3].*

Example 2 suggests that if the radius γ of the neighborhood is chosen adaptively based on the available training data, then our proposed estimator coincides with the k-nearest neighbor estimator, and hence consistency is inherited in a straightforward manner. The robust estimator with an ambiguity size $\rho > 0$ and an adaptive neighborhood radius γ can thus be considered as a robustification of the k-nearest neighbor, which is obtained in a systematic way using the DRO framework.

It is desirable to provide a descriptive connection between the distributionally robust estimator vis-à-vis some popular statistical quantities. For the local conditional mean estimation, our estimate $\beta^\star$ coincides with the conditional mean of the distribution with the highest conditional variance. This insight culminates in the next proposition and bolsters the explainability of this class of estimators.

Proposition 5 (Conditional mean estimate) *Suppose that $\mathcal{Y} = \mathbb{R}$, $\ell(y, \beta) = (y - \beta)^2$ and $\mathbb{D}_{\mathcal{Y}}(\cdot, \widehat{y})$ is convex, coercive for any $\widehat{y}$. For any $\rho \geq \min_{i\in[N]} \kappa_{i,\gamma}$, define $\mathbb{Q}^\star$ as*

$$\mathbb{Q}^\star = \arg \max_{\mathbb{Q} \in \mathbb{B}_\rho^\infty, \mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) > 0} \textit{Variance}_{\mathbb{Q}}(Y | X \in \mathcal{N}_\gamma(x_0)),$$

then $\beta^\star = \mathbb{E}_{\mathbb{Q}^\star}[Y | X \in \mathcal{N}_\gamma(x_0)]$ is the optimal estimate that solves problem (2).

4 Numerical experiment

In this section we compare the quality of our proposed Distributionally Robust Conditional Mean Estimator (DRCME) to k-nearest neighbour (k-NN), Nadaraya-Watson (N-W), and Nadaraya-Epanechnikov (N-E) estimators, together with the robust k-NN approach in [5] (BertEtAl) using a synthetic and the MNIST datasets.

4.1 Conditional mean estimation with synthetic data

In this section, we conducted 500 independent experiments where the training set contains $N = 100$ i.i.d. samples of (X, Y) in each experiment. The marginal distribution of X has piecewise constant density function $p(x)$, which is chosen as $p(x) = 100/72$ if $x \in [0, 0.3] \cup [0.7, 1]$ and $p(x) = 30/72$ if $x \in (0.3, 0.7)$. Given X, the distribution of Y is determined by $Y = f(X) + \varepsilon$, where $f = \sin(10 \cdot x)$ and ε is i.i.d. Gaussian noise independent of X with mean 0 and variance 0.01. The conditional mean estimation problem is challenging when x_0 is close to the jump points of the density function $p(x)$, that is at $x_0 = 0.3$ or $x_0 = 0.7$, because the data are gathered unequally in the neighborhoods. Thus, to test the robustness of all the estimators, we employ all the five estimators to estimate the conditional mean $\mathbb{E}_{\mathbb{P}}[Y|X = x_0]$, for $x_0 = 0.2, 0.21, \ldots, 0.4$ around the jump point $x_0 = 0.3$. We select $\mathbb{D}_{\mathcal{X}}(x, x') = |x - x'|$ and $\mathbb{D}_{\mathcal{Y}}(y, y') = |y - y'|$. The hyperparameters of all the estimators, whose range and selection are given in Appendix A, are chosen by leave-one-out cross validation.

Figure 2 displays the average of the mean estimation errors taken over 500 independent runs for different values $x_0 \in [0.2, 0.4]$. One can observe from the figure that DRCME uniformly outperforms k-NN, BertEtAl for all x_0 of interest. When compared with N-W and N-E, we remark that DRCME is the most accurate estimator around the jump point of $p(x)$. As x_0 moves away from the location 0.3, the performance of DRCME decays and becomes slightly worse than N-W as x_0 goes far from the jump point. Figure 3 presents the cumulative distribution of the estimation errors when $x_0 \in [0.28, 0.32]$. The empirical error distribution of DRCME is stochastically smaller than that of other estimators, which reinforces that DRCME outperforms around the jump point in a strong sense.

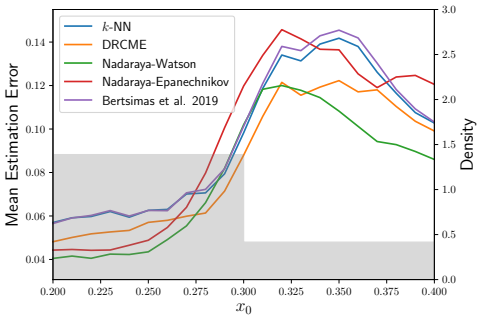


Figure 2: Comparison of the mean absolute errors of conditional mean estimators for synthetic data. The gray shade shows the density of X.

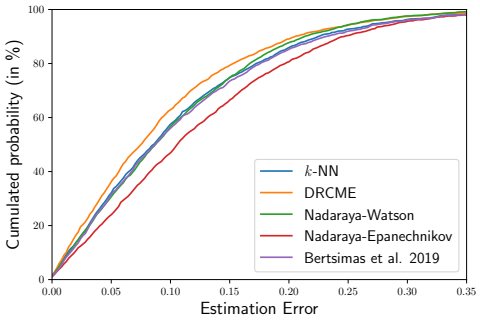


Figure 3: Comparison of the distributions of absolute estimation errors of conditional mean estimators for synthetic data.

4.2 Digit estimation with MNIST database

In this section, we compare the quality of the estimators on a digit estimation problem using the MNIST database [24]. While to this date most studies have focused on out-of-sample classification performances for this dataset, here we shift our attention to the task of estimation of digits as **cardinal** quantities and are especially interested in performance at a low-data regime. Treating the labels as cardinal quantities allows us to assess the distinctive features of DRCME in its most simplistic form (i.e. univariate conditional mean estimation of a real random variable). Mean estimation might in fact be more relevant than classification when trying to recognize handwritten measurements where confusing a 0 with a 6 is more damaging than with a 3.

We executed 100 experiments where training and test sets were randomly drawn without replacement from the 60,000 training examples of this dataset. Training set sizes were $N = 50$, 100, or 500 while test sets' size remained at 100. Each (x, y) pair is composed of the normalized vector, in $\mathbb{R}^{28^2}$ of grayscale intensities normalized so that $\|x\|_1 = 1$. For simplicity, we let $\mathbb{D}_{\mathcal{X}}(x, \widehat{x}) = \|x - \widehat{x}\|_2$ and $\mathbb{D}_{\mathcal{Y}}(y, \widehat{y}) = \alpha|y - \widehat{y}|$. In each experiment, the hyper-parameters of all four methods were chosen based

on a leave-one-out cross validation process. In the case of DRCME , we adapt the radius of the neighborhood γ and ρ locally at x_0 to account for the non-uniform density of X.[1] Table 1 presents the median choice of hyper parameters for each estimator.

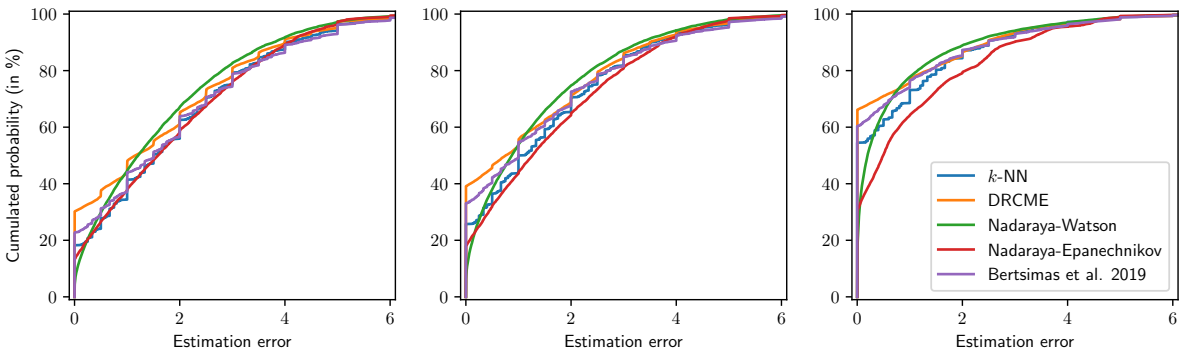
Method	H.P.	N=50	N=100	N=500
k-NN	k	3	4	4
N-W	h	0.022	0.019	0.015
N-E	h	0.087	0.078	0.068
BertEtAl	k	3	4	5
	ρ	0.712	1.313	1.313
	γ	$h^{\gamma}_{1.3}(\cdot)$	$h^{\gamma}_{1.3}(\cdot)$	$h^{\gamma}_{1.6}(\cdot)$
DRCME	ρ	0.13γ	0.13γ	0.06γ
	α	0.004	0.002	0.001

Table 1: Median of hyper-parameters (H.P.) obtained with cross-validation.

Method	N=50	N=100	N=500
k-NN	24 ± 2	35 ± 2	60 ± 1
N-W	30 ± 2	38 ± 2	65 ± 1
N-E	26 ± 1	32 ± 1	50 ± 1
BertEtAl	29 ± 2	41 ± 2	67 ± 1
DRCME	36 ± 2	46 ± 2	71 ± 1

Table 2: Comparison of expected out-of-sample classification accuracy (in % with 90% confidence intervals) from rounded estimates.

Figure 4 presents the out-of-sample estimation error distribution of all four conditional estimators. One can quickly remark that the DRCME outperforms BertEtAl , k-NN, and N-E estimators, especially for low-data regime. In particular, for all three training set sizes, the distribution of error for DRCME stochastically dominates the three other distributions. In particular, one even notices in (c) that DRCME has the largest chance of reaching an exact estimation: 66% compared to 60%, 55%, 30%, and 8% for the other estimators. This explains why DRCME is also the most accurate estimator when rounding it to the nearest integer as reported in Table 2: with a margin greater than 4% from all estimators across all N's. It is worth noting that while N-W does not produce high accuracy estimate, it however has less chances of producing estimation with large errors. This is also apparent when comparing the expected type-p deviation of the estimation error, i.e. $(\mathbb{E}[|y - \hat{y}|^p])^{1/p}$, for each estimator. Specifically, N-W slightly outperforms DRCME for deviation metrics of type $p \geq 1$, e.g. with a root mean square error of 1.32 compared to 1.41 when $N = 500$. On the other hand, DRCME significantly outperforms N-W when $p < 1$ where high precision estimators are encouraged. We refer the reader to Appendix A for further details.



(a) N=50 (b) N=100 (c) N=500

Figure 4: Comparison of the distributions of out-of-sample absolute estimation errors of conditional mean estimators for the MNIST database under different training set sizes.

Finally, we report on an experiment that challenges the capacity of both N-W and DRCME estimators to be resilient to adversarial corruption of the test images. This is done by exposing the two

[1] Specifically, we let $\gamma = h^{\gamma}_i(x_0) := \kappa_{\lfloor \lfloor i \rfloor \rfloor,0} + (i - \lfloor i \rfloor)(\kappa_{\lfloor \lceil i \rceil \rfloor,0} - \kappa_{\lfloor \lfloor i \rfloor \rfloor,0})$, where $\lfloor j \rfloor$ refers to the j-th smallest element while $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ refer to the floor and ceil operations, i.e. the radius is set to the linear interpolation between the distance of the $\lfloor i \rfloor$-th and $\lfloor i \rfloor + 1$-th closest members of the training set to x_0. We further let ρ be proportional to γ. This lets DRCME reduce to k-NN when $\gamma = h^{\gamma}_k(x_0)$, $\rho = 0$, and $\alpha = 1$.

estimators to images from the training set ($N = 100$) that have been corrupted in a way that makes them resemble the closest differently-labeled image in the set.[2] Figure 5 presents several visual examples of the progressively corrupted images and the resulting N-W and DRCME estimations. Overall, one quickly notices how the estimation produced by DRCME is less sensitive to such attacks, "sticking" to the original label until there is substantial evidence of a new label. More examples are in Appendix A.

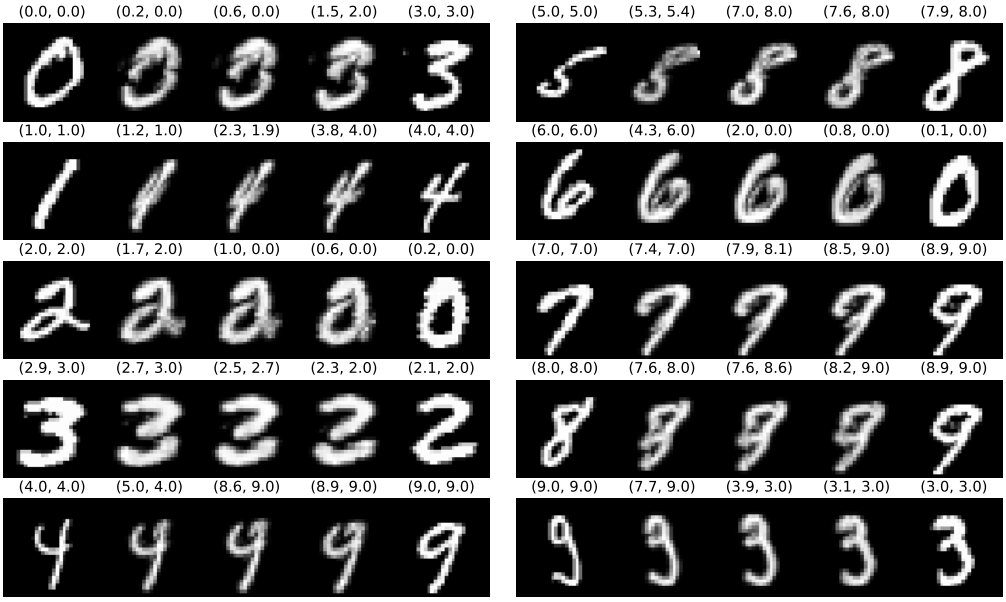


Figure 5: Comparison of estimations from N-W and DRCME on entropic regularized Wasserstein barycenters of pairs of images from the training set. Estimations are presented above each image in the format "(N-W, DRCME)".

A Additional experiment results

A.1 Conditional mean estimation with synthetic data

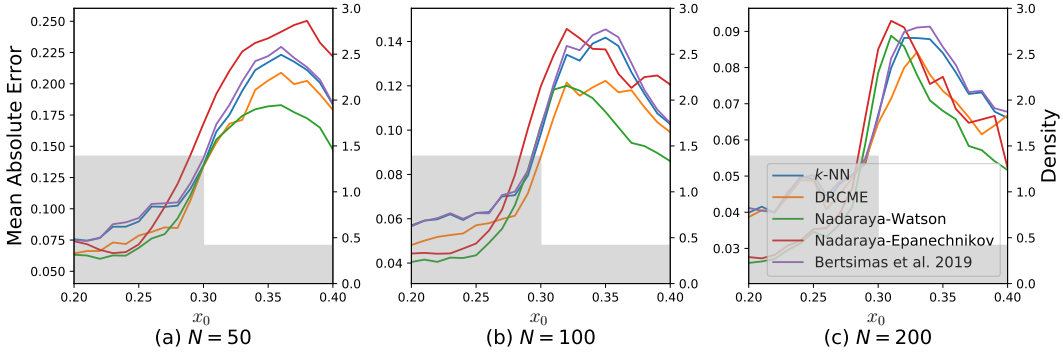
We report in Figure A.1 the plot of mean estimation errors versus x_0 for different training set sizes $N = 50, 100, 200$. In Figure A.2 we present the plot of the distribution of absolute estimation errors for $x_0 \in [0.28, 0.32]$. For comparison, we also include the results of training set size $N = 100$ that are already reported in Figure 2 and 3. We remark that the estimation error of all the estimators becomes smaller when training set size is larger, and DRCME has best estimation performance among all the estimators around the jump point $x = 0.3$ for all different training set sizes.

We report the hyper-parameters selected by cross-validation in Table A.1.

Table A.1: Median of hyper-parameters (H.P.) for synthetic data experiment obtained with cross-validation.

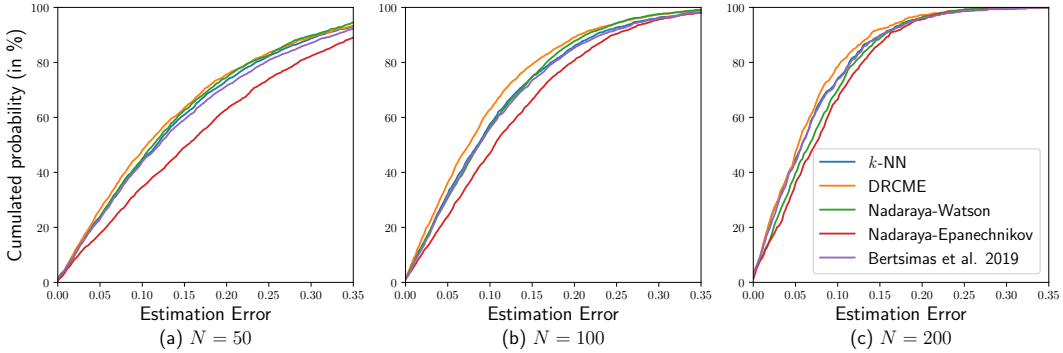
Method	H.P.	N=50	N=100	N=200
k-NN	k	1	3	5
N-W	h	0.026	0.019	0.018
N-E	h	0.078	0.055	0.038
BertEtAl	k	1	3	5
	ρ	0.063	0.016	0.000
	γ	$h_1^\gamma(\cdot)$	$h_2^\gamma(\cdot)$	$h_3^\gamma(\cdot)$
DRCME	ρ	0.031γ	0.063γ	0.063γ

[2] Implementation wise, we exploit the Python Optimal Transport toolbox [14] to compute different entropic regularized Wasserstein barycenters of the two normalized images treated as distributions.



(a) $N = 50$ (b) $N = 100$ (c) $N = 200$

Figure A.1: Comparison of the mean absolute errors of conditional mean estimators for synthetic data under different training set sizes. The gray shade shows the density of X.

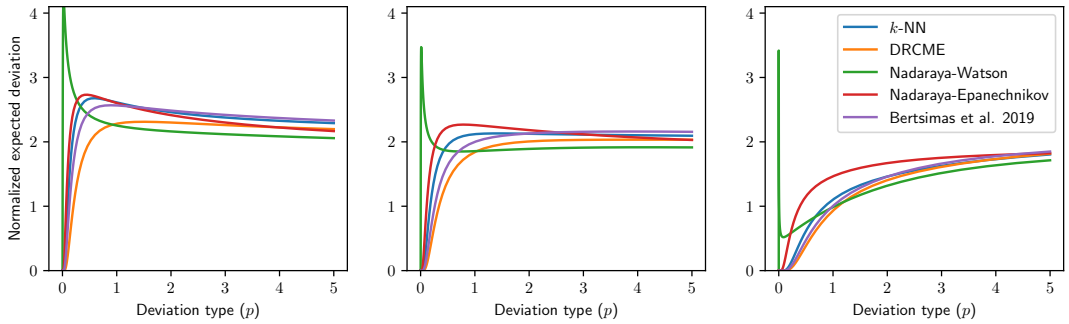


(a) $N = 50$ (b) $N = 100$ (c) $N = 200$

Figure A.2: Comparison of the distributions of absolute estimation errors of conditional mean estimators for synthetic data under different training set sizes.

A.2 Digit estimation with MNIST database

The distinction between N-W and DRCME is also apparent in Figure A.3 which presents the normalized expected type-p deviation of the estimation error for each estimator, i.e. $\sqrt{2/p}(\mathbb{E}[|y - \hat{y}|^p])^{1/p}$. Specifically, N-W slightly outperforms DRCME for deviation metrics of type $p \geq 1$, e.g. with a root mean square error of 1.34 compared to 1.45 when $N = 500$. On the other hand, DRCME significantly outperforms N-W when $p < 1$ where high precision estimators are encouraged.



(a) N=50 (b) N=100 (c) N=500

Figure A.3: Comparison of normalized expected type-p deviation of the out-of-sample error of four non-parametric conditional mean estimation methods for the MNIST database under different training set sizes. E.g., at $p = 2$ is presented the root-mean square error.

We also include in Figure A.4 some additional examples of labels from DRCME and N-W. On the other hand, Figure A.5 compares the labels from DRCME and BertEtAl .

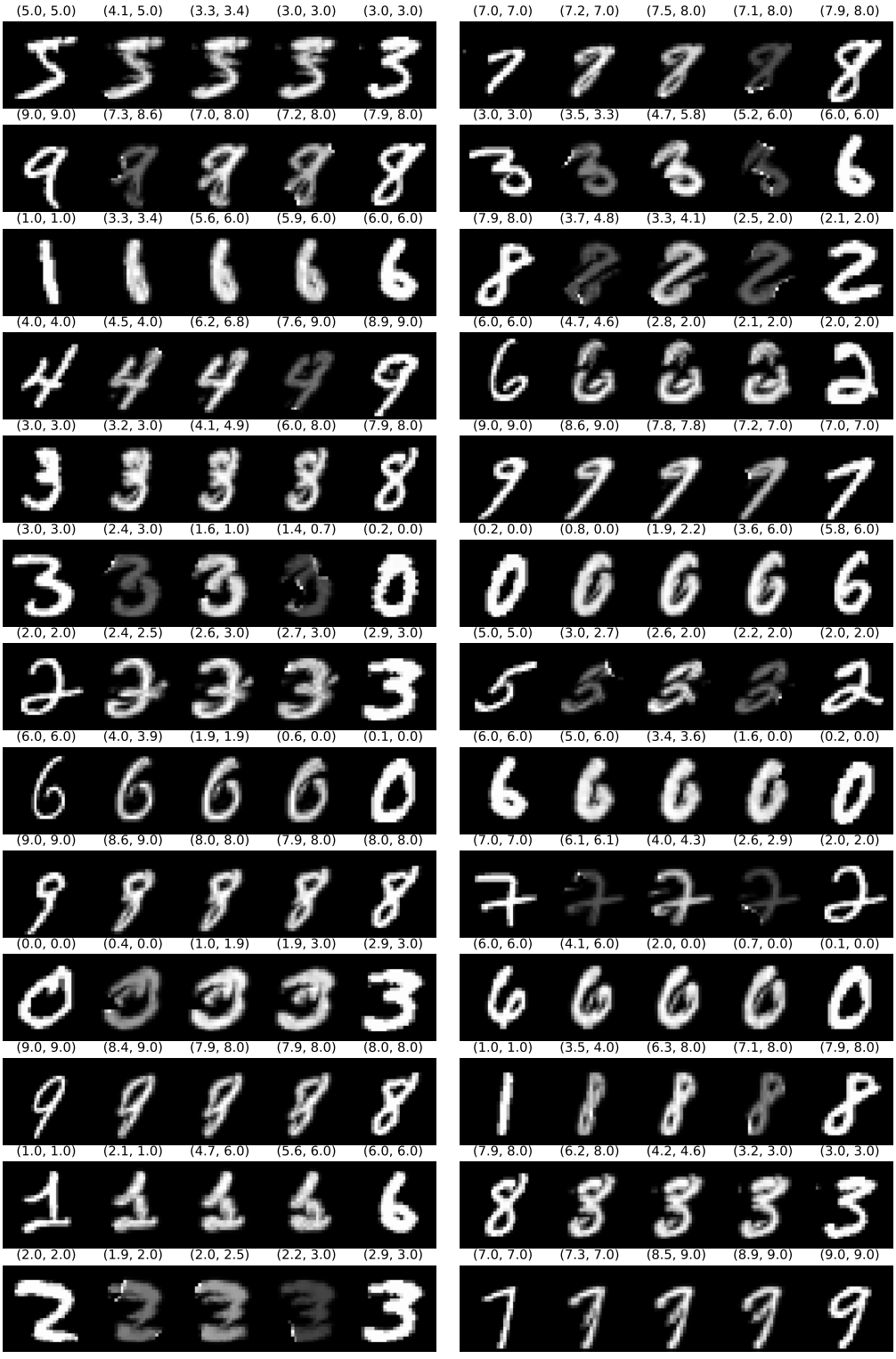


Figure A.4: Comparison of estimations from N-W and DRCME on entropic regularized Wasserstein barycenters of pairs of images from the training set. Estimations are presented above each image in the format "(N-W, DRCME)".

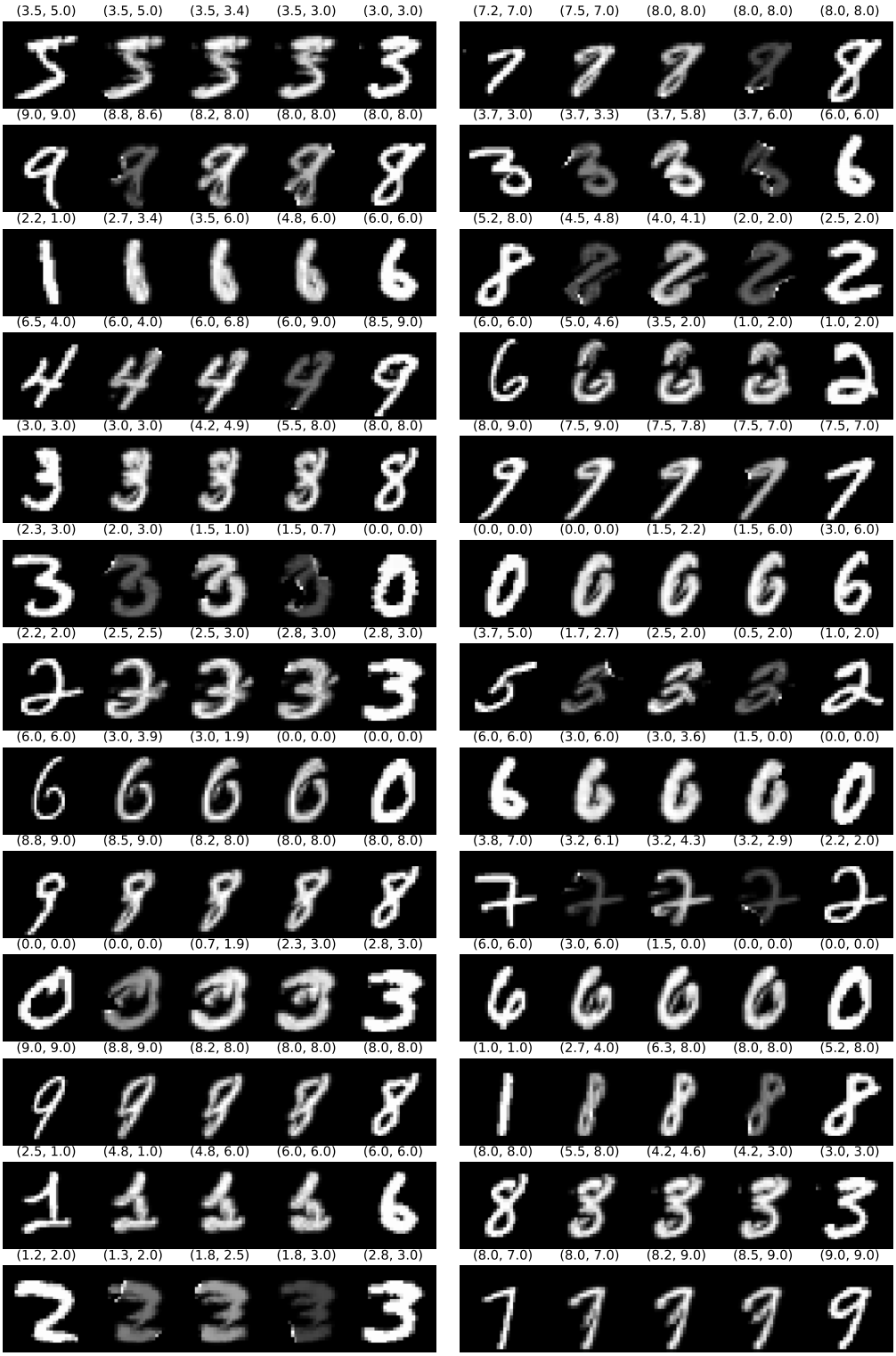


Figure A.5: Comparison of estimations from BertEtAl and DRCME on entropic regularized Wasserstein barycenters of pairs of images from the training set. Estimations are presented above each image in the format "(BertEtAl , DRCME)".

B Proofs

This section contains the proofs of all technical results presented in the main paper.

B.1 Proofs of Section 2

Proof of Proposition 1. Using the definition of the type-∞ Wasserstein distance, we can re-express the ambiguity set $\mathbb{B}_\rho^\infty$ as

$$\begin{aligned}\mathbb{B}_\rho^\infty &= \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X}\times\mathcal{Y}) : \begin{array}{l} \exists \pi \in \Pi(\mathbb{Q}, \widehat{\mathbb{P}}) \text{ such that} \\ \operatorname{ess\,sup}_{\pi} \{\mathbb{D}_{\mathcal{X}}(x,x') + \mathbb{D}_{\mathcal{Y}}(y,y')\} \leq \rho \end{array} \right\} \\ &= \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X}\times\mathcal{Y}) : \begin{array}{l} \exists \pi_i \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})\ \forall i \in [N] \text{ such that } \mathbb{Q} = \frac{1}{N}\sum_{i\in[N]} \pi_i \\ \operatorname*{ess\,sup}_{\frac{1}{N}\sum_{i\in[N]} \pi_i \otimes \delta_{(\widehat{x}_i,\widehat{y}_i)}} \left\{\mathbb{D}_{\mathcal{X}}(x,x') + \mathbb{D}_{\mathcal{Y}}(y,y')\right\} \leq \rho \end{array} \right\},\end{aligned}$$

where in the second equality we exploit the fact that $\widehat{\mathbb{P}}$ is an empirical measure and thus any joint probability measure $\pi \in \Pi(\mathbb{Q}, \widehat{\mathbb{P}})$ can be written as $\pi = N^{-1}\sum_{i\in[N]} \pi_i \otimes \delta_{(\widehat{x}_i,\widehat{y}_i)}$, where each π_i is a probability measure supported on $\mathcal{X}\times\mathcal{Y}$. The last constraint can now be written as

$$\mathbb{D}_{\mathcal{X}}(x,\widehat{x}_i) + \mathbb{D}_{\mathcal{Y}}(y,\widehat{y}_i) \leq \rho \quad \forall (x,y) \in \operatorname{supp}(\pi_i) \quad \forall i \in [N],$$

where $\operatorname{supp}(\pi_i)$ denotes the support of the probability measure π_i [1, Page 441]. We thus have

$$\mathbb{B}_\rho^\infty = \left\{ \mathbb{Q} \in \mathcal{M}(\mathcal{X}\times\mathcal{Y}) : \begin{array}{l} \exists \pi_i \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})\ \forall i \in [N] \text{ such that } \mathbb{Q} = \frac{1}{N}\sum_{i\in[N]} \pi_i \\ \mathbb{D}_{\mathcal{X}}(x,\widehat{x}_i) + \mathbb{D}_{\mathcal{Y}}(y,\widehat{y}_i) \leq \rho \quad \forall (x,y) \in \operatorname{supp}(\pi_i) \quad \forall i \in [N] \end{array} \right\}.$$

Suppose that $\rho < \min_{i\in[N]} \kappa_{i,\gamma}$, then this implies by the last constraint of the feasible set that $\pi_i(\mathcal{N}_\gamma(x_0)\times\mathcal{Y}) = 0$ for all $i \in [N]$. As a consequence, any $\mathbb{Q} \in \mathbb{B}_\rho^\infty$ should satisfy

$$\mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) = \sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0)\times\mathcal{Y}) = 0.$$

Hence $\mathbb{B}_\rho^\infty \cap \{\mathbb{Q} \in \mathcal{M}(\mathcal{X}\times\mathcal{Y}) : \mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) > 0\} = \emptyset$.

Suppose on the contrary that $\rho \geq \min_{i\in[N]} \kappa_{i,\gamma}$. Let $i^\star = \arg\min_{i\in[N]} \kappa_{i,\gamma}$, and consider the following set of probability measures

$$\forall i \in [N]: \qquad \pi_i = \begin{cases} \delta_{(\widehat{x}_i^p,\widehat{y}_i)} & \text{if } i = i^\star, \\ \delta_{(\widehat{x}_i,\widehat{y}_i)} & \text{otherwise,} \end{cases}$$

and set $\mathbb{Q} = \frac{1}{N}\sum_{i\in[N]} \pi_i$. It is easy to verify that $\mathbb{Q} \in \mathbb{B}_\rho^\infty$, and that

$$\mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) \geq \frac{1}{N}\pi_{i^\star}(X \in \mathcal{N}_\gamma(x_0)) = \frac{1}{N} > 0.$$

This observation completes the proof. □

The proof of Theorem 1 relies on the following result.

Lemma 1 (Optimal solution of a fractional linear program) *Let d be an strictly positive integer. The linear fractional program*

$$\min\left\{ \frac{c + \sum_{i=1}^K v_i\alpha_i}{d + \sum_{i=1}^K \alpha_i} : \alpha \in [0,1]^K \right\}$$

admits the optimal solution

$$\forall i \in [K]: \qquad \alpha_i^\star = \begin{cases} 1 & \text{if } v_i > \dfrac{c + \sum_{j:v_j>v_i} v_j}{d + |\{j : v_j > v_i\}|}, \\ 0 & \text{otherwise.} \end{cases}$$

Proof of Lemma 1. Without loss of generality assume that v_i are ordered decreasingly. Because the objective function is pseudolinear, the optimal solution is at some binary vertex [21, Lemma 3.3]. Consider the equivalent problem

$$\max_{k,\alpha}\left\{\frac{c+\sum_{i=1}^{K} v_i\alpha_i}{d+\sum_{i=1}^{K}\alpha_i} : \alpha\in\{0,1\}^K,\ \sum_{i=1}^{K}\alpha_i = k,\ k\in[K]\right\}.$$

For any value $k\in[K]$, the corresponding optimal value of α dependent on k is

$$\alpha_i^\star(k) = \begin{cases} 1 & \text{if } i\le k,\\ 0 & \text{otherwise,}\end{cases}$$

where we exploit the fact that v_i are ordered decreasingly. The above optimization problem can be simplified to

$$\max_{k}\left\{\frac{c+\sum_{i=1}^{k} v_i}{d+k} : k\in[K]\right\}. \tag{6}$$

Now we need to show that the objective function $g(k)\triangleq(c+\sum_{i=1}^{k} v_i)/(d+k)$ becomes non-increasing once it starts decreasing. Indeed, the incremental improvement in the objective value of (6) at k can be written as

$$\begin{aligned}\Delta_g(k) = g(k+1)-g(k) &= \frac{c+\sum_{i=1}^{k+1} v_i}{d+k+1} - g(k)\\ &= \frac{(d+k)g(k)+v_{k+1}}{d+k+1} - g(k)\\ &= \frac{v_{k+1}-g(k)}{d+k+1}.\end{aligned}$$

If $\Delta_g(k)<0$, this implies that $v_{k+1}<g(k)$. We also know that $v_{k+2}\le v_{k+1}$. So we can show that:

$$\begin{aligned}\Delta_g(k+1) = g(k+2)-g(k+1) &= \frac{v_{k+2}-g(k+1)}{d+k+2}\\ &= \frac{(d+k+1)v_{k+2}-(d+k)g(k)-v_{k+1}}{(d+k+2)(d+k+1)}\\ &\le \frac{(d+k+1)v_{k+1}-(d+k)g(k)-v_{k+1}}{(d+k+2)(d+k+1)}\\ &= \frac{(d+k)(v_{k+1}-g(k))}{(d+k+2)(d+k+1)} < 0.\end{aligned}$$

Moreover, the above line of arguments also reveals that if $v_{k+2}=v_{k+1}$ then both $\Delta_g(k)$ and $\Delta_g(k+1)$ have the same sign. Thus, the value $k^\star$ that maximizes (6) is also the solution of

$$\max\{k : \Delta_g(k-1)\ge 0\}.$$

Leveraging on the formula of $\alpha_i^\star(k)$, the solution $\alpha^\star$ of the original fractional linear program has the form

$$\begin{aligned}\forall i: \qquad \alpha_i^\star &= \begin{cases} 1 & \text{if } v_i > \dfrac{c+\sum_{j:j<i} v_i}{d+|\{j:j<i\}|},\\ 0 & \text{otherwise,}\end{cases}\\ &= \begin{cases} 1 & \text{if } v_i > \dfrac{c+\sum_{j:v_j>v_i} v_j}{d+|\{j:v_j>v_i\}|},\\ 0 & \text{otherwise,}\end{cases}\end{aligned}$$

where the second equality comes from the ordering of v_i. This observation completes the proof. □

Proof of Theorem 1. A conditional measure μ_0 of Y given $X \in \mathcal{N}_\gamma(x_0)$ induced by a probability measure $\mathbb{Q}$ satisfying $\mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) > 0$ can be written as

$$\mathbb{Q}(\mathcal{N}_\gamma(x_0) \times A) = \mu_0(A)\mathbb{Q}(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) \qquad \forall A \subseteq \mathcal{Y} \text{ measurable.}$$

One can rewrite the worst-case conditional expected loss $f(\beta)$ as

$$f(\beta) = \left\{ \begin{array}{ll} \sup & \displaystyle\int_{\mathcal{Y}} \ell(y,\beta)\ \mu_0(\mathrm{d}y) \\ \text{s.t.} & \mathbb{Q} \in \mathbb{B}_\rho^\infty,\ \mathbb{Q}(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) > 0 \\ & \mathbb{Q}(\mathcal{N}_\gamma(x_0) \times A) = \mu_0(A)\mathbb{Q}(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) \qquad \forall A \subseteq \mathcal{Y} \text{ measurable.} \end{array} \right.$$

By decomposing the measure $\mathbb{Q}$ using the set of probability measures π_i and exploiting the definition of the type-∞ Wasserstein distance as in the proof of Proposition 1, we have

$$f(\beta) = \left\{ \begin{array}{ll} \sup & \displaystyle\int_{\mathcal{Y}} \ell(y,\beta)\ \mu_0(\mathrm{d}y) \\ \text{s.t.} & \mu_0 \in \mathcal{M}(\mathcal{Y}),\ \pi_i \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \quad \forall i \in [N] \\ & \displaystyle\sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) > 0 \\ & \displaystyle\sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times A) = \mu_0(A) \sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) \quad \forall A \subseteq \mathcal{Y} \text{ measurable} \\ & \mathbb{D}_{\mathcal{X}}(x,\widehat{x}_i) + \mathbb{D}_{\mathcal{Y}}(y,\widehat{y}_i) \le \rho \quad \forall (x,y) \in \operatorname{supp}(\pi_i) \qquad \forall i \in [N]. \end{array} \right.$$

For any set of feasible solutions $\{\pi_i\}_{i\in[N]}$, we have $\sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) > 0$. We can thus re-express $\mu_0(A)$ for any Borel measurable set $A \subseteq \mathcal{Y}$ as

$$\mu_0(A) = \frac{\sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times A)}{\sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times \mathcal{Y})} \quad \forall A \subseteq \mathcal{Y} \text{ measurable.}$$

Thus, we can eliminate the variables μ_0 from the above optimization problem to obtain the equivalent representation

$$f(\beta) = \left\{ \begin{array}{lll} \sup & \displaystyle\frac{1}{\sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times \mathcal{Y})} \sum_{i\in[N]} \int_{\mathcal{Y}} \ell(y,\beta)\ \pi_i(\mathcal{N}_\gamma(x_0) \times \mathrm{d}y) & \\ \text{s.t.} & \pi_i \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) & \forall i \in [N] \\ & \mathbb{D}_{\mathcal{X}}(x,\widehat{x}_i) + \mathbb{D}_{\mathcal{Y}}(y,\widehat{y}_i) \le \rho \quad \forall (x,y) \in \operatorname{supp}(\pi_i) & \forall i \in [N] \\ & \sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) > 0. & \end{array} \right. \tag{7}$$

We now show that problem (7) now can be written as

$$f(\beta) = \left\{ \begin{array}{ll} \sup & \displaystyle\frac{1}{\sum_{i\in[N]} \alpha_i} \sum_{i\in[N]} \alpha_i v_i^\star(\beta) \\ \text{s.t.} & \alpha \in [0,1]^N \\ & \alpha_i = 1 \text{ if } \mathbb{D}_{\mathcal{X}}(x_0,\widehat{x}_i) + \rho \le \gamma \\ & \alpha_i = 0 \text{ if } \mathbb{D}_{\mathcal{X}}(x_0,\widehat{x}_i) > \rho + \gamma \\ & \sum_{i\in[N]} \alpha_i > 0, \end{array} \right. \tag{8}$$

where the value $v_i^\star(\beta)$ is calculated as

$$v_i^\star(\beta) = \sup\left\{\ell(y_i,\beta)\ : y_i \in \mathcal{Y},\ \mathbb{D}_{\mathcal{Y}}(y_i,\widehat{y}_i) \le \rho - \mathbb{D}_{\mathcal{X}}(\widehat{x}_i^p,\widehat{x}_i)\right\}.$$

The equivalence between the supremum problems (7) and (8) can be shown in two steps. First, for (7) $\le$ (8), given any feasible solution of (7), one can construct a feasible solution of (8) using $\alpha_i = \pi_i(\mathcal{N}_\gamma(x_0) \times \mathcal{Y})$. For this candidate we have

$$\frac{\sum_{i\in[N]} \int_{\mathcal{Y}} \ell(y,\beta)\ \pi_i(\mathcal{N}_\gamma(x_0) \times \mathrm{d}y)}{\sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times \mathcal{Y})} \le \frac{\sum_{i\in[N]} \alpha_i \ell(y_i^\star,\beta)}{\sum_{i\in[N]} \alpha_i}.$$

Alternatively, given a feasible solution for (8), one can construct the following feasible solution for (7): for any $\epsilon > 0$, let $y_i^\epsilon \in \mathcal{Y}$ be such that $\mathbb{D}_\mathcal{Y}(y_i^\epsilon, \widehat{y}_i) \leq \rho - \mathbb{D}_\mathcal{X}(x_0, \widehat{x}_i)$ and $\ell(y_i^\epsilon, \beta) \geq v_i^\star(\beta) - \epsilon$, and let

$$\forall i \in [N]: \quad \pi_i^\epsilon = \begin{cases} \delta_{(\widehat{x}_i^p, y_i^\epsilon)} & \text{if } \mathbb{D}_\mathcal{X}(x_0, \widehat{x}_i) + \rho \leq \gamma, \\ \alpha_i \delta_{(\widehat{x}_i^p, y_i^\epsilon)} + (1-\alpha_i)\delta_{(x_i^r, \widehat{y}_i)} & \text{if } \mathbb{D}_\mathcal{X}(x_0, \widehat{x}_i) > \rho + \gamma, \\ \delta_{(\widehat{x}_i, \widehat{y}_i)} & \text{otherwise,} \end{cases}$$

where x_i^r is any point such that $\mathbb{D}_\mathcal{X}(x_i^r, \widehat{x}_i) \leq \rho$ and $x_i^r \notin \mathcal{N}_\gamma(x_0)$. Again, this candidate is feasible in (7) and we have that

$$\begin{aligned} f(\beta) &\geq \sup_{\epsilon > 0} \frac{\sum_{i\in[N]} \int_\mathcal{Y} \ell(y, \beta)\ \pi_i^\epsilon(\mathcal{N}_\gamma(x_0) \times \mathrm{d}y)}{\sum_{i\in[N]} \pi_i^\epsilon(\mathcal{N}_\gamma(x_0) \times \mathcal{Y})} \\ &\geq \sup_{\epsilon > 0} \frac{\sum_{i\in[N]} \alpha_i(\ell(y_i^\star, \beta) - \epsilon)}{\sum_{i\in[N]} \alpha_i} \\ &= \frac{\sum_{i\in[N]} \alpha_i \ell(y_i^\star, \beta)}{\sum_{i\in[N]} \alpha_i} = \frac{\sum_{i\in[N]} \alpha_i v_i^\star(\beta)}{\sum_{i\in[N]} \alpha_i}. \end{aligned}$$

Let $\mathcal{I}$ and $\mathcal{I}_1$ be the index sets defined as in (4a)–(4b), the value $f(\beta)$ is equal to the optimal value of a fractional linear program

$$f(\beta) = \max \left\{ \frac{\sum_{i\in\mathcal{I}} v_i^\star(\beta)\alpha_i}{\sum_{i\in\mathcal{I}} \alpha_i} : \alpha \in [0,1]^N,\ \alpha_i = 1\ \forall i \in \mathcal{I}_1,\ \sum_{i\in\mathcal{I}} \alpha_i > 0 \right\} \tag{9a}$$

$$= \max \left\{ \frac{\sum_{i\in\mathcal{I}_1} v_i^\star(\beta) + \sum_{i\in\mathcal{I}_2} v_i^\star(\beta)\alpha_i}{|\mathcal{I}_1| + \sum_{i\in\mathcal{I}_2} \alpha_i} : \alpha \in [0,1]^N,\ \alpha_i = 1\ \forall i \in \mathcal{I}_1,\ |\mathcal{I}_1| + \sum_{i\in\mathcal{I}_2} \alpha_i > 0 \right\}. \tag{9b}$$

Notice that the objective function and the constraints of (9b) depend only on α_i for $i \in \mathcal{I}$. Suppose that $\mathcal{I}_1 \neq \emptyset$, Lemma 1 indicates that the optimal solution $\alpha^\star$ that solves (9b) is

$$\forall i \in \mathcal{I}: \quad \alpha_i^\star = \begin{cases} 1 & \text{if } i \in \mathcal{I}_1, \\ 1 & \text{if } v_i^\star(\beta) > \dfrac{\sum_{i\in\mathcal{I}_1} v_i^\star(\beta) + \sum_{j: v_j^\star(\beta) > v_i^\star(\beta)} v_j^\star(\beta)}{|\mathcal{I}_1| + |\{j : v_j^\star(\beta) > v_i^\star(\beta)\}|}, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Suppose that $\mathcal{I}_1 = \emptyset$, then the optimal solution of problem (9b) is

$$\forall i \in \mathcal{I}: \quad \alpha_i^\star = \begin{cases} 1 & \text{if } v_i^\star(\beta) \geq \max_{j\in\mathcal{I}_2} v_j^\star(\beta), \\ 0 & \text{otherwise.} \end{cases}$$

Combining the above two cases, we can rewrite the optimal value of α that solves (9b) as in the statement of the theorem. This completes the proof. □

Proof of Corollary 1. Because $\mathbb{D}_\mathcal{Y}$ is an absolute distance, we have

$$\{y_i \in \mathcal{Y} : |y_i - \widehat{y}_i| \leq \rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i)\} = [\max\{a, \widehat{y}_i - \rho + \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i)\}, \min\{b, \widehat{y}_i + \rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i)\}],$$

where the equality follows from $\mathcal{Y} = [a, b]$. Because both the $\|\cdot\|_2^2$ and the quantile loss functions are convex, the value $v_i^\star(\beta)$ is thus attained at the extreme points of the interval. Calculating the value of $\ell(\cdot, \beta)$ at these two endpoints and taking the maximum between them completes the proof. □

Before proving Proposition 2, we need the following two results which asserts the analytical optimal value of maximizing a convex quadratic functions over a norm ball. These results can be found in the literature, the proof is included here for completeness.

Lemma 2 (Convex quadratic maximization over a norm ball) *For any $\beta \in \mathbb{R}^m$, $\widehat{y} \in \mathbb{R}^m$ and $r \in \mathbb{R}_+$, the following assertions hold.*

(i) Over a $\|\cdot\|_2$ ball, we have

$$\sup\left\{\|y-\beta\|_2^2 : \ \|y-\widehat{y}\|_2^2 \leq r^2\right\} = (r + \|\widehat{y}-\beta\|_2)^2.$$

(ii) Over a $\|\cdot\|_\infty$ ball, we have

$$\sup\left\{\|y-\beta\|_2^2 : \ \|y-\widehat{y}\|_\infty \leq r\right\} = \sum_{j\in[m]} \max\left\{(\widehat{y}_j - \beta_j - r)^2, (\widehat{y}_j - \beta_j + r)^2\right\},$$

where β_j and $\widehat{y}_j$ denote the j-th element of the vector β and $\widehat{y}$, respectively.

Proof of Lemma 2. We first prove Assertion (i). First, the optimal value is upper bounded by $(r + \|\widehat{y}-\beta\|_2)^2$ because

$$\|y-\beta\|_2 \leq \|y-\widehat{y}\|_2 + \|\widehat{y}-\beta\|_2 \leq r + \|\widehat{y}-\beta\|_2$$

by triangle inequality. Yet, it is equal to that amount since that amount is attained when $y = \widehat{y} + r(\widehat{y}-\beta)/\|\widehat{y}-\beta\|_2$.

Consider now Assertion (ii). Using a change of variables $z \leftarrow y - \beta$ and a change of parameters $w \leftarrow \widehat{y} - \beta$, we find

$$\sup\left\{\|y-\beta\|_2^2 : \ \|y-\widehat{y}\|_\infty \leq r\right\} = \max\left\{\|z\|_2^2 : \|z-w\|_\infty \leq r\right\}, \tag{11}$$

where the maximization operators are justified by Weierstrass' maximum value theorem [1, Theorem 2.43] because the feasible set is compact and the objective function is continuous. By extending the norm constraint into the vector form, we have the equivalence

$$\max\left\{\|z\|_2^2 : w - r\mathbb{1}_m \leq z \leq w + r\mathbb{1}_m\right\},$$

where the inequalities in the constraints are understood as element-wise inequalities, and $\mathbb{1}_m$ is an m-dimensional vector of ones. This maximization problem is separable in the decision variables and can be decomposed into m independent univariate subproblems of the form

$$\max\left\{z_j^2 : w_j - r \leq z_j \leq w_j + r\right\}$$

for each $j \in [m]$. It is easy to verify that the optimal value of each univariate subproblem is equal to

$$\max\left\{(w_j - r)^2, (w_j + r)^2\right\},$$

and summing up the optimal values over j completes the proof. □

We are now ready to prove Proposition 2.

Proof of Proposition 2. Following from equation (9a) in the proof of Theorem 1, we have

$$f(\beta) = \max\left\{\frac{\sum_{i\in\mathcal{I}} v_i^\star(\beta)\alpha_i}{\sum_{i\in\mathcal{I}}\alpha_i} : \alpha \in [0,1]^N,\ \alpha_i = 1\ \forall i \in \mathcal{I}_1,\ \sum_{i\in\mathcal{I}}\alpha_i > 0\right\}$$

By applying the Charnes-Cooper transformation [10] with

$$z_i = \frac{\alpha_i}{\sum_{i\in\mathcal{I}} \alpha_i}, \quad \text{and} \quad t = \frac{1}{\sum_{i\in\mathcal{I}} \alpha_i}$$

to reformulate this fractional linear problem, we have

$$f(\beta) = \left\{ \begin{array}{lll} \max & \sum_{i\in\mathcal{I}} v_i^\star(\beta) z_i & \\ \text{s.t.} & \sum_{i\in\mathcal{I}} z_i = 1,\ t \geq 0 & \\ & z_i - t = 0 & \forall i \in \mathcal{I}_1 \\ & 0 \leq z_i \leq t & \forall i \in \mathcal{I}_2. \end{array} \right.$$
$$= \left\{ \begin{array}{ll} \min & \lambda \\ \text{s.t.} & \lambda \in \mathbb{R},\ u_i \in \mathbb{R}\ \forall i \in \mathcal{I}_1,\ u_i \in \mathbb{R}_+\ \forall i \in \mathcal{I}_2 \\ & \lambda + u_i \geq v_i^\star(\beta) \quad \forall i \in \mathcal{I} \\ & \sum_{i\in\mathcal{I}} u_i \leq 0, \end{array} \right.$$

where the second equality follows from linear programming duality. Using the last minimization reformulation of $f(\beta)$, problem (2) is now equivalent to

$$\min_\beta\ f(\beta) = \left\{ \begin{array}{ll} \min & \lambda \\ \text{s.t.} & \beta \in \mathbb{R}^m,\ \lambda \in \mathbb{R},\ u_i \in \mathbb{R}\ \forall i \in \mathcal{I}_1,\ u_i \in \mathbb{R}_+\ \forall i \in \mathcal{I}_2 \\ & \lambda + u_i \geq v_i^\star(\beta) \quad \forall i \in \mathcal{I} \\ & \sum_{i\in\mathcal{I}} u_i \leq 0, \end{array} \right.$$

When $\mathbb{D}_\mathcal{Y}$ is a 2-norm, each value $v_i^\star(\beta)$ calculated from (5) becomes

$$v_i^\star(\beta) = \sup\left\{ \|y-\beta\|_2^2 :\ \|y-\widehat{y}_i\|_2 \leq \rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i) \right\} \qquad \forall i \in [N].$$

For any $i \in \mathcal{I}$, the value $v_i^\star(\beta)$ is finite and $v_i^\star(\beta)$ can be re-expressed by exploiting Lemma 2(i) as

$$v_i^\star(\beta) = \left(\rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i) + \|\widehat{y}_i - \beta\|_2\right)^2.$$

Problem (2) is now equivalent to

$$\begin{array}{ll} \min & \lambda \\ \text{s.t.} & \beta \in \mathbb{R}^m,\ \lambda \in \mathbb{R},\ u_i \in \mathbb{R}\ \forall i \in \mathcal{I}_1,\ u_i \in \mathbb{R}_+\ \forall i \in \mathcal{I}_2 \\ & \lambda + u_i \geq \left(\rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i) + \|\widehat{y}_i - \beta\|_2\right)^2 \quad \forall i \in \mathcal{I} \\ & \sum_{i\in\mathcal{I}} u_i \leq 0. \end{array} \tag{12}$$

To obtain a second-order cone program formulation, it now suffices to add the hypergraph formulation $t_i \geq \|\widehat{y}_i - \beta\|_2$ with $t_i \geq 0$, and reformulate the quadratic constraint into a second-order cone constraint using results from [2, Section 2]. This completes the proof for claim (i).

We now proceed to prove claim (ii). When $\mathbb{D}_\mathcal{Y}$ is the ∞-norm, each value $v_i^\star(\beta)$ becomes

$$v_i^\star(\beta) = \sup\left\{ \|y-\beta\|_2^2 :\ \|y-\widehat{y}_i\|_\infty \leq \rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i) \right\} \qquad \forall i \in [N].$$

For any $i \in \mathcal{I}$, the value $v_i^\star(\beta)$ is finite and $v_i^\star(\beta)$ can be re-expressed using Lemma 2(ii) as

$$v_i^\star(\beta) = \sum_{j\in[m]} \max\left\{ (\widehat{y}_{ij} - \beta_j - \rho + \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i))^2, (\widehat{y}_{ij} - \beta_j + \rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i))^2 \right\}.$$

By adding auxiliary variables T_{ij} with the constraints

$$(\widehat{y}_{ij} - \beta_j - \rho + \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i))^2 \leq T_{ij}^2, \quad \text{and} \quad (\widehat{y}_{ij} - \beta_j + \rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i^p, \widehat{x}_i))^2 \leq T_{ij}^2,$$

problem (2) is now equivalent to

$$\begin{array}{ll}\min & \lambda \\ \text{s.t.} & \beta \in \mathbb{R}^m,\ \lambda \in \mathbb{R},\ T \in \mathbb{R}_+^{|\mathcal{I}|\times m},\ u_i \in \mathbb{R}\ \forall i \in \mathcal{I}_1,\ u_i \in \mathbb{R}_+\ \forall i \in \mathcal{I}_2 \\ & \sum_{i\in\mathcal{I}} u_i \leq 0 \\ & \lambda + u_i \geq \sum_{j\in[m]} T_{ij}^2 \quad \forall i \in \mathcal{I} \\ & (\widehat{y}_{ij} - \beta_j - \rho + \mathbb{D}_{\mathcal{X}}(\widehat{x}_i^p, \widehat{x}_i))^2 \leq T_{ij}^2 & \forall (i,j) \in \mathcal{I}\times[m] \\ & (\widehat{y}_{ij} - \beta_j + \rho - \mathbb{D}_{\mathcal{X}}(\widehat{x}_i^p, \widehat{x}_i))^2 \leq T_{ij}^2 & \forall (i,j) \in \mathcal{I}\times[m].\end{array}$$

The last two constraints can be re-expressed as linear constraints of the form

$$\begin{array}{ll} -T_{ij} \leq \widehat{y}_{ij} - \beta_j - \rho + \mathbb{D}_{\mathcal{X}}(\widehat{x}_i^p, \widehat{x}_i) \leq T_{ij} & \forall (i,j) \in \mathcal{I}\times[m] \\ -T_{ij} \leq \widehat{y}_{ij} - \beta_j + \rho - \mathbb{D}_{\mathcal{X}}(\widehat{x}_i^p, \widehat{x}_i) \leq T_{ij} & \forall (i,j) \in \mathcal{I}\times[m].\end{array}$$

Formulating the quadratic constraint $\lambda + u_i \geq \sum_{j\in[m]} T_{ij}^2$ using [2, Section 2] completes the proof. □

Proof of Proposition 3. For the purpose of this proof, define the following sets

$$\mathcal{Y}_i \triangleq \{y_i \in \mathcal{Y} : \mathbb{D}_{\mathcal{Y}}(y_i, \widehat{y}_i) \leq \rho - \mathbb{D}_{\mathcal{X}}(\widehat{x}_i^p, \widehat{x}_i)\} \qquad \forall i \in \mathcal{I}.$$

Because $\mathbb{D}_{\mathcal{Y}}$ is coercive and continuous, each set $\mathcal{Y}_i$ is compact. Because the loss function is continuous, there thus exists $y_i^\star$ satisfying $y_i^\star \in \mathcal{Y}_i$ and $\ell(y_i^\star, \beta) = v_i^\star(\beta)$ for any $i \in \mathcal{I}$. Following from Equation (9a) in the proof of Theorem 1, we have

$$\begin{aligned} f(\beta) &= \max\ \left\{ \frac{\sum_{i\in\mathcal{I}} v_i^\star(\beta)\alpha_i}{\sum_{i\in\mathcal{I}} \alpha_i} : \alpha \in [0,1]^N,\ \alpha_i = 1\ \forall i \in \mathcal{I}_1,\ \sum_{i\in\mathcal{I}} \alpha_i > 0 \right\} \\ &= \max\ \left\{ \frac{\sum_{i\in\mathcal{I}} \ell(y_i, \beta)\alpha_i}{\sum_{i\in\mathcal{I}} \alpha_i} : \alpha \in [0,1]^N,\ \alpha_i = 1\ \forall i \in \mathcal{I}_1,\ \sum_{i\in\mathcal{I}} \alpha_i > 0,\ y_i \in \mathcal{Y}_i\ \forall i \in \mathcal{I} \right\}. \end{aligned}$$

If $\mathcal{I}_1 = \emptyset$, then we have

$$f(\beta) = \ell(y_{i^\star}, \beta) \quad \forall i^\star \in \arg\max_{i\in\mathcal{I}_2} v_i^\star(\beta),$$

and a subgradient of f is $\partial f(\beta) = \partial_\beta \ell(y_{i^\star}, \beta)$ for any $i^\star \in \arg\max_{i\in\mathcal{I}_2} v_i^\star(\beta)$. By incorporating the optimal value of α in the statement of Theorem 1, we have $\partial f(\beta) = \alpha_i \partial_\beta \ell(y_i^\star, \beta)$.

If $\mathcal{I}_1 \neq \emptyset$, then we have

$$\begin{aligned} f(\beta) &= \max\ \left\{ \frac{\sum_{i\in\mathcal{I}_1} v_i^\star(\beta) + \sum_{i\in\mathcal{I}_2} v_i^\star(\beta)\alpha_i}{|\mathcal{I}_1| + \sum_{i\in\mathcal{I}_2} \alpha_i} : \alpha \in [0,1]^N,\ \alpha_i = 1\ \forall i \in \mathcal{I}_1 \right\} \\ &= \max\ \left\{ \frac{\sum_{i\in\mathcal{I}_1} \ell(y_i, \beta) + \sum_{i\in\mathcal{I}_2} \ell(y_i, \beta)\alpha_i}{|\mathcal{I}_1| + \sum_{i\in\mathcal{I}_2} \alpha_i} : \alpha \in [0,1]^N,\ \alpha_i = 1\ \forall i \in \mathcal{I}_1,\ y_i \in \mathcal{Y}_i\ \forall i \in \mathcal{I} \right\} \end{aligned}$$

Notice that the function

$$\beta \mapsto \frac{\sum_{i\in\mathcal{I}_1} \ell(y_i, \beta) + \sum_{i\in\mathcal{I}_2} \ell(y_i, \beta)\alpha_i}{|\mathcal{I}_1| + \sum_{i\in\mathcal{I}_2} \alpha_i}$$

is convex for any feasible value of (α, y) in the above optimization problem. Moreover, by Tychonoff's theorem [1, Theorem 2.61], the feasible set of the above optimization problem is a compact set in the product topology. One can now apply [3, Proposition A.22] to conclude that a subgradient of f in this case is

$$\partial f(\beta) = \frac{\sum_{i\in\mathcal{I}_1} \partial_\beta \ell(y_i, \beta) + \sum_{i\in\mathcal{I}_2} \partial_\beta \ell(y_i, \beta)\alpha_i}{|\mathcal{I}_1| + \sum_{i\in\mathcal{I}_2} \alpha_i}.$$

Combining the two cases, we have the postulated result. □

B.2 Proofs of Section 3

Proof of Proposition 4. Under the conditions of the proposition, we have $\mathbb{P}(X \in \mathcal{N}_\gamma(x_0)) > 0$ because $\mathbb{P}$ admits a density, and that $\mathcal{N}_\gamma(x_0) \cap \mathcal{X}$ is a set with non-empty interior for any $\gamma > 0$. The proof now follows trivially from [16, Theorem 1.1]. Indeed, under the conditions of the proposition, with probability of at least $1 - O(N^{-c})$, we have $\mathbb{P} \in \mathbb{B}_\rho^\infty$, and hence the bound follows. □

Proof of Example 1. For the purpose of this proof, we let $\mathbb{P}^\infty = \mathbb{P} \otimes \mathbb{P} \otimes \cdots$ be the joint distribution of $(\widehat{x}_1, \widehat{y}_1), (\widehat{x}_2, \widehat{y}_2), \cdots$. The selection of parameter $\gamma = 0$ implies that $\mathcal{I} = \mathcal{I}_2$, and for any fixed $\rho > 0$ we have $\mathbb{P}^\infty \left(\lim_{N\to\infty} |\mathcal{I}| = +\infty\right) = 1$ by Borel-Cantelli lemma. In this example, the DRO problem is feasible if $\mathcal{I}$ is nonempty, and we have an explicit optimal solution

$$\beta_N^\star = \frac{1}{2} \min_{i\in\mathcal{I}} \{\widehat{y}_i - \rho + \mathbb{D}_\mathcal{X}(\widehat{x}_i, x_0)\} + \frac{1}{2} \max_{i\in\mathcal{I}} \{\widehat{y}_i + \rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i, x_0)\}$$

Notice that with probability 1 we have

$$\min_{i\in\mathcal{I}} \{\widehat{y}_i - \rho + \mathbb{D}_\mathcal{X}(\widehat{x}_i, x_0)\} \geq -\rho \text{ and } \max_{i\in\mathcal{I}} \{\widehat{y}_i + \rho - \mathbb{D}_\mathcal{X}(\widehat{x}_i, x_0)\} \geq \max_{i\in\mathcal{I}} \{\widehat{y}_i\}.$$

Consequently we have $\beta_N^\star \geq \frac{1}{2} \max_{i\in\mathcal{I}} \{\widehat{y}_i\} - \frac{1}{2}\rho$. For all $y > 0$, we have

$$\begin{aligned}\mathbb{P}^\infty \left(\lim_{N\to\infty} \beta_N^\star > y\right) &\geq \mathbb{P}^\infty \left(\lim_{N\to\infty} \max_{i\in\mathcal{I}} \{\widehat{y}_i\} > 2y + \rho\right) = \lim_{N\to\infty} \mathbb{P}^\infty \left(\max_{i\in\mathcal{I}} \{\widehat{y}_i\} > 2y + \rho\right) \\ &= \lim_{N\to\infty} 1 - \mathbb{P}(Y \leq 2y + \rho)^{|\mathcal{I}|} = 1.\end{aligned}$$

Let y tend to infinity concludes the proof. □

Before proving Proposition 5, we first present the following minimax result.

Lemma 3 (Minimax result) *Suppose that $\ell(y, \cdot)$ is convex and coercive for any $y \in \mathcal{Y}$, and that $\mathbb{D}_\mathcal{Y}(\cdot, \widehat{y})$ is convex and coercive for any $\widehat{y}$. For any $\rho \geq \min_{i\in[N]} \kappa_{i,\gamma}$, we have*

$$\begin{aligned}\min_{\beta\in\mathbb{R}^m} \sup_{\mathbb{Q}\in\mathbb{B}_\rho^\infty, \mathbb{Q}(X\in\mathcal{N}_\gamma(x_0))>0} &\mathbb{E}_\mathbb{Q}\left[\ell(Y, \beta) | X \in \mathcal{N}_\gamma(x_0)\right] \\ &= \sup_{\mathbb{Q}\in\mathbb{B}_\rho^\infty, \mathbb{Q}(X\in\mathcal{N}_\gamma(x_0))>0} \min_{\beta\in\mathbb{R}^m} \mathbb{E}_\mathbb{Q}\left[\ell(Y, \beta) | X \in \mathcal{N}_\gamma(x_0)\right].\end{aligned}$$

To facilitate the proof of Lemma 3, we define the following conditional ambiguity set induced by $\mathbb{B}_\rho^\infty$ as

$$\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty) \triangleq \left\{\mu_0 \in \mathcal{M}(\mathcal{Y}) : \begin{array}{l}\exists \mathbb{Q} \in \mathbb{B}_\rho^\infty,\ \mathbb{Q}(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) > 0 \\ \mathbb{Q}(\mathcal{N}_\gamma(x_0) \times A) = \mu_0(A)\, \mathbb{Q}(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) \quad \forall A \subseteq \mathcal{Y} \text{ measurable}\end{array}\right\}, \quad (13)$$

where the last constraint defining the set $\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)$ is from the dis-integration of the joint measure into a marginal distribution and the corresponding conditional distributions [37, Theorem 9.2.2].

The proof of Lemma 3 relies on the following two results which assert the convexity of the joint ambiguity set $\mathbb{B}_\rho^\infty$ and its induced conditional ambiguity set $\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)$.

Lemma 4 (Convexity of $\mathbb{B}_\rho^\infty$) *The ambiguity set $\mathbb{B}_\rho^\infty$ is convex.*

Proof of Lemma 4. Because the nominal probability measure is an empirical measure, the ambiguity set $\mathbb{B}_\rho^\infty$ can be represented as

$$\mathbb{B}_\rho^\infty = \left\{\mathbb{Q} \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) : \begin{array}{l}\exists \pi_i \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})\ \forall i \in [N] \text{ such that :} \\ \mathbb{Q} = N^{-1} \sum_{i\in[N]} \pi_i,\ \ \sum_{i\in[N]} \pi_i(\mathcal{N}_\gamma(x_0) \times \mathcal{Y}) > 0 \\ \mathbb{D}_\mathcal{X}(x, \widehat{x}_i) + \mathbb{D}_\mathcal{Y}(y, \widehat{y}_i) \leq \rho \quad \forall (x, y) \in \operatorname{supp}(\pi_i) \quad \forall i \in [N]\end{array}\right\}.$$

Pick any arbitrary $\mathbb{Q}^0$ and $\mathbb{Q}^1$ from $\mathbb{B}_\rho^\infty$. Associated with $\mathbb{Q}^j$, $j \in \{0,1\}$ is a collection of probability measures $\{\pi_i^j\} \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})^N$ satisfying

$$\begin{cases} \mathbb{Q}^j = N^{-1}\sum_{i\in[N]}\pi_i^j, \ \sum_{i\in[N]}\pi_i^j(\mathcal{N}_\gamma(x_0)\times\mathcal{Y}) > 0 \\ \mathbb{D}_\mathcal{X}(x,\widehat{x}_i)+\mathbb{D}_\mathcal{Y}(y,\widehat{y}_i)\le\rho \quad \forall(x,y)\in\mathrm{supp}(\pi_i^j) \quad \forall i\in[N]. \end{cases}$$

Consider any convex combination $\mathbb{Q}^\lambda = \lambda\mathbb{Q}^1 + (1-\lambda)\mathbb{Q}^0$ for $\lambda\in(0,1)$. It is easy to verify that the joint measure $\pi_i^\lambda = \lambda\pi_i^1 + (1-\lambda)\pi_i^0$ for any $i\in[N]$ satisfies

$$\begin{cases} \mathbb{Q}^\lambda = N^{-1}\sum_{i\in[N]}\pi_i^\lambda, \ \sum_{i\in[N]}\pi_i^\lambda(\mathcal{N}_\gamma(x_0)\times\mathcal{Y}) > 0 \\ \mathbb{D}_\mathcal{X}(x,\widehat{x}_i)+\mathbb{D}_\mathcal{Y}(y,\widehat{y}_i)\le\rho \quad \forall(x,y)\in\mathrm{supp}(\pi_i^\lambda) \quad \forall i\in[N], \end{cases}$$

where the last constraint is satisfied by noticing that $\mathrm{supp}(\pi_i^\lambda) = \mathrm{supp}(\pi_i^0)\cup\mathrm{supp}(\pi_i^1)$. This observation implies that $\mathbb{Q}^\lambda\in\mathbb{B}_\rho^\infty$. □

Lemma 5 (Convexity of $\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)$) *The conditional ambiguity set $\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)$ is convex.*

Proof of Lemma 5. Let $\mu_0^0,\mu_0^1\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)$ be two arbitrary probability measures. Associated with each μ_0^j, $j\in\{0,1\}$, is a corresponding joint measure $\mathbb{Q}^j\in\mathcal{M}(\mathcal{X}\times\mathcal{Y})$ such that

$$\mathbb{Q}^j(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})>0 \quad \text{and} \quad \frac{\mathbb{Q}^j(\mathcal{N}_\gamma(x_0)\times A)}{\mathbb{Q}^j(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})} = \mu_0^j(A) \quad \forall A\subseteq\mathcal{Y}\text{ measurable}.$$

Select any $\lambda\in(0,1)$. We proceed to show that $\mu_0^\lambda = \lambda\mu_0^1+(1-\lambda)\mu_0^0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)$. Indeed, consider the joint measure

$$\mathbb{Q}^\lambda = \theta\mathbb{Q}^1 + (1-\theta)\mathbb{Q}^0$$

with θ being defined as

$$\theta = \frac{\lambda\mathbb{Q}^0(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})}{\lambda\mathbb{Q}^0(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})+(1-\lambda)\mathbb{Q}^1(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})}\in[0,1].$$

By definition, we have $\mathbb{Q}^\lambda(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})>0$, and by convexity of $\mathbb{B}_\rho^\infty$ from Lemma 4, we have $\mathbb{Q}^\lambda\in\mathbb{B}_\rho^\infty$. Moreover, we have for any set $A\subseteq\mathcal{Y}$ measurable,

$$\begin{aligned} \frac{\mathbb{Q}^\lambda(\mathcal{N}_\gamma(x_0)\times A)}{\mathbb{Q}^\lambda(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})} &= \frac{\theta\mathbb{Q}^1(\mathcal{N}_\gamma(x_0)\times A)+(1-\theta)\mathbb{Q}^0(\mathcal{N}_\gamma(x_0)\times A)}{\theta\mathbb{Q}^1(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})+(1-\theta)\mathbb{Q}^0(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})} \\ &= \frac{\lambda\mathbb{Q}^0(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})\mathbb{Q}^1(\mathcal{N}_\gamma(x_0)\times A)+(1-\lambda)\mathbb{Q}^1(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})\mathbb{Q}^0(\mathcal{N}_\gamma(x_0)\times A)}{\mathbb{Q}^0(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})\mathbb{Q}^1(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})} \\ &= \frac{\lambda\mathbb{Q}^1(\mathcal{N}_\gamma(x_0)\times A)}{\mathbb{Q}^1(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})}+\frac{(1-\lambda)\mathbb{Q}^0(\mathcal{N}_\gamma(x_0)\times A)}{\mathbb{Q}^0(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})} \\ &= \lambda\mu_0^1(A)+(1-\lambda)\mu_0^0(A), \end{aligned}$$

where the second equality follows from the definition of θ. This implies that $\mu_0^\lambda\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)$, and further implies the convexity of $\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)$. □

We are now ready to prove Lemma 3.

Proof of Lemma 3. By the definition of the conditional ambiguity set $\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)$, it suffices to prove the equivalence

$$\min_{\beta\in\mathbb{R}^m}\sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)}\mathbb{E}_{\mu_0}[\ell(Y,\beta)] = \sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}_\rho^\infty)}\min_{\beta\in\mathbb{R}^m}\mathbb{E}_{\mu_0}[\ell(Y,\beta)].$$

First, consider the mapping $\beta \mapsto \sup_{\mu_0 \in \mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)} \mathbb{E}_{\mu_0}[\ell(Y,\beta)]$. The properties of ℓ implies that this mapping is lower semi-continuous and coercive. As a consequence, without loss of optimality, we can restrict the feasible set β to some convex, compact ball $\mathcal{S} \triangleq \{\beta : \|\beta\|_2 \leq R\}$ for some radius $R \in \mathbb{R}_{++}$ sufficiently big.

We now consider the mapping $\mu_0 \mapsto \mathbb{E}_{\mu_0}[\ell(Y,\beta)]$ parametrized by β. For any β, it is a linear function of μ_0, and hence it is concave. It is also weakly continuous. To see this, notice that when $\mathbb{D}(\cdot,\widehat{y})$ is coercive, the set

$$\mathcal{A} \triangleq \bigcup_{i\in[N]} \{y : \mathbb{D}_{\mathcal{Y}}(y,\widehat{y}_i) \leq \rho\},$$

being a finite union of bounded sets, is bounded. Pick any $\mathbb{Q} \in \mathbb{B}^\infty_\rho$, by the definition of the type-∞ Wasserstein distance, we have $\mathbb{Q}(\mathcal{A}) = 1$. Consider the conditional measure $\mu_0^{\mathbb{Q}}$ induced by $\mathbb{Q}$, then we have

$$\mu_0^{\mathbb{Q}}(\mathcal{A}\cap\mathcal{Y}) = \frac{\mathbb{Q}(\mathcal{N}_\gamma(x_0)\times(\mathcal{A}\cap\mathcal{Y}))}{\mathbb{Q}(\mathcal{N}_\gamma(x_0)\times\mathcal{Y})} \geq \frac{\mathbb{Q}(\mathcal{N}_\gamma(x_0)\times(\mathcal{A}\cap\mathcal{Y}))}{\mathbb{Q}(\mathcal{N}_\gamma(x_0)\times(\mathcal{A}\cap\mathcal{Y}))} = 1,$$

which implies that $\mu_0^{\mathbb{Q}}$ has a bounded support. This implies that $\mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho) \subseteq \mathcal{M}(\mathcal{A})$, where $\mathcal{M}(\mathcal{A})$ is the set of all probability measures supported on a bounded set $\mathcal{A}$. Because $\ell(\cdot,\beta)$ is continuous, there exists a bound $U \in \mathbb{R}_{++}$ such that $|\ell(y,\beta)| \leq U$ for every $y \in \mathcal{A}$. Define now the function $\ell_U(\cdot,\beta) = \max\{-U, \min\{\ell(\cdot,\beta), U\}\}$, which is continuous and bounded. Consider any sequence of conditional measures $\{\mu_0^k\} \in \mathcal{M}(\mathcal{A})$ that weakly converges to μ_0^∞, we have

$$\lim_{k\uparrow\infty} \mathbb{E}_{\mu_0^k}[\ell(Y,\beta)] = \lim_{k\uparrow\infty} \mathbb{E}_{\mu_0^k}[\ell_U(Y,\beta)] = \mathbb{E}_{\mu_0^\infty}[\ell_U(Y,\beta)] = \mathbb{E}_{\mu_0^\infty}[\ell(Y,\beta)],$$

which implies that the function $\mu_0 \mapsto \mathbb{E}_{\mu_0}[\ell(Y,\beta)]$ is weakly continuous over $\mathcal{M}(\mathcal{A})$.

This line of argument suggests that

$$\begin{aligned}
\min_{\beta} \sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)} \mathbb{E}_{\mu_0}[\ell(Y,\beta)] &= \min_{\beta:\|\beta\|_2\leq R} \sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)} \mathbb{E}_{\mu_0}[\ell(Y,\beta)] \\
&= \sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)} \min_{\beta:\|\beta\|_2\leq R} \mathbb{E}_{\mu_0}[\ell(Y,\beta)] \qquad (14a)\\
&= \sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)} \min_{\beta} \mathbb{E}_{\mu_0}[\ell(Y,\beta)], \qquad (14b)
\end{aligned}$$

where equality (14b) follows from the coercivity of the loss function, thus the constraint on β can be dropped for R sufficiently big. Equality (14a) holds by Sion's minimax theorem [35]. This finishes the proof. □

Proof of Proposition 5. Because the loss function is coercive and convex in β, we have

$$\begin{aligned}
\min_{\beta\in\mathbb{R}} \sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)} \mathbb{E}_{\mu_0}[(Y-\beta)^2] &= \sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)} \min_{\beta\in\mathbb{R}} \mathbb{E}_{\mu_0}[(Y-\beta)^2] \\
&= \sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)} \mathbb{E}_{\mu_0}[(Y-\mathbb{E}_{\mu_0}[Y])^2] \\
&= \text{Variance}_{\mu_0^\star}(Y),
\end{aligned}$$

where the first equality follows from Lemma 3, the second equality follows from the fact that for any $\mu_0 \in \mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)$, the estimate $\beta^\star(\mu_0) = \mathbb{E}_{\mu_0}[Y]$ minimizes the objective $\mathbb{E}_{\mu_0}[(Y-\beta)^2]$. The last equality follows from the definition of $\mu_0^\star$.

Let $\beta^\star$ be the optimal estimate that solves (2), we now have

$$\begin{aligned}
\text{Variance}_{\mu_0^\star}(Y) &= \sup_{\mu_0\in\mathcal{B}_{x_0,\gamma}(\mathbb{B}^\infty_\rho)} \mathbb{E}_{\mu_0}[(Y-\beta^\star)^2] \\
&\geq \mathbb{E}_{\mu_0^\star}[(Y-\beta^\star)^2] = \text{Variance}_{\mu_0^\star}(Y) + (\beta^\star - \mathbb{E}_{\mu_0^\star}[Y])^2,
\end{aligned}$$

where the last equality follows from the bias-variance decomposition. This implies that $\beta^\star = \mathbb{E}_{\mu_0^\star}[Y]$ and completes the proof. □

C Golden-section search for univariate conditional estimate

We elaborate here on the procedure of applying a golden-section search to solve a one-dimensional local conditional estimation with a convex loss function ℓ. We suppose that $\mathcal{Y} = [a, b]$ for some finite values $-\infty < a < b < \infty$, that $\ell(y, \cdot)$ is convex for every y and that we have access to an oracle that solves (5). Given any β, the worst-case conditional expected loss $f(\beta)$ can be computed using Theorem 1. Algorithm 1 can be used to find the optimal conditional estimate $\beta^\star$ to any arbitrary precision.

Algorithm 1 Golden-section Search Algorithm

Input: Range $[a, b] \in \mathbb{R}$, tolerance $\epsilon \in \mathbb{R}_{++}$
Initialization: Set $r \leftarrow 0.618$, $\beta_1 \leftarrow a$, $\beta_4 \leftarrow b$
while $|\beta_4 - \beta_1| > \epsilon$ **do**
  Set $\beta_2 \leftarrow r\beta_1 + (1-r)\beta_4$, $\beta_3 \leftarrow (1-r)\beta_1 + r\beta_4$
  if $f(\beta_2) \leq f(\beta_3)$ **then** Set $\beta_4 \leftarrow \beta_3$ **else** Set $\beta_1 \leftarrow \beta_2$ **endif**
end while
Set $\beta^\star \leftarrow (\beta_1 + \beta_4)/2$
Output: $\beta^\star$

References

[1] C. D. Aliprantis and K. C. Border. Infinite Dimensional Analysis: A Hitchhiker's Guide. Springer, 2006.

[2] F. Alizadeh and D. Goldfarb. Second-order cone programming. Mathematical Programming, 95:3–51, 2003.

[3] D. P. Bertsekas. Control of Uncertain Systems with a Set-Membership Description of Uncertainty. PhD thesis, Massachusetts Institute of Technology, 1971.

[4] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. Mathematical Programming, 167(2):235–292, 2018.

[5] D. Bertsimas, C. McCord, and B. Sturt. Dynamic optimization with side information. arXiv preprint arXiv:1907.07307, 2019.

[6] D. Bertsimas, S. Shtern, and B. Sturt. Two-stage sample robust optimization. arXiv preprint arXiv:1907.07142, 2019.

[7] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. Mathematics of Operations Research, 44(2):565–600, 2019.

[8] L. Breiman. Random forests. Machine Learning, 45:5–32, 2001.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Wadsworth and Brooks, 1984.

[10] A. Charnes and W. W. Cooper. Programming with linear fractional functionals. Naval Research Logistics Quarterly, 9(3–4):181–186, 1962.

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

[12] L. Devroye. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. IEEE Transactions on Information Theory, 24(2):142–151, 1978.

[13] V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications, 14(1):153–158, 1969.

[14] R. Flamary and N. Courty. Pot python optimal transport library, 2017.

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

[16] N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. Canadian Journal of Mathematics, 67(6):1358–1383, 2015.

[17] C. Givens and R. Shortt. A class of Wasserstein metrics for probability distributions. The Michigan Mathematical Journal, 31(2):231–240, 1984.

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

[19] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

[20] P. J. Huber. Robust estimation of a location parameter. In Breakthroughs in statistics, pages 492–518. Springer, 1992.

[21] S. Kruk and H. Wolkowicz. Pseudolinear programming. SIAM Review, 41(4):795–805, 1999.

[22] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. INFORMS TutORials in Operations Research, pages 130–166, 2019.

[23] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In Proceedings of the Fifth International Conference on Learning Representations, 2017.

[24] Y. LeCun and C. Cortes. The MNIST Database of Handwritten Digits, 1998 (accessed May 28, 2020).

[25] X. Li, Y. Chen, Y. He, and H. Xue. Advknn: Adversarial attacks on k-nearest neighbor classifiers with approximate gradients. arXiv preprint arXiv:1911.06591, 2019.

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

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

[28] MOSEK ApS. MOSEK Optimizer API for Python 9.2.10, 2019.

[29] E. A. Nadaraya. On estimating regression. Theory of Probability & Its Applications, 9(1):141–142, 1964.

[30] H. Namkoong and J. C. Duchi. Variance-based regularization with convex objectives. In Advances in Neural Information Processing Systems 30, pages 2971–2980, 2017.

[31] V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. arXiv preprint arXiv:1805.07194, 2018.

[32] A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. In International Conference on Learning Representations, 2018.

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

[34] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In International Conference on Learning Representations, 2018.

[35] M. Sion. On general minimax theorems. Pacific Journal of Mathematics, 8(1):171–176, 1958.

[36] C. J. Stone. Consistent nonparametric regression. Annals of Statistics, 5(4):595–620, 1977.

[37] D. Stroock. Probability Theory: An Analytic View. Cambridge University Press, 2011.

[38] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. D. McDaniel. Ensemble adversarial training: Attacks and defenses. In Proceedings of the Sixth International Conference on Learning Representations, 2018.

[39] G. S. Watson. Smooth regression analysis. Sankhyā: The Indian Journal of Statistics, Series A, pages 359–372, 1964.

[40] W. Xie. Tractable reformulations of distributionally robust two-stage stochastic programs with ∞-Wasserstein distance. arXiv preprint arXiv:1908.08454, 2019.

[41] G. Zhao and Y. Ma. Robust nonparametric kernel regression estimator. Statistics & Probability Letters, 116:72–79, 2016.