
Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Wasserstein Distributionally Robust Optimization and Variation Regularization

Rui Gao, Xi Chen, Anton J. Kleywegt

**To cite this article:**
Rui Gao, Xi Chen, Anton J. Kleywegt (2022) Wasserstein Distributionally Robust Optimization and Variation Regularization. Operations Research

Published online in Articles in Advance 01 Nov 2022

. https://doi.org/10.1287/opre.2022.2383

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
*Articles in Advance*, pp. 1–15
ISSN 0030-364X (print), ISSN 1526-5463 (online)



Methods

# Wasserstein Distributionally Robust Optimization and Variation Regularization

**Rui Gao,**[a,*] **Xi Chen,**[b] **Anton J. Kleywegt**[c]

[a] Department of Information, Risk and Operations Management, University of Texas at Austin, Austin, Texas 78712; [b] Stern School of Business, New York University, New York, New York 10012; [c] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332
*Corresponding author

**Contact:** rui.gao@mccombs.utexas.edu, https://orcid.org/0000-0003-0145-8577 (RG); xchen3@stern.nyu.edu, https://orcid.org/0000-0002-9049-9452 (XC); anton@isye.gatech.edu, https://orcid.org/0000-0002-1849-7276 (AJK)

**Received:** November 10, 2020
**Revised:** February 4, 2022
**Accepted:** August 21, 2022
**Published Online in Articles in Advance:** November 1, 2022

**Area of Review:** Optimization

**https://doi.org/10.1287/opre.2022.2383**



**Abstract.** Wasserstein distributionally robust optimization (DRO) is an approach to optimization under uncertainty in which the decision maker hedges against a set of probability distributions, specified by a Wasserstein ball, for the uncertain parameters. This approach facilitates robust machine learning, resulting in models that sustain good performance when the data are to some extent different from the training data. This robustness is related to the well-studied effect of regularization. The connection between Wasserstein DRO and regularization has been established in several settings. However, existing results often require restrictive assumptions, such as smoothness or convexity, that are not satisfied by many important problems. In this paper, we develop a general theory for the *variation regularization* effect of the Wasserstein DRO—a new form of regularization that generalizes total-variation regularization, Lipschitz regularization, and gradient regularization. Our results cover possibly nonconvex and nonsmooth losses and losses on non-Euclidean spaces and highlight the *bias-variation tradeoff* intrinsic in the Wasserstein DRO, which balances between the empirical mean of the loss and the *variation of the loss*. Example applications include multi-item newsvendor, linear prediction, neural networks, manifold learning, and intensity estimation for Poisson processes. We also use our theory of variation regularization to derive new generalization guarantees for adversarial robust learning.

**Funding:** X. Chen is supported by the National Science Foundation [Grant IIS-1845444].
**Supplemental Material:** The online appendix is available at https://doi.org/10.1287/opre.2022.2383.

**Keywords:** distributionally robust optimization • data-dependent regularization • Wasserstein metric • adversarial attack

## 1. Introduction

Wasserstein distributionally robust optimization (DRO) (Wozabal 2014, Mohajerin Esfahani and Kuhn 2018, Zhao and Guan 2018, Blanchet and Murthy 2019, Gao and Kleywegt 2022) is a framework for decision making under uncertainty, including learning, in which the decision maker has limited knowledge of the data-generating mechanism. The approach uses a minimax robust optimization problem:

$$\min_{f\in\mathcal{F}} \sup_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho} \mathbb{E}_{z\sim\mathbb{P}}[f(z)], \qquad (\mathsf{P})$$

where $f(z)$ represents the loss as a function of the unknown data $z$. The inner supremum finds the worst-case expected loss among a ball of distributions with radius $\rho$, containing all distributions that are close, in $p$-Wasserstein distance $\mathcal{W}_p$, to the empirical distribution $\mathbb{P}_n$ based on a sample of size $n$. Wasserstein DRO has been applied to problems in machine learning, including (semi)-supervised learning (Chen and Paschalidis 2018, Blanchet and Kang 2020), adversarial learning (Staib and Jegelka 2017, Sinha et al. 2018, Najafi et al. 2019, Levine and Feizi 2020), reinforcement learning (Abdullah et al. 2019, Smirnova et al. 2019, Derman and Mannor 2020), and transfer learning (Lee and Raginsky 2018, Volpi et al. 2018, Duchi et al. 2020). Kuhn et al. (2019) provide a recent survey.

The robustness of solutions produced by Wasserstein DRO can be related to the well-studied effect of regularization. The connection between Wasserstein DRO and regularization has been established in various settings. Shafieezadeh-Abadeh et al. (2015), Mohajerin Esfahani and Kuhn (2018), Chen and Paschalidis (2018), Blanchet et al. (2019), and Shafieezadeh-Abadeh et al. (2019), and provide equivalence results when $p = 1$, and Gao et al. (2017), Volpi et al. (2018), Blanchet et al. (2019), Bartl et al. (2021), and Blanchet et al. (2022), and provide asymptotic equivalence results when $p \in (1, \infty)$. Nonetheless, all the results mentioned previously are based on restrictive assumptions that limit their application to important classes of problems in operations research and machine learning.

For example, equivalence results for 1-Wasserstein DRO (Mohajerin Esfahani and Kuhn 2018, Shafieezadeh-Abadeh et al. 2019) require unbounded support of distributions and convexity of loss functions, whereas the distributions used in real-world problems have bounded support, and many loss functions used in machine learning are not convex. Also, asymptotic equivalence results (Gao et al. 2017, Volpi et al. 2018, Blanchet et al. 2019, Bartl et al. 2021) for $p$-Wasserstein DRO require loss functions to be smooth, whereas some widely used loss functions are merely piecewise smooth, including newsvendor cost, least absolute loss, and the rectified linear unit (ReLU) neural network and its variants. Therefore, it is clear that the current theory of the regularization effect of Wasserstein DRO is not complete.

In this paper, we aim to close this gap by providing a general connection between Wasserstein DRO and regularization. To this end, we develop a new concept, called the *variation of loss* (see Definition 1), denoted as $\mathcal{V}(f)$, that measures the magnitude of change of the expected loss as the data distribution is perturbed. It generalizes total variation for real-valued functions and reduces to the homogeneous Lipschitz norm for Lipschitz continuous functions and to weighted empirical gradient norm for differentiable functions. Intuitively, when the variation of loss is controlled, small perturbations of random data would have little impact on the expected loss and thus would not deteriorate the solution quality much. We develop results that show that Wasserstein DRO (P) is closely related to a *variation regularization* problem:

$$\min_{f\in\mathcal{F}} \mathbb{E}_{z\sim\mathbb{P}_n}[f(z)] + \rho\mathcal{V}(f). \tag{V}$$

Our results illustrate the variation regularization effect that is intrinsically associated with Wasserstein DRO. More specifically, we establish the following results:

(I) For $p$-Wasserstein DRO, $p \in (1,\infty]$, we show that for a broad class of loss functions, possibly nonconvex and nonsmooth, with high probability, Wasserstein DRO (P) is asymptotically equivalent to variation regularization (V) up to a higher order $O(\rho^{2\wedge p})$ remainder (Theorem 1).

For $p < 2$, the bound $\rho^p$ is tight (Example 2), indicating a qualitative disparity among different Wasserstein orders.

For $p = 2$ (one of the most popular choices of Wasserstein order), we demonstrate our results with multi-item newsvendor (Example 4) and gradient regularization for leaky ReLU networks (Example 5). Moreover, our results hold for general non-Euclidean metric spaces, illustrated with Laplacian regularization for manifold learning (Example 9), and score function regularization for intensity estimation of point processes (Example 10).

For $p = \infty$, we apply our results to adversarial robust learning and establish its equivalence to empirical total-variation regularization (Example 11).

(II) For 1-Wasserstein DRO ($p = 1$), we show that the asymptotic equivalence between Wasserstein DRO (P) and variation regularization (V) may not hold in general (Example 6). For this setting we prove a sandwich theorem (Theorem 2 and Corollary 1), that shows that with high probability, (P) with radius $\rho$ is upper bounded by (V) with a tuning parameter $\rho$ and lower bounded by (V) with a tuning parameter $\eta\rho$, where $\eta \in (0,1]$. This establishes an approximate equivalence between control of the Wasserstein robust loss and control of the variation of the loss function. As applications, we consider linear prediction with Lipschitz losses (Examples 7 and 8), and we extend the existing equivalence results (Mohajerin Esfahani and Kuhn 2018, Shafieezadeh-Abadeh et al. 2019) to a more general class of nonconvex functions (Corollary 2).

(III) In addition to understanding the regularization effect of Wasserstein DRO, our new results enable us to develop new generalization guarantees for adversarial robust learning that quantify the gap between the empirical adversarial risk and population adversarial risk (Theorem 3). We show that in the adversarial setting, the generalization behavior of a machine learning model is affected not only by the complexity of the loss function class as in classical empirical risk minimization, but also by the complexity of the slope of the loss function class.

In essence, our analysis is based on Taylor expansions of the loss function on each data point, sharing the same spirit as several existing works on smooth losses. Nevertheless, the main challenges for nonsmooth loss functions in a data-driven setting is that perturbing data points results in a random and nonsmooth change on the loss function values. Consequently, the existing arguments cannot be simply adopted, and new probabilistic analysis is needed to analyze the remainder of the Taylor expansion. We refer to the next section for more detailed comparisons with the literature.

### 1.1. Related Work

The relation between robust optimization and regularization has been explored for various settings, dating back to the pioneering work of Xu et al. (2008, 2009). They established an equivalence between data-driven robust optimization and norm regularization for LASSO and support vector machines, which was generalized to linear and matrix regression in Bertsimas and Copenhaver (2018), among others. Given the close relationship between Wasserstein DRO and data-driven robust optimization (Gao and Kleywegt 2022), it is expected that Wasserstein DRO would also exhibit a regularization effect. Indeed, the equivalence between 1-Wasserstein DRO and norm regularization has been established for piecewise-linear convex losses (Mohajerin Esfahani and Kuhn 2018), logistic regression (Shafieezadeh-Abadeh et al. 2015, Blanchet et al. 2019), support vector machines (Blanchet et al. 2019), and linear regression and classification and their kernelization (Chen and Paschalidis 2018, Shafieezadeh-Abadeh et al. 2019). All these results require convexity

and unboundedness of the data space. Blanchet et al. (2019) established the connection between $p$-Wasserstein DRO, $p \in (1,\infty]$, and norm regularization for certain settings, and studied the optimal selection of Wasserstein radius. The previous version of this work (Gao et al. 2017) established an asymptotic equivalence between $p$-Wasserstein DRO and gradient regularization for smooth loss functions and was generalized by Bartl et al. (2021) under weaker assumptions. Blanchet et al. (2022) established a finer analysis of the asymptotic equivalence for 2-Wasserstein DRO, and Volpi et al. (2018) developed an asymptotic equivalence for its Lagrangian relaxation. All these results require differentiability of the loss functions, which facilitates the Taylor expansion on each data point with a remainder that can be bounded easily. The only exception is Bartl et al. (2021), who also considered weakly differentiable functions in their remark 11. However, their analysis relies crucially on a continuous nominal distribution for which the nondifferentiable points are negligible, which is not the case when the nominal distribution is random and discrete, for which each nondifferentiable point leads on a random and nonsmooth change on the worst-case value. For $p = \infty$, an equivalent form of Wasserstein DRO has been studied extensively in the context of adversarial robust learning (Goodfellow et al. 2015, Lyu et al. 2015, Shaham et al. 2018). Recently, generalization bounds for adversarial robust learning have been studied in Yin et al. (2019), Attias et al. (2019), and Awasthi et al. (2020), and generalization bounds for other finite $p$-Wasserstein DRO have been investigated in Sinha et al. (2018), Lee and Raginsky (2018), Shafieezadeh-Abadeh et al. (2019), Najafi et al. (2019), and Gao (2022).

In the DRO literature, besides Wasserstein DRO, other choices of distributional uncertainty sets (ambiguity sets) have been explored (Scarf 1958, Žáčková 1966, Shapiro and Kleywegt 2002, Calafiore and El Ghaoui 2006, Erdoğan and Iyengar 2006, Popescu 2007, Delage and Ye 2010, Goh and Sim 2010, Ben-Tal et al. 2013, Wiesemann et al. 2014, Bayraksan and Love 2015, Jiang and Guan 2016, Wang et al. 2016). In particular, the asymptotic equivalence of $\phi$-divergence DRO and variance regularization has been established in Lam (2016), Gotoh et al. (2018), and Duchi and Namkoong (2019). Other connections between regularization and various DRO formulations have been discussed in Gotoh et al. (2020) and Anderson and Philpott (2022). Rahimian and Mehrotra (2019) give a recent survey on distributionally robust optimization.

The paper proceeds as follows. We briefly review some results for Wasserstein DRO in Section 2 and define variation regularization in Section 3. We establish the connection between Wasserstein DRO and variation regularization in Section 4 for $p > 1$ and Section 5 for $p = 1$. Discussion and extension of our results are provided in Section 6. As an application of our theory, in Section 7, we study Wasserstein DRO in the context of adversarial robust learning and derive new generalization bounds. We conclude the paper in Section 8. Proofs of our results are deferred to the online appendices.

## 2. Wasserstein Distributionally Robust Optimization

In this section, we introduce notation and provide some results for Wasserstein distributionally robust optimization.

Throughout this paper, we let $p \in [1,\infty]$, and we let $q$ denote its Hölder conjugate, that is, $1/p + 1/q = 1$. Let $\mathcal{Z}$ denote a metric space with metric $\mathsf{d}(\cdot,\cdot)$, measuring the difference between data points. The distance between a point $z \in \mathcal{Z}$ and a set $\mathcal{D} \subset \mathcal{Z}$ is defined as $\mathsf{d}(z,\mathcal{D}) := \inf_{\tilde{z}\in\mathcal{D}} \mathsf{d}(\tilde{z},z)$. The interior and closure of a set $A$ are denoted by $\mathrm{int}(A)$ and $\mathrm{cl}(A)$, respectively. The diameter of metric space $(\mathcal{Z},\mathsf{d})$ is defined as $\mathrm{diam}(\mathcal{Z}) := \sup_{\tilde{z},z\in\mathcal{Z}} \mathsf{d}(\tilde{z},z)$. Let $\limsup_{\tilde{z}\to z} h(\tilde{z}) := \lim_{\delta\downarrow 0} \sup\{h(\tilde{z}) : 0 < \mathsf{d}(\tilde{z},z) < \delta\}$. The sup-norm of a function $h : \mathcal{Z} \mapsto \mathbb{R}$ is denoted by $\|h\|_\infty := \sup_{z\in\mathcal{Z}} |h(z)|$, and the homogeneous Lipschitz norm of a Lipschitz continuous function $h : \mathcal{Z} \mapsto \mathbb{R}$ is denoted by $\|h\|_{\mathrm{Lip}} := \sup_{\tilde{z},z\in\mathcal{Z}} [f(\tilde{z}) - f(z)]/\mathsf{d}(\tilde{z},z)$. When $\mathcal{Z}$ is a normed space with norm $\|\cdot\|$, let $\|\cdot\|_*$ denote the dual norm, and let $\langle\cdot,\cdot\rangle$ denote the associated bilinear form. We denote $a \wedge b := \min\{a,b\}$, $a \vee b := \max\{a,b\}$, and $a_+ := \max\{a,0\}$. The support of a distribution $\mathbb{Q}$ is denoted by $\mathrm{supp}\,\mathbb{Q}$. We use $O$ and $O_p$ to represent the big $O$ and the big $O$ in probability notations, respectively, and use $\tilde{O}$ when we omit the polylog term. Let $\mathcal{P}(\mathcal{Z})$ denote the set of all Borel probability measures on $\mathcal{Z}$. For any $p \in [1,\infty)$ and $\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$, the $\mathcal{L}^p(\mathbb{Q})$-norm of a $\mathbb{Q}$-measurable function $h : \mathcal{Z} \mapsto \mathbb{R}$ is denoted by $\|h\|_{\mathbb{Q},p} := \left(\int_{\mathcal{Z}} h^p d\mathbb{Q}\right)^{1/p}$, and $\|h\|_{\mathbb{Q},\infty} := \mathbb{Q}\text{-ess sup}_{z\in\mathcal{Z}}\, h(z)$.

The Wasserstein distance of order $p$ between distributions $\mathbb{P},\mathbb{Q} \in \mathcal{P}(\mathcal{Z})$ is defined as

$$\mathcal{W}_p(\mathbb{P},\mathbb{Q}) := \begin{cases} \inf\limits_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \left(\mathbb{E}_{(\tilde{z},z)\sim\gamma}\left[\mathsf{d}(\tilde{z},z)^p\right]\right)^{1/p}, & \text{if } p \in [1,\infty), \\ \inf\limits_{\gamma\in\Gamma(\mathbb{P},\mathbb{Q})} \gamma\text{-ess sup}_{\mathcal{Z}\times\mathcal{Z}}\, \mathsf{d}(z,\tilde{z}), & \text{if } p = \infty, \end{cases}$$

where the minimization is over the set $\Gamma(\mathbb{P},\mathbb{Q})$ of all Borel probability distributions on $\mathcal{Z}\times\mathcal{Z}$ with marginal distributions $\mathbb{P}$ and $\mathbb{Q}$. For any $p \in [1,\infty)$ and $z_0 \in \mathcal{Z}$, let $\mathcal{P}_p(\mathcal{Z}) := \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : \mathbb{E}_{z\sim\mathbb{Q}}[\mathsf{d}(z,z_0)^p] < \infty\}$ denote the subset of $\mathcal{P}(\mathcal{Z})$ with finite $p$th moment. To ease notation, we adopt the convention that $\mathcal{P}_\infty(\mathcal{Z}) := \mathcal{P}(\mathcal{Z})$.

Given a family $\mathcal{F}$ of loss functions $f : \mathcal{Z} \mapsto \mathbb{R}$, a nominal distribution $\mathbb{Q} \in \mathcal{P}_p(\mathcal{Z})$, and a radius $\rho \geq 0$, the corresponding Wasserstein DRO problem is

$$\min_{f\in\mathcal{F}} \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})} \{\mathbb{E}_{z\sim\mathbb{P}}[f(z)] : \mathcal{W}_p(\mathbb{P},\mathbb{Q}) \leq \rho\}. \tag{P}$$

The dual problem of the inner supremum in (P) is defined as

$$\begin{cases} \min\limits_{\lambda\geq 0}\left\{\lambda\rho^p + \mathbb{E}_{z\sim\mathbb{Q}}\left[\sup\limits_{\tilde{z}\in\mathcal{Z}}\{f(\tilde{z}) - \lambda \mathsf{d}(\tilde{z},z)^p\}\right]\right\}, & p\in[1,\infty), \\ \mathbb{E}_{z\sim\mathbb{Q}}\left[\sup\limits_{\tilde{z}\in\mathcal{Z}}\{f(\tilde{z}) : \mathsf{d}(\tilde{z},z)\leq\rho\}\right], & p=\infty. \end{cases} \quad (\mathbf{D})$$

An important result in Wasserstein distributionally robust optimization is that strong duality holds under quite general conditions and in particular for $\mathbb{Q}=\mathbb{P}_n$ and the setup described previously (see Lemma EC.1 in Online Appendix EC.1 for more details).

We define the *Wasserstein regularizer* $\mathcal{R}$ as the difference between the Wasserstein robust loss and the nominal loss

$$\mathcal{R}_{\mathbb{Q},p}(\rho;f) := \sup_{\mathbb{P}\in\mathcal{P}(\mathcal{Z})}\{\mathbb{E}_{z\sim\mathbb{P}}[f(z)] : \mathcal{W}_p(\mathbb{P},\mathbb{Q})\leq\rho\} - \mathbb{E}_{\mathbb{Q}}[f].$$

Often we consider a data-driven problem with data $z_i^n$, $i=1,\ldots,n$. For some analysis, it is assumed that the data $z_i^n$, $i=1,\ldots,n$ are independent and identically distributed with distribution $\mathbb{P}_{\mathsf{true}}$, although the Wasserstein DRO approach makes sense also when the data do not satisfy such an assumption. Thus, we consider a setting in which the empirical distribution $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{z_i^n}$ is chosen as the nominal distribution $\mathbb{Q}$, where $\boldsymbol{\delta}_z$ denotes the Dirac point mass on $z$. We use $\mathbb{P}_\otimes$ or $\mathbb{E}_\otimes$ to indicate that the probability or expectation is evaluated with respect to the sampling distribution, namely the $n$-fold product distribution $\otimes_{i=1}^n\mathbb{P}_{\mathsf{true}}$ over $\mathcal{Z}^n$. By definition, when $\mathbb{Q}=\mathbb{P}_n$, the Wasserstein regularizer $\mathcal{R}_{\mathbb{P}_n,p}(\rho;f)$ can be viewed as a data-dependent regularizer of the loss $f$. Proposition 1 is a consistency-type result under a growth condition, which shows that $\mathcal{R}_{\mathbb{P}_n,p}(\rho;f)$ converges to zero as the radius shrinks.

**Proposition 1** (Consistency). *Let $p\in[1,\infty]$. Assume that $f$ is upper semicontinuous for all $f\in\mathcal{F}$, and that the following growth condition is satisfied when $p<\infty$:*

$$\exists\, z_0\in\mathcal{Z} \text{ such that } \sup_{f\in\mathcal{F}}\ \limsup_{\mathsf{d}(\tilde{z},z_0)\to\infty}\frac{[f(\tilde{z})-f(z_0)]_+}{\mathsf{d}(\tilde{z},z_0)^p}<\infty, \quad (\mathbf{G})$$

*where we use the convention that the ratio is zero if* $\mathrm{diam}(\mathcal{Z})<\infty$. *Then,*

$$\lim_{\rho\to 0}\mathcal{R}_{\mathbb{P}_n,p}(\rho;f) = 0 = \mathcal{R}_{\mathbb{P}_n,p}(0;f).$$

The growth condition (G) means that the loss functions should grow no faster than a polynomial of order $p$ uniformly in $\mathcal{F}$. The assumptions in Proposition 1 are minimal in the following sense. The upper semicontinuity of $f$ is necessary to ensure that $\lim_{\rho\to 0}\mathcal{R}_{\mathbb{P}_n,p}(\rho;f)=0$, and the growth condition is necessary to ensure that the Wasserstein robust loss is finite (Gao and Kleywegt 2022). Proposition 1 generalizes theorem 3.6(i) in Mohajerin Esfahani and Kuhn (2018) by relaxing the convergence condition on the radius $\rho$. We remark that under additional conditions on the radius, theorem 3.6 in Mohajerin Esfahani and Kuhn (2018) proves that the optimal value of the Wasserstein DRO converges from above to the optimal value of the true stochastic program, and the Wasserstein DRO solution is asymptotically consistent. An important goal of this paper is to study the *convergence rate* of the Wasserstein regularizer $\mathcal{R}_{\mathbb{P}_n,p}(\rho;f)$ as $\rho\to 0$.

## 3. Variation Regularization

To study the regularization effect of Wasserstein DRO, in this section, we introduce a new concept, the *variation* of a function, inspired from the total variation of a real-valued function. This is built on the notion of *slope* adopted from Cheeger (1999) and Ambrosio et al. (2008), which measures the modulus of continuity of a function on a metric space without isolated points.

**Definition 1** (Slopes and Variation). The local slope $|\partial f|(z)$ and global slope $\mathfrak{l}_f(z)$ of a function $f:\mathcal{Z}\to\mathbb{R}$ at $z\in\mathcal{Z}$ is defined as

$$|\partial f|(z) := \limsup_{\tilde{z}\to z}\frac{(f(\tilde{z})-f(z))_+}{\mathsf{d}(\tilde{z},z)},$$
$$\mathfrak{l}_f(z) := \sup_{\tilde{z}\neq z}\frac{(f(\tilde{z})-f(z))_+}{\mathsf{d}(\tilde{z},z)}.$$

The variation of a function f with respect to a distribution $\mathbb{Q}$ is defined as

$$\mathcal{V}_{\mathbb{Q},q}(f) := \begin{cases} \||\partial f|\|_{\mathbb{Q},q}, & q\in[1,\infty), \\ \|\mathfrak{l}_f\|_{\mathbb{Q},\infty}, & q=\infty. \end{cases}$$

This definition of slope generalizes the slope for univariate functions. The local slope $|\partial f|(z)$ measures the magnitude of the change of the loss when perturbing $z$ locally, whereas the global slope $\mathfrak{l}_f(z)$ measures the largest magnitude of the loss when perturbing $z$ to any point in $\mathcal{Z}$. Obviously, we have $|\partial f|\leq\mathfrak{l}_f$. Slopes are well defined for a very broad class of continuous but not necessarily differentiable loss functions on any metric space without isolated points. The variation $\mathcal{V}_{\mathbb{Q},q}(f)$ is essentially a weighted average of slopes over all data $z\in\mathrm{supp}\,\mathbb{Q}$. In particular, when $f$ is univariate and differentiable, $\mathcal{V}_{\mathbb{Q},1}(f)$ reduces to the usual representation of total variation of a function $\int_{\mathbb{R}}|f'(z)|dz$. The next example demonstrates our definition when $\mathcal{Z}$ is a Banach space.

**Example 1.** Suppose $\mathcal{Z}$ is a Banach space $(\mathcal{B},\|\cdot\|)$ and $f:\mathcal{B}\to\mathbb{R}$.

(I) When $f$ is differentiable, by definition and Cauchy-Schwarz inequality, we have

$$|\partial f|(z) = \limsup_{\tilde{z}\to z}\frac{f(\tilde{z})-f(z)}{\|\tilde{z}-z\|} = \limsup_{\tilde{z}\to z}\frac{\langle\nabla f(z),\tilde{z}-z\rangle}{\|\tilde{z}-z\|}$$
$$= \|\nabla f(z)\|_*.$$

(II) When $f$ is Lipschitz, by definition $|\partial f|(z) \le \mathrm{I}_f(z) \le \|f\|_{\mathrm{Lip}}$. Thus,

$$\|\mathrm{I}_f\|_{\mathbb{Q},\infty} = \mathbb{Q}-\operatorname*{ess\,sup}_{z\in\mathcal{Z}} \mathrm{I}_f(z) = \mathbb{Q}-\operatorname*{ess\,sup}_{z\in\mathcal{Z}} \sup_{\tilde{z}\neq z} \frac{(f(\tilde{z})-f(z))_+}{\mathsf{d}(\tilde{z},z)}.$$

Thus, by Example 1, if $\nabla f$ exists $\mathbb{Q}$-almost everywhere, then $\||\partial f|(z)\|_{\mathbb{Q},q} = \|\|\nabla f\|_*\|_{\mathbb{Q},q}$ for $q\in[1,\infty)$; if $f$ is Lipschitz continuous, then $\mathcal{V}_{\mathbb{Q},\infty}(f) \le \|f\|_{\mathrm{Lip}}$, and if in addition $\operatorname{supp}\mathbb{Q} = \mathcal{Z}$, then $\mathcal{V}_{\mathbb{Q},\infty}(f) = \|f\|_{\mathrm{Lip}}$.

As promised, we will bound $\mathcal{R}_{\mathbb{Q},p}(\rho;f)$ using $\mathcal{V}_{\mathbb{Q},q}(f)$. In particular, if $\mathcal{R}_{\mathbb{Q},p}(\rho;f)$ is shown to be lower bounded by $\mathcal{V}_{\mathbb{Q},q}(f)$, then minimizing the Wasserstein robust loss controls the variation of the loss. In Sections 4 and 5, we show that the variation is a natural quantity characterizing the convergence rate of the Wasserstein regularizer $\mathcal{R}_{\mathbb{P}_n,p}(f;\rho)$ as $\rho\to 0$ by separating the cases of $p>1$ and $p=1$. Quite often, we focus on a radius selection rule $\rho_n = \rho_0/\sqrt{n}$ for some $\rho_0>0$, which has been empirically used in practice and also theoretically investigated in Blanchet et al. (2019), Shafieezadeh-Abadeh et al. (2019), Blanchet et al. (2022), and Gao (2022).

## 4. Variation Regularization Effect of $p$-Wasserstein DRO ($p>1$)

In this section, we consider $p\in(1,\infty]$, and the study of $p=1$ is relegated to the next section, as they have essential differences. To ease the exposition, in this section, we first present our results for piecewise smooth losses on a Banach space $\mathcal{Z}$, and extension to general loss functions on a metric space is postponed to Section 6.2. Throughout this section, we impose the following two assumptions.

**Assumption 1** (Piecewise Smoothness). *For every $f\in\mathcal{F}$, there exists a partition $\mathcal{Z} = \cup_{1\le k\le K_f}\mathcal{Z}_{f,k}$, where $\mathcal{Z}_{f,j}\cap\mathcal{Z}_{f,k} = \emptyset$ for all $j\neq k$ such that $f$ is continuously differentiable on $\operatorname{int}(\mathcal{Z}_{f,k})$, $1\le k\le K_f$. Moreover, there exists $H\in\mathcal{L}^{\frac{p}{p-2}}(\mathbb{P}_{\mathsf{true}})$ when $p\in(2,\infty]$ and $H\in\mathcal{L}^{\infty}(\mathcal{Z})$ when $p\in(1,2]$ such that for any $\epsilon>0$, there exists $\delta>0$ such that for all $f\in\mathcal{F}$, $1\le k\le K_f$ and $\tilde{z},z\in\operatorname{int}(\mathcal{Z}_{f,k})$ with $\|\tilde{z}-z\|\le\delta$,*

$$\frac{\|\nabla f(\tilde{z})-\nabla f(z)\|_*}{\|\tilde{z}-z\|} \le H(z)+\epsilon. \tag{S}$$

*We denote by*

$$\mathcal{D}_f := \bigcup_{1\le k\neq j\le K_f} \mathrm{cl}(\mathcal{Z}_{f,j})\cap\mathrm{cl}(\mathcal{Z}_{f,k})$$

*the union set of intersections of pieces, which is assumed to be a $\mathbb{P}_{\mathsf{true}}$-null set for every $f\in\mathcal{F}$.*

By definition, all nondifferentiable points of $f$ are contained in $\mathcal{D}_f$. Although $\mathcal{D}_f$ has $\mathbb{P}_{\mathsf{true}}$-measure zero, $\cup_{f\in\mathcal{F}}\mathcal{D}_f$ may have positive $\mathbb{P}_{\mathsf{true}}$-measure.

**Assumption 2** (Growth and Jump of Gradient).

(I) *When $p\in(1,\infty)$, assume there exist constants $M,L\ge 0$ such that for every $f\in\mathcal{F}$ and $\tilde{z},z\in\mathcal{Z}\setminus\mathcal{D}_f$,*

$$\|\nabla f(\tilde{z})-\nabla f(z)\|_* \le M + L\|\tilde{z}-z\|^{p-1}.$$

(II) *When $p=\infty$, assume there exist constants $M\ge 0$ and $\delta_0>0$ such that for every $f\in\mathcal{F}$ and $\tilde{z},z\in\mathcal{Z}\setminus\mathcal{D}_f$ with $\|\tilde{z}-z\|<\delta_0$,*

$$\|\nabla f(\tilde{z})-\nabla f(z)\|_* \le M.$$

Assumption 2 imposes a growth condition on the gradient norm when $p\in(1,\infty)$, consistent with the growth condition (G) on the loss; as well as a bounded jump condition on $f$, namely, the gap of gradient norms around a nondifferentiable point is at most $M$. For smooth loss functions, we have $M=0$.

Assumption 1 and the bounded jump condition in Assumption 2 can be viewed as an extension of twice differentiability for smooth losses, which together bound the change of the losses when the data are perturbed within in a small neighbor. Assumptions 1 and 2 imply a weaker Assumption 5 in Section 6.2, which serves as the foundation for Taylor expansion-type analysis for nonsmooth loss functions.

In Section 4.1, we start by establishing upper and lower bounds for smooth loss functions. For nonsmooth losses, the probability of observing a nondifferentiable sample point is zero for a single loss function but may be strictly positive for a family of loss functions, as the uncountable union of measure-zero sets can have a positive measure. Furthermore, it is likely to observe a sample point that is near the nondifferentiable region, resulting in a nonsmooth change of the loss when it is adversarially perturbed. Therefore, to ensure a probabilistic guarantee on the remainder of Taylor expansion of nonsmooth loss functions, additional assumptions on the function class $\mathcal{F}$ and the underlying true distribution $\mathbb{P}_{\mathsf{true}}$ are needed. In Section 4.2, we study a simple example of piecewise linear loss to motivate proper assumptions, and then we develop the result for general piecewise smooth functions in Theorem 1, which is further generalized in Section 6.2.

### 4.1. Smooth Losses

We first establish a result demonstrating the gradient regularization effect for smooth losses, whose detailed proof is given in Online Appendix EC.2.2.

**Lemma 1.** *Let $p\in(1,\infty]$ and $\rho_n = \rho_0/\sqrt{n}$. Assume Assumption 1 holds with $K_f=1$ for all $f\in\mathcal{F}$ and Assumption*

2(I) *holds. Then there exist constants* $\bar{\rho}, C > 0$ *such that for all* $\rho_0 < \bar{\rho}$ *and* $f \in \mathcal{F}$,

$$\left|\mathcal{R}_{\mathbb{P}_n,p}(\rho_n;f) - \rho_n \mathcal{V}_{\mathbb{P}_n,q}(f)\right| \le \rho_n^{2\wedge p}\left(C + \|H\|_{\mathbb{P}_n,\frac{p}{p-2}}\mathbf{1}\{p>2\}\right).$$

We remark that $\left(\|H\|_{\mathbb{P}_n,\frac{p}{p-2}} - \|H\|_{\mathbb{P}_{\text{true}},\frac{p}{p-2}}\right)_+$ is of the order $O_p(n^{-1/2})$ under mild conditions (such as $H$ has bounded variance so that Chebyshev's inequality holds or $H$ has bounded moment generating function in a neighborhood of zero so that Chernoff bound applies). As a result, when $\rho_n = O(n^{-1/2})$, Lemma 1 gives a first-order Taylor expansion for the Wasserstein regularizer $\mathcal{R}_{\mathbb{P}_n,p}(\rho;f)$ of smooth losses for $p \in [2,\infty]$ with a remainder $O_p(n^{-1})$ uniformly for all $f \in \mathcal{F}$. The cases for $p = 2$ and $p \in (2,\infty)$ in normed vector spaces have been developed in Volpi et al. (2018, section 3.2) and Bartl et al. (2021, remark 8) respectively. When $p \in (1,2)$, the order of the remainder $O(n^{-p/2})$ cannot be improved in general, as can be seen from the following example.

**Example 2.** Consider $f(z) = z^p$, where $p \in (1,2)$ and $z \in \mathcal{Z} = (\mathbb{R}_+, |\cdot|)$. Suppose $\mathbb{P}_{\text{true}} = \boldsymbol{\delta}_0$, which implies $\mathbb{P}_n = \boldsymbol{\delta}_0$ almost surely. Observe that

$$\sup_{\tilde{z}\in\mathbb{R}_+}\{f(\tilde{z}) - f(0) - \lambda|\tilde{z}-0|^p\} = \begin{cases} +\infty, & \forall \lambda < 1, \\ 0, & \forall \lambda \ge 1. \end{cases}$$

Thus, by (D), we have

$$\begin{aligned}\mathcal{R}_{\mathbb{P}_n,p}(\rho;f) &= \min_{\lambda\ge 0}\left\{\lambda\rho^p + \sup_{\tilde{z}\in\mathbb{R}_+}\{f(\tilde{z}) - f(0) - \lambda|\tilde{z}-0|^p\}\right\} \\ &= \rho^p.\end{aligned}$$

On the other hand, $\mathcal{V}_{\mathbb{P}_n,q}(f) = f'(0) = 0$ almost surely. Hence, $\mathcal{R}_{\mathbb{P}_n,p}(\rho;f) - \rho\mathcal{V}_{\mathbb{P}_n,q}(f) = \rho^p$ for all $\rho \ge 0$.

### 4.2. Nonsmooth Losses

Next, we consider nonsmooth losses. Proofs of the results in this section can be found in Online Appendix EC.2.3. We start with an illustrating example of piecewise linear functions.

**Example 3.** Let $\mathcal{Z} = [0,1] \subset (\mathbb{R}, |\cdot|)$. Suppose $\rho_n = \rho_0/\sqrt{n}$ for some $\rho_0 > 0$. Consider

$$f_\theta(z) = \theta z \wedge 1, \text{where } \theta \ge 0,$$

which is illustrated in Figure 1. Then $f'_\theta(z)$ is $\theta$ when $z \in [0, 1/\theta)$ and is zero when $z \in (1/\theta, 1]$. Thus, whenever $1/\theta \notin \text{supp } \mathbb{P}_n$, which holds almost surely when $\mathbb{P}_{\text{true}}$ is continuous, we have

$$\mathcal{V}_{\mathbb{P}_n,1}(f_\theta) = \theta\mathbb{E}_{\mathbb{P}_n}[\mathbf{1}\{z < 1/\theta\}].$$

Using the dual form (D),

$$\begin{aligned}\mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n;f_\theta) &= \mathbb{E}_{\mathbb{P}_n}\left[\sup_{0\le\tilde{z}\le 1}\{f_\theta(\tilde{z}) - f_\theta(z) : |\tilde{z}-z| \le \rho_n\}\right] \\ &= \mathbb{E}_{\mathbb{P}_n}[\rho_n\theta \cdot \mathbf{1}\{z \le 1/\theta - \rho_n\} + (1-\theta z) \\ &\quad \cdot \mathbf{1}\{1/\theta - \rho_n < z < 1/\theta\}] \\ &= \rho_n\theta\mathbb{E}_{\mathbb{P}_n}[\mathbf{1}\{z < 1/\theta\}] - \theta\mathbb{E}_{\mathbb{P}_n} \\ &\quad [(z - (1/\theta - \rho_n))\mathbf{1}\{1/\theta - \rho_n < z < 1/\theta\}],\end{aligned}$$

where the second term indicates that for a point $z$ close to the nondifferentiable point $1/\theta$ of $f_\theta(z)$, perturbing $z$ leads to a change of loss by at most $(1-\theta z)$, which is less than $\rho_n\theta$. It follows that

$$\begin{aligned}&\rho_n\mathcal{V}_{\mathbb{P}_n,\infty}(f_\theta) - \mathcal{R}_{\mathbb{P}_n,\infty}(\rho_n;f_\theta) \\ &= \theta\mathbb{E}_{\mathbb{P}_n}[(z - (1/\theta - \rho_n))\mathbf{1}\{1/\theta - \rho_n < z < 1/\theta\}] \\ &= \theta\mathbb{E}_{\mathbb{P}_{\text{true}}}[(z - (1/\theta - \rho_n))\mathbf{1}\{1/\theta - \rho_n < z < 1/\theta\}] + \theta\epsilon_n,\end{aligned}$$

where

$$\begin{aligned}\epsilon_n &= \mathbb{E}_{\mathbb{P}_n}[(z - (1/\theta - \rho_n))\mathbf{1}\{1/\theta - \rho_n < z < 1/\theta\}] \\ &\quad - \mathbb{E}_{\mathbb{P}_{\text{true}}}[(z - (1/\theta - \rho_n))\mathbf{1}\{1/\theta - \rho_n < z < 1/\theta\}] \\ &\le \rho_n|\mathbb{P}_n\{1/\theta - \rho_n < z < 1/\theta\} - \mathbb{P}_{\text{true}}\{1/\theta - \rho_n \\ &\quad < z < 1/\theta\}| = O_p(1/n).\end{aligned}$$

Then the gap would be of the order $O_p(1/n)$ if

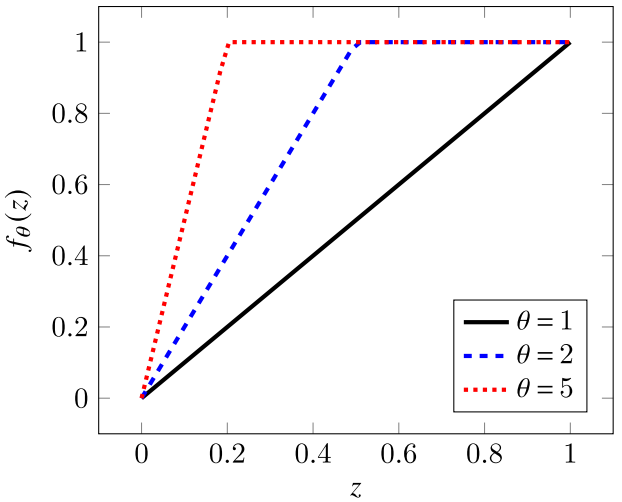
$$\mathbb{E}_{\mathbb{P}_{\text{true}}}[(z - (1/\theta - \rho_n))\mathbf{1}\{1/\theta - \rho_n < z < 1/\theta\}] = O(1/n).$$

The left-hand side equals

$$\int_{1/\theta-\rho_n}^{1/\theta}(z - (1/\theta - \rho_n))d\mathbb{P}_{\text{true}}(z),$$

which is $O(1/n)$ if $\mathbb{P}_{\text{true}}$ has a bounded density on $[1/\theta - \rho_n, 1/\theta]$.

**Figure 1.** (Color online) Plots of $f_\theta(z) = \theta z \wedge 1$, $z \in [0,1]$



Motivated by Example 3, we impose the following continuity assumption around nonsmooth points for the underlying data-generating distribution. Recall from Assumption 1 that $\mathcal{D}_f$ is the union of intersections of pieces of $f$.

**Assumption 3** (Bounded Density).

$$\limsup_{\delta \downarrow 0} \sup_{f \in \mathcal{F}: \mathcal{D}_f \neq \varnothing} \frac{\mathbb{P}_{\text{true}}\{z : 0 < \mathsf{d}(z, \mathcal{D}_f) < \delta\}}{\delta} < \infty.$$

A sufficient condition to ensure Assumption 3 is that $\mathbb{P}_{\text{true}}$ has a bounded density everywhere.

When $p \in (1, \infty)$, we impose an additional assumption on the normalized gradient norm $\frac{\|\nabla f(z)\|_*}{\|\|\nabla f\|_*\|_{\mathbb{P}_{\text{true}},q}}$.

**Assumption 4** (Growth of Normalized Gradient Norm). *Let $p \in (1, \infty)$. For $z \in \mathcal{Z} \setminus \mathcal{D}_f$, define $w_f(z) := \left(\frac{\|\nabla f(z)\|_*}{\|\|\nabla f\|_*\|_{\mathbb{P}_{\text{true}},q}}\right)^{\frac{1}{p-1}}$. Assume there exist constants $c_1, c_2, c_3 > 0$ such that for all $f \in \mathcal{F}$ with $\|\|\nabla f\|_*\|_{\mathbb{P}_{\text{true}},q} > 0$ and $\mathcal{D}_f \neq \varnothing$ and for all $z \in \mathcal{Z} \setminus \mathcal{D}_f$,*

$$c_3 \le w_f(z)^{p-1} \le c_1 + c_2 \mathsf{d}(z, \mathcal{D}_f)^{p-1}.$$

This technical assumption specifies the growth and jump deviated from nondifferentiable points for the normalized gradient norm $w_f$. The upper bound is similar to the growth and jump Assumption 2 of $\|\nabla f\|_*$ but imposed on the normalized gradient norm $w_f$. Whenever the upper bound of Assumption 4 holds, $M$ in Assumption 2 can be replaced with $\Delta\|\|\nabla f\|_*\|_{\text{true},q}$ for some $\Delta > 0$ (see Lemma EC.8 in Online Appendix EC.2.3). Whenever the lower bound of Assumption 4 holds, $\|\|\nabla f\|_*\|_{\text{true},q}$ and its empirical counterpart have a bounded ratio. Practical examples satisfying this condition will be provided in Section 4.3.

We define classes of functions:

$$\mathcal{I}_\rho := \{z \mapsto \mathbf{1}\{\mathsf{d}(z, \mathcal{D}_f) < \rho\} : f \in \mathcal{F}, \mathcal{D}_f \neq \varnothing\}. \quad (1)$$

$$\mathcal{E} := \{\mathsf{d}(\cdot, \mathcal{D}_f) : f \in \mathcal{F} \text{ with } \mathcal{D}_f \neq \varnothing\}. \quad (2)$$

Intuitively, the set $\mathcal{I}_\rho$ is the class of indicator functions for samples that falls within the margin of the nondifferentiable regions, and the set $\mathcal{E}$ is the class of distance functions describing the distance of samples to the nondifferentiable regions. Both are empty for smooth classes.

Recall the *Rademacher complexity* of a function class $\mathcal{H}$ with respect to a sample $\{z_i^n\}_{i=1}^n$ is defined as $\mathfrak{R}_n(\mathcal{H}) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \boldsymbol{\sigma}_i h(z_i^n)\right]$, where $\boldsymbol{\sigma}_i$ s are independently and identically distributed Rademacher random variables with $\mathbb{P}\{\boldsymbol{\sigma}_i = \pm 1\} = \frac{1}{2}$. The Rademacher complexity of the function class $\mathcal{H}$ with respect to $\mathbb{P}_{\text{true}}$ for sample size $n$ is defined as $\mathbb{E}_{\otimes}[\mathfrak{R}_n(\mathcal{H})]$. Recall also that the *covering number* $\mathcal{N}(\epsilon; \mathcal{H}, \mathsf{d}_{\mathcal{H}})$ of a function class $\mathcal{H}$ with respect to a metric $\mathsf{d}_{\mathcal{H}}$ is defined as the smallest cardinality of an $\epsilon$-cover of $\mathcal{H}$; here $\mathcal{H}_\epsilon$ is an *$\epsilon$-cover* of $\mathcal{H}$ if for each $h \in \mathcal{H}$, there exists $\tilde{h} \in \mathcal{H}_\epsilon$ such that $\mathsf{d}_{\mathcal{H}}(\tilde{h}, h) \le \epsilon$.

Now we are ready to state the main result in this section.

**Theorem 1** (*p*-Wasserstein DRO).

(I) *Let $p = \infty$. Assume Assumptions 1 and 2 are in force. Then there exists $\bar{\rho} > 0$ such that for all $\rho < \bar{\rho}$ and $f \in \mathcal{F}$,*

$$|\mathcal{R}_{\mathbb{P}_n,\infty}(\rho; f) - \rho \mathcal{V}_{\mathbb{P}_n,1}(f)| \le \rho^2 \|H\|_{\mathbb{P}_n,1} + M \mathbb{E}_{\mathbb{P}_n}[(\rho - \mathsf{d}(z, \mathcal{D}_f))_+].$$

(II) *Let $p \in (1, \infty)$ and $\rho_n = \rho_0/\sqrt{n}$. Assume Assumptions 1, 2, and 4 are in force. Then there exist constants $\bar{\rho}, C_1, C_2 > 0$ such that for all $\rho_0 < \bar{\rho}$ and $f \in \mathcal{F}$,*

$$|\mathcal{R}_{\mathbb{P}_n,p}(\rho_n; f) - \rho_n \mathcal{V}_{\mathbb{P}_n,q}(f)| \le \rho_n^{2 \wedge p}(\|H\|_{\mathbb{P}_n,\frac{p}{p-2}} \mathbf{1}\{p > 2\} + C_1) + M \mathbb{E}_{\mathbb{P}_n}[(C_2 \rho_n - \mathsf{d}(z, \mathcal{D}_f))_+].$$

(III) *The second term in (I) or (II) can be bounded as follows. Assume Assumption 3 holds. Let $t > 0$. Then there exist constants $\bar{\rho}, C > 0$ such that for all $\rho < \bar{\rho}$, with probability at least $1 - e^{-t}$, for every $f \in \mathcal{F}$,*

$$\mathbb{E}_{\mathbb{P}_n}[(\rho - \mathsf{d}(z, \mathcal{D}_f))_+] \le C\rho^2 + 2\rho \mathbb{E}_{\otimes}[\mathfrak{R}_n(\mathcal{I}_\rho)] + \rho\sqrt{\frac{t}{2n}},$$

*and with probability at least $1 - e^{-t}$, for every $f \in \mathcal{F}$,*

$$\mathbb{E}_{\mathbb{P}_n}[(\rho - \mathsf{d}(z, \mathcal{D}_f))_+] \le 2C\rho^2 + \frac{24\rho}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\epsilon\rho; \mathcal{E}, \|\cdot\|_{\mathbb{P}_n,2})}\, d\epsilon + \rho\sqrt{\frac{t}{2n}}.$$

(I) and (II) bound the gap between Wasserstein regularizer and variation regularizer for $p = \infty$ and $1 < p < \infty$, respectively, and (III) provides a further probabilistic bound on the gap. Comparing (I) and (II), $\infty$-Wasserstein DRO requires fewer assumptions on the loss functions and imposes no assumption on the scaling of the radius than the $p$-Wasserstein DRO with $p \in (1, \infty)$. This is largely because $\infty$-Wasserstein only allows local perturbations of data as seen from the hard distance constraints their dual problems (D), and thus the gap between the Wasserstein regularizer and the variation regularizer is small for sufficiently small radius, regardless of the sample size. In contrast, for $p \in (1, \infty)$, we need to impose the scaling $\rho_n = O(1/\sqrt{n})$ to control how far a sample can be perturbed in the worst case, so that the gap can be properly bounded by exploiting the assumptions.

Compared with the smooth case (Lemma 1), the major difference in the nonsmooth case is that the gap involves a probabilistic term $\mathbb{E}_{\mathbb{P}_n}[(\rho - \mathsf{d}(z, \mathcal{D}_f))_+]$, dependent on how the sample falls into the $\rho$-margin of nondifferentiable regions. When $\rho_n = O(1/\sqrt{n})$, (III) indicates that this term is $O_p(1/n)$ as long as the Rademacher complexity $\mathbb{E}_{\otimes}[\mathfrak{R}_n(\mathcal{I}_{\rho_n})] = O(1/\sqrt{n})$ or the entropy integral $\int_0^1 \sqrt{\log \mathcal{N}(\epsilon\rho_n; \mathcal{E}, \|\cdot\|_{\mathbb{P}_n,2})}\, d\epsilon = O_p(1)$. Examples of such will be given in Section 4.3, for which we rigorously prove the bound. Whenever $\rho = \rho_n = O(1/\sqrt{n})$ and $\mathbb{E}_{\mathbb{P}_n}[(\rho_n - \mathsf{d}(z, \mathcal{D}_f))_+] = O_p(1/n)$, an immediate consequence

of Theorem 1 is that the $p$-Wasserstein DRO (P), $p \in (1,\infty]$, is asymptotically equivalent to the empirical variation regularization problem (V):

$$\min_{f\in\mathcal{F}} \sup_{\mathbb{P}:\mathcal{W}_p(\mathbb{P},\mathbb{P}_n)\le\rho_n} \mathbb{E}_{z\sim\mathbb{P}}[f(z)] = \min_{f\in\mathcal{F}}\{\mathbb{E}_{z\sim\mathbb{P}_n}[f(z)] + \rho_n \mathcal{V}_{\mathbb{P}_n,q}(f)\} + O_p(n^{-1\wedge\frac{p}{2}}).$$

### 4.3. Applications

In this section, we instantiate our results using various examples. For each example, we discuss why the assumptions for the corresponding result hold and provide more detailed verification of the assumptions in Online Appendix EC.2.4. We illustrate the case $p \in (1,\infty)$ in multi-item newsvendor (Example 4) and leaky ReLU neural networks (Example 5), and for $p = \infty$, we will demonstrate it in Section 7 for adversarial learning (Example 11).

**4.3.1. Multi-Item Newsvendor.** We start with the classical newsvendor problem with piecewise linear objective that makes use of Theorem 1.

**Example 4** (Multi-Item Newsvendor). Consider a newsvendor problem in which the decision maker needs to decide the ordering quantities $\theta \in \mathbb{R}^d_+$ for $d$ products before their random demands $z$ are realized. Let $h = (h_1,\dots,h_d)$ and $b = (b_1,\dots,b_d)$ be, respectively, the holding cost vector and back-order cost vector. The overall cost is given by

$$f_\theta(z) = \sum_{j=1}^d h_j(\theta_j - z_j)_+ + b_j(z_j-\theta_j)_+.$$

Suppose $\mathcal{Z} \subset (\mathbb{R}^d_+, \|\cdot\|_2)$ and $\Theta \subset \{\theta \in \mathbb{R}^d_+ : \|\theta\|_\infty \le B\}$. Assume each marginal distribution of $\mathbb{P}^{z_j}_{\text{true}}$ has continuous density on $\mathbb{R}_+$ bounded by $\mu$.

Let us verify the assumptions required by Theorem 1. First, $f_\theta$ has $2^d$ pieces determined by the sign of $z_j - \theta_j$, $j = 1,\dots,d$ and each piece is linear, thus Assumption 1 is satisfied with $H = 0$. Second, Assumption 2 is satisfied with $M = \sum_{j=1}^d |h_j| + |b_j|$ and $L = 0$. Third, $\mathsf{d}(z,\mathcal{D}_{f_\theta}) = \min_{1\le j\le d}|z_j - \theta_j|$, thereby Assumption 3 holds because

$$\frac{1}{\delta}\mathbb{P}_{\text{true}}\left\{\min_{1\le j\le d}|z_j-\theta_j| < \delta\right\} \le \frac{1}{\delta}\sum_{j=1}^d \mathbb{P}^{z_j}_{\text{true}}\{|z_j-\theta_j|<\delta\} \le d\mu.$$

Fourth, Assumption 4 is verified in Lemma EC.12 in Online Appendix EC.2.4.1.

Let $p \in (1,\infty)$ and $t > 0$. Using Lemma EC.13 in Online Appendix EC.2.4.1, $\int_0^1 \sqrt{\log \mathcal{N}(\epsilon\rho_n;\mathcal{E},\|\cdot\|_{\mathbb{P}_n,2})}d\epsilon \le \sqrt{d\log(1+B/\rho_n)} + \sqrt{d\pi}$; thus, by Theorem 1, we have with probability at least $1-e^{-t}$, for all $\theta \in \Theta$,

$$\left|\mathcal{R}_{\mathbb{P}_n,p}(\rho_n;f_\theta) - \rho_n\left(\mathbb{E}_{\mathbb{P}_n}\left[\sum_{j=1}^d |h_j|^q \mathbf{1}\{z_j<\theta_j\} + |b_j|^q\mathbf{1}\{\theta_j>z_j\}\right]\right)^{\frac{1}{q}}\right| \le C_1\rho_n^{2\wedge p} + \frac{C_2\rho_n}{\sqrt{n}}\left(\sqrt{d\log(1+B/\rho_n)} + \sqrt{d\pi}\right) + M\rho_n\sqrt{\frac{t}{2n}}.$$

**4.3.2. Neural Networks.** In this section, we consider a two-layer network with leaky ReLU activations $\sigma(z) = z$ if $z \ge 0$ and $\sigma(z) = az$ if $z < 0$, where $a > 0$. As before, we consider a $K$-class classification. Let $z = (x,y) \in \mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y}$ is the probability simplex in $\mathbb{R}^K$. Suppose $\mathsf{d}(\tilde z, z) = \|\tilde x - x\|_2 + \infty\mathbf{1}\{\tilde y \ne y\}$.

**Example 5** (Leaky ReLU Network). Let $\theta = (W_1, W_2)$, where $W_1 \in \mathbb{R}^{d_1\times d}$ and $W_2 \in \mathbb{R}^{K\times d_1}$ are weight matrices. Define a two-layer ReLU network with cross-entropy loss:

$$f_\theta(z) := \ell(W_2\sigma(W_1x), y) = -\log\frac{\sum_{k=1}^K y_k \exp(W_{2,k}\sigma(W_1x))}{\sum_{k=1}^K \exp(W_{2,k}\sigma(W_1x))},$$

where $W_{2,k}$ is the $k$th row of $W_2$. Denote by $\sigma'(x)$ the diagonal matrix whose $j$th diagonal equals $\mathbf{1}\{x_j \ge 0\}$. Using the chain rule, at differentiable point, we have

$$\|\nabla f_\theta(z)\|_2 = \|\nabla\ell(W_2\sigma(W_1x),y)^\top W_2\sigma'(W_1x)W_1\|_2,$$

where we have adopted the convention that the gradient is a row vector. Assume $\mathcal{X}$ is compact; $\|W_2\|_{op}\|W_1\|_{op} \le 1$, where $\|\cdot\|_{op}$ denotes the matrix operator norm; and the marginal distribution of $\mathbb{P}^x_{\text{true}}$ on $\mathcal{X}$ is continuous.

Now we verify the assumptions required by Theorem 1. Assumption 1 is satisfied with $L = 0$ because of the Lipschitz continuity of $\ell$ and piecewise linearity of the ReLU activation function. Assumption 2 is satisfied due to the Lipschitz continuity of $\ell$ and $\sigma$. Because $\mathbb{P}_{\text{true}}$ is continuous, $\mathcal{D}_{f_\theta}$ is bounded, and $\Theta$ is compact, the conditional density $d\mathbb{P}_{\text{true}}(\mathcal{D}_{f_\theta})$ is uniformly bounded over $\theta \in \Theta$; hence, Assumption 3 is satisfied. Finally, Assumption 4 is verified in Online Appendix EC.2.4.2.

Let $t > 0$ and $\rho_n = \rho_0/\sqrt{n}$. By Lemma EC.15 in Online Appendix EC.2.4.2, $\mathbb{E}_\otimes[\mathfrak{R}_n(\mathcal{I}_\rho)] \le \sqrt{\frac{Cdd_1\log d_1\log(n+1)}{n}}$; hence, using Theorem 1, there exist constants $\bar\rho, C > 0$ such that for all $\rho_0 < \bar\rho$, with probability at least $1-e^{-t}$, for every $\theta \in \Theta$,

$$\begin{aligned}&|\mathcal{R}_{\mathbb{P}_n,2}(\rho_n;f_\theta) - \rho_n\|\|\nabla f_\theta\|_*\|_{\mathbb{P}_n,2}| \\ &\le C_1\rho_n^2 + 2M\rho_n\sqrt{\frac{Cdd_1\log d_1\log(n+1)}{n}} + M\rho_n\sqrt{\frac{t}{2n}} \\ &= \tilde O(dd_1/n).\end{aligned}$$

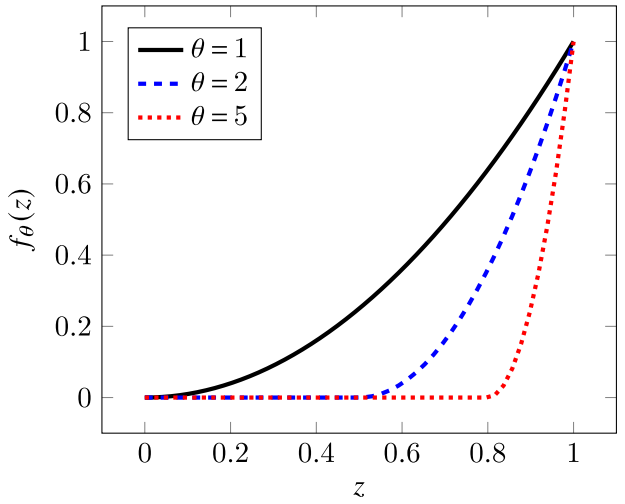
## 5. Variation Regularization Effect of 1-Wasserstein DRO

In this section, we study $p = 1$, which turns out to be qualitatively different from $p > 1$, as hinted from the remainder term of Lemma 1. We study a simple example in Section 5.1 that motivates our main result in Section 5.2 and exemplifies our results in Section 5.3.

### 5.1. Motivating Example

**Example 6.** Let $\mathcal{Z} = [0,1]$ with $\mathsf{d}(\tilde z, z) = |\tilde z - z|$ and $\mathbb{Q} = \text{Uniform}(0,1)$. Consider

**Figure 2.** (Color online) Plots of $f_\theta(z) = (\theta z - \theta + 1)_+^2$, $z \in [0,1]$




$$f_\theta(z) = (\theta z - \theta + 1)_+^2, \quad \theta \geq 1,$$

illustrated in Figure 2. Then $f_\theta(\cdot)$ is differentiable and $|\partial f_\theta|(z) = |f'_\theta(z)| = 2\theta(\theta z - \theta + 1)_+$. By Definition 1, we have $\mathcal{V}_{\mathbb{Q},\infty}(f_\theta) = \||f_\theta|\|_{\mathbb{Q},\infty} = |f'_\theta(1)| = 2\theta$. Let $\rho \in (0, 1/2)$. We claim that the worst-case distribution $\mathbb{P}_*$ has the form

$$\mathbb{P}_* = \mathbb{Q}_{|\{z<z_*\}} + \mathbb{Q}\{z \geq z_*\} \cdot \boldsymbol{\delta}_1, \quad z_* \in [0,1],$$

where $\mathbb{Q}_{|\{z<z_*\}}$ is the restriction of $\mathbb{Q}$ on $\{z < z_*\}$. To see this, observe from the convexity of $f$ that, for any $\lambda \geq 0$, $\sup_{\tilde{z}\in\mathcal{Z}}\{f_\theta(\tilde{z}) - f_\theta(z) - \lambda|\tilde{z} - z|\}$ attains its maximum at $\tilde{z} = 1$, and there exists $z_* \in [0,1]$ such that

$$\sup_{\tilde{z}\in\mathcal{Z}} \{f_\theta(\tilde{z}) - f_\theta(z) - \lambda|\tilde{z} - z|\} \begin{cases} > 0, & \forall z > z_*, \\ = 0, & \forall z = z_*. \end{cases}$$

Consequently, using the dual formulation (D) and the structure of the worst-case distribution (Gao and Kleywegt 2022), we prove the claim. Solving for $\rho = \mathbb{E}_\mathbb{Q}[(1-z)\mathbf{1}\{z > z_*\}]$ yields $z_* = 1 - \sqrt{2\rho}$. It follows that

$$\begin{aligned} \mathcal{R}_{\mathbb{Q},1}(\rho; f_\theta) &= \mathbb{E}_\mathbb{Q}[(1 - (\theta z - \theta + 1)_+^2)\mathbf{1}\{z > z_*\}] \\ &= \begin{cases} 2\rho\theta - (2\rho)^{\frac{3}{2}}\theta^2/3, & \theta \leq 1/\sqrt{2\rho}, \\ \dfrac{2}{3\theta} + \sqrt{2\rho} - 1/\theta, & \theta > 1/\sqrt{2\rho}. \end{cases} \end{aligned}$$

Therefore,

$$\rho\mathcal{V}_{\mathbb{Q},\infty}(f_\theta) - \mathcal{R}_{\mathbb{Q},1}(\rho; f_\theta) = \begin{cases} (2\rho)^{\frac{3}{2}}\theta^2/3, & \theta \leq 1/\sqrt{2\rho}, \\ 2\rho\theta - \dfrac{2}{3\theta} - \sqrt{2\rho} + 1/\theta, & \theta > 1/\sqrt{2\rho}. \end{cases}$$

This shows that the remainder $\rho\mathcal{V}_{\mathbb{Q},\infty}(f_\theta) - \mathcal{R}_{\mathbb{Q},1}(\rho; f_\theta)$ may not be of the desired order $O(\rho^2)$ and can even be *linear* in $\rho$ for $\theta > 1/\sqrt{2\rho}$. We remark that for any fixed $\theta$, $\mathcal{R}_{\mathbb{Q},1}(\cdot; f_\theta)$ is not Lipschitz at zero, so there is no functional $\mathcal{V}$ such that the expansion $\mathcal{R}_{\mathbb{Q},1}(\rho; f_\theta) \simeq \rho\mathcal{V}(f_\theta)$ holds.

In Example 6, the worst-case distribution perturbs points in $[z_*, 1]$, which have large slopes, to the boundary point 1 to maximize the loss. Recall from Definition 1 that the maximum rate of change of loss by perturbing a point $z$ equals $\sup_{\tilde{z}\neq z}(f(\tilde{z}) - f(z))_+/\|\tilde{z} - z\|$ and that $\mathcal{V}_{\mathbb{Q},\infty}(f)$ measures the largest possible change of loss by perturbation among all $z \in \operatorname{supp}\mathbb{Q}$. The gap between these two quantities can lead to a remainder with an undesired order. Specifically, in Example 6, let us consider a fixed $\theta$ and a sufficiently small $\rho$. By definition of the global slope, we have $f_\theta(1) - f_\theta(z) = 1 - f_\theta(z) \leq \mathcal{V}_{\mathbb{Q},\infty}(f)(1-z)$ and thus $\mathcal{R}_{\mathbb{Q},1}(\rho; f_\theta) = \mathbb{E}_\mathbb{Q}[(1 - f_\theta(z))\mathbf{1}\{z > z_*\}] \leq \rho\mathcal{V}_{\mathbb{Q},\infty}(f_\theta)$. Thereby, to have an $O(\rho^2)$ remainder, we have to have

$$1 - \frac{\mathcal{R}_{\mathbb{Q},1}(\rho; f_\theta)}{\rho\mathcal{V}_{\mathbb{Q},\infty}(f_\theta)} = 1 - \mathbb{E}_\mathbb{Q}\left[\left(\frac{(1 - f_\theta(z))/(1-z)}{\rho\mathcal{V}_{\mathbb{Q},\infty}(f_\theta)}\right) \cdot (1-z)\mathbf{1}\{z > z_*\}\right] = O(\rho^2).$$

This means for $\mathbb{Q}$-almost every perturbed point $z \in \mathcal{Z}$, we need

$$1 - \frac{(1 - f_\theta(z))/\|1 - z\|}{\mathcal{V}_{\mathbb{Q},\infty}(f_\theta)} = O(\rho). \tag{3}$$

However, for $z > z_*$, we have

$$\begin{aligned} \frac{\sup_{\tilde{z}\neq z}(f_\theta(\tilde{z}) - f_\theta(z))_+/\|\tilde{z} - z\|}{\mathcal{V}_{\mathbb{Q},\infty}(f_\theta)} &= \frac{(1 - f_\theta(z))/(1-z)}{\mathcal{V}_{\mathbb{Q},\infty}(f_\theta)} \\ &= \frac{1 - (\theta z - \theta + 1)^2}{(1-z)\cdot 2\theta} = 1 - \frac{\theta(1-z)}{2}. \end{aligned}$$

As a consequence, perturbing points close to $z_* = 1 - \sqrt{2\rho}$ do not provide sufficient change of the loss compared with $\mathcal{V}_{\mathbb{Q},\infty}(f)$, leading to a large remainder. We remark that Condition (3) can be even more difficult to satisfy in the high-dimensional counterpart of this example. More specifically, consider $\mathcal{Z} \subset [0,1]^d$ and $f_\theta = (\theta\|z\|_2/\sqrt{d} - \theta + 1)_+^2$. In this case, the worst-case distribution would perturb data points to the all-one vector in $\mathbb{R}^d$, denoted as $z_{\max}$. Then only points in the neighborhood $\{z : \|z - z_{\max}\| \leq c\rho\}$ satisfy Condition (3). Suppose $\mathbb{Q}$ is continuous with bounded density, then this set has $\mathbb{Q}$-measure only $O(\rho^d)$, where $d$ is the dimension of $\mathcal{Z}$. Hence, most points being perturbed by the worst-case distribution would not satisfy (3).

The previous discussion suggests that for $p = 1$, the remainder $\rho\mathcal{V}_{\mathbb{Q},\infty}(f) - \mathcal{R}_{\mathbb{Q},1}(\rho; f)$ cannot be of the desired order $O(\rho^2)$ in general. Fortunately, as will be formalized in the next section, one can show that $\mathcal{R}_{\mathbb{Q},1}(\rho; f)$ achieves a fraction of $\rho\mathcal{V}_{\mathbb{Q},\infty}(f)$ uniformly for all $f \in \mathcal{F}$ under mild conditions. Thereby, the variation of loss is still under control by minimizing the Wasserstein robust loss.

### 5.2. Sandwich Theorem

Motivated by the discussion in the previous section, particularly Condition (3), we develop Theorem 2, which is instantiated under two important situations (Corollaries 1 and 2). The proofs are given in Online Appendix EC.3.1.

Define

$$\mathsf{d}_{\mathcal{F}}(f,\tilde{f}) := \max(\|f-\tilde{f}\|_\infty, \|\|f\|_{\mathrm{Lip}} - \|\tilde{f}\|_{\mathrm{Lip}}|).$$

**Theorem 2** (1-Wasserstein DRO). *Let $p=1$. Assume every $f\in\mathcal{F}$ is Lipschitz continuous. Assume further that there exist constants $\varepsilon>0$, $\delta_n,\eta\in(0,1]$ such that for every $f\in\mathcal{F}$, there exist a set $\mathcal{Z}_f\subset\mathcal{Z}$ and a measurable map $\mathcal{T}_f:\mathcal{Z}_f\to\mathcal{Z}$ such that with probability at least $1-\delta_n$,*

$$f(\mathcal{T}_f(z))-f(z)\ge\eta(\|f\|_{\mathrm{Lip}}-\varepsilon)\|\mathcal{T}_f(z)-z\|,\quad\forall z\in\mathcal{Z}_f,$$
$$\mathbb{E}_{\mathbb{P}_n}[\|\mathcal{T}_f(z)-z\|\mathbf{1}\{z\in\mathcal{Z}_f\}]>0. \qquad (\mathbf{T})$$

*Suppose $\rho\le\min_{f\in\mathcal{F}}\mathbb{E}_{\mathbb{P}_n}[\|\mathcal{T}_f(z)-z\|\mathbf{1}\{z\in\mathcal{Z}_f\}]$. Then with probability at least $1-\mathcal{N}(\frac{1}{n};\mathcal{F},\mathsf{d}_{\mathcal{F}})\cdot\delta_n$,*

$$\eta\rho\mathcal{V}_{\mathbb{P}_n,\infty}(f)-\rho\varepsilon-(1+\rho)/n\le\mathcal{R}_{\mathbb{P}_n,1}(\rho;f)\le\rho\mathcal{V}_{\mathbb{P}_n,\infty}(f).$$

This theorem shows that the Wasserstein regularizer $\mathcal{R}_{\mathbb{P}_n,1}(\rho;f)$ is sandwiched by $\rho\mathcal{V}_{\mathbb{P}_n,\infty}(f)$ and its *fraction* $\eta\rho\mathcal{V}_{\mathbb{P}_n,\infty}(f)$, which, according to the discussion in Section 5.1, is generally the best one can hope for. Assumption (T) means that every point $z\in\mathcal{Z}_f$ can be perturbed to some point $\mathcal{T}_f(z)$, resulting in an increment no less than a fixed fraction of $\mathcal{V}_{\mathbb{P}_n,\infty}(f)$ and that the total perturbations $\mathbb{E}_{\mathbb{P}_n}[\mathsf{d}(\mathcal{T}_f(z),z)\mathbf{1}\{z\in\mathcal{Z}_f\}]$ have a positive lower bound uniformly for all $f\in\mathcal{F}$. There is a tradeoff between $\eta$ and $\delta_n$: One can increase $\eta$ at the cost of a smaller $\delta_n$. For loss functions that are spurious, that is, with large probability, $\|\nabla f(z)\|_*$ is much smaller than $\mathcal{V}_{\mathbb{P}_n,\infty}(f)$, we would like to reduce the fraction $\eta$ to increase the probability bound $\delta_n$ so that it has a mild dependence on the dimension of $\mathcal{Z}$. Here we provide two important situations where the condition (T) can be satisfied either probabilistically or deterministically.

**Corollary 1** (Data-Driven 1-Wasserstein DRO). *Assume every $f\in\mathcal{F}$ is Lipschitz continuous. Assume every $f\in\mathcal{F}$ is $\hbar$-semiconvex, that is, there exists $\hbar\in\mathbb{R}$ such that*

$$f(\tilde{z})-f(z)\ge g^\top(\tilde{z}-z)-\hbar\|\tilde{z}-z\|^2,\quad\forall z,\tilde{z}\in\mathcal{Z},$$

*where $g$ is any element in the subdifferential $\partial f(z)$. Assume further that there exists $\eta\in(0,1]$ such that*

$$\alpha:=\inf_{f\in\mathcal{F}}\mathbb{P}_{\mathsf{true}}\Big\{z:\sup_{g\in\partial f(z)}\|g\|_*\ge\eta\|f\|_{\mathrm{Lip}}\Big\}\in(0,1).$$

*Let $c<\alpha$. Denote by $H(a\|b):=a\log\frac{a}{b}+(1-a)\log\frac{1-a}{1-b}$. Then the condition (T) is satisfied by setting*

$$\varepsilon=c\hbar\rho^2,\quad\mathcal{Z}_f=\Big\{z:\sup_{g\in\partial f(z)}\|g\|_*\ge\eta\|f\|_{\mathrm{Lip}}\Big\},$$
$$\delta_n=\exp(-nH(c\|\alpha)).$$

*In addition, with probability at least $1-\exp(-nH(c\|\alpha)+\log\mathcal{N}(\frac{1}{n};\mathcal{F},\mathsf{d}_{\mathcal{F}}))$,*

$$\eta\rho\mathcal{V}_{\mathbb{P}_n,\infty}(f)-c\hbar\rho^2-(1+\rho)/n\le\mathcal{R}_{\mathbb{P}_n,1}(\rho;f)\le\rho\mathcal{V}_{\mathbb{P}_n,\infty}(f).$$

An illustration of this result for linear prediction with Lipschitz loss on a bounded domain is given in Example 7 in Section 5.3.

**Corollary 2** (Lipschitz Regularization). *Assume every $f\in\mathcal{F}$ is Lipschitz continuous. Suppose $\mathrm{diam}(\mathcal{Z})=\infty$ and there exists $z_0\in\mathcal{Z}$ such that*

$$\limsup_{\|\tilde{z}-z_0\|\to\infty}\frac{f(\tilde{z})-f(z_0)}{\|\tilde{z}-z_0\|}=\|f\|_{\mathrm{Lip}},\quad\forall f\in\mathcal{F}, \qquad (\mathbf{L})$$

*then (T) is satisfied for any $\varepsilon>0$ with $\eta=1$ and $\delta=0$. In addition, for all $\rho\ge0$ and $f\in\mathcal{F}$,*

$$\mathcal{R}_{\mathbb{P}_n,1}(\rho;f)=\rho\mathcal{V}_{\mathbb{P}_n,\infty}(f)=\rho\|f\|_{\mathrm{Lip}}.$$

This provides a situation of exact equivalence between Wasserstein DRO and regularization. As detailed in the proof, if (L) holds for some $z_0\in\mathcal{Z}$ then it holds for every $z\in\mathcal{Z}$. Hence, Condition (L) means that the Lipschitz norm is attained approximately between $z\in\mathrm{supp}\,\mathbb{P}_n$ and some distant point $\tilde{z}$: for any $\epsilon>0$ and $r>0$, there exists $\tilde{z}=:\mathcal{T}_f(z)$ such that $\|\mathcal{T}_f(z)-z\|>r$ and $f(\mathcal{T}_f(z))-f(z)\ge(\|f\|_{\mathrm{Lip}}-\epsilon)\|\mathcal{T}_f(z)-z\|$. The (approximately) worst-case distribution perturbs some point $z_{i_0}\in\mathrm{supp}\,\mathbb{P}_n$ to $\mathcal{T}_f(z_{i_0})$ with tiny probability $\delta/n$, where $\delta\in(0,1)$, and therefore has the form

$$\frac{1}{n}\sum_{i\ne i_0}\boldsymbol{\delta}_{z_i}+\frac{1-\delta}{n}\boldsymbol{\delta}_{z_{i_0}}+\frac{\delta}{n}\boldsymbol{\delta}_{\mathcal{T}_f(z_{i_0})}.$$

Condition (L) can be satisfied when $f$ is convex and Lipschitz, which has been considered in Mohajerin Esfahani and Kuhn (2018) and Shafieezadeh-Abadeh et al. (2019). It also holds for nonconvex losses; a one-dimensional example is the inverse $S$-shaped curve plotted in Figure 3. We illustrate this corollary in Example 8 in Section 5.3 for linear prediction with Lipschitz loss on an unbounded domain.

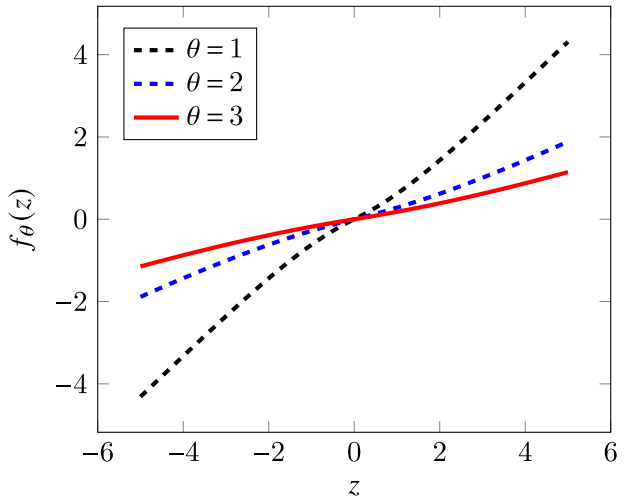
### 5.3. Applications

In this section, we consider two examples on linear prediction, covering two particular cases of $p=1$ (Corollaries 1 and 2). Let $z=(x,y)$, where $x\in\mathcal{X}\subset(\mathbb{R}^d,\|\cdot\|)$, and $y\in\mathbb{R}$ for regression, whereas for classification, $y\in\{\pm1\}$. To ease the exposition, we assume $\mathsf{d}(z,\tilde{z})=\|x-\tilde{x}\|+\infty\cdot\mathbf{1}\{y\ne\tilde{y}\}$, thereby we can omit the $y$-component when we compute $\|\nabla f(z)\|_*$.

**Example 7** (Lipschitz Loss on a Bounded Domain, $p=1$). Suppose the loss function has a form

$$f_\theta(z):=l(\theta^\top x,y):=\begin{cases}\ell(\theta^\top x-y), & \text{regression},\\ y\ell(\theta^\top x), & \text{binary classification},\end{cases} \qquad (4)$$

**Figure 3.** (Color online) Plots of $f_\theta(z) = \mathrm{sgn}(z)\ln((1+\exp(z/\theta))/2)$, $z \in \mathbb{R}$




where $\theta \in \Theta \subset \{\theta \in \mathbb{R}^d : \|\theta\|_* \le B\}$ for some $B > 0$ and $\ell : \mathbb{R} \to \mathbb{R}$ is $L_\ell$-Lipschitz. Denote by $l'(\cdot, y)$ the derivative of $l$ with respect to its first argument. Assume $\ell$ has $\hbar_\ell$-Lipschitz gradient. Assume there exists $\eta \in (0,1]$ such that

$$0 < \alpha := \begin{cases} \inf\limits_{\theta\in\Theta} \mathbb{P}_{\text{true}}\{(x,y) : |\ell'(\theta^\top x - y)| \ge \eta L_\ell\}, \\ \qquad\qquad \text{regression}, \\ \inf\limits_{\theta\in\Theta} \mathbb{P}_{\text{true}}\{(x,y) : |\ell'(\theta^\top x)| \ge \eta L_\ell\}, \\ \qquad\qquad \text{binary classification}. \end{cases} \tag{5}$$

Let us verify the assumptions in Corollary 1. We have $\|\nabla f_\theta(z)\|_* = \|\theta\|_* l'(\theta^\top x, y)$; thus, $\|f_\theta\|_{\text{Lip}} = \|\theta\|_* \sup_{(x,y)\in\mathcal{Z}} l'(\theta^\top x, y) \le L_\ell \|\theta\|_*$ and $f_\theta$ is $\hbar_\ell B^2$-semiconvex. Moreover, the constraints in (5) are equivalent to $\|\nabla f_\theta(z)\|_* \ge \eta \|f_\theta\|_{\text{Lip}}$; thereby, $\alpha$ in Corollary 1 is well defined. By Online Appendix EC.3.2, $\log \mathcal{N}(\epsilon; \mathcal{F}, \mathsf{d}_\mathcal{F}) \le 1 + d \log\left(\frac{L_\ell \mathrm{diam}(\mathcal{X}) \vee (L_\ell + B\hbar_\ell \mathrm{diam}(\mathcal{X}))}{\epsilon}\right)$. Let $c < \alpha$. Then using Corollary 1, with probability at least

$$1 - d\exp(-nH(c\|\alpha)) + \log(1 + nL_\ell \mathrm{diam}(\mathcal{X}) \vee n(L_\ell + B\hbar_\ell \mathrm{diam}(\mathcal{X})))),$$

it holds for all $\theta \in \Theta$ that

$$\eta\rho_n\|\theta\|_* \max_{1\le i\le n} |l'(\theta^\top x_i^n, y_i^n)| - c\hbar_\ell B^2 \rho_n^2 - (1+\rho_n)/n \le \mathcal{R}_{\mathbb{P}_n,1}(\rho_n; f_\theta) \le \rho_n \|\theta\|_* \max_{1\le i\le n} |l'(\theta^\top x_i^n, y_i^n)|.$$

**Example 8** (Lipschitz Loss on an Unbounded Domain). Consider the loss function defined in (4). Suppose $\mathcal{X} = \mathbb{R}^d$. Assume additionally $\limsup_{|t|\to\infty} \frac{\ell(t)}{|t|} = L_\ell$. Examples of $\ell(t)$ include convex losses such as hinge loss $(1-t)_+$, softplus (logistic) loss $\log(1+e^t)$, and nonconvex losses such as inverse S-shaped curve $\mathrm{sgn}(t)\log\left(\frac{1}{2}(1+e^t)\right)$.

For classification, assume further that there exists $(x_0, y_0) \in \mathrm{supp}\ \mathbb{P}_n$ with $y_0 = 1$. Then $f_\theta$ is Lipschitz continuous with constant bounded by $L_\ell B$ and $\limsup_{\|x\|\to\infty} l(\theta^\top x, y) = L_\ell \|\theta\|_* = \|f_\theta\|_{\text{Lip}}$; thus, (L) in Corollary 2 is satisfied, and we have

$$\mathcal{R}_{\mathbb{P}_n,1}(\rho; f_\theta) = \rho \cdot \|f_\theta\|_{\text{Lip}} = \rho \cdot L_\ell \|\theta\|_*, \quad \forall \theta \in \Theta.$$

We remark that this result relaxes the convexity assumption in the equivalence results derived in Mohajerin Esfahani and Kuhn (2018) and Shafieezadeh-Abadeh et al. (2019).

## 6. Discussions

### 6.1. Comparison Among Wasserstein Orders

Comparing Theorems 1 and 2, we observe that as the Wasserstein order $p$ decreases, stronger assumptions are needed to obtain the asymptotic equivalence between the Wasserstein DRO and variation regularization, which sheds light on the modeling choice of Wasserstein order $p$. Specifically, when $p = \infty$, only local assumptions on the continuity and jump are needed (Assumption 5(I)); when $p \in [1, \infty)$, global growth condition is required (Assumptions 5, (II) and (III)), and the cases $p \ge 2$, $p < 2$ have different orders of gap $O(\rho_n^{p\wedge 2})$; when $p = 1$, the lower and upper bound in Theorem 2 cannot be matched in general, but only a sandwich inequality is available.

As shown in the proof, this can be explained from the qualitative differences in the worst-case distribution among $p = 1$, $p \in (1,2)$, and $p \in [2, \infty]$. For $p \ge 2$, the largest distance of perturbation is bounded for all empirical points with high probability when $\rho_n = O(1/\sqrt{n})$, whereas for $p \in (1,2)$, the worst-case distribution tends to perturb the empirical points with a large distance, resulting a lower order of the remainder $O(\rho_n^p)$; when $p = 1$, the worst-case distribution can even perturb the empirical points to arbitrarily far with a tiny probability (see the comment after Corollary 2).

### 6.2. Extension to General Losses on a Metric Space

In previous sections, we primarily focus on losses on a Banach space. In this section, we show that the results hold for a general metric space without isolated point.

Define

$$G_f(\delta, z) := \sup_{\tilde{z}\in\mathcal{Z}:\, \mathsf{d}(\tilde{z}, z)\le\delta} f(\tilde{z}) - f(z), \quad \delta \ge 0,\ z \in \mathcal{Z}. \tag{6}$$

We impose the following assumption.

**Assumption 5** (Growth, Continuity, and Jump).

(I) *When $p = \infty$, assume there exists $\delta_0, M \ge 0$, and $H \in \mathcal{L}^1(\mathbb{P}_{\text{true}})$ such that for all $f \in \mathcal{F}$ and all $z \in \mathcal{Z}$,*

$$|G_f(\delta, z) - |\partial f|(z)\delta| \le H(z)\delta^2 + M(\delta - \mathsf{d}(z, \mathcal{D}_f))_+, \quad \forall \delta < \delta_0.$$

(II) *When* $p \in (2,\infty)$, *assume there exists* $\delta_0, L, M \geq 0$, *and* $H \in \mathcal{L}^{\frac{p}{p-2}}(\mathbb{P}_{\text{true}})$ *such that for all* $f \in \mathcal{F}$ *and all* $z \in \mathcal{Z}$,

$$G_f(\delta,z) - |\partial f|(z)\delta \leq H(z)\delta^2 + L\delta^p + M(\delta - \mathsf{d}(z,\mathcal{D}_f))_+, \ \forall \delta \geq 0,$$
$$G_f(\delta,z) - |\partial f|(z)\delta \geq -H(z)\delta^2 - M(\delta - \mathsf{d}(z,\mathcal{D}_f))_+, \quad \forall \delta \leq \delta_0.$$

(III) *When* $p \in (1,2]$, *assume there exists* $L, M \geq 0$ *such that for all* $f \in \mathcal{F}$ *and all* $z \in \mathcal{Z}$,

$$-L\delta^p - M(\delta - \mathsf{d}(z,\mathcal{D}_f))_+ \leq G_f(\delta,z) - |\partial f|(z)\delta \leq L\delta^p + M(\delta - \mathsf{d}(z,\mathcal{D}_f))_+, \quad \forall \delta \geq 0.$$

In Lemma EC.8 in Online Appendix EC.2.2, we will show that a sufficient condition ensuring Assumption 5 is by assuming that Assumptions 1 and 2 hold. The case of $M = 0$ corresponds to smooth losses. By replacing Assumptions 1 and 2 with Assumption 5 and substituting $\|\cdot - \cdot\|$ by $\mathsf{d}(\cdot,\cdot)$, Theorems 1 and 2 remain to hold. In fact, in the Online Appendix, Theorem 1 is proved by assuming Assumption 5, and the proof of Theorem 2 applies to the general setting directly.

Next, we illustrate our results for manifold regularization (Example 9) and intensity estimation of point processes (Example 10).

**Example 9** (Manifold Regularization). Suppose $\mathcal{Z} \subset \mathbb{R}^d$ is a Riemannian manifold and $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable with bounded Hessian. Then $|\partial f|(z) = \text{Exp}(\nabla f)$, where Exp denotes the exponential map (Carmo 1992). For example, when $\mathcal{Z}$ is the unit sphere $\{z \in \mathbb{R}^d : \|z\|_2 = 1\}$, then $\|\nabla f\|_*(z) = \text{Exp}(\nabla f) = \|(I_d - zz^\top)\ \nabla f(z)\|_2$, where $I_d$ denotes the $d$-dimensional identity matrix. By Theorem 1 for the general metric space, there exists $a, C > 0$ such that for all $\rho_0 < a$, $n \in \mathbb{N}_{\geq 1}$ and $f \in \mathcal{F}$,

$$|\mathcal{R}_{\mathbb{P}_n,2}(\rho_n;f) - \rho_n \mathcal{V}_{\mathbb{P}_n,2}(f)| \leq C\rho_n^2,$$

where $\mathcal{V}_{\mathbb{P}_n,2}(f) = \|\,\|\text{Exp}(\nabla f)\|_2\,\|_{\mathbb{P}_n,2}$. This establishes a connection between Wasserstein DRO and Laplacian regularization in manifold optimization (Belkin et al. 2006).

As another example, we consider the case where the distance $\mathsf{d}(\tilde{z},z)$ is defined through another Wasserstein distance, in which each sample point $z$ is viewed as a measure on a metric space $(\Xi,\mathsf{d}_\Xi)$. We define the metric $\mathsf{d}$ on $\mathcal{Z}$ as a 2-Wasserstein metric

$$\mathsf{d}(\tilde{z},z) = \mathcal{W}_\Xi(\tilde{z},z) := \inf_{\gamma \in \Gamma(\tilde{z},z)} \|\mathsf{d}_\Xi\|_{\gamma,2},$$

where $\Gamma(\tilde{z},z)$ represents the set of Borel measures on $\Xi^2$ with marginal measures $\tilde{z}$ and $z$. This setup occurs in various applications. For instance, let $\mathcal{Z}$ be the space of nonhomogeneous Poisson processes on $\Xi = [0,T]$. Then each $z \in \mathcal{Z}$ can be viewed as a distribution of sample paths on $\Xi$. Each sample path is identified with a counting measure on $\Xi$, and the distance $\mathsf{d}(\tilde{z},z)$ between two sample paths $\tilde{z}$ and $z$ is measured by the Wasserstein distance between counting measures on $\Xi$. This is called nested Wasserstein distance in Gao and Kleywegt (2022, section 4.2). As another example, let $\mathcal{Z}$ be the space of black-and-white images with fixed resolution $r \times r$. Then each image $z \in \mathcal{Z}$ can be viewed as a two-dimensional histogram on the space of pixels $\Xi = \{1,\ldots,r\}^2$, with each pixel representing a bin. The distance $\mathsf{d}(\tilde{z},z)$ between two images $\tilde{z}$ and $z$ is measured by the Wasserstein distance between two-dimensional histograms on $\Xi$. This is called Wasserstein of Wasserstein loss in Dukler et al. (2019).

**Example 10** (Intensity Estimation for Point Processes). Let $\Xi \subset (\mathbb{R}^d, \|\cdot\|)$. Consider the problem of estimating the intensity function $f : \Xi \to \mathbb{R}$ of a point process. Suppose the negative log-likelihood of a sample path $z_i^n = \sum_{m=1}^{M_i} \boldsymbol{\delta}_{\xi_{i,m}}$ has the form

$$\int_\Xi f(\xi)d\xi - \sum_{m=1}^{M_i} \log f(\xi_{i,m}) = \int_\Xi f(\xi)d\xi - \mathbb{E}_{\xi\sim z_i^n}[\log f(\xi)],$$

which holds for, for example, the inhomogeneous Poisson process. Then the distributionally robust negative log-likelihood function is

$$\mathcal{L}_n^{\text{rob}}(f;\rho_n) = \sup_{\mathbb{P}:\mathcal{W}_2(\mathbb{P},\mathbb{P}_n)\leq\rho_n} \left\{\int_\Xi f(\xi)d\xi - \mathbb{E}_{z\sim\mathbb{P}}[\mathbb{E}_{\xi\sim z}[\log f(\xi)]]\right\}.$$

Assume $\log f$ has Lipschitz gradient bounded by $\hbar > 0$. Then in Online Appendix EC.4, we show that

$$|\mathcal{L}_n^{\text{rob}}(f;\rho_n) - \mathcal{L}_n^{\text{reg}}(f;\rho_n)| \leq \frac{C}{n},$$

where

$$\mathcal{L}_n^{\text{reg}}(f;\rho_n) := \int_\Xi f(\xi)d\xi - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\xi\sim z_i^n}[\log f(\xi)] + \rho_n\left(\frac{1}{n}\sum_{i=1}^n \sum_{m=1}^{M_i} \|\nabla_\xi \log f(\xi_{i,m})\|_2^2\right)^{\frac{1}{2}},$$

where $\nabla_\xi \log f(\xi_{i,m})$ is also known as the *score function* in statistics. Therefore, this example demonstrates that 2-Wasserstein DRO penalizes the norm of the score function.

## 7. Generalization Guarantees for Adversarial Robust Learning

In this section, we study *adversarial robust learning*, as an application of our developed theory on variation regularization.

Recent studies (Szegedy et al. 2013, Goodfellow et al. 2015) have shown that machine learning models are vulnerable to adversarial attacks. For example, by adding a small perturbation adversarially to an image, a well-trained classification model may make a wrong

prediction, even when such perturbation is imperceptible to human eyes. To improve the robustness and generalization of machine learning models, one popular approach is the following adversarial robust learning framework, which considers the following *empirical adversarial risk minimization* problem

$$\min_{f\in\mathcal{F}}\left\{\mathcal{A}_n(\rho;f):=\frac{1}{n}\sum_{i=1}^{n}\sup_{x\in\mathcal{X}:\|x-x_i^n\|\le\rho}\ell(f(x),y_i^n)\right\},\tag{7}$$

where $\ell:\mathbb{R}\times\{\pm1\}\to[0,1]$ is a classification loss function such as cross-entropy, $\mathcal{F}$ is the hypothesis family on $\mathcal{X}$, and $\rho>0$ is a small real number. Note that (7) is the dual formulation (D) of $\infty$-Wasserstein DRO when $\mathbb{Q}=\mathbb{P}_n$. The *population adversarial risk minimization* corresponding to (7) is

$$\min_{f\in\mathcal{F}}\left\{\mathcal{A}(\rho;f):=\mathbb{E}_{(x,y)\sim\mathbb{P}_{\text{true}}}\left[\sup_{\tilde{x}\in\mathcal{X}:\|\tilde{x}-x\|\le\rho}\ell(f(\tilde{x}),y)\right]\right\},$$

which is the dual formulation (D) of $\infty$-Wasserstein DRO when $\mathbb{Q}=\mathbb{P}_{\text{true}}$. One fundamental question that this minimax formulation raises is to characterize the generalization capability of the adversarial risk, that is, the gap between the empirical adversarial risk and the population adversarial risk (Attias et al. 2019, Yin et al. 2019, Awasthi et al. 2020).

An immediate consequence of Theorem 1 is the following.

**Example 11** (Adversarial Robust Learning and Total Variation Regularization). Assume $\ell$ is smooth and every $f\in\mathcal{F}$ is piecewise smooth, which is satisfied by cross-entropy loss and the ReLU network. Assume $\mathbb{P}_{\text{true}}$ is a continuous distribution on a compact set $\mathcal{X}\times\{\pm1\}$. Then Assumptions 1–3 are satisfied immediately. Thereby, Theorem 1 shows that with probability at least $1-e^{-t}$, Problem (7) is equivalent to an empirical total variation regularization problem

$$\min_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i^n),y_i^n)+\rho\cdot\mathbb{E}_{\mathbb{P}_n}[\ell'(f(x),y)\|\nabla f(x)\|_*]\right\}+\epsilon_n,$$

where the remainder $\epsilon_n=\rho^2(C+\|H\|_{\mathbb{P}_n,1})+2\rho\mathbb{E}_\otimes[\mathfrak{R}_n(\mathcal{J}_\rho)]+\rho\sqrt{\frac{t}{2n}}$ with terms defined in Theorem 1 and its assumptions, and $\mathcal{J}_\rho:=\{x\mapsto\mathbf{1}\{\mathsf{d}(x,\mathcal{D}_f)<\rho\}:f\in\mathcal{F},\mathcal{D}_f\neq\varnothing\}$.

We develop an upper bound on the generalization gap $\mathcal{A}(\rho;f)-\mathcal{A}_n(\rho;f)$, whose proof is given in Online Appendix EC.5. Define $|\partial\mathcal{F}|=\{|\partial f|:f\in\mathcal{F}\}$ and $|\partial(-\mathcal{F})|=\{|\partial(-f)|:f\in\mathcal{F}\}$, recalling $|\partial f|$ is the local slope of $f$ defined in Definition 1.

**Theorem 3.** *Under the setting of Example 11, assume $\ell$ is $L_\ell$-Lipschitz, and each piece of $f\in\mathcal{F}$ has gradient bounded by $L>0$. Let $t>0$. Then there exists $\bar{\rho},C,M>0$ such that for all $\rho<\bar{\rho}$, with probability at least $1-3e^{-t}$, for every $f\in\mathcal{F}$,*

$$\begin{aligned}\mathcal{A}(\rho;f)-\mathcal{A}_n(\rho;f)\le{}&2L_\ell(\mathbb{E}_\otimes[\mathfrak{R}_n(\mathcal{F})]\\&+\rho\mathbb{E}_\otimes[\mathfrak{R}_n(|\partial\mathcal{F}|\cup|\partial(-\mathcal{F})|)]\\&+C\rho\mathbb{E}_\otimes[\mathfrak{R}_n(\mathcal{J}_\rho)])\\&+C(1+(L+1)L_\ell\rho)\sqrt{\frac{t}{2n}}+C(L_\ell+1)\rho^2.\end{aligned}$$

Theorem 3 unveils that, apart from $\mathfrak{R}_n(\mathcal{F})$ that appears in the generalization bound for the empirical risk minimization, the Rademacher complexity of the local slope $\mathfrak{R}_n(|\partial\mathcal{F}|\cup|\partial(-\mathcal{F})|)$ plays a crucial role in controlling the generalization gap in adversarial robust learning. When $\rho=0$, our bound reduces to the usual generalization bound for empirical risk minimization. When $\mathcal{F}$ is a family of smooth losses, the bound in Theorem 3 can be simplified to

$$\begin{aligned}\mathcal{A}(\rho;f)-\mathcal{A}_n(\rho;f)\le{}&2L_\ell(\mathbb{E}_\otimes[\mathfrak{R}_n(\mathcal{F})]+\rho\mathbb{E}_\otimes[\mathfrak{R}_n(\|\nabla\mathcal{F}\|_*)])\\&+(1+LL_\ell\rho)\sqrt{\frac{t}{2n}}+L_\ell C\rho^2,\end{aligned}$$

where $\|\nabla\mathcal{F}\|_*=\{\|\nabla f\|_*:f\in\mathcal{F}\}$. When $\mathcal{F}$ is a family of linear losses $\mathcal{F}=\{f_\theta=\theta^\top x:\theta\in\Theta\}$, the bound becomes

$$\begin{aligned}\mathcal{A}(\rho;f_\theta)-\mathcal{A}_n(\rho;f_\theta)\le{}&2L_\ell(\mathbb{E}_\otimes[\mathfrak{R}_n(\mathcal{F})]\\&+\rho\mathbb{E}_\otimes[\mathfrak{R}_n(\{\|\theta\|_*:\theta\in\Theta\})])+(1+LL_\ell\rho)\sqrt{\frac{t}{2n}},\end{aligned}$$

which leads to the bounds developed in Awasthi et al. (2020, theorem 4 and lemma 2). The $\mathfrak{R}_n(\|\nabla\mathcal{F}\|_*)$ factor appears to be new in the literature but should make intuitive sense. Indeed, if the complexity of the gradient norm functions is small, the model is more robust to the adversarial perturbations and thus tends to generalize better.

## 8. Concluding Remarks

Regularization is at the core of many learning and decision-making tasks in the world of big data. In this paper, we introduce a new family of regularization schemes, termed as variation regularization, and develop a framework connecting Wasserstein DRO and variation regularization. The general theory developed in this paper expands the connection between robustness and regularization from simple or smooth losses on Euclidean space to general possibly nonsmooth losses on a metric space, which greatly enlarge its practicality. Thereby, it fills the gap between the empirical success of Wasserstein DRO and the theoretical understanding of its regularization effect and helps to explain why Wasserstein DRO works from a regularization perspective or why variation regularization in deep learning works

from a robust perspective. Moreover, the developed theory makes a step toward the understanding of the generalization capability of robust learning. For example, as we illustrate in Section 7, our theory helps to develop new generalization bounds for adversarial robust learning, which is an interesting phenomenon in deep learning but does not yet have a full theoretical understanding. In the follow-up work (An and Gao 2021, Gao 2022), our theory serves as an important building block for proving the finite-sample performance guarantees for Wasserstein DRO.

## Acknowledgments

The authors thank Zhen Yang for assisting with the proof of the strong duality for $\infty$-Wasserstein DRO.

## References

Abdullah MA, Ren H, Ammar HB, Milenkovic V, Luo R, Zhang M, Wang J (2019) Wasserstein robust reinforcement learning. Preprint, submitted July 30, https://arxiv.org/abs/1907.13196.

Ambrosio L, Fusco N, Pallara D (2000) *Functions of Bounded Variation and Free Discontinuity Problems* (Oxford University Press, Oxford, UK).

Ambrosio L, Gigli N, Savare G (2008) *Gradient Flows: In Metric Spaces and in the Space of Probability Measures* (Birkhäuser, Basel, Switzerland).

An Y, Gao R (2021) Generalization bounds for (Wasserstein) robust optimization. *Adv. Neural Inform. Processing Systems* 34:10382–10392.

Anderson E, Philpott A (2022) Improving sample average approximation using distributional robustness. *INFORMS J. Optim.* 4(1):90–124.

Attias I, Kontorovich A, Mansour Y (2019) Improved generalization bounds for robust learning. *Algorithmic Learning Theory* (PMLR), 162–183.

Aubin JP, Frankowska H (2009) *Set-Valued Analysis* (Springer Science & Business Media, Boston, MA).

Awasthi P, Frank N, Mohri M (2020) Adversarial learning guarantees for linear hypotheses and neural networks. *Proc. Internat. Conf. on Machine Learn.* (PMLR), 431–441.

Bartl D, Drapeau S, Obłój J, Wiesel J (2021) Sensitivity analysis of Wasserstein distributionally robust optimization problems. *Proc. Royal Soc. A* 477(2256):20210176.

Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. *INFORMS Tutorials in Operations Research* (INFORMS), 1–19.

Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Machine Learn. Res.* 7(11):2399–2434.

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

Bertsimas D, Copenhaver MS (2018) Characterization of the equivalence of robustification and regularization in linear and matrix regression. *Eur. J. Oper. Res.* 270(3):931–942.

Blanchet J, Kang Y (2020) Semi-supervised learning based on distributionally robust optimization. *Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods.* 5:1–33.

Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* 44(2):565–600.

Blanchet J, Kang Y, Murthy K (2019) Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probability* 56(3):830–857.

Blanchet J, Murthy K, Si N (2022) Confidence regions in Wasserstein distributionally robust estimation. *Biometrika* 109(2):295–315.

Calafiore GC, El Ghaoui L (2006) On distributionally robust chance-constrained linear programs. *J. Optim. Theory Appl.* 130(1):1–22.

Carmo MPD (1992) *Riemannian Geometry* (Birkhäuser, Basel, Switzerland).

Cheeger J (1999) Differentiability of Lipschitz functions on metric measure spaces. *Geometric Functional Anal.* 9(3):428–517.

Chen R, Paschalidis IC (2018) A robust learning approach for regression models based on distributionally robust optimization. *J. Machine Learn. Res.* 19(1):517–564.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

Derman E, Mannor S (2020) Distributional robustness and regularization in reinforcement learning. Preprint, submitted March 5, https://arxiv.org/abs/2003.02894.

Duchi J, Namkoong H (2019) Variance-based regularization with convex objectives. *J. Machine Learn. Res.* 20(1):2450–2504.

Duchi J, Hashimoto T, Namkoong H (2020) Distributionally robust losses for latent covariate mixtures. Preprint, submitted July 28, https://arxiv.org/abs/2007.13982.

Dukler Y, Li W, Lin A, Montufar G (2019) Wasserstein of Wasserstein loss for learning generative models. Chaudhuri K, Salakhutdinov R, eds. *Proc. 36th Internat. Conf. on Machine Learn.*, vol. 97, (PMLR, Long Beach, CA), 1716–1725.

Erdoğan E, Iyengar G (2006) Ambiguous chance constrained problems and robust optimization. *Math. Programming* 107(1–2):37–61.

Gao R (2022) Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Oper. Res.* Forthcoming.

Gao R, Kleywegt AJ (2022) Distributionally robust stochastic optimization with Wasserstein distance. *Math. Oper. Res.* Forthcoming.

Gao R, Chen X, Kleywegt AJ (2017) Wasserstein distributionally robust optimization and variation regularization. Preprint, submitted December 17, https://arxiv.org/abs/1712.06050.

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Oper. Res.* 58(4-part-1):902–917.

Goodfellow I, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. *Proc. Internat. Conf. on Learn. Representations*.

Gotoh JY, Kim MJ, Lim AE (2018) Robust empirical optimization is almost the same as mean–variance optimization. *Oper. Res. Lett.* 46(4):448–452.

Gotoh JY, Kim MJ, Lim A (2020) Worst-case sensitivity. Preprint, submitted October 21, https://arxiv.org/abs/2010.10794.

Guide AH (2006) *Infinite Dimensional Analysis* (Springer, Berlin).

Jiang R, Guan Y (2016) Data-driven chance constrained stochastic program. *Math. Programming* 158(1):291–327.

Jylhä H (2015) The $L^\infty$ optimal transport: Infinite cyclical monotonicity and the existence of optimal transport maps. *Calc. Var. Partial Differential Equations.* 52(1):303–326.

Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research & Management Science in the Age of Analytics* (INFORMS), 130–166.

Lam H (2016) Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.* 41(4):1248–1275.

Ledoux M, Talagrand M (2013) *Probability in Banach Spaces: Isoperimetry and Processes* (Springer Science & Business Media, Boston).

Lee J, Raginsky M (2018) Minimax statistical learning with Wasserstein distances. *Adv. Neural Inform. Processing Systems* 31:2692–2701.

Levine A, Feizi S (2020) Wasserstein smoothing: Certified robustness against Wasserstein adversarial attacks. *Proc. Internat. Conf. on Artificial Intelligence and Statist.* 3938–3947.

Lyu C, Huang K, Liang HN (2015) A unified gradient regularization family for adversarial examples. *Proc. IEEE Internat. Conf. on Data Mining* (IEEE, New York), 301–309.

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

Najafi A, Maeda SI, Koyama M, Miyato T (2019) Robustness to adversarial perturbations in learning from incomplete data. *Adv. Neural Inform. Processing Systems* 32:5541–5551.

Popescu I (2007) Robust mean-covariance solutions for stochastic optimization. *Oper. Res.* 55(1):98–112.

Rahimian H, Mehrotra S (2019) Distributionally robust optimization: A review. Preprint, submitted August 13, https://arxiv.org/abs/1908.05659.

Scarf H (1958) A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production* (Stanford University Press, Redwood City, CA), 201–209.

Shafieezadeh-Abadeh S, Mohajerin Esfahani PM, Kuhn D (2015) Distributionally robust logistic regression. *Adv. Neural Inform. Processing Systems* 28:1576–1584.

Shafieezadeh-Abadeh S, Kuhn D, Esfahani PM (2019) Regularization via mass transportation. *J. Machine Learn. Res.* 20(103):1–68.

Shaham U, Yamada Y, Negahban S (2018) Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing* 307:195–204.

Shalev-Shwartz S, Ben-David S (2014) *Understanding Machine Learning: From Theory to Algorithms* (Cambridge University Press, Cambridge, UK).

Shapiro A, Kleywegt A (2002) Minimax analysis of stochastic problems. *Optim. Methods Software* 17(3):523–542.

Sinha A, Namkoong H, Duchi J (2018) Certifying some distributional robustness with principled adversarial training. *Proc. Internat. Conf. on Learn. Representations*.

Smirnova E, Dohmatob E, Mary J (2019) Distributionally robust reinforcement learning. Preprint, submitted February 23, https://arxiv.org/abs/1902.08708.

Staib M, Jegelka S (2017) Distributionally robust deep learning as a generalization of adversarial training. *Proc. NIPS Workshop on Machine Learning and Computer Security*.

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. Preprint, submitted December 21, https://arxiv.org/abs/1312.6199.

Volpi R, Namkoong H, Sener O, Duchi JC, Murino V, Savarese S (2018) Generalizing to unseen domains via adversarial data augmentation. *Adv. Neural Inform. Processing Systems* 31:5334–5344.

Wainwright MJ (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48 (Cambridge University Press, Cambridge, UK).

Wang Z, Glynn PW, Ye Y (2016) Likelihood robust optimization for data-driven problems. *Comput. Management Sci.* 13(2):241–261.

Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Oper. Res.* 62(6):1358–1376.

Wozabal D (2014) Robustifying convex risk measures for linear portfolios: A nonparametric approach. *Oper. Res.* 62(6):1302–1315.

Xu H, Caramanis C, Mannor S (2008) Robust regression and lasso. *Adv. Neural Inform. Processing Systems* 21:1801–1808.

Xu H, Caramanis C, Mannor S (2009) Robustness and regularization of support vector machines. *J. Machine Learn. Res.* 10(Jul):1485–1510.

Yin D, Kannan R, Bartlett P (2019) Rademacher complexity for adversarially robust generalization. *Proc. Internat. Conf. on Machine Learn.* (PMLR), 7085–7094.

Žáčková J (1966) On minimax solutions of stochastic linear programming problems. *Časopis Pěstování Matematiky* 91(4):423–430.

Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.* 46(2):262–267.

**Rui Gao** is an assistant professor in the McCombs School of Business at the University of Texas at Austin. His main research studies data-driven decision making under uncertainty and prescriptive data analytics.

**Xi Chen** is an associate professor in the Department of Technology, Operations, and Statistics at the Stern School of Business, New York University. He is also an affiliated professor in the Department of Computer Science and Center for Data Science at New York University. His research interests include machine learning, high-dimensional statistics, large-scale stochastic optimization, and data-driven operations management.

**Anton J. Kleywegt** is an associate professor in the Stewart School of Industrial and Systems Engineering at Georgia Tech. He conducts research in optimization and stochastic modeling with applications in transportation, distribution, and logistics.