B9145: Topics in Trustworthy AI **Spring 2025**

Lecture 2: Distributionally Robust Optimization

Lecturer: Hongseok Namkoong *Scribe: Samuel Deng and Rachitesh Kumar*

2.1 Distributionally Robust Optimization (DRO)

Let $\Theta \subseteq \mathbb{R}^d$ be the model class or decision space, $\mathcal{Z}$ be the data domain, and $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ be a loss function representing statistical prediction error. Let P be the data generating distribution over $\mathcal{Z}$, so $Z \sim P$ is a draw of random data. Our typical learning problem minimizes *average risk* (or, henceforth, just *risk*):

$$\min_{\theta \in \Theta} R(\theta) := \min_{\theta \in \Theta} \mathbb{E}_P[\ell(\theta; Z)]. \tag{2.1}$$

If the data-generating distribution P is sufficiently representative of the population of interest, (2.1) is effective. However, this requirement is frequently violated, as in the following real-world examples:

- Data is often collected from a particular set of geospatial locations, and may not be representative of the entire population of interest. For instance, Figure 2.1 plots the demographic compositions of low-income adults in Oregon and Texas.
- Even small shifts in the environment (as in image classification in ImageNet) degrade the performance of state-of-the-art models, up to 11-14% for average-case risk.
- Machine learning systems deterioriate on subpopulations and underrepresented user groups in datasets (possibly reflecting societal biases) in applications as varied as: speech reconigtion, facial reconition, video captioning, language identification, and recommender systems.

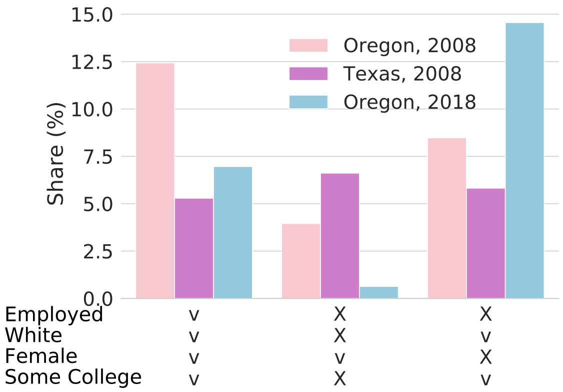

Figure 2.1: Demographics of low-income adults.

Instead of taking the average-case approach of (2.1), we consider a worst-case approach. Given a set $\mathcal{Q}$ of probability distributions, we minimize the worst-case expected loss over distributions $Q \in \mathcal{Q}$:

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[\ell(\theta; Z)]. \tag{2.2}$$

This is *distributionally robust optimization (DRO).* For a fixed model $\theta \in \Theta$, we interpret this as an adversary playing the worst distribution Q for that model. Of course, we should ask the question: what does the set $\mathcal{Q}$ of distributions include?

In this lecture, we will explore distributional robustness in a neighborhood around the data-generating distribution P. This is a natural goal for prediction problems where we are interested in learning models θ that perform uniformly well across small perturbations to the data-generating distribution. We define what we mean by a "neighborhood" in the sequel.

Before we move on, we also consider briefly what it means to optimize the *sample risk*, the empirical estimate of (2.1):

$$\min_{\theta \in \Theta} \widehat{R}(\theta) := \sum_{i=1}^{n} \frac{1}{n} \ell(\theta; Z_i), \tag{2.3}$$

where $Z_1, \ldots, Z_n$ are i.i.d. data drawn from P. We may view the $\frac{1}{n}$ as a uniform weighting over the losses of each example, $\ell(\theta; Z_i)$. Instead, in the DRO framework, we might generalize this to assign different weights to each $\ell(\theta; Z_i)$, as follows:

$$\min_{\theta \in \Theta} \sup_{p \in \mathcal{P}_n} \sum_{i=1}^{n} p_i \ell(\theta; Z_i), \tag{2.4}$$

where $\mathcal{P}_n$ is an appropriately chosen set of n-vectors. We will explore (2.4) in the sequel.

2.2 f-divergence DRO

We begin by defining the notion of an f-divergence, a notion of closeness between two distributions using a convex function f.

Definition 1 (f-divergence). *Let $f : \mathbb{R} \to \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ be a convex function satisfying $f(1) = 0$ and $f(t) = +\infty$ for any $t < 0$. Then, the f-divergence between distributions Q and P is:*

$$D_f\left(Q \| P\right) := \int f\left(\frac{dQ}{dP}\right) dP.$$

This is a general notion that, for different choices of f, give us familiar notions of distance between distributions. For example:

- $f(t) = t \log t$ gives KL-divergence.
- $f(t) = |t - 1|$ gives total variation distance.
- $f(t) = (t - 1)^2$ gives χ^2 divergence.

One should think of f-divergences in terms of the simple picture in Figure 2.2. When distributions are equal, the f divergence always evaluates to 0. When the ratio between dQ and dP increases, i.e. when Q is sufficiently different from P, the f-divergence blows up.

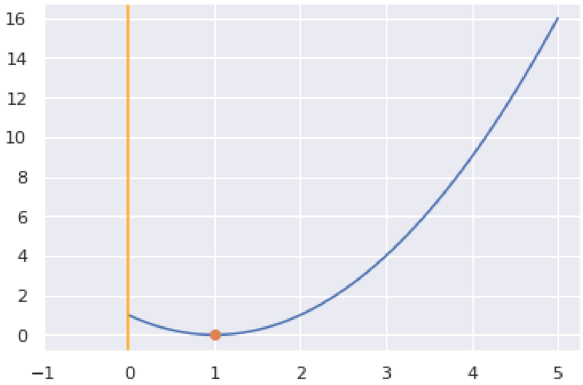

Figure 2.2. $f(t) = (t-1)^2$, the convex function that χ^2 divergence is based on. To the left of $t = 0$ (indicated by the yellow line), the function takes value $f(t) = \infty$.

Using the tool of f-divergence, we may now properly define the set $\mathcal{Q}$ of "close" distributions and the "neighborhood" discussed in Section 2.1. The *distributionally robust optimization* problem minimizes, for some fixed distribution P, f-divergence, and radius ρ:

$$\min_{\theta\in\Theta} \left\{ R_f(\theta; P) := \sup_{Q\ll P} \{\mathbb{E}_Q[\ell(\theta; Z)] : D_f(Q\|P) \le \rho\} \right\}. \tag{2.5}$$

Note that, here, there are two main parameters defining the collection of distributions of interest, $\mathcal{Q}$: (1) ρ, the radius/magnitude for the unknown distribution shift and (2) f, the choice of distance between distributions. In practice, these are nontrivial to choose, for nailing down *all* distributions of interest in real-world problems is difficult (and not even clearly captured by f and ρ). A couple other remarks:

- f-divergence limits us to using distributions Q that have the same support as P. In future lecture, we will relax this with a different family of distances (Wasserstein).
- (2.5) has the effect of upweighting regions of $\mathcal{Z}$ with high losses $\ell(\theta; Z)$. That is, it optimizes the performance of θ on the tails or the "hard" examples. (2.4) above suggests this, and the duality derivation in Section 2.2.1 makes this clear as well.

2.2.1 Dual reformulation

By itself, the inner optimization in (2.5) seems intractable, so we will reformulate the problem through duality. This gives us the following important proposition.

Proposition 1 (Duality of DRO). *Let P be a probability measure on $\mathcal{Z}$ and $\rho > 0$. Let f^* be the Fenchel conjugate of f,*

$$f^*(s) := \sup_t \{st - f(t)\}.$$

Then,

$$R_f(\theta; P) = \inf_{\lambda\ge 0, \eta\in\mathbb{R}} \left\{ \lambda \mathbb{E}_P \left[f^* \left(\frac{\ell(\theta; Z) - \eta}{\lambda} \right) \right] + \lambda\rho + \eta \right\} \tag{2.6}$$

for all θ. Moreover, if the supremum on the left hand side is finite, there are finite $\lambda(\theta) \ge 0$ and $\eta(\theta) \in \mathbb{R}$ attaining the infimum on the right hand side.

Proof. Fix some $\theta \in \Theta$ and distribution P. We redefine (2.5) in terms of the likelihood ratio, $L(Z) := \frac{dQ(Z)}{dP(Z)}$. This gives us:

$$R_f(\theta; P) = \sup_{L \geq 0} \left\{ \mathbb{E}_P[L(Z)\ell(\theta; Z)] : \mathbb{E}_P[f(L(Z))] \leq \rho, \mathbb{E}_P[L(Z)] = 1 \right\}. \tag{2.7}$$

From Lagrangian duality, we will assign $\lambda \geq 0$ to the constraint $\mathbb{E}_P[f(L(Z))] \leq \rho$ and $\eta \in \mathbb{R}$ to $\mathbb{E}_P[L(Z)] = 1$, so:

$$= \sup_{L \geq 0} \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \{\mathbb{E}_P[L(Z)\ell(\theta; Z)] + \lambda(\rho - \mathbb{E}_P[f(L(Z))] - \eta(\mathbb{E}_P[L(Z)] - 1)\}$$

Taking $L \equiv 1$ gives us $\mathbb{E}_P[f(L)] = 0$ and $\mathbb{E}_P[L] = 1$ so the extended Slater condition holds and we can switch the order of the inf and sup . After doing this and rearranging, we obtain:

$$\begin{aligned} &= \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \sup_{L \geq 0} \{\mathbb{E}_P[L(Z)\ell(\theta; Z) - \lambda f(L(Z)) - \eta L(Z)]\} + \lambda\rho + \eta \\ &= \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \sup_{L \geq 0} \left\{ \lambda \mathbb{E}_P \left[\frac{L(Z)(\ell(\theta; Z) - \eta)}{\lambda} - f(L(Z)) \right] \right\} + \lambda\rho + \eta. \end{aligned} \tag{2.8}$$

Notice that $L : \mathcal{Z} \to \mathbb{R}_+$ is an arbitrary nonnegative measurable function that we are taking a supremum over. This allows us to interchange the inner supremum and the integral, giving us the Fenchel conjugate:

$$\begin{aligned} \sup_{L \geq 0} \mathbb{E}_P \left[\frac{L(Z)(\ell(\theta; Z) - \eta)}{\lambda} - f(L(Z)) \right] &= \sup_{L \geq 0} \int_{\mathcal{Z}} \frac{L(Z)(\ell(\theta; Z) - \eta)}{\lambda} - f(L(Z)) dP \\ &= \int_{\mathcal{Z}} \sup_{L \geq 0} \left\{ L \left(\frac{\ell(\theta; Z) - \eta}{\lambda} \right) - f(L) \right\} dP \\ &= \int_{\mathcal{Z}} f^* \left(\frac{\ell(\theta; Z) - \eta}{\lambda} \right) dP \\ &= \mathbb{E}_P \left[f^* \left(\frac{\ell(\theta; Z) - \eta}{\lambda} \right) \right]. \end{aligned}$$

Plugging this back into (2.8) gives us our desired result. □

As a concrete example of Proposition 1, consider divergences that look like t^k. Specifically, for $k \in (1, \infty)$, we consider the Cressie-Read family of divergences, defined as:

$$f_k(t) := \begin{cases} t^k - 1 & \text{if } t \geq 0 \\ \infty & \text{otherwise} \end{cases} \tag{2.9}$$

Denoting $k_* = \frac{k}{k+1}$ and $(s)_+ = \max(s, 0)$, we obtain the following Fenchel conjugate,

$$f_k^*(s) = k^{-k_*}(k-1)(s)_+^{k_*} + 1. \tag{2.10}$$

Applying Proposition 1, we obtain:

$$R_f(\theta; P) = \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ \lambda^{1-k_*} k^{-k_*}(k-1) \mathbb{E}_P \left[(\ell(\theta; Z) - \eta)_+^{k_*} \right] + \lambda(\rho + 1) + \eta \right\}$$

Now, optimizing over $\lambda \geq 0$ and denoting $c_k(\rho) := (1 + \rho)^{\frac{1}{k}}$, we obtain the final dual form:

$$R_k(\theta; P) = \inf_{\eta \in \mathbb{R}} \left\{ c_k(\rho) \mathbb{E}_P \left[(\ell(\theta; Z) - \eta)_+^{k_*} \right]^{\frac{1}{k^*}} + \eta \right\} \tag{2.11}$$

Thus, in this particular case of Cressie-Read f-divergences, the simplified form above shows that distributional robustness is equivalent to optimizing the tail performance of the model. η is the threshold at which we care about the loss of an example; for examples with loss less than η, the expectation in (2.11) vanishes. Thus, the harder examples (with loss greater than η) are emphasized by a power of k_*.

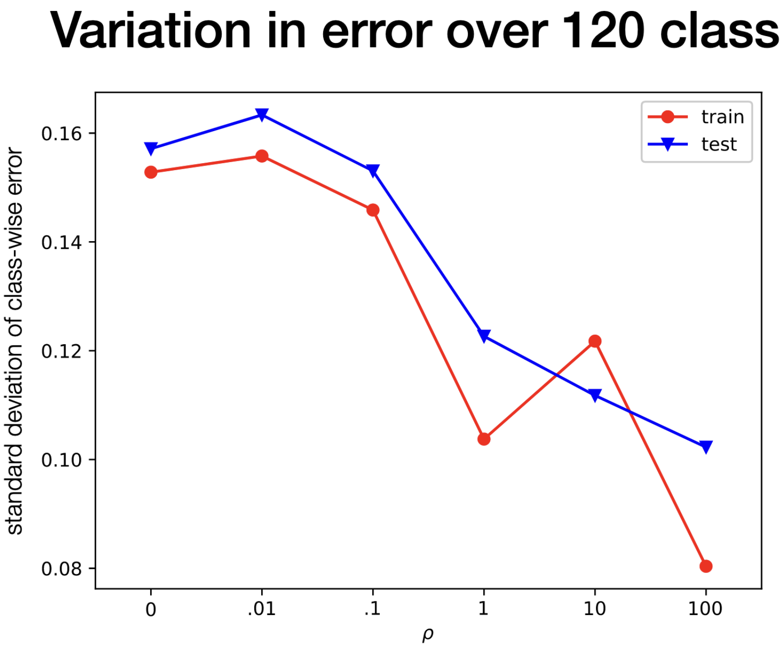

Figure 2.3. Variation in error over all the classes. As ρ increases (DRO hedges against further distributions), we see that the variation in class-wise error becomes smaller; the error rates become more uniform through the classes.

2.2.2 Optimization

The following section of notes are at a higher level – in this portion of the lecture, we discussed how DRO works in practice and its pros and cons. First, to actually operationalize (2.5) in practice, we of course do not have access to population distributions, so we must use an empirical plug-in. Suppose we have a finite sample of data $\{Z_i\}_{i=1}^n$. The empirical plug-in formulation of the primal is:

$$\sup_q \left\{ \sum_{i=1}^n q_i \ell(\theta; Z_i) : D_f \left(q \| \mathbf{1}/n\right) \leq \rho, q^\top \mathbf{1} = 1, q_i \geq 0 \right\}. \tag{2.12}$$

To optimize this, we can play a 2-player stochastic game (adversary plays q and the player plays $\theta \in \Theta$), or just do batch gradient descent on the whole thing. Alternatively, we can optimize the empirical formulation of the dual in (2.6) using a standard solver.

However, how does DRO perform in practice? In lecture, we looked at an experiment for fine-grained recognition. The taks was to classify images of dogs to dog breeds (with 120 breeds/classes total). In the dataset, there was no underrepresentation: each breed/class had the same number of images. The experiment used DRO with χ^2 divergence, parametrized with $\rho > 0$ as the magnitude of distance between distributions.

In Figure 2.3, we see that performing DRO with a significant ρ parameter decreases the variation in error across the classes. In Figure 2.4, we actually see that DRO in this specific experiment actually consistently does better than the standard average-case risk minimization, *even* when we are accounting for every single class. This suggests that, at least in some cases, DRO is useful even for the original goal of average-case risk minimization (2.1). This motivates Section 2.2.3

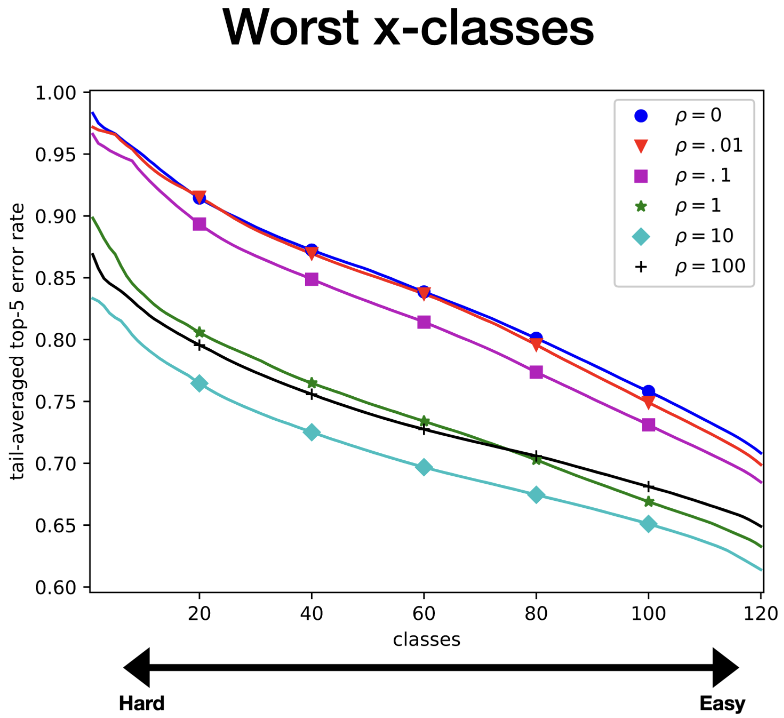

Figure 2.4. Top-5 error rate across x classes. On the far left is top-5 error rate on the hardest 20 classes; on the far right is the top-5 error rate over all classes. Each curve represents DRO at level $\rho \geq 0$; $\rho = 0$ is standard average-case risk minimization.

2.2.3 Connections to variance regularization

In Section 2.2.2, we observed that, at least in this specific experiment, applying DRO actually helped with the original goal of standard average-case risk minimization. Recall the classical notion of (average-case) risk for a data-generating distribution P, which we rewrite here for convenience:

$$R(\theta) = \mathbb{E}_P[\ell(\theta; Z)] \tag{2.13}$$

for $Z \sim P$.

This raises the question: does DRO provide any guarantees for our *original* (classical) goal of minimizing average-case risk (2.13)? In this section, forget all about the goal of robust optimization to distribution shifts; we return to our classical goal of minimizing standard average-risk. We will show a theoretical result that DRO helps in this classical goal.

Fix some data-generating distribution P. The standard approach for minimizing risk in statistical learning is ERM. That is, we find the model $\hat{\theta}^{\text{erm}} \in \Theta$ that minimizes the average-case sample risk:

$$\hat{\theta}^{\text{erm}} \in \operatorname*{argmin}_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; Z_i),$$

for a dataset $\{Z_i\}_{i=1}^n$ all drawn i.i.d. from P. The hope is that $\widehat{R}_n(\theta)$ is a good approximation of $R(\theta)$.

From the empirical Bernstein's inequality, with probability $1 - \delta$,

$$R(\theta) = \mathbb{E}_P[\ell(\theta; Z)] \leq \underbrace{\widehat{R}_n(\theta)}_{\text{bias}} + \underbrace{\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(\ell(\theta; Z))}{n}}}_{\text{variance}} + \frac{C \log \frac{1}{\delta}}{n}. \tag{2.14}$$

Above, $\mathrm{Var}_{\widehat{P}_n}$ is empirical variance. Any estimator has a bias and variance, and (2.14) makes this explicit. We might hope to use this upper bound directly (optimally trading off between bias and variance) by solving this *variance-regularized sample risk.*

$$\hat{\theta}^{\text{var}} \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ \widehat{R}_n(\theta) + \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(\ell(\theta; Z))}{n}} \right\}. \tag{2.15}$$

The problem is that we cannot solve 2.15 through standard optimization methods because it is non-convex. However, skipping to the punchline, it will happen that using DRO with a specific setup will allow us to minimize 2.15.

We will setup the DRO problem now. Fix some P, and draw a dataset of n examples $\{Z_i\}_{i=1}^n$ i.i.d. from P. Let $\widehat{P}_n$ be the empirical distribution from the n examples. Consider the χ^2 divergence, where we use the function $f(t) = \frac{1}{2}(t-1)^2$ as the f-divergence, following Definition 1. We denote the f-divergence as $D_{\chi^2}(Q \| P)$. For some $\rho > 0$, define the class of distributions $\mathcal{P}_{n,\rho}$ that we want to be robust against as:

$$\mathcal{P}_{n,\rho} := \left\{ \text{Distribution } P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \leq \frac{\rho}{n} \right\} \tag{2.16}$$

Notice that, in this collection of "close" distributions, as $n \to \infty$, we get closer and closer to our original distribution P. We are sticking very close to our original distribution, and, for $n \to \infty$ at the limit, this is just standard optimizing for the original distribution P (which we can just do via ERM). Maximizing over this collection gives us the *empirical likelihood upper confidence bound*:

$$R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_P[\ell(\theta; Z)] = \max_{p : D_{\chi^2}\left(P \| \widehat{P}_n\right) \leq \frac{\rho}{n}} \sum_{i=1}^{n} p_i \ell(\theta; Z_i). \tag{2.17}$$

Solving for the best model that optimizes (2.17) is just a DRO problem. We'll denote this best DRO model as:

$$\hat{\theta}^{\text{rob}} \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) = \max_{p : D_{\chi^2}(P \| \widehat{P}_n) \le \frac{\rho}{n}} \sum_{i=1}^{n} p_i \ell(\theta; Z_i) \right\}. \tag{2.18}$$

The difference between $\hat{\theta}^{\text{var}}$ from (2.15) and $\hat{\theta}^{\text{rob}}$ from (2.18) is that $\hat{\theta}^{\text{rob}}$ can actually be efficiently solved for using DRO methods! The problem is convex, and we can use any of the methods in Section 2.2.2 to solve it. It turns out that (2.17) actually converges to the term being optimized in (2.15), which connects DRO to the direct variance-regularized risk.

Theorem 2 (Equivalence of DRO and Variance-Regularized Risk)**.** *For general f-divergences and bounded loss $\ell(\theta; Z) \le M$,*

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \operatorname{Var}_{\widehat{P}_n}(\ell(\theta; Z))}{n}} + \operatorname{Rem}_n(\theta). \tag{2.19}$$

Let $\sigma^2(\theta) := \operatorname{Var}(\ell(\theta; Z))$. Also, $\operatorname{Rem}_n(\theta) \le \frac{\sqrt{12}\rho M}{n}$ and $\operatorname{Rem}_n(\theta) = 0$ with probability at least $1 - \exp\left(-\frac{n\sigma^2(\theta)}{36M^2}\right)$.

We will prove Theorem 2 below. Using Theorem 2, we get the following guarantee on the true risk of the DRO model which optimizes $R_n(\theta, \mathcal{P}_{n,\rho})$. Recall the definition of $\hat{\theta}^{\text{rob}}$ from (2.18).

Theorem 3. *Let $\rho = \log \frac{1}{\delta} + d \log n$. Then, with probability at least $1 - \delta$,*

$$\begin{aligned} R(\hat{\theta}^{\text{rob}}) &\le R_n(\hat{\theta}^{\text{rob}}, \mathcal{P}_{n,\rho}) + \frac{crM\rho}{n} \\ &\le \min_{\theta \in \Theta} \left\{ R(\theta) + \sqrt{\frac{2\rho \operatorname{Var}_{\widehat{P}_n}(\ell(\theta; Z))}{n}} \right\} + \frac{crM\rho}{n}, \end{aligned}$$

where $R(\theta)$ is the standard average-case risk from (2.13) and $\mathcal{P}_{n,\rho}$ is from (2.16).

The thrust of Theorem 3 is that running DRO allows us to achieve the optimal bias-variance tradeoff we wanted in (2.14) with respect to the standard measure of risk, $R(\theta)$. Further, this actually beats ERM! Denote $R(\theta^*) := \inf_{\theta \in \Theta} R(\theta)$, the true minimizer of standard risk. ERM gives us the following guarantee:

$$R(\hat{\theta}^{\text{erm}}) \le R(\theta^*) + \sqrt{\frac{2\rho M R(\theta^*)}{n}} + \frac{CM\rho}{n}. \tag{2.20}$$

If $\operatorname{Var}(\ell(\theta; X)) \ll MR(\theta^*)$, then the bound in Theorem 3 is actually *tighter* than that of (2.20). This shows DRO beating ERM, which provides a possible theoretical explanation for Figure 2.4 in Section 2.2.2.

Finally, we provide the proof (sketch) of Theorem 2.

Proof. Denote $z_i := \ell(\theta, Z_i)$ and denote $u_i = p_i - \frac{1}{n}$ and denote $\overline{z}$ and s_n^2 the sample mean and sample variance, respectively. With this notation, the empirical likelihood upper confidence bound becomes:

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \max_p \left\{ \langle p, z \rangle : D_{\chi^2}(p \| \mathbf{1}/n) \le \frac{\rho}{n} \right\}$$

Using the definition of χ^2 divergence,

$$= \max_p \left\{ \langle p, z \rangle : \frac{1}{n} \sum_{i=1}^{n} (np_i - 1)^2 \le \frac{\rho}{n}, p^\top \mathbf{1} = 1, p \ge 0 \right\}.$$

Using the change of variable from $u_i := p_i - \frac{1}{n}$, we get:

$$\begin{aligned}
&= \overline{z} + \max_u \left\{ \langle u, z - \overline{z} \rangle : \|u\|^2 \leq \frac{\rho}{n^2}, u^\top \mathbf{1} = 1, u \geq -\frac{\mathbf{1}}{n} \right\} \\
&\leq \overline{z} + \frac{\sqrt{2\rho}}{n} \|z - \overline{z}\|_2 = \overline{z} + \sqrt{\frac{2\rho}{n} s_n^2} \qquad \text{(by Cauchy-Schwarz).}
\end{aligned}$$

The final inequality is tight if, for all i,

$$u_i = \frac{1}{n} \sqrt{\frac{2\rho}{n s_n^2}} (z_i - \overline{z}) \geq -\frac{1}{n}.$$

□

2.3 Wasserstein Distributionaly Robust Optimization

The f-divergence takes value ∞ whenever a perturbed distribution Q has support outside of that of P. This may be limiting when there is a natural geometry in the data space. In this case, instead of reweighting data, we may consider directly perturbing data values according to this geometry. For example, this is appropriate for adversarial attacks that perturb pixels of images by an amount imperceptible to humans.

Wasserstein distances uses the geometry of the underlying space to define a notion of closeness between distributions. Let $\mathcal{Z} \subset \mathbb{R}^m$, and let $(\mathcal{Z}, \mathcal{A}, P)$ be a probability space. Let the transportation cost $c : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ be nonnegative, lower semi-continuous, and satisfy $c(z, z) = 0$. For probability measures P and Q supported on $\mathcal{Z}$, let $\Pi(P, Q)$ denote their couplings, meaning measures π on $\mathcal{Z}^2$ with $\pi(A, \mathcal{Z}) = P(A)$ and $\pi(\mathcal{Z}, A) = Q(A)$ for all $A \subset \mathcal{Z}$. The Wasserstein distance between P and Q is

$$W_c(Q, P) := \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi[c(Z, Z')].$$

This infimization problem is known as the optimal transport problem, where we wish to transport mass away from P to Q, where $c(z, z')$ represents the unit cost of transporting mass from z to z'.

2.3.1 Dual reformulation

We can perform distributionally robust optimization (DRO) w.r.t. the Wasserstein distance. For $\rho \geq 0$ and distribution P_0, we let $\mathcal{Q} = \{Q : W_c(Q, P) \leq \rho\}$, the Wasserstein DRO problem is given by

$$\underset{\theta \in \Theta}{\text{minimize}} \left\{ \mathcal{R}_c(\theta; P) := \sup_Q \left\{ \mathbb{E}_Q[\ell(\theta; Z)] : W_c(Q, P) \leq \rho \right\} \right\}. \tag{2.21}$$

In particular, the Wasserstein ball allows for distributions Q that have a different support to P, so long as the cost of transporting mass from P to Q is not too high.

The following proposition gives a duality result for Wasserstein DRO (2.21). We assume $\mathbb{E}_P[\ell(\theta; Z)] < \infty$ throughout.

Proposition 4. *Fix any $\theta \in \Theta$. Let $z \mapsto \ell(\theta; z)$ be upper semi-continuous. Let $\phi_\lambda(\theta; z_0) = \sup_{z \in \mathcal{Z}} \{\ell(\theta; z) - \lambda c(z, z_0)\}$ be the robust surrogate. For any distribution Q and any $\rho > 0$,*

$$\sup_{Q : W_c(Q,P) \leq \rho} \mathbb{E}_Q[\ell(\theta; Z)] = \inf_{\lambda \geq 0} \left\{ \lambda\rho + \mathbb{E}_P[\phi_\lambda(\theta; Z)] \right\}. \tag{2.22}$$

The dual form makes crisp how the optimal transport problem plays a role in defining worst-case perturbations. The supremum inside the expectation considers a perturbation z to the data Z, such that it makes the loss $\ell(\theta; z)$ bigger, while being penalized by the cost of moving mass from Z to z. Comparing this to the f-divergence dual that upweighted examples with higher loss, we see that Wasserstein DRO (2.21) considers the geometry of the inputs by using the cost function c.

The computational cost of considering probabilities whose support may differ from P is steep. The dual formulation (2.22) has reformulated an infinite-dimensional problem over probabilities to computing the robust surrogate ϕ_λ, but even evaluating the robust surrogate is computationally intractable in general. The maximization problem $\phi_\lambda(\theta; Z) = \sup_z \ell(\theta; z) - \lambda c(Z, z)$ is almost always non-concave, even for simple linear models. Furthermore, a naive analysis of the statistical estimation of Wasserstein DRO yields nonparametric rates. Identifying structured scenarios with alleviated computational and statistical difficulties is an area of active research.

Before proceeding with the proof of Proposition 4, we consider an example.

Example 1. Consider the cost function $c(z, z') = \frac{1}{2}\|z - z'\|_2^2$, and the corresponding robust surrogate function

$$\phi_\lambda(\theta; Z) = \sup_{z' \in \mathcal{Z}} \left\{ \ell(\theta; z') - \frac{\lambda}{2}\|z' - z\|_2^2 \right\} .$$

Plugging the first order approximation $\ell(\theta; z') \approx \ell(\theta; z) + \nabla_z \ell(\theta; z)^\top (z' - z)$ into the robust surrogate yields

$$\phi_\lambda(\theta; z) \approx \sup_{z' \in \mathcal{Z}} \left\{ \ell(\theta; z) + \nabla_z \ell(\theta; z)^\top (z' - z) - \frac{\lambda}{2}\|z' - z\|_2^2 \right\} .$$

First order condition of optimality implies that the supremum is attained at $z' \in \mathcal{Z}$ such that

$$\nabla_z \ell(\theta; z) = \lambda \cdot (z' - z) \qquad \equiv \qquad z' = z + \frac{1}{\lambda} \cdot \nabla_z \ell(\theta; z) .$$

Note that the worst-case perturbation z' is simply a gradient-ascent step from z along $\nabla_z \ell(\theta; z)$. Therefore, we can approximate the robust surrogate as

$$\phi_\lambda(\theta; z) \approx \ell(\theta; z) + \frac{1}{2\lambda} \cdot \|\nabla_z \ell(\theta; z)\|_2^2 .$$

Plugging this into the dual for the empirical distribution $\widehat{P}_n$ (which is simply the uniform distribution on samples $\{Z_1, \ldots, Z_n\}$) yields

$$\begin{aligned}
\inf_{\lambda \geq 0} \left\{ \lambda\rho + \frac{1}{n} \cdot \sum_{i=1}^n \phi_\lambda(\theta; Z_i) \right\} &= \inf_{\lambda \geq 0} \left\{ \lambda\rho + \frac{1}{n} \cdot \sum_{i=1}^n \left\{ \ell(\theta; Z_i) + \frac{1}{2\lambda} \cdot \|\nabla_z \ell(\theta; Z_i)\|_2^2 \right\} \right\} \\
&= \frac{1}{n} \cdot \sum_{i=1}^n \ell(\theta; Z_i) + \inf_{\lambda \geq 0} \left\{ \lambda\rho + \frac{1}{2\lambda} \cdot \frac{1}{n} \cdot \sum_{i=1}^n \|\nabla_z \ell(\theta; Z_i)\|_2^2 \right\} \\
&= \frac{1}{n} \cdot \sum_{i=1}^n \ell(\theta; Z_i) + \inf_{\lambda \geq 0} \left\{ \lambda\rho + \frac{1}{2\lambda} \cdot \mathbb{E}_{\widehat{P}_n} \left[\|\nabla_z \ell(\theta; Z)\|_2^2 \right] \right\} \\
&= \frac{1}{n} \cdot \sum_{i=1}^n \ell(\theta; Z_i) + \sqrt{2\rho} \cdot \left(\mathbb{E}_{\widehat{P}_n} \left[\|\nabla_z \ell(\theta; Z)\|_2^2 \right] \right)^{1/2}
\end{aligned}$$

where the third equality follows from the fact that the infimum is attained at $\lambda = \frac{1}{\rho}$. Therefore, under first-order approximations, Wasserstein DRO amounts to a regularization that makes $\|\nabla_z \ell(\theta; Z)\|$ small and guards against data perturbations. ◇

Proof of Proposition 4. Although the proof of Proposition 4 is involved, we can gain basic intuition by considering a substantially simplified scenario. Consider a discrete sample space

$$\mathcal{Z} := \{z_1, \ldots, z_k\}.$$

The definition of the Wasserstein distance can then be simplified to

$$\min_{\pi(z_i,z_j)\geq 0} \left\{ \sum_{i,j} \pi(z_i,z_j) c(z_i,z_j) : \sum_i \pi(z_i,z_j) = q(z_j), \;\; \sum_j \pi(z_i,z_j) = p(z_i), \;\; \sum_{i,j} \pi(z_i,z_j) = 1 \right\}.$$

Then, $\mathcal{R}_c(\theta; P)$, the Wasserstein distributionally robust objective (2.21) can be written as

$$\max_{\pi(z_i,z_j)\geq 0} \left\{ \sum_{i,j} \pi(z_i,z_j) \ell(\theta; z_j) : \sum_j \pi(z_i,z_j) = p(z_i), \;\; \sum_{i,j} \pi(z_i,z_j) = 1, \;\; \sum_{i,j} \pi(z_i,z_j) c(z_i,z_j) \leq \rho \right\}.$$

Now, use Lagrangian duality to note that

$$\mathcal{R}_c(\theta; P) = \min_{\lambda \geq 0} \max_{\pi \geq 0} \left\{ \lambda\rho + \sum_{i,j} \pi(z_i,z_j)(\ell(\theta; z_j) - \lambda c(z_i,z_j)) : \sum_j \pi(z_i,z_j) = p(z_i), \;\; \sum_{i,j} \pi(z_i,z_j) = 1 \right\}.$$

The inner maximum problem is evidently attained at

$$\pi(z_i,z_j) = \begin{cases} p(z_i) & \text{if } j \text{ is the smallest index in } \operatorname{argmax}_j \{\ell(\theta; z_j) - \lambda c(z_i,z_j)\} \\ 0 & \text{otherwise} \end{cases}.$$

We conclude that

$$\mathcal{R}_c(\theta; P) = \min_{\lambda \geq 0} \left\{ \lambda\rho + \sum_i p(z_i) \max_j \{\ell(\theta; z_j) - \lambda c(z_i,z_j)\} \right\},$$

which is the desired result (2.22) for discrete sample spaces. □

2.3.2 Connection to regularization

By choosing the regularizer, we can show that Wasserstein DRO is equivalent to classical regularizers.

Proposition 5 (Regression)**.** *Consider the cost function* $c((x,y),(x',y')) = \|(x,y) - (x',y')\|_k^2$ *for some* $k \in [0,\infty)$*. Then,*

$$\sup_{Q: W_c(Q,\widehat{P}_n) \leq \rho} \mathbb{E}_Q\left[(Y - \theta^\top X)^2\right] = \left(\left(\frac{1}{n} \cdot \sum_{i=1}^n (Y_i - \theta^\top X_i)^2 \right)^{1/2} + \sqrt{\rho} \cdot \|[\theta, -1]\|_{k_*} \right)^2,$$

where $k^* = k/(k-1)$ *and satisfies* $\frac{1}{k} + \frac{1}{k^*} = 1$.

Proof. To simplify notation, set $Z = (X, Y)$ and $\bar{\theta} = [\theta, -1] \in \mathcal{R}^{d+1}$. From the duality result for Wasserstein DRO (Proposition 4), we get

$$\sup_{Q: W_c(Q,\widehat{P}_n) \leq \rho} \mathbb{E}_Q\left[(Y - \theta^\top X)^2\right] = \inf_{\lambda \geq 0} \left\{ \lambda\rho + \mathbb{E}_{\widehat{P}_n} \sup_{z'} \left\{ (\bar{\theta}^\top z')^2 - \lambda \|Z - z'\|_k^2 \right\} \right\}.$$

First, we simplify the surrogate loss $\phi_\lambda(\bar{\theta}, Z) = \sup_{z'} \left\{(\bar{\theta}^\top z')^2 - \lambda\|Z - z'\|_k^2\right\}$. Doing a change of variable $\Delta = Z - z'$ yields

$$
\begin{aligned}
\phi_\lambda(\bar{\theta}, Z) &= \sup_{\Delta} \left\{ \left(\bar{\theta}^\top Z - \bar{\theta}^\top \Delta\right)^2 - \lambda\|\Delta\|_k^2 \right\} \\
&= \sup_{\Delta} \left\{ \left(\bar{\theta}^\top Z + \operatorname{sign}(\bar{\theta}^\top Z) \cdot |\bar{\theta}^\top \Delta|\right)^2 - \lambda\|\Delta\|_k^2 \right\} && \text{(sup attained when signs match)} \\
&= \sup_{\Delta} \left\{ \left(|\bar{\theta}^\top Z| + |\bar{\theta}^\top \Delta|\right)^2 - \lambda\|\Delta\|_k^2 \right\} \\
&= \sup_{\Delta} \left\{ \left(|\bar{\theta}^\top Z| + \|\bar{\theta}\|_{k_*}\|\Delta\|_k\right)^2 - \lambda\|\Delta\|_k^2 \right\} && \text{(sup is attained when Holder's ineq. is tight)} \\
&= \left(\bar{\theta}^\top Z\right)^2 + \sup_{\Delta} \left\{ -(\lambda - \|\bar{\theta}\|_{k_*}^2) \cdot \|\Delta\|_k^2 + 2 \cdot |\bar{\theta}^\top Z| \cdot \|\bar{\theta}\|_{k_*}\|\Delta\|_k \right\} \\
&= \begin{cases} \frac{\lambda}{\lambda - \|\bar{\theta}\|_{k_*}^2} \cdot (\bar{\theta}^\top Z)^2 & \text{if } \lambda > \|\bar{\theta}\|_{k_*}^2 \\ \infty & \text{otherwise} \end{cases}
\end{aligned}
$$

This allows us to conclude

$$
\sup_{Q : W_c(Q, \widehat{P}_n) \le \rho} \mathbb{E}_Q\left[(Y - \theta^\top X)^2\right] = \inf_{\lambda > \|\bar{\theta}\|_{k_*}^2} \left\{ \lambda\rho + \frac{\lambda}{\lambda - \|\bar{\theta}\|_{k_*}^2} \cdot \frac{1}{n}\sum_{i=1}^n (\bar{\theta}^\top Z_i)^2 \right\}.
$$

First order condition of optimality implies that the infimum is attained at $\lambda = \|\bar{\theta}\|_{k_*}^2 + \left(\frac{\|\bar{\theta}\|_{k_*}^2}{\rho} \cdot \frac{1}{n}\sum_{i=1}^n (\bar{\theta}^\top Z_i)^2\right)^{1/2}$, which yields

$$
\sup_{Q : W_c(Q, \widehat{P}_n) \le \rho} \mathbb{E}_Q\left[(Y - \theta^\top X)^2\right] = \left(\left(\frac{1}{n} \cdot \sum_{i=1}^n (\bar{\theta}^\top Z_i)^2 \right)^{1/2} + \sqrt{\rho} \cdot \|\bar{\theta}\|_{k_*} \right)^2,
$$

as required. □

A similar equivalence can be shown for

- **Regression under Covariate Shift:** If the cost function c is

$$
c((x, y), (x', y')) = \begin{cases} \|x - x'\|_k^2 & \text{if } y = y' \\ \infty & \text{if } y \neq y' \end{cases}
$$

  then

$$
\sup_{Q : W_c(Q, \widehat{P}_n) \le \rho} \mathbb{E}_Q\left[(Y - \theta^\top X)^2\right] = \left(\left(\frac{1}{n} \cdot \sum_{i=1}^n (Y_i - \theta^\top X_i)^2 \right)^{1/2} + \sqrt{\rho} \cdot \|\theta\|_{k_*} \right)^2
$$

- **Logistic Loss:** If $Y \in \{-1, +1\}$ and the cost function c is

$$
c((x, y), (x', y')) = \begin{cases} \|x - x'\|_k & \text{if } y = y' \\ \infty & \text{if } y \neq y' \end{cases}
$$

  then

$$
\sup_{Q : W_c(Q, \widehat{P}_n) \le \rho} \mathbb{E}_Q\left[\log\left(1 + e^{-Y \cdot \theta^\top X}\right)\right] = \frac{1}{n} \cdot \sum_{i=1}^n \log\left(1 + e^{-Y_i \cdot \theta^\top X_i}\right) + \rho \cdot \|\theta\|_{k_*}
$$

- **Support Vector Machine:** If $Y \in \{-1, +1\}$ and the cost function c is

$$c((x,y),(x',y')) = \begin{cases} \|x - x'\|_k & \text{if } y = y' \\ \infty & \text{if } y \neq y' \end{cases}$$

then

$$\sup_{Q : W_c(Q, \widehat{P}_n) \leq \rho} \mathbb{E}_Q \left[\left(1 - Y \cdot \theta^\top X\right)_+ \right] = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(1 - Y_i \cdot \theta^\top X_i\right)_+ + \rho \cdot \|\theta\|_{k_*}$$