**B9145: Reliable Statistical Learning** **Spring 2023**

# Lecture 6: Wasserstein DRO

*Lecturer: Hongseok Namkoong* *Scribe: Rachitesh Kumar*

## 6.1 Wasserstein Distance

The $f$-divergence takes value $\infty$ 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 $\pi$ 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 $z'$.

## 6.2 Wasserstein Distributionaly Robust Optimization

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

$$\underset{\theta \in \Theta}{\text{minimize}} \left\{ \mathcal{R}_c(\theta; P) := \sup_Q \{\mathbb{E}_Q[\ell(\theta; Z)] : W_c(Q, P) \leq \rho\} \right\}. \tag{6.1}$$

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 (6.1). We assume $\mathbb{E}_P[\ell(\theta; Z)] < \infty$ throughout.

**Proposition 1.** *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 $\rho > 0$,*

$$\sup_{Q: W_c(Q,P) \leq \rho} \mathbb{E}_Q[\ell(\theta; Z)] = \inf_{\lambda \geq 0} \left\{ \lambda\rho + \mathbb{E}_P[\phi_\lambda(\theta; Z)] \right\}. \tag{6.2}$$

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 (6.1) 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 (6.2) has reformulated an infinite-dimensional problem over probabilities to computing the robust surrogate $\phi_\lambda$, 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 1, 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

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

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 1.* Although the proof of Proposition 1 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 (6.1) 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 (6.2) for discrete sample spaces. □

### 6.2.1 Connection to Regularization

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

**Proposition 2** (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 1), 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

$$\begin{aligned}
\phi_\lambda(\bar{\theta}, Z) &= \sup_{\Delta} \left\{ \left(\bar{\theta}^\top Z - \bar{\theta}^\top \Delta\right)^2 - \lambda \|\Delta\|_k^2 \right\} \\
&= \sup_{\Delta} \left\{ \left(\bar{\theta}^\top Z + \operatorname{sign}(\bar{\theta}^\top Z) \cdot |\bar{\theta}^\top \Delta|\right)^2 - \lambda \|\Delta\|_k^2 \right\} \qquad \text{(sup attained when signs match)} \\
&= \sup_{\Delta} \left\{ \left(|\bar{\theta}^\top Z| + |\bar{\theta}^\top \Delta|\right)^2 - \lambda \|\Delta\|_k^2 \right\} \\
&= \sup_{c>0} \sup_{\Delta : \|\Delta\|_k = c} \left\{ \left(|\bar{\theta}^\top Z| + |\bar{\theta}^\top \Delta|\right)^2 - \lambda c^2 \right\},
\end{aligned}$$

where in the final line we separated the optimization problem into concentric circles of radius $r$ in the $\|\cdot\|_\cdot$. Using Holder's inequality, the preceding display is bounded by

$$\sup_{c>0} \left\{ \left(|\bar{\theta}^\top Z| + \|\bar{\theta}\|_{k_*} c\right)^2 - \lambda c^2 \right\},$$

but since there is always a $\Delta$ satisfying $\|\Delta\|_k = c$ for which Holder's inequality is tight, the bound is in fact an equality. Hence, the surrogate loss can be rewritten as follows

$$\begin{aligned}
&\left(\bar{\theta}^\top Z\right)^2 + \sup_{c>0} \left\{ -(\lambda - \|\bar{\theta}\|_{k_*}^2) \cdot c^2 + 2 \cdot |\bar{\theta}^\top Z| \cdot \|\bar{\theta}\|_{k_*} c \right\} \\
&= \begin{cases} \frac{\lambda}{\lambda - \|\bar{\theta}\|_{k_*}^2} \cdot (\bar{\theta}^\top Z)^2 & \text{if } \lambda > \|\bar{\theta}\|_{k_*}^2 \\ \infty & \text{otherwise} \end{cases}
\end{aligned}$$

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

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

- **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

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