Lecture 4: Distributional Robustness

Hongseok Namkoong

October 5, 2020

As before, we consider a loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ representing monetary cost or (a surrogate) statistical prediction error. Let P be the data-generating distribution. In previous lectures, we studied standard average-case optimization problems of the form

$$\underset{\theta \in \Theta}{\text{minimize}}\, \mathbb{E}_P[\ell(\theta; Z)], \tag{1}$$

where $\Theta \subseteq \mathbb{R}^d$ is the model class (or decision space), and $Z \sim P$ is the random data. This simple notation abstracts the complexities of real-world data-collection. The data collection distribution P is an artifact made out of many different systems; it is the result of careful study (or product) design, database management, data cleaning, and feature engineering.

The average-case formulation (1) is only effective when the data-generating distribution P is representative of the overall population of interest. However, this requirement is frequently violated. Data is often collected from a particular set of geospatial locations, and may not represent the population of interest. As a basic illustration, Figure 1 plots the demographic compositions of low-income adults in Oregon and Texas. Over different points in space and time, compositions vary up to fivefold. Even state-of-the-art models with high average test accuracy severely underperform under small shifts in the environment; for example, average error of state-of-the-art models deteriorates by 11-14% on a new test set for ImageNet [5].

Modern applications involve heterogeneous subpopulations across which we want uniformly good performance. For example, in natural language processing (NLP), large-scale corpora is collected over different domains, each with different difficulty levels. Almost all modern applications of learning involve a diverse array of user groups across different demographics, but data collection systems necessarily embody the societal biases we see throughout society. Statistical models that optimize average-case performance (1) often perform poorly on minority groups underrepresented in the dataset. Performance of speech recognition systems deteriorate on people with minority accents. Similar significant performance fluctuations across demographic groupings such as race, gender, or age have been observed in facial recognition, automatic video captioning, language identification, academic recommender systems.

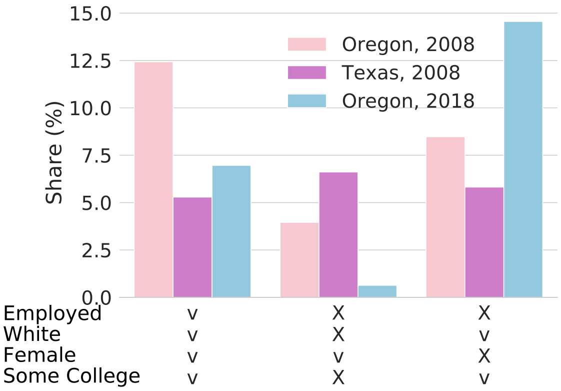

Figure 1: Demographics of low-income adults

1 Distributional Robustness

Instead of taking an *average-case* approach (1) over the random observations $Z \sim P$, we may consider a *deterministic worst-case* approach: for some uncertainty set $U \subset \mathcal{Z}$ representing "plausible" values of Z, we can optimize for the worst-case scenario over $z \in U$

$$\underset{\theta \in \Theta}{\text{minimize}} \ \ \sup_{z \in U} \ell(\theta; z) \tag{2}$$

This deterministic worst-case formulation (2) is called a robust optimization problem; see Ben-Tal et al. [1] or Bertsimas et al. [2] for an extensive overview of solution methods, and their various applications. The deterministic robust optimization formulation (2) tends to be overly conservative, and the choice of the uncertainty set is usually driven by availability of efficient solution methods.

To explicitly incorporate the statistical nature of the random vector $Z \sim P$, we can formulate a distributionally robust problem that bridges the above two dichotomous frameworks. Given a set $\mathcal{P}$ of probability distributions, we minimize the worst-case expected loss over probabilities $Q \in \mathcal{Q}$

$$\underset{\theta \in \Theta}{\text{minimize}} \ \ \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[\ell(\theta; Z)]. \tag{3}$$

By setting $\mathcal{Q}$ to be the set of point masses on the set U, any robust optimization formulation can be represented as a distributionally robust optimization problem. The formulation (3) finds models or decisions θ that maintains good performance over distributional shifts in $\mathcal{Q}$. As a general desiderata, the choice of the set $\mathcal{Q}$ should represent realistic distributional shifts, while allowing efficient solution methods for the minimax problem (3).

In this lecture, we 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. To make the inner worst-case problem (3) over infinite-dimensional probability distributions tractable, we will derive a dual reformulation. We will formulate the entire problem (3) as a single minimization problem over models $\theta \in \Theta$ and dual variables relating to the inner worst-case problem.

The most important hyperparameter in these formulations are 1) the choice of distance—for probability distributions—that defines the neighborhood around P, and 2) the radius of this neighborhood. Both are nontrivial to choose, and a principled understanding of related trade-offs remain an open problem.

1.1 f-divergences

First, we consider f-divergences, which define a notion of closeness between two distributions using a convex function f. 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 the f-divergence between Q and P is

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

For $f(t) = t \log t$, we get KL-divergences. For $f(t) = |t - 1|$, we get the total variation distance. For $f(t) = (t-1)^2$, we get the χ^2 divergence.

[Draw picture visualizing f-divergences]

Rather than minimizing the average loss $\mathbb{E}_{P_0}[\ell(\theta; X)]$, consider the *distributionally robust* problem over f-divergence balls around P

$$\underset{\theta \in \Theta}{\text{minimize}} \ \ \left\{ \mathcal{R}_f(\theta; P) := \sup_{Q \ll P} \left\{ \mathbb{E}_Q[\ell(\theta; Z)] : D_f\left(Q \| P\right) \le \rho \right\} \right\}, \tag{4}$$

where the hyperparameter $\rho > 0$ modulates the magnitude of the unknown distributional shift.

The worst-case risk (4) upweights regions of $\mathcal{X}$ with high losses $\ell(\theta; Z)$, and thus formulation (4) optimizes performance on the tails, as measured by the loss on "hard" examples. So long as the perturbed distribution Q remains ρ-close to the data-generating distribution P, the model $\theta^\star \in \Theta$ optimizing the distributionally robust formulation (4) guarantees $\mathbb{E}_Q[\ell(\theta^\star; Z)] \le \mathcal{R}_f(\theta^\star; P_0)$, providing the smallest such bound. The main limitation of f-divergence balls is that it only allows distributions Q with the same support as P; by definition, this is what is required for the likelihood ratio $\frac{dQ}{dP}$ to be well defined. In particular, the empirical estimator of the worst-case formulation (4) adaptively upweights hard examples—at the current model θ.

We now derive a dual reformulation of the worst-case objective (4). We may use the likelihood ratio $L(Z) := dQ(Z)/dP(Z)$ to reformulate our distributionally robust problem (4) via

$$\mathcal{R}_f(\theta; P) = \sup_{L \ge 0} \left\{ \mathbb{E}_P[L(Z)\ell(\theta; Z)] \mid \mathbb{E}_P[f(L(Z))] \le \rho, \mathbb{E}_P[L(Z)] = 1 \right\}, \tag{5}$$

where the supremum is over measurable functions. Let f^* be the Fenchel conjugate of f

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

Proposition 1. *Let P be a probability measure on $\mathcal{Z}$ and $\rho > 0$. Then*

$$\mathcal{R}_f(\theta; P) = \inf_{\lambda \ge 0, \eta \in \mathbb{R}} \left\{ \mathbb{E}_P \left[\lambda f^* \left(\frac{\ell(\theta; Z) - \eta}{\lambda} \right) \right] + \lambda \rho + \eta \right\} \tag{6}$$

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.

For convex losses $\theta \mapsto \ell(\theta; Z)$, the dual form (6) is jointly convex in (θ, η, λ).

Sketch of Proof Fix any $\theta \in \Theta$ and let $B(z) = \ell(\theta; z)$ to simplify notation. Let us consider the likelihood ratio formulation (5). Introducing Lagrange multiplier $\lambda \ge 0$ for the constraint $\int f(L) dP \le \rho$ and $\eta \in \mathbb{R}$ for $\mathbb{E}_P[L] = 1$, we obtain the Lagrangian

$$\mathcal{L}(L, \lambda, \eta) = \int_{\mathcal{Z}} \left[(B(z) - \eta) L(z) - \lambda f(L(z)) \right] dP(x) + \lambda \rho + \eta.$$

Then taking $L \equiv 1$, we have that $\int f(L) dP = 0$ and $\mathbb{E}_P[L] = 1$, so the extended Slater condition holds. Thus we have (see, e.g., Luenberger [4, Theorem 8.6.1 and Problem 8.7]) that

$$\begin{aligned}
&\sup_{Q \ll P} \left\{ \mathbb{E}_Q[W] : D_f(Q \| P) \le \rho \right\} \\
&= \inf_{\lambda \ge 0, \eta \in \mathbb{R}} \sup_{L \ge 0} \left\{ \int_{\mathcal{X}} \left[(B(z) - \eta) L(x) - \lambda f(L(x)) \right] dP(x) + \lambda \rho + \eta \right\}.
\end{aligned} \tag{7}$$

Next, we wish to interchange the inner supremum over all (measurable) nonnegative functions $L : \mathcal{Z} \to \mathbb{R}_+$ and the integral in the dual (7). In this case, we have $\sup_{\ell \ge 0} \{ \frac{z - \eta}{\lambda} \ell - f(\ell) \} = f^*(\frac{z - \eta}{\lambda})$.

The reason this is a proof sketch is because we need to carefully deal with measurability issues and corner cases. □

As a concrete illustration, consider divergences that look like t^k. The For $k \in (1, \infty)$, $k_* = \frac{k}{k-1}$, the Cressie-Read family of f-divergences [3] is given by

$$f_k(t) := t^k - 1 \quad \text{so} \quad f_k^*(s) := k^{-k_*}(k-1)\, (s)_+^{k_*} + 1 \tag{8}$$

We let $f_k(t) = +\infty$ for $t < 0$. As k becomes smaller, we have a more conservative DRO formulation.

By minimizing out $\lambda \ge 0$ in the above dual, we obtain a simplified formulation for this family of divergences. Let $c_k(\rho) := (1 + \rho)^{\frac{1}{k}}$ to ease notation.

Lemma 1. *For any probability P on $\mathcal{Z}$,*

$$\mathcal{R}_k(\theta; P) = \inf_{\eta \in \mathbb{R}} \left\{ c_k(\rho) \mathbb{E}_P \left[(\ell(\theta; Z) - \eta)_+^{k_*} \right]^{\frac{1}{k_*}} + \eta \right\}. \tag{9}$$

This simplified form shows that distributional robustness is equivalent to optimizing the tail-performance of a model. Equivalently, we are only considering harder examples in your data, those with loss above the threshold η; tail inputs are emphasized by a power of k_*. From the lemma, our final distributionally robust problem for the Cressie-Read family becomes

$$\operatorname*{minimize}_{\theta \in \Theta, \eta \in \mathbb{R}} \left\{ c_k(\rho) \mathbb{E}_P \left[(\ell(\theta; Z) - \eta)_+^{k_*} \right]^{\frac{1}{k_*}} + \eta \right\},$$

which is jointly convex in (θ, η) if $\theta \mapsto \ell(\theta; z)$ is convex.

1.2 Wasserstein distances

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 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'.

For $\rho \geq 0$ and distribution P_0, we let $\mathcal{Q} = \{Q : W_c(Q, P) \leq \rho\}$, considering the Wasserstein distributionally robust optimization (DRO) problem

$$\operatorname*{minimize}_{\theta \in \Theta} \quad \left\{ \mathcal{R}_c(\theta; P) := \sup_{Q \ll P} \{ \mathbb{E}_Q[\ell(\theta; Z)] : W_c(Q, P) \leq \rho \} \right\}. \tag{10}$$

In particular, the Wasserstein ball allows 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 duality result gives a duality result for Wasserstein DRO (10). We assume $\mathbb{E}_P[\ell(\theta; Z)] < \infty$ throughout.

Proposition 2. *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 $\rho > 0$,*

$$\sup_{Q : W_c(Q,P) \leq \rho} \mathbb{E}_P[\ell(\theta; Z)] = \inf_{\lambda \geq 0} \left\{ \lambda \rho + \mathbb{E}_P[\phi_\lambda(\theta; Z)] \right\}. \tag{11}$$

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 (6) that upweighted examples with higher loss, we see that Wasserstein DRO (10) 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 (11) 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.

Although the proof of Proposition 2 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 (10) 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 (11) for discrete sample spaces.

References

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

[2] D. Bertsimas, D. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[3] N. Cressie and T. R. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, pages 440–464, 1984.

[4] D. Luenberger. *Optimization by Vector Space Methods.* Wiley, 1969.

[5] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, 2019.