Distributionally Robust Optimization

Lecturer: Kamalika Chaudhuri

April 24, 2020

We begin with a general framework that is capable of addressing departures from the statistical learning framework – distributionally robust optimization (DRO). Although this framework has some limitations, it is fairly general, and gives us some insight into how departures from the statistical learning framework may be addressed.

1 The General Framework

Recall that in the statistical learning framework, there is an underlying data distribution D over labeled examples, and the goal is to find a classifier $\theta \in \Theta$ that minimizes the quantity:

$$\theta^* = \text{argmin}_{\theta \in \Theta} \mathbb{E}_{(X,Y) \sim D}[\ell(\theta, X, Y)],$$

where ℓ is a smooth, convex loss. In Distributionally Robust Optimization, the goal is to find instead a $\theta \in \Theta$ that minimizes:

$$\theta^{DRO} = \text{argmin}_{\theta \in \Theta} \sup_{P : d(P,D) \leq \epsilon} \mathbb{E}_{(X,Y) \sim P}[\ell(\theta, X, Y)],$$

where P is a distribution, d measures the difference between two distributions, and ϵ is a radius parameter. Thus, DRO produces a θ that minimizes the worst case expected loss with the worst case being over distributions that are close to D. Closeness is measured by the measure d, and close means that $d(P, D) \leq \epsilon$. This allows us small (upto ϵ in measure d) departures from the statistical learning framework.

Observe that the DRO framework is very general – by varying d, we can get a lot of different kinds of departures from the We now provide two concrete and popular examples of the DRO.

1.1 Bounded f-Divergence

Suppose f is a convex function such that $f(1) = 0$. The f-divergence D_f between two distributions P and Q such that P is absolutely continuous with respect to Q is defined as:

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

The f-divergences cover a whole range of highly popular divergences – such as the KL divergence (which translates to $f(t) = t \log t$), the total variation distance (which translates to $f(t) = \frac{1}{2}|t - 1|$) and α-divergences (which translate to $f_\alpha(t) = \frac{t^\alpha - 1}{\alpha(\alpha-1)}$.

In Distributionally Robust Optimization with bounded f-divergence, we set d to be a suitable f-divergence D_f. Thus, the problem is to find a θ that minimizes:

$$\theta^f = \text{argmin}_{\theta \in \Theta} \sup_{P : D_f(P,D) \leq \epsilon} \mathbb{E}_{(X,Y)\sim P}[\ell(\theta, X, Y)], \tag{1}$$

Of course, since D is unknown, this cannot be done directly. Hence, given training data $(x_1, y_1), \ldots, (x_n, y_n)$, we minimize instead the following empirical objective:

$$\widehat{\theta}^f = \text{argmin}_{\theta \in \Theta} \max_{p : D_f(p, 1/n) \leq \epsilon} \frac{1}{n} \sum_{i=1}^{n} p_i \cdot \ell(\theta, x_i, y_i), \tag{2}$$

where p is a probability vector with n elements, and $1/n$ is the vector $[1/n, \ldots, 1/n]$.

To solve the optimization problem 2, we can rewrite it as a constrained optimization as follows:

$$\begin{aligned} \min_{\theta} \max_{p} \frac{1}{n} \sum_{i=1}^{n} p_i \ell(\theta, x_i, y_i) \\ \text{subject to:} \\ \forall i, p_i \geq 0 \\ \sum_{i=1}^{n} p_i = 1 \\ \sum_{i=1}^{n} \frac{1}{n} \cdot f(np_i) \leq \epsilon \end{aligned} \tag{3}$$

Here, the first two constraints state that p is a probability vector, and the last one bounds the f divergence between p and the uniform distribution over the training samples. Observe that this is a minimax game between two players – the p player, who tries to choose a p to maximize the objective, and the θ player who seeks to minimize the objective. As such, it can be solved by an alternating minimization algorithm – either a full-gradient or stochastic version.

1.2 Bounded Wasserstein Distance

A second popular use of Distributionally Robust Optimization is when d is the Wasserstein distance. Suppose P and Q are distributions over points in a metric space Δ. Then, the Wasserstein distance of order p between P and Q is defined as:

$$W_p(P, Q) = \inf_{\gamma \in \Gamma(P,Q)} (\mathbb{E}_{(X,Y)\sim\gamma}[\Delta(X, Y)^p])^{1/p},$$

where $\Gamma(P, Q)$ is the set of all joint distributions with marginals P and Q. When $p = 1$, the Wasserstein Distance can be thought of as a sort of transportation cost for transferring probability masses between P and Q – where the cost of moving an unit mass is proportional to the distance moved.

In this case, the DRO objective becomes

$$\min_{\theta} \sup_{P : W_p(P,D) \leq \epsilon} \mathbb{E}_{(X,Y)\sim P}[\ell(\theta, X, Y)]$$

To solve it, we write it in its Lagrangian form:

$$\min_{\theta} \sup_{P} \mathbb{E}_{(X,Y)\sim P}[\ell(\theta, X, Y)] - \lambda W_p(P, D)^p,$$

which in turn can be written as:

$$\min_{\theta} \mathbb{E}_{(X,Y)\sim D}[\sup_{(x',y')} [\ell(\theta, x', y') - \lambda \Delta((x, y), (x', y'))^p]]$$

Since D is unknown we solve the empirical form of this objective function:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} [\sup_{(x',y')} [\ell(\theta, x', y') - \lambda \Delta((x_i, y_i), (x', y'))^p]]$$

It has been shown that DRO with bounded Wasserstein Distance can be used to model adversarial robustness; for example, if the order of the Wasserstein distance $p = \infty$, then the supremum is over all distributions where each point in D can be moved by a radius of at most ϵ. Of course the disadvantage of using W_∞ is that the DRO problem becomes discontinuous and ill-behaved, and hence p is relaxed to be smaller for ease of optimization.

2 Advantages and Challenges

The main advantage of the DRO framework is its extreme generality – if we can find a suitable distance d, then we can express a problem in this general framework. In fact, a number of instantiations of this framework has been used in practice – to ensure robustness against adversarial perturbations, distribution shift, and even fairness.

The extreme generality of DRO also comes with a cost – finding d that is suitable to any particular application is not easy. Added to this is the computational challenge; even in the simple case of bounded f divergences or bounded Wasserstein distances, DRO is significantly more computationally challenging than plain empirical risk minimization – it is a minimax game, and convergence is slow. Introducing arbitrary difference measures into the equation can make the problem significantly more computationally intensive, and can potentially result in trivial solutions. Thus, while overall DRO is a nice and general framework and often a nice way of formulating problems, in practice, we may still need specific solutions for specific problems.

3 Bibliographic Notes

Distributionally Robust Optimization (DRO) has been around for a while, and has its roots in the robust optimization literature. The bounded f-divergence formulation is taken from [DGN16] and the algorithm from [ND16]. To the best of my knowledge, the Wasserstein DRO formulation and its dual derivation is due to [GK16], and the application to adversarial robustness is due to [SND17]. The potential to lead to trivial solutions was shown by [HNSS16].

References

[DGN16] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[GK16] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[HNSS16] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? *arXiv preprint arXiv:1611.02041*, 2016.

[ND16] Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in neural information processing systems*, pages 2208–2216, 2016.

[SND17] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.