Asymptotic Normality and Confidence Regions in Wasserstein Distributionally Robust Optimization

Nian Si
Joint work with Jose Blanchet and Karthyek Murthy

INFORMS Annual Meeting 2019

Stanford University

October 24, 2019

Road map

Stanford University

1. Introduction to DRO and optimal transport

2. Asymptotic behaviors and confidence regions of DRO estimators

Motivation

Stanford University

Stochastic optimization problem:

$$\inf_{\beta \in \mathbb{R}^d} \mathbb{E}_{P_*}[\ell(X;\beta)],$$

P_* : Ground truth distribution, ☹ usually unknown in practice.

Motivation

Stanford University

Stochastic optimization problem:

$$\inf_{\beta \in \mathbb{R}^d} \mathbb{E}_{P_*}[\ell(X;\beta)],$$

P_* : Ground truth distribution, ☹ usually unknown in practice.

$\Downarrow$

Empirical risk minimization (ERM):

$$\inf_{\beta \in \mathbb{R}^d} \mathbb{E}_{P_n}[\ell(X;\beta)],$$

P_n : Empirical distribution.

☹ : Overfitting $\Rightarrow$ poor out-of-sample performance.

☹ : Non-robustness $\Rightarrow$ adversarial examples [Goodfellow et al., 2014].

Motivation

Stanford University

Stochastic optimization problem:

$$\inf_{\beta \in \mathbb{R}^d} \mathbb{E}_{P_*}[\ell(X;\beta)],$$

P_* : Ground truth distribution, ☹ usually unknown in practice.

⇓

Empirical risk minimization (ERM):

$$\inf_{\beta \in \mathbb{R}^d} \mathbb{E}_{P_n}[\ell(X;\beta)],$$

P_n : Empirical distribution.

☹ : Overfitting ⇒ poor out-of-sample performance.

☹ : Non-robustness ⇒ adversarial examples [Goodfellow et al., 2014].

⇓

Robust data-driven framework.

DRO formulation

Stanford University

Distributionally Robust Optimization (DRO):

$$\inf_{\beta \in \mathbb{R}^d} \underbrace{\sup_{P \in \mathcal{U}} \mathbb{E}_P[\ell(X;\beta)]}_{\text{worst case expectation}},$$

$\mathcal{U}$: distributional uncertainty set.

DRO formulation

Stanford University

Distributionally Robust Optimization (DRO):

$$\inf_{\beta \in \mathbb{R}^d} \underbrace{\sup_{P \in \mathcal{U}} \mathbb{E}_P[\ell(X;\beta)]}_{\text{worst case expectation}},$$

$\mathcal{U}$: distributional uncertainty set.

Construction of distributional uncertainty set $\mathcal{U}$:

$$\mathcal{U} = \mathcal{U}_\delta(P_n) = \{P \in \mathcal{P}(S) : D(P, P_n) \leq \delta\}$$

Choices of $D(\cdot,\cdot)$: $f-$divergence, optimal transport cost

DRO formulation

Stanford University

Distributionally Robust Optimization (DRO):

$$\inf_{\beta \in \mathbb{R}^d} \underbrace{\sup_{P \in \mathcal{U}} \mathbb{E}_P[\ell(X;\beta)]}_{\text{worst case expectation}},$$

$\mathcal{U}$: distributional uncertainty set.

Construction of distributional uncertainty set $\mathcal{U}$:

$$\mathcal{U} = \mathcal{U}_\delta(P_n) = \{P \in \mathcal{P}(S) : D(P, P_n) \leq \delta\}$$

Choices of $D(\cdot,\cdot)$: $f-$divergence, optimal transport cost

Literatures on DRO

Stanford University

- **f-divergence:** [Bagnell, 2005; Ben-Tal et al., 2013; Bertsimas, Gupta & Kallus, 2013; Hu & Hong 2013; Lam, 2013; 2016; Wang, Glynn & Ye, 2014; Bayrakskan & Love, 2015; Duchi, Glynn & Namkoong, 2016; Duchi & Namkoong, 2016; 2017]
- **Optimal transport:** [Esfahani & Kuhn, 2018; Blanchet & Murthy, 2019; Gao & Kleywegt, 2016; Blanchet, Kang & Murthy, 2016; Gao, Chen & Kleywegt, 2017; Sinha, Namkoong & Duchi, 2017; Nguyen, Kuhn & Esfahani, 2018; Nguyen et al., 2018; **Blanchet et al., 2019**]

Optimal transport

Stanford University

- Let $P \in \mathcal{P}(S)$ and $Q \in \mathcal{P}(S)$ be two probability distributions defined on a space S; $c : S \times S \to [0, \infty]$ is a cost function.
- Optimal transport cost:

$$D_c(P, Q) = \inf_{\pi} \left\{ \mathbb{E}_\pi[c(U, V)] \mid \pi \in \mathcal{P}(S \times S), \pi_U = P, \pi_V = Q \right\}$$

Optimal transport

Stanford University

- Let $P \in \mathcal{P}(S)$ and $Q \in \mathcal{P}(S)$ be two probability distributions defined on a space S; $c : S \times S \to [0, \infty]$ is a cost function.
- Optimal transport cost:

$$D_c(P, Q) = \inf_{\pi} \left\{ \mathbb{E}_\pi[c(U, V)] \mid \pi \in \mathcal{P}(S \times S), \pi_U = P, \pi_V = Q \right\}$$

- Advantages:
  - P and Q are not required to have the same support;
  - Continuous distributions are included;
  - General enough to cover popular distances used in practice,
    $c(u, v) = \|u - v\|^\rho \implies D_c^{1/\rho}$: ρ-Wasserstein distance;
    $c(u, v) = \mathbf{1}\{u \neq v\} \implies D_c$: total variation distance.

DRO estimators

Stanford University

- **Square-root LASSO** [Belloni, Chernozhukov and Wang 2011]:

$$\ell((x,y);\beta) = \|y - \beta^T x\|_2^2$$

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}(dx,dy)$$

$$c((x,y),(x',y')) = \|x - x'\|_q^2 + \infty \cdot \mathbf{1}\{y \neq y'\}$$

DRO estimators

Stanford University

- **Square-root LASSO** [Belloni, Chernozhukov and Wang 2011]:

$$\ell((x,y);\beta) = \|y - \beta^T x\|_2^2$$

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}(dx,dy)$$

$$c((x,y),(x',y')) = \|x - x'\|_q^2 + \infty \cdot \mathbf{1}\{y \neq y'\}$$

DRO is equivalent to the square-root LASSO [Blanchet, Kang and Murthy, 2016],
$(1/p+1/q = 1)$

$$\sup_{P: D_c(P,P_n)\le\delta} \mathbb{E}_P\left[\ell\left((X,Y);\beta\right)\right] = \left(\sqrt{\mathbb{E}_{P_n}[\ell((X,Y);\beta)]} + \sqrt{\delta}\|\beta\|_p\right)^2.$$

DRO estimators

Stanford University

- **Square-root LASSO** [Belloni, Chernozhukov and Wang 2011]:

$$\ell((x,y);\beta) = \|y-\beta^T x\|_2^2$$

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{(X_i,Y_i)}(dx,dy)$$

$$c((x,y),(x',y')) = \|x-x'\|_q^2 + \infty \cdot \mathbf{1}\{y \neq y'\}$$

DRO is equivalent to the square-root LASSO [Blanchet, Kang and Murthy, 2016],
$(1/p+1/q=1)$

$$\sup_{P:D_c(P,P_n)\leq\delta} \mathbb{E}_P\left[\ell\left((X,Y);\beta\right)\right] = \left(\sqrt{\mathbb{E}_{P_n}[\ell((X,Y);\beta)]} + \sqrt{\delta}\|\beta\|_p\right)^2.$$

- Regularized logistic regression, SVMs...

Road map

Stanford University

- **The asymptotic behaviors of DRO estimators?**
  Suppose $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P_*$,

$$\begin{aligned} \beta_n^{ERM} &\in arg\min_{\beta} \mathbb{E}_{P_n}\left[\ell(X;\beta)\right], \\ \beta_n^{DRO}(\delta) &\in \arg\min_{\beta} \sup_{P\in\mathcal{U}_\delta(P_n)} \mathbb{E}_{P_n}\left[\ell(X;\beta)\right], \\ \beta_* &= \arg\min_{\beta} \mathbb{E}_{P_*}\left[\ell(X;\beta)\right]. \end{aligned}$$

  We want to study the joint limit of
  $\left(n^{1/2}(\beta_n^{ERM} - \beta_*), n^{?}(\beta_n^{DRO}(\delta_n) - \beta_*)\right)$ with the correct scaling rate.

- **The asymptotic behaviors of DRO estimators?**
  Suppose $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P_*$,

$$\begin{aligned} \beta_n^{ERM} &\in arg\min_{\beta} \mathbb{E}_{P_n}\left[\ell(X;\beta)\right], \\ \beta_n^{DRO}(\delta) &\in \arg\min_{\beta} \sup_{P\in\mathcal{U}_\delta(P_n)} \mathbb{E}_{P_n}\left[\ell(X;\beta)\right], \\ \beta_* &= \arg\min_{\beta} \mathbb{E}_{P_*}\left[\ell(X;\beta)\right]. \end{aligned}$$

  We want to study the joint limit of
  $\left(n^{1/2}(\beta_n^{ERM} - \beta_*), n^{?}(\beta_n^{DRO}(\delta_n) - \beta_*)\right)$ with the correct scaling rate.
- **The suitable confidence regions in DRO problems?**
  We want to find a confidence region Λ_n that

$$\beta_n^{ERM} \in \Lambda_n,\ \beta_n^{DRO}(\delta_n) \in \Lambda_n \text{ and } \lim_{n\to\infty} \mathbf{P}\left(\beta_* \in \Lambda_n\right) = 1 - \alpha.$$

"Compatible" set

Stanford University

- Define "Compatible" set as

$$\Lambda_{\delta_n}(P_n) := \left\{ \beta \in \mathbb{R}^d : \beta \in \arg\min_{\beta} \mathbb{E}_P \left[\ell(X;\beta) \right] \text{ for a } P \in \mathcal{U}_{\delta_n}(P_n) \right\}.$$

- $\Lambda_{\delta_n}(P_n)$ denotes the set of choices of $\beta \in \mathbb{R}^d$ that are "compatible" with the distributional uncertainty region, in the sense that for every $\beta \in \Lambda_{\delta_n}(P_n)$, there exists a probability distribution $P \in \mathcal{U}_{\delta_n}(P_n)$ for which β is optimal.
- $\Lambda_{\delta_n}(P_n)$ naturally serves as a good candidate of confidence regions.

Preliminaries

Stanford University

- We consider the cost function with the form $c(u, w) = \|u - w\|_q^2$.
- Let $h(x, \beta) := D_\beta \ell(x, \beta)$ be the gradient of the loss function and $C := \mathbb{E}\left[D_\beta h(X, \beta_*)\right] \succ \mathbf{0}$.
- Define

$$\varphi(\xi) := \frac{1}{4}\mathbb{E}_{P_*}\left(\left\|\left(D_x h(X, \beta_*)\right)^T \xi\right\|_p^2\right),$$

where $1/p + 1/q = 1$ and its convex conjugate:

$$\varphi^*(\zeta) := \sup_{\xi \in \mathbb{R}^d}\left\{\xi^T \zeta - \varphi(\xi)\right\}.$$

- Define

$$S(\beta) := \sqrt{\mathbb{E}_{P_*}\|D_x \ell(X;\beta)\|_p^2}.$$

Main asymptotic theorem

Stanford University

Theorem (Main theorem)

Suppose $\ell(x,\cdot)$ *is convex and* $\ell(\cdot)$ *satisfies mild regularity conditions. Let* $\delta_n = n^{-\gamma}\eta$ *for* $\gamma,\eta \in (0,\infty)$, *and* $H \sim \mathcal{N}(\mathbf{0}, Cov[h(X,\beta_*)])$. *Then,*

$$\left(n^{1/2}(\beta_n^{ERM} - \beta_*),\ n^{\bar{\gamma}/2}(\beta_n^{DRO}(\delta_n) - \beta_*), n^{1/2}\left(\Lambda_{\delta_n}(P_n) - \beta_*\right)\right)$$
$$\Rightarrow \left(C^{-1}H,\ C^{-1}f_{\eta,\gamma}(H),\ \Lambda_{\eta,\gamma} + C^{-1}H\right),$$

where $\bar{\gamma} := \min\{\gamma, 1\}$ *and* $f_{\eta,\gamma}(x), \Lambda_{\eta,\gamma}$ *will be defined later according to* γ.

Main asymptotic theorem : Remarks

Stanford University

This theorem works for every scaling rate $\delta_n = \eta/n^\gamma, \gamma > 0$. However, only $\delta_n = \eta/n$ gives the non-trivial limits.

Main asymptotic theorem : Remarks

Stanford University

This theorem works for every scaling rate $\delta_n = \eta/n^\gamma, \gamma > 0$. However, only $\delta_n = \eta/n$ gives the non-trivial limits.

- $\gamma > 1$: Lack of robustness. β_n^{DRO} and β_n^{ERM} are asymptotically indistinguishable,

$$n^{1/2}\left((\beta_n^{ERM} - \beta_*), (\beta_n^{DRO}(\delta_n) - \beta_*), (\Lambda_{\delta_n}(P_n) - \beta_*)\right) \Rightarrow \left(C^{-1}H, C^{-1}H, \{C^{-1}H\}\right).$$

Main asymptotic theorem : Remarks

Stanford University

This theorem works for every scaling rate $\delta_n = \eta/n^{\gamma}, \gamma > 0$. However, only $\delta_n = \eta/n$ gives the non-trivial limits.

- $\gamma > 1$: Lack of robustness. β_n^{DRO} and β_n^{ERM} are asymptotically indistinguishable,

$$n^{1/2}\left((\beta_n^{ERM} - \beta_*), (\beta_n^{DRO}(\delta_n) - \beta_*), (\Lambda_{\delta_n}(P_n) - \beta_*)\right) \Rightarrow \left(C^{-1}H, C^{-1}H, \{C^{-1}H\}\right).$$

- $\gamma < 1$: Excessive robustness. Slow convergence rate and an asymptotically bias,

$$\left(n^{\gamma/2}(\beta_n^{DRO}(\delta_n) - \beta_*), n^{1/2}\left(\Lambda_{\delta_n}(P_n) - \beta_*\right)\right) \Rightarrow \left(-\sqrt{\eta}C^{-1}D_\beta S(\beta_*), \mathbb{R}^d\right).$$

Main asymptotic theorem : $\gamma = 1$

Stanford University

- $\gamma = 1$: non-trivial limits.

$$n^{1/2}\left((\beta_n^{ERM} - \beta_*),\ (\beta_n^{DRO}(\delta_n) - \beta_*), (\Lambda_{\delta_n}(P_n) - \beta_*)\right)$$
$$\Rightarrow \left(C^{-1}H,\ C^{-1}H - \sqrt{\eta}C^{-1}D_\beta S(\beta_*), \{u : \varphi^*(Cu) \leq \eta\} + C^{-1}H\right).$$

- Here, $\Lambda_{\eta,1}$ is defined by

$$\Lambda_{\eta,1} = \{u : \varphi^*(Cu) \leq \eta\}.$$

Confidence regions: $\delta_n = \eta/n$

Stanford University

- DRO solution is inside the "compatible" set $(\beta_n^{DRO}(\delta_n) \in \Lambda_{\delta_n}(P_n))$, because of the proposition below.

Proposition (Blanchet et.al., 2016)

If $\ell(x, \cdot)$ is convex, we have for any $\delta > 0$,

$$\inf_{\beta} \sup_{P: D(P_n, P) \leq \delta} \mathbb{E}_P \left[\ell(X; \beta) \right] = \sup_{P: D(P_n, P) \leq \delta} \inf_{\beta} \mathbb{E}_P \left[\ell(X; \beta) \right].$$

Confidence regions: $\delta_n = \eta / n$

Stanford University

- DRO solution is inside the "compatible" set ($\beta_n^{DRO}(\delta_n) \in \Lambda_{\delta_n}(P_n)$), because of the proposition below.

Proposition (Blanchet et.al., 2016)

If $\ell(x, \cdot)$ is convex, we have for any $\delta > 0$,

$$\inf_{\beta} \sup_{P:D(P_n,P)\le\delta} \mathbb{E}_P\left[\ell(X;\beta)\right] = \sup_{P:D(P_n,P)\le\delta} \inf_{\beta} \mathbb{E}_P\left[\ell(X;\beta)\right].$$

- $\Lambda_{\delta_n}(P_n)$ has exact asymptotic coverage.

$$\begin{aligned}\lim_{n\to\infty} \mathbf{P}\left(\beta_* \in \Lambda_{\delta_n}(P_n)\right) &= \mathbf{P}(-C^{-1}H \in \{u : \varphi^*(Cu) \le \eta_\alpha\}) \\ &= \mathbf{P}(\varphi^*(H) \le \eta) = 1 - \alpha.\end{aligned}$$

where η_α is the $(1-\alpha)$-quantile of the random variable $\varphi^*(H)$.

Approximation of confidence regions

Stanford University

- $\Lambda_{\delta_n}(P_n)$ is generally challenging to compute. Here we provide an approximation of $\Lambda_{\delta_n}(P_n)$ based on the following corollary.

Corollary (informal)

Under the assumptions of main theorem, we have (omitting γ in $\Lambda_{\eta,\gamma}$)

$$\Lambda_{\delta_n}(P_n) \approx \beta_n^{ERM} + n^{-1/2}\Lambda_\eta \approx \beta_n^{ERM} + n^{-1/2}\Lambda_\eta^n.$$

where $\Lambda_\eta^n := \{u : \varphi_n^*(C_n u) \leq \eta\}$ and $\varphi_n(\cdot)$, C_n are the empirical analogs of $\varphi(\cdot)$, C.

Computation of confidence regions

Stanford University

- Evaluating convex conjugate is generally time-consuming. We give a computationally efficient algorithm of Λ_η using support function.

Computation of confidence regions

Stanford University

- Evaluating convex conjugate is generally time-consuming. We give a computationally efficient algorithm of Λ_η using support function.
- For any $u_1, \ldots, u_m \in \mathbb{R}^d$, with $\|u_i\|_2 = 1$ we have

$$\Lambda_\eta = \cap_u \{v : u \cdot v \le h_{\Lambda_\eta}(u)\} \subset \cap_{u_1,\ldots u_m} \{v : u_i \cdot v \le h_{\Lambda_\eta}(u_i)\}.$$

We can sample directions $u_1, \ldots, u_m$ to obtain a tight envelope of Λ_η.

Computation of confidence regions

- Evaluating convex conjugate is generally time-consuming. We give a computationally efficient algorithm of Λ_η using support function.
- For any $u_1, ..., u_m \in \mathbb{R}^d$, with $\|u_i\|_2 = 1$ we have

$$\Lambda_\eta = \cap_u \{v : u \cdot v \leq h_{\Lambda_\eta}(u)\} \subset \cap_{u_1, \ldots u_m} \{v : u_i \cdot v \leq h_{\Lambda_\eta}(u_i)\}.$$

  We can sample directions $u_1, ..., u_m$ to obtain a tight envelope of Λ_η.
- $h_{\Lambda_\eta}(v)$ is the support function of the convex set Λ_η, defined as

$$h_{\Lambda_\eta}(x) := \sup_a \{x \cdot a : a \in \Lambda_\eta\} = 2\sqrt{\eta\varphi(C^{-1}v)},$$

Computation of confidence regions

Stanford University

- Evaluating convex conjugate is generally time-consuming. We give a computationally efficient algorithm of Λ_η using support function.
- For any $u_1, ..., u_m \in \mathbb{R}^d$, with $\|u_i\|_2 = 1$ we have

$$\Lambda_\eta = \cap_u \{v : u \cdot v \leq h_{\Lambda_\eta}(u)\} \subset \cap_{u_1,\ldots u_m} \{v : u_i \cdot v \leq h_{\Lambda_\eta}(u_i)\}.$$

  We can sample directions $u_1, ..., u_m$ to obtain a tight envelope of Λ_η.
- $h_{\Lambda_\eta}(v)$ is the support function of the convex set Λ_η, defined as

$$h_{\Lambda_\eta}(x) := \sup_a \{x \cdot a : a \in \Lambda_\eta\} = 2\sqrt{\eta\varphi(C^{-1}v)},$$

- A completely analogous method can be used to estimate Λ_η^n.

Confidence regions of square-root LASSO

Stanford University

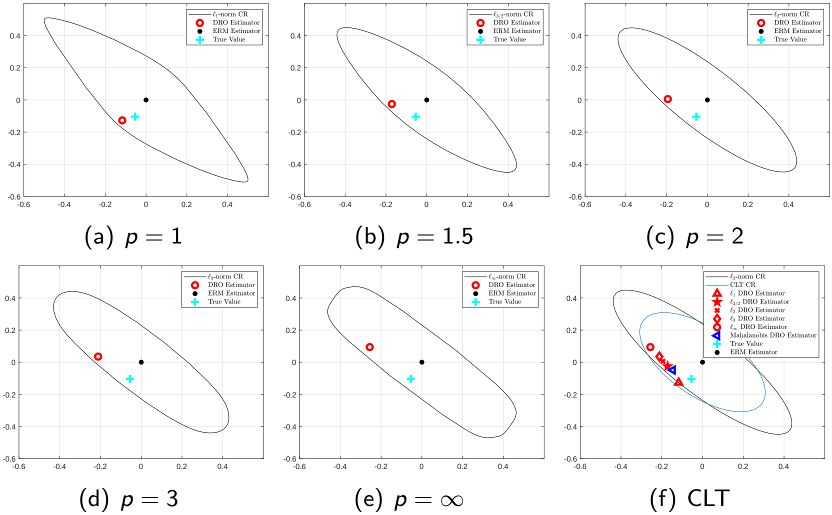

Figure: Confidence regions for different norms centered at the ERM solution

Contributions

Stanford University

- Asymptotic normality of Wasserstein-DRO estimators: arbitrary scaling of uncertainty size.
- Suitable confidence regions for DRO problems: coverage, approximation and computation.

Reference

Stanford University

Blanchet, J., Murthy, K., & **Si, N.** (2019). Confidence Regions in Wasserstein Distributionally Robust Estimation. arXiv preprint arXiv:1906.01614.

Thanks!