Decision Making and Inference Under Model Misspecification

Jose Blanchet.

Stanford University (Management Science and Engineering), and Institute for Computational and Mathematical Engineering).

Goals

- Distributional robust decisions (portfolio choice).

Goals

- Distributional robust decisions (portfolio choice).
- Statistical guarantees.

Goals

- Distributional robust decisions (portfolio choice).
- Statistical guarantees.
- Constraints (e.g. market signals).

Goals

- Distributional robust decisions (portfolio choice).
- Statistical guarantees.
- Constraints (e.g. market signals).
- Optimization algorithms.

Portfolio Optimization "101"

- Suppose that R is a random vector of returns and w are portfolio weights.

Portfolio Optimization "101"

- Suppose that R is a random vector of returns and w are portfolio weights.
- Mean-variance portfolio selection can be expressed as:

$$
\begin{aligned}
& \min_{w^T 1=1,\alpha} E_{P_*}[\left(w^T R-\alpha\right)^2]-\lambda E_{P^*}\left(w^T R\right) \\
= & \min_{w^T 1=1} Var_{P_*}\left(w^T R-\alpha\right)-\lambda E_{P^*}\left(w^T R\right).
\end{aligned}
$$

Portfolio Optimization "101"

- Suppose that R is a random vector of returns and w are portfolio weights.
- Mean-variance portfolio selection can be expressed as:

$$
\begin{aligned}
& \min_{w^T 1=1,\alpha} E_{P_*}[\left(w^T R-\alpha\right)^2]-\lambda E_{P^*}\left(w^T R\right) \\
= & \min_{w^T 1=1} Var_{P_*}\left(w^T R-\alpha\right)-\lambda E_{P^*}\left(w^T R\right).
\end{aligned}
$$

- The notation $E_{P_*}(\cdot)$ means using the probability model P_*.

Generalization

- Suppose that X is a random vector with distribution P_* we want to solve
$$\min_{\theta \in \Theta} E_{P_*}\left[L\left(X, \theta\right)\right].$$

Generalization

- Suppose that X is a random vector with distribution P_* we want to solve
$$\min_{\theta \in \Theta} E_{P_*}\left[L\left(X,\theta\right)\right].$$
- Important special case: affine decision rules
$$L\left(X,\theta\right) = l\left(\theta^T X\right).$$

Generalization

- Suppose that X is a random vector with distribution P_* we want to solve

$$\min_{\theta \in \Theta} E_{P_*}\left[L\left(X, \theta\right)\right].$$

- Important special case: affine decision rules

$$L\left(X, \theta\right) = l\left(\theta^T X\right).$$

- Affine decision rules includes portfolio selection + generalized linear models.

Generalization

- Suppose that X is a random vector with distribution P_* we want to solve

$$\min_{\theta\in\Theta} E_{P_*}\left[L\left(X,\theta\right)\right].$$

- Important special case: affine decision rules

$$L\left(X,\theta\right)=l\left(\theta^T X\right).$$

- Affine decision rules includes portfolio selection + generalized linear models.
- **Problem: Don't have access to P_*...**

Distributionally Robust Optimization

- Choose a proxy model, P_0 (e.g. $P_n =$ empirical measure).

Distributionally Robust Optimization

- Choose a proxy model, P_0 (e.g. $P_n =$ empirical measure).
- Choose a distributional uncertainty region around $\mathcal{U}_\delta\left(P_0\right)$.

Distributionally Robust Optimization

- Choose a proxy model, P_0 (e.g. $P_n =$ empirical measure).
- Choose a distributional uncertainty region around $\mathcal{U}_\delta\left(P_0\right)$.
- Solve

$$\min_{\theta} \max_{P \in \mathcal{U}_\delta(P_0)} E_P\left(L\left(X,\theta\right)\right).$$

Distributionally Robust Optimization

- Choose a proxy model, P_0 (e.g. $P_n =$ empirical measure).
- Choose a distributional uncertainty region around $\mathcal{U}_\delta\left(P_0\right)$.
- Solve

$$\min_{\theta} \max_{P \in \mathcal{U}_\delta(P_0)} E_P\left(L\left(X, \theta\right)\right).$$

- Say, $\mathcal{U}_\delta\left(P_0\right) = \{P : D\left(P, P_0\right) \leq \delta\}$, how to choose D?

Distributionally Robust Optimization

- Choose a proxy model, P_0 (e.g. $P_n =$ empirical measure).
- Choose a distributional uncertainty region around $\mathcal{U}_\delta\left(P_0\right)$.
- Solve

$$\min_\theta \max_{P \in \mathcal{U}_\delta(P_0)} E_P\left(L\left(X, \theta\right)\right).$$

- Say, $\mathcal{U}_\delta\left(P_0\right) = \{P : D\left(P, P_0\right) \leq \delta\}$, how to choose D?
- What does this mean? What's the intuition?

Distributionally Robust Optimization

- Choose a proxy model, P_0 (e.g. $P_n =$ empirical measure).
- Choose a distributional uncertainty region around $\mathcal{U}_\delta\left(P_0\right)$.
- Solve

$$\min_{\theta} \max_{P \in \mathcal{U}_\delta(P_0)} E_P\left(L\left(X, \theta\right)\right).$$

- Say, $\mathcal{U}_\delta\left(P_0\right) = \left\{P : D\left(P, P_0\right) \leq \delta\right\}$, how to choose D?
- What does this mean? What's the intuition?
- What about $\delta =$ size of uncertainty?

Distributionally Robust Optimization (DRO)

- Solve

$$\min_{\theta} \max_{P \in \mathcal{U}_{\delta}(P_0)} E_P \left(l \left(X, \theta \right) \right).$$

Distributionally Robust Optimization (DRO)

- Solve

$$\min_{\theta} \max_{P \in \mathcal{U}_{\delta}(P_0)} E_P \left(l \left(X, \theta \right) \right).$$

- How to compute θ optimally?

Distributionally Robust Optimization (DRO)

- Solve
$$\min_{\theta} \max_{P \in \mathcal{U}_\delta(P_0)} E_P\left(l\left(X, \theta\right)\right).$$
- How to compute θ optimally?
- Structure of the worst case distribution?

Distributionally Robust Optimization (DRO)

- Solve

$$\min_{\theta} \max_{P \in \mathcal{U}_{\delta}(P_0)} E_P \left(l \left(X, \theta \right) \right).$$

- How to compute θ optimally?
- Structure of the worst case distribution?
- Is there a Nash equilibrium?

Distributionally Robust Optimization (DRO)

- Solve

$$\min_{\theta} \max_{P \in \mathcal{U}_{\delta}(P_0)} E_P \left(l \left(X, \theta \right) \right).$$

- How to compute θ optimally?
- Structure of the worst case distribution?
- Is there a Nash equilibrium?
- How does this relate to stats theory etc?

Distributionally Robust Optimization (DRO)

- Solve

$$\min_{\theta} \max_{P \in \mathcal{U}_{\delta}(P_0)} E_P \left(l \left(X, \theta \right) \right).$$

- How to compute θ optimally?
- Structure of the worst case distribution?
- Is there a Nash equilibrium?
- How does this relate to stats theory etc?
- How does this approach work in portfolio optimization?

Quick Answers + References

- $D(P, P_0)$ using optimal transport: General duality - applicable even to control problems (B. & Murthy (2019) - https://pubsonline.informs.org/doi/10.1287/moor.2018.0936).

Quick Answers + References

- $D(P, P_0)$ using optimal transport: General duality - applicable even to control problems (B. & Murthy (2019) - https://pubsonline.informs.org/doi/10.1287/moor.2018.0936).
- Meaning: Recovers exactly (sqrt-Lasso + many other classical ML estimators): B., Murthy & Kang (2019) - "Robust Wasserstein Profile Inference (RWPI)"-https://doi.org/10.1017/jpr.2019.49

Quick Answers + References

- $D(P, P_0)$ using optimal transport: General duality - applicable even to control problems (B. & Murthy (2019) - https://pubsonline.informs.org/doi/10.1287/moor.2018.0936).
- Meaning: Recovers exactly (sqrt-Lasso + many other classical ML estimators): B., Murthy & Kang (2019) - "Robust Wasserstein Profile Inference (RWPI)"-https://doi.org/10.1017/jpr.2019.49
- Choose δ optimally using a projection criterion (RWPI) - recovers high-dimensional stats prescriptions.

Quick Answers + References

- $D(P, P_0)$ using optimal transport: General duality - applicable even to control problems (B. & Murthy (2019) - https://pubsonline.informs.org/doi/10.1287/moor.2018.0936).
- Meaning: Recovers exactly (sqrt-Lasso + many other classical ML estimators): B., Murthy & Kang (2019) - "Robust Wasserstein Profile Inference (RWPI)"-https://doi.org/10.1017/jpr.2019.49
- Choose δ optimally using a projection criterion (RWPI) - recovers high-dimensional stats prescriptions.
- Computing θ efficiently: Optimal iteration complexity for affine decision rules B., Murthy, Zhang (2021) https://pubsonline-informs-org.stanford.idm.oclc.org/doi/abs/10.1287/moor.2021.1178

Quick Answers + References

- $D(P, P_0)$ using optimal transport: General duality - applicable even to control problems (B. & Murthy (2019) - https://pubsonline.informs.org/doi/10.1287/moor.2018.0936).
- Meaning: Recovers exactly (sqrt-Lasso + many other classical ML estimators): B., Murthy & Kang (2019) - "Robust Wasserstein Profile Inference (RWPI)"-https://doi.org/10.1017/jpr.2019.49
- Choose δ optimally using a projection criterion (RWPI) - recovers high-dimensional stats prescriptions.
- Computing θ efficiently: Optimal iteration complexity for affine decision rules B., Murthy, Zhang (2021) https://pubsonline-informs-org.stanford.idm.oclc.org/doi/abs/10.1287/moor.2021.1178
- Structure of the worst case distribution - also in B, Murthy, Zhang (2021).

Quick Answers + References

- Is there a Nash equilibrium? B., Murthy, Si (2021) https://academic.oup.com/biomet/advance-article/doi/10.1093/biomet/asac001/6537610

Quick Answers + References

- Is there a Nash equilibrium? B., Murthy, Si (2021) https://academic.oup.com/biomet/advance-article/doi/10.1093/biomet/asac001/6537610
- CLT for decisions (under convexity) also in B., Murthy, Si (2021)

Quick Answers + References

- Is there a Nash equilibrium? B., Murthy, Si (2021) https://academic.oup.com/biomet/advance-article/doi/10.1093/biomet/asac001/6537610
- CLT for decisions (under convexity) also in B., Murthy, Si (2021)
- Portfolio optimization: B., Chen, Zhou (2021) https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2021.4155

Quick Answers + References

- Is there a Nash equilibrium? B., Murthy, Si (2021) https://academic.oup.com/biomet/advance-article/doi/10.1093/biomet/asac001/6537610
- CLT for decisions (under convexity) also in B., Murthy, Si (2021)
- Portfolio optimization: B., Chen, Zhou (2021) https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2021.4155
- Martingale constraints: Zhou, B., Glynn - https://arxiv.org/abs/2106.07191

Quick Answers + References

- Is there a Nash equilibrium? B., Murthy, Si (2021) https://academic.oup.com/biomet/advance-article/doi/10.1093/biomet/asac001/6537610
- CLT for decisions (under convexity) also in B., Murthy, Si (2021)
- Portfolio optimization: B., Chen, Zhou (2021) https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2021.4155
- Martingale constraints: Zhou, B., Glynn - https://arxiv.org/abs/2106.07191
- Other market constraints (implied volatility) - also in B., Murthy, Zhang (2021)

Tutorials + Surveys

- Kuhn, Esfahani, Nguyen, Shafieezadeh-Abadeh: https://pubsonline.informs.org/doi/10.1287/educ.2019.0198

Tutorials + Surveys

- Kuhn, Esfahani, Nguyen, Shafieezadeh-Abadeh: https://pubsonline.informs.org/doi/10.1287/educ.2019.0198
- Rahimian and Mehrotra (2019): **https://arxiv.org/abs/1908.05659**.

Tutorials + Surveys

- Kuhn, Esfahani, Nguyen, Shafieezadeh-Abadeh: https://pubsonline.informs.org/doi/10.1287/educ.2019.0198
- Rahimian and Mehrotra (2019): **https://arxiv.org/abs/1908.05659**.
- B., Murthy, Nguyen (2022) - https://pubsonline.informs.org/doi/abs/10.1287/educ.2021.02337191

Related Literature

- **General RO & Divergence-DRO:** Dupuis, James & Peterson '00; Hansen & Sargent '01, '08; Nilim & El Ghaoui '02, '03; Iyengar '05; A. Ben-Tal, L. El Ghaoui, & A. Nemirovski '09; Bertsimas & Sim '04; Bertsimas, Brown, Caramanis '13; Lim & Shanthikumar '04; Lam '13, '17; Csiszár & Breuer '13; Jiang & Guan '12; Hu & Hong '13; Wang, Glynn & Ye '14; Bayrakskan & Love '15; Duchi, Glynn & Namkoong '16; Bandi and Bertsimas '15; Bertsimas, Gupta & Kallus '13.

Related Literature

- **General RO & Divergence-DRO:** Dupuis, James & Peterson '00; Hansen & Sargent '01, '08; Nilim & El Ghaoui '02, '03; Iyengar '05; A. Ben-Tal, L. El Ghaoui, & A. Nemirovski '09; Bertsimas & Sim '04; Bertsimas, Brown, Caramanis '13; Lim & Shanthikumar '04; Lam '13, '17; Csiszár & Breuer '13; Jiang & Guan '12; Hu & Hong '13; Wang, Glynn & Ye '14; Bayrakskan & Love '15; Duchi, Glynn & Namkoong '16; Bandi and Bertsimas '15; Bertsimas, Gupta & Kallus '13.
- **Wasserstein/OT-DRO & Moments:** Scarf '58; Shapiro '15; Delage & Ye '10; Hampel '73; Huber '81; Pflug & Wozabal '07; Delage & Ye '10; Mehrotra & Zhang '14; Esfahani & Kuhn '15; Blanchet & Murthy '16; Gao & Kleywegt '16; Duchi & Namkoong '17.

What Is the Wasserstein Distance /OT Cost?

- Formally, given $c(x,y) \geq 0$ lower semicontinuous with $c(x,x) = 0$,

$$
\begin{aligned}
D(P,Q) &= \min \int c(x,y)\,\pi(dx,dy) \\
\text{s.t. } \int \pi(dx,dy) &= P(dx) \\
\int \pi(dx,dy) &= Q(dy) \\
\pi(dx,dy) &\geq 0.
\end{aligned}
$$

What Is the Wasserstein Distance /OT Cost?

- Formally, given $c(x,y) \geq 0$ lower semicontinuous with $c(x,x) = 0$,

$$
\begin{aligned}
D(P,Q) &= \min \int c(x,y)\,\pi(dx,dy) \\
\text{s.t.} \int \pi(dx,dy) &= P(dx) \\
\int \pi(dx,dy) &= Q(dy) \\
\pi(dx,dy) &\geq 0.
\end{aligned}
$$

- Wasserstein distance $= c(x,y) = \|x-y\|$.

What Is the Wasserstein Distance /OT Cost?

- Formally, given $c(x,y) \geq 0$ lower semicontinuous with $c(x,x) = 0$,

$$\begin{aligned} D(P,Q) &= \min \int c(x,y)\, \pi(dx,dy) \\ \text{s.t. } \int \pi(dx,dy) &= P(dx) \\ \int \pi(dx,dy) &= Q(dy) \\ \pi(dx,dy) &\geq 0. \end{aligned}$$

- Wasserstein distance $= c(x,y) = \|x-y\|$.
- **Consider type 2 Wasserstein distance** $c(x,y) = \|x-y\|^2$.

Why Care about OT DRO?

- **A way to naturally deal with out-of-sample impact...**

Why Care about OT DRO?

- **A way to naturally deal with out-of-sample impact...**
- Can interpret $\mathcal{U}_\delta\left(P_n\right)$ in Wasserstein DRO as perturbing: $X_i \rightarrow X_i + \Delta_i$ such that

$$\frac{1}{n}\sum_{i=1}^{n} c\left(X_i, X_i + \Delta_i\right) \leq \delta.$$

Why Care about OT DRO?

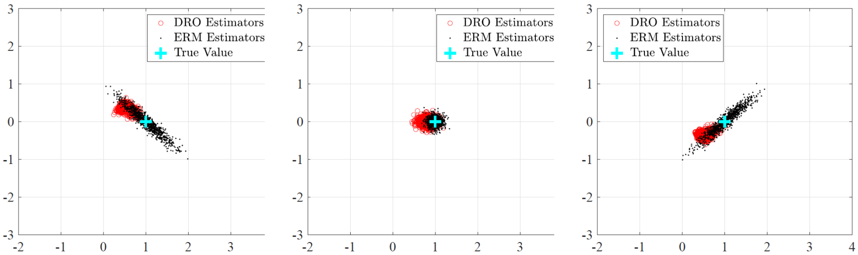
- **A way to naturally deal with out-of-sample impact...**
- Can interpret $\mathcal{U}_{\delta}\left(P_n\right)$ in Wasserstein DRO as perturbing: $X_i \rightarrow X_i + \Delta_i$ such that

$$\frac{1}{n}\sum_{i=1}^{n} c\left(X_i, X_i + \Delta_i\right) \leq \delta.$$

- Wasserstein DRO estimator is best response when perturbing each data point subject to an average budget δ.

Why Care about Wasserstein DRO?

- **A way to provide statistical regularization**



(a) $\rho = 0.95$ (b) $\rho = 0$ (c) $\rho = -0.95$

FIGURE 3. Scatter plots of β_n^{ERM} (black circles) and β_n^{DRO} (red circles) for $\beta_0 = [1.0, 0.0]^T$.

Why Care about Wasserstein DRO?

- **A way to provide statistical regularization**

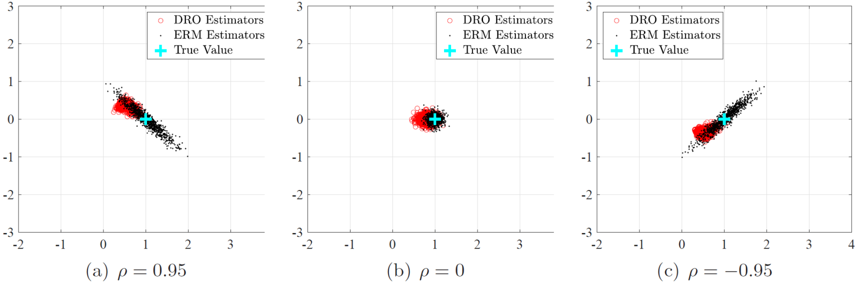


FIGURE 3. Scatter plots of β_n^{ERM} (black circles) and β_n^{DRO} (red circles) for $\beta_0 = [1.0, 0.0]^{\mathrm{T}}$.

- Linear regression with close to co-linear covariates (1000 experiments) showing Empirical Risk Minimization (i.e. $\delta = 0$) vs DRO with **optimal choice of** δ.

Key Duality Results

- B. & Murthy (2019) - https://pubsonline.informs.org/doi/10.1287/moor.2018.0936

Primal form:

$$\inf_{\theta\in\Theta} \sup_{P:D_c(P,P_n)\leq\delta} E_P\left[\ell(X,\theta)\right]$$

$$\Updownarrow$$

under mild conditions

Dual form:

$$\inf_{\theta\in\Theta} \inf_{\lambda\geq 0} \lambda\delta + E_{P_n}[\max_z \ell(z,\theta) - \lambda c(z,X)]$$

Example 1

$$\min_{\theta} \sup_{P:D_c(P,P_n)\leq \delta} E_P[(Y-\theta^\intercal X)^2]$$

$$\Updownarrow \quad c\big((x,y),(x',y')\big) = \|x-x'\|_q^2 + \infty\,|y-y'|^2$$

$$\frac{1}{q}+\frac{1}{p}=1$$

$$\min_{\theta} E_{P_n}\left[(Y-\theta^\intercal X)^2\right]^{1/2} + \delta^{1/2}\|\theta\|_p$$

Example 2

$$\min \left\{ \sup_{P:D_c(P,P_n)\leq \delta} \theta^\intercal \mathrm{Cov}_P(X)\theta : \; \theta^\intercal 1 = 1, \min_{P:D_c(P,P_n)\leq \delta} E_P\left[\theta^\intercal X\right] \geq t \right\}$$

$$c(x,x') = \|x - x'\|_q^2, \quad \frac{1}{q} + \frac{1}{p} = 1$$

$$\min \left[\sqrt{\theta^\intercal \mathrm{Cov}_{P_n}(X)\theta} + \delta^{\frac{1}{2}} \|\theta\|_p \right]^2$$

$$\text{s.t. } \theta^\intercal 1 = 1, \; \theta^\intercal E_{P_n}[X] \geq t + \delta^{\frac{1}{2}} \|\theta\|_p$$

Comments

- Include side information: For example

$$c\left(X_i, X_i'\right) = \left(X_i - X_i'\right)^T A_i \left(X_i - X_i'\right),$$

where A_i is calibrated to market data.

Comments

- Include side information: For example

$$c\left(X_i, X_i'\right) = \left(X_i - X_i'\right)^T A_i \left(X_i - X_i'\right),$$

  where A_i is calibrated to market data.
- A_i is inversely proportional to implied volatility (more volatility –> cheaper transport –> adversary focus on risky asset) - B., Murthy, Zhang (2021).

Comments

- Include side information: For example

$$c\left(X_i, X_i'\right) = \left(X_i - X_i'\right)^T A_i \left(X_i - X_i'\right),$$

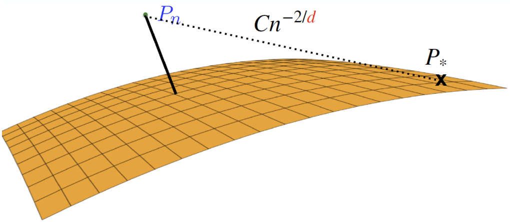
  where A_i is calibrated to market data.
- A_i is inversely proportional to implied volatility (more volatility −> cheaper transport −> adversary focus on risky asset) - B., Murthy, Zhang (2021).
- Worst case adversarial distribution.

Comments

- Include side information: For example
$$c\left(X_i, X_i'\right) = \left(X_i - X_i'\right)^T A_i \left(X_i - X_i'\right),$$
where A_i is calibrated to market data.
- A_i is inversely proportional to implied volatility (more volatility $->$ cheaper transport $->$ adversary focus on risky asset) - B., Murthy, Zhang (2021).
- Worst case adversarial distribution.
- Insight on choosing δ (without time-consuming cross validation).

Choosing Size of Uncertainty

DRO: $\min_{\theta\in\Theta} \max_{P:D_c(P,P_n)\leq\delta} E_P\left[\ell(X,\theta)\right]$



From concentration inequalities:

select δ large enough s.t.

$$D_c(P_*,P_n) \leq \delta$$

with high probability:

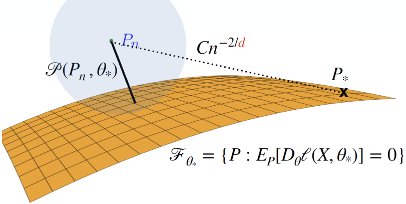
$$\delta > Cn^{-2/d}$$

=> need 2^d x more samples
to reduce error by 1/2

Choosing Size of Uncertainty Optimally

- B., Kang, Murthy (2019): https://doi.org/10.1017/jpr.2019.49

DRO: $\min_{\theta \in \Theta} \max_{P: D_c(P, P_n) \leq \delta} E_P \left[\ell(X, \theta)\right]$



Compatible uncertainty in decisions:

$$\Lambda_\delta(P_n) = \left\{\theta \in \Theta : \theta \text{ is optimal for some } P \in \mathscr{U}_\delta(P_n)\right\}$$

Question:

What is the smallest δ s.t. a true optimal solution θ_* lies in $\Lambda_\delta(P_n)$?

Note:

$\theta_* \in \Lambda_\delta(P_n) \iff$ projection $\in \mathscr{U}_\delta(P_n)$

$\iff \delta > \mathscr{P}(P_n, \theta_*)$

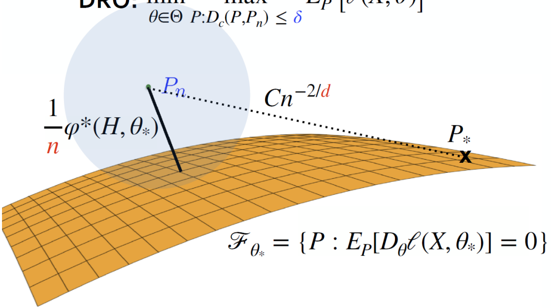
Optimal Choice of Uncertainty

- B., Kang, Murthy (2019): https://doi.org/10.1017/jpr.2019.49

Optimal Choice of Uncertainty

- B., Kang, Murthy (2019): https://doi.org/10.1017/jpr.2019.49
- **Optimal choice of δ:** χ/n where χ is (say 95%) quantile of the distribution a certain (sometimes *chi*-squared).

DRO: $\min_{\theta \in \Theta} \max_{P : D_c(P, P_n) \leq \delta} E_P\left[\ell(X, \theta)\right]$



Theorem

$$n \times \mathscr{P}(P_n, \theta_*) \Rightarrow \varphi^*(H, \theta_*)$$

Linear Regression Example and Sqrt-Lasso

Minimizing least squares: $E[\underbrace{(Y - \theta_*^T X)X}_{h(X,\theta_*)}] = 0$

For the example with unit error variance and cov(X) = $\mathbb{I}_d$, $p = 2$,

$$\varphi(\xi, \theta) = \frac{\|\xi\|_2^2}{4(1 + \|\theta_*\|_2^2)}$$

Then $\varphi(\xi, \theta) \sim \dfrac{\chi_d^2}{4(1 + \|\theta_*\|_2^2)}$,

and $\delta = \dfrac{\eta_{1-\alpha}}{n} = \dfrac{O(\log d)}{n}$

What About Central Limit Theorem for DRO Estimator?

- Choose $\delta = \chi / n$ for $c(x, y) = \|x - y\|_p^2$:

What About Central Limit Theorem for DRO Estimator?

- Choose $\delta = \chi/n$ for $c(x, y) = \|x - y\|_p^2$:
- The transportation distance is then $O\left(1/n^{1/2}\right)$ and thus Lagrange multiplier $O\left(n^{1/2}\right)$.

What About Central Limit Theorem for DRO Estimator?

- Choose $\delta = \chi / n$ for $c(x, y) = \|x - y\|_p^2$:
- The transportation distance is then $O\left(1/n^{1/2}\right)$ and thus Lagrange multiplier $O\left(n^{1/2}\right)$.
- Thus: $\Delta = \bar{\Delta}/n^{1/2}$, $\lambda = \bar{\lambda} n^{1/2}$ if $\delta = \chi / n$

$$
\begin{aligned}
& \min_{\theta} \max_{\lambda} \left\{ \frac{\bar{\lambda}\chi}{n^{1/2}} + E_{P_n} \max_{\Delta} \left\{ l\left(X + \frac{\bar{\Delta}}{n^{1/2}}, \theta \right) - \frac{\bar{\lambda}}{n^{1/2}} \left\| \bar{\Delta} \right\|_p^2 \right\} \right\} \\
\approx \quad & \min_{\theta} \left\{ E_{P_n} l(X, \theta) + n^{-1/2} \eta^{1/2} E_{P_n}^{1/2} \left\| D_x l(X, \theta) \right\|_q^2 \right\}
\end{aligned}
$$

What About Central Limit Theorem for DRO Estimator?

- Choose $\delta = \chi/n$ for $c(x, y) = \|x - y\|_p^2$:
- The transportation distance is then $O\left(1/n^{1/2}\right)$ and thus Lagrange multiplier $O\left(n^{1/2}\right)$.
- Thus: $\Delta = \bar{\Delta}/n^{1/2}$, $\lambda = \bar{\lambda} n^{1/2}$ if $\delta = \chi/n$

$$
\begin{aligned}
& \min_{\theta} \max_{\lambda} \left\{ \frac{\bar{\lambda}\chi}{n^{1/2}} + E_{P_n} \max_{\Delta} \left\{ l\left(X + \frac{\bar{\Delta}}{n^{1/2}}, \theta \right) - \frac{\bar{\lambda}}{n^{1/2}} \|\bar{\Delta}\|_p^2 \right\} \right\} \\
\approx \quad & \min_{\theta} \{ E_{P_n} l(X, \theta) + n^{-1/2} \eta^{1/2} E_{P_n}^{1/2} \|D_x l(X, \theta)\|_q^2 \}
\end{aligned}
$$

- So, $\bar{\Delta}_{opt}(X_i)$ is aligned (in l_p) to $D_x l(X, \theta)$ & $\|\Delta_{opt}(X_i)\|_p = \|D_x l(X, \theta)\|_q / (2\lambda)$.

Insights

- So, we get that
$$X_i^* \to X_i + \Delta_{opt}\left(X_i\right)/n^{1/2}.$$

Insights

- So, we get that
$$X_i^* \rightarrow X_i + \Delta_{opt}\left(X_i\right)/n^{1/2}.$$
- From the form
$$\min_{\theta}\left\{E_{P_n} l\left(X,\theta\right) + n^{-1/2} v\left(\theta\right)\right\}$$
it is not difficult to see that if H is the Hessian at θ_* then
$$\theta_n^{DRO} = \theta_n^{ERM} - n^{-1/2} H^{-1} \nabla v\left(\theta_*\right) + o\left(n^{-1/2}\right),$$
where θ_n^{ERM} is the case $\delta = 0$ and
$$v\left(\theta\right) = \eta^{1/2} E_{P_*}^{1/2}\left(\left\| D_x l\left(X,\theta\right)\right\|_q^2\right).$$

Insights

- So, we get that

$$X_i^* \to X_i + \Delta_{opt}\left(X_i\right)/n^{1/2}.$$

- From the form

$$\min_{\theta}\left\{E_{P_n} l\left(X,\theta\right) + n^{-1/2} v\left(\theta\right)\right\}$$

it is not difficult to see that if H is the Hessian at θ_* then

$$\theta_n^{DRO} = \theta_n^{ERM} - n^{-1/2} H^{-1} \nabla v\left(\theta_*\right) + o\left(n^{-1/2}\right),$$

where θ_n^{ERM} is the case $\delta = 0$ and

$$v\left(\theta\right) = \eta^{1/2} E_{P_*}^{1/2}\left(\left\|D_x l\left(X,\theta\right)\right\|_q^2\right).$$

- This reduces the asymptotic normality of θ_n^{DRO} to that of the (standard) θ_n^{ERM}.

Conclusions

- Optimal Transport DRO -> gradient norm regularization.

Conclusions

- Optimal Transport DRO -> gradient norm regularization.
- Recovers many estimators (exactly).

Conclusions

- Optimal Transport DRO -> gradient norm regularization.
- Recovers many estimators (exactly).
- Can use market aware regularization.

Conclusions

- Optimal Transport DRO -> gradient norm regularization.
- Recovers many estimators (exactly).
- Can use market aware regularization.
- Intuitive worst case adversarial structure + CLTs.

Conclusions

- Optimal Transport DRO -> gradient norm regularization.
- Recovers many estimators (exactly).
- Can use market aware regularization.
- Intuitive worst case adversarial structure + CLTs.
- Many references to key questions: algorithms, optimal regularization, Nash equilibrium...