Decision Making and Inference Under Model Misspecification

Jose Blanchet.

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

Goal:

Goals: a) Introduce optimal transport methods popular applications and properties, then b) use these results for robust peformance analysis and finally c) also show how optimal transport applied to statistical estimation.

Agenda

- Day 1: Introduction to Optimal Transport (Primals and Duals)

Agenda

- Day 1: Introduction to Optimal Transport (Primals and Duals)
- Day 2: Distributionally robust performance analysis and optimization.

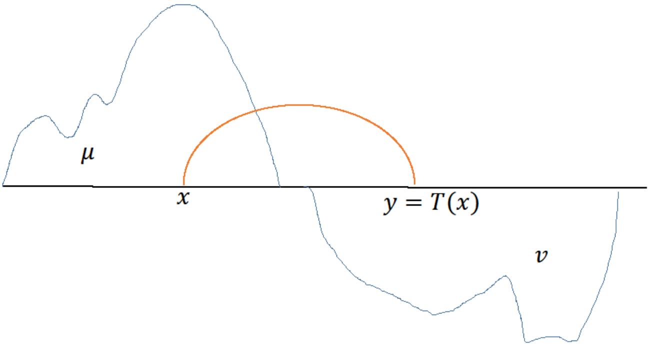
Agenda

- Day 1: Introduction to Optimal Transport (Primals and Duals)
- Day 2: Distributionally robust performance analysis and optimization.
- Day 3: Statistical properties of estimators.

Monge-Kantorovich Problem & Duality
(see e.g. C. Villani's 2008 textbook)

Monge Problem

- What's the cheapest way to transport a pile of sand to cover a sinkhole?



Monge Problem

- What's the cheapest way to transport a pile of sand to cover a sinkhole?

$$\min_{T(\cdot):T(X)\sim v} E_{\mu}\left\{c\left(X, T\left(X\right)\right)\right\},$$

Monge Problem

- What's the cheapest way to transport a pile of sand to cover a sinkhole?

$$\min_{T(\cdot):T(X)\sim v} E_{\mu}\left\{c\left(X,T\left(X\right)\right)\right\},$$

- where $c\left(x,y\right)\geq 0$ is the cost of transporting x to y.

Monge Problem

- What's the cheapest way to transport a pile of sand to cover a sinkhole?

$$\min_{T(\cdot):T(X)\sim v} E_{\mu}\left\{c\left(X,T\left(X\right)\right)\right\},$$

- where $c\left(x,y\right)\geq 0$ is the cost of transporting x to y.
- $T\left(X\right)\sim v$ means $T\left(X\right)$ follows distribution $v\left(\cdot\right)$.

Monge Problem

- What's the cheapest way to transport a pile of sand to cover a sinkhole?

$$\min_{T(\cdot):T(X)\sim v} E_{\mu}\left\{c\left(X, T\left(X\right)\right)\right\},$$

- where $c\left(x, y\right) \geq 0$ is the cost of transporting x to y.
- $T\left(X\right) \sim v$ means $T\left(X\right)$ follows distribution $v\left(\cdot\right)$.
- Problem is highly non-linear, not much progress for about 160 yrs!

Kantorovich Relaxation: Primal Problem

- Let $\Pi(\mu, v)$ be the class of joint distributions π of random variables (X, Y) such that

$$\pi_X = \text{marginal of } X = \mu, \ \pi_Y = \text{marginal of } Y = v.$$

Kantorovich Relaxation: Primal Problem

- Let $\Pi(\mu, v)$ be the class of joint distributions π of random variables (X, Y) such that

$$\pi_X = \text{marginal of } X = \mu, \;\; \pi_Y = \text{marginal of } Y = v.$$

- Solve

$$\min\{E_\pi [c(X, Y)] : \pi \in \Pi(\mu, v)\}$$

Kantorovich Relaxation: Primal Problem

- Let $\Pi(\mu, v)$ be the class of joint distributions π of random variables (X, Y) such that

$$\pi_X = \text{marginal of } X = \mu, \ \pi_Y = \text{marginal of } Y = v.$$

- Solve

$$\min\{E_\pi[c(X, Y)] : \pi \in \Pi(\mu, v)\}$$

- Linear programming (infinite dimensional):

$$\begin{aligned} D_c(\mu, v) \quad : \quad = & \min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \pi(dx, dy) \\ & \int_{\mathcal{Y}} \pi(dx, dy) = \mu(dx), \int_{\mathcal{X}} \pi(dx, dy) = v(dy). \end{aligned}$$

Kantorovich Relaxation: Primal Problem

- Let $\Pi(\mu, v)$ be the class of joint distributions π of random variables (X, Y) such that

$$\pi_X = \text{marginal of } X = \mu, \;\; \pi_Y = \text{marginal of } Y = v.$$

- Solve

$$\min\{E_\pi[c(X, Y)] : \pi \in \Pi(\mu, v)\}$$

- Linear programming (infinite dimensional):

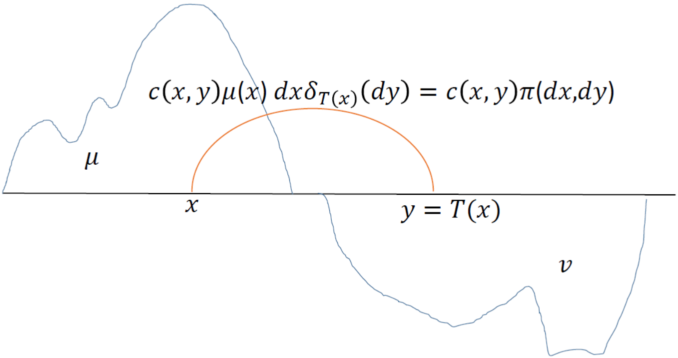
$$\begin{aligned} D_c(\mu, v) \;\; : \;\; &= \min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \pi(dx, dy) \\ & \int_{\mathcal{Y}} \pi(dx, dy) = \mu(dx), \int_{\mathcal{X}} \pi(dx, dy) = v(dy). \end{aligned}$$

- If $c(x, y) = d(x, y)$ (d-metric) then $D_c(\mu, v)$ is a metric <– We'll check this later (this is Wasserstein distance).

Illustration of Optimal Transport Costs

- Monge's solution would take the form

$$\pi^* (dx, dy) = \delta_{\{T(x)\}} (dy) \, \mu (dx) \, .$$



Warm up exercise to practice primal interpretation...

Warm up exercise: Check that $D_c(\cdot)$ is a metric if $c(x,y) = d(x,y)$ where $d(\cdot)$ is a metric.

i) $D_d(\mu, v) = D_d(v, \mu)$

ii) $D_d(\mu, v) \geq 0$ and $D_d(\mu, v) = 0$ if and only if $\mu = v$.

iii) $D_d(\mu, w) \leq D_d(\mu, v) + D_d(v, w)$.

Kantorovich Relaxation: Primal Problem

- Keep in mind primal:

$$D_c(\mu, v) \quad := \quad \min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} d(x,y)\,\pi(dx,dy)$$

$$\int_{\mathcal{Y}} \pi(dx,dy) = \mu(dx)\,,\ \int_{\mathcal{X}} \pi(dx,dy) = v(dy)\,.$$

Kantorovich Relaxation: Primal Problem

- Keep in mind primal:

$$
\begin{aligned}
D_c(\mu, v) \quad : \quad &= \min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} d(x,y)\,\pi(dx,dy) \\
&\int_{\mathcal{Y}} \pi(dx,dy) = \mu(dx), \int_{\mathcal{X}} \pi(dx,dy) = v(dy).
\end{aligned}
$$

- Primal always has a solution (if c is lower semicontinuous) $->$ easy to see if $\mathcal{Y}$ and $\mathcal{X}$ are compact.

Kantorovich Relaxation: Primal Problem

- Keep in mind primal:

$$
\begin{aligned}
D_c(\mu, v) \quad : \quad = \min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} d(x,y)\, \pi(dx,dy) \\
\int_{\mathcal{Y}} \pi(dx,dy) = \mu(dx), \int_{\mathcal{X}} \pi(dx,dy) = v(dy).
\end{aligned}
$$

- Primal always has a solution (if c is lower semicontinuous) $->$ easy to see if $\mathcal{Y}$ and $\mathcal{X}$ are compact.
- If $D_d(\mu, v) = 0$, then $E_{\pi^*}(d(X,Y)) = 0$, then $X = Y$-π^* a.s. so $\mu(A) = \pi(X \in A) = \pi(Y \in A) = v(A)$.

Kantorovich Relaxation: Primal Problem

- Now verify triangle inequality

$$D_d(\mu, w) \leq D_d(\mu, v) + D_d(v, w).$$

Kantorovich Relaxation: Primal Problem

- Now verify triangle inequality

$$D_d(\mu, w) \leq D_d(\mu, v) + D_d(v, w).$$

- Pick X, Y, Z so that $X \sim \mu$, $Y \sim v$ and $Z \sim w$. Sample $Y \sim v$ and then $X|Y = y$ from the optimal coupling solving $D_d(\mu, v)$. Also, sample $Z|Y = y$ using optimal coupling for computing $D_d(v, w)$.

Kantorovich Relaxation: Primal Problem

- Now verify triangle inequality

$$D_d(\mu, w) \le D_d(\mu, v) + D_d(v, w).$$

- Pick X, Y, Z so that $X \sim \mu$, $Y \sim v$ and $Z \sim w$. Sample $Y \sim v$ and then $X|Y = y$ from the optimal coupling solving $D_d(\mu, v)$. Also, sample $Z|Y = y$ using optimal coupling for computing $D_d(v, w)$.
- Previous construction gives a coupling for X and Z, which is not necessarily optimal for computing $D_d(\mu, w)$.

Kantorovich Relaxation: Primal Problem

- Now verify triangle inequality

$$D_d(\mu, w) \leq D_d(\mu, v) + D_d(v, w).$$

- Pick X, Y, Z so that $X \sim \mu$, $Y \sim v$ and $Z \sim w$. Sample $Y \sim v$ and then $X|Y = y$ from the optimal coupling solving $D_d(\mu, v)$. Also, sample $Z|Y = y$ using optimal coupling for computing $D_d(v, w)$.
- Previous construction gives a coupling for X and Z, which is not necessarily optimal for computing $D_d(\mu, w)$.
- On the other hand, $d(X, Z) \leq d(X, Y) + d(Y, Z)$ because $d(\cdot)$ is a metric.

Kantorovich Relaxation: Primal Problem

- Now verify triangle inequality

$$D_d(\mu, w) \leq D_d(\mu, v) + D_d(v, w).$$

- Pick X, Y, Z so that $X \sim \mu$, $Y \sim v$ and $Z \sim w$. Sample $Y \sim v$ and then $X|Y = y$ from the optimal coupling solving $D_d(\mu, v)$. Also, sample $Z|Y = y$ using optimal coupling for computing $D_d(v, w)$.
- Previous construction gives a coupling for X and Z, which is not necessarily optimal for computing $D_d(\mu, w)$.
- On the other hand, $d(X, Z) \leq d(X, Y) + d(Y, Z)$ because $d(\cdot)$ is a metric.
- Thus $D_d(\mu, w) \leq E(d(X, Z)) \leq D_d(\mu, v) + D_d(v, w)$.

Towards the Dual Problem

It is always natural to study the dual of a linear programming problem...

Kantorovich Relaxation: Dual Problem

- Primal:

$$\min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} d\left(x,y\right)\pi\left(dx,dy\right)$$

$$\int_{\mathcal{Y}} \pi\left(dx,dy\right) = \mu\left(dx\right), \int_{\mathcal{X}} \pi\left(dx,dy\right) = v\left(dy\right).$$

Kantorovich Relaxation: Dual Problem

- Primal:

$$\min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} d\left(x,y\right)\pi\left(dx,dy\right)$$

$$\int_{\mathcal{Y}} \pi\left(dx,dy\right) = \mu\left(dx\right), \int_{\mathcal{X}} \pi\left(dx,dy\right) = v\left(dy\right).$$

- Dual:

$$\sup_{\alpha,\beta} \int_{\mathcal{X}} \alpha\left(x\right)\mu\left(dx\right) + \int_{\mathcal{Y}} \beta\left(y\right)v\left(dy\right)$$

$$\alpha\left(x\right) + \beta\left(y\right) \leq c\left(x,y\right) \ \ \forall\left(x,y\right) \in \mathcal{X}\times\mathcal{Y}.$$

Kantorovich Relaxation: Dual Problem

- Primal:

$$\min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} d\left(x,y\right)\pi\left(dx,dy\right)$$

$$\int_{\mathcal{Y}} \pi\left(dx,dy\right) = \mu\left(dx\right), \int_{\mathcal{X}} \pi\left(dx,dy\right) = v\left(dy\right).$$

- Dual:

$$\sup_{\alpha,\beta} \int_{\mathcal{X}} \alpha\left(x\right)\mu\left(dx\right) + \int_{\mathcal{Y}} \beta\left(y\right) v\left(dy\right)$$

$$\alpha\left(x\right) + \beta\left(y\right) \leq c\left(x,y\right) \ \ \forall\left(x,y\right) \in \mathcal{X}\times\mathcal{Y}.$$

- Here α and β can be taken continuous

Kantorovich Relaxation: Primal Interpretation

- Martin wants to remove of a pile of sand, $\mu(\cdot)$.

Kantorovich Relaxation: Primal Interpretation

- Martin wants to remove of a pile of sand, $\mu(\cdot)$.
- Henry wants to cover a sinkhole, $v(\cdot)$.

Kantorovich Relaxation: Primal Interpretation

- Martin wants to remove of a pile of sand, $\mu(\cdot)$.
- Henry wants to cover a sinkhole, $v(\cdot)$.
- Cost for Martin and Henry to transport the sand to cover the sinkhole is
$$D_c(\mu, v) = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \pi^*(dx, dy).$$

Kantorovich Relaxation: Primal Interpretation

- Martin wants to remove of a pile of sand, $\mu(\cdot)$.
- Henry wants to cover a sinkhole, $v(\cdot)$.
- Cost for Martin and Henry to transport the sand to cover the sinkhole is
$$D_c(\mu, v) = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \pi^*(dx, dy).$$
- Now comes Victoria, who has a business...

Kantorovich Relaxation: Primal Interpretation

- Martin wants to remove of a pile of sand, $\mu(\cdot)$.
- Henry wants to cover a sinkhole, $v(\cdot)$.
- Cost for Martin and Henry to transport the sand to cover the sinkhole is

$$D_c(\mu, v) = \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \pi^*(dx, dy).$$

- Now comes Victoria, who has a business...
- Vicky promises to transport on behalf of Martin and Henry the whole amount.

Kantorovich Relaxation: Primal Interpretation

- Vicky charges John $\alpha(x)$ per-unit of mass at x (similarly to Peter, $\beta(y)$).

Kantorovich Relaxation: Primal Interpretation

- Vicky charges John $\alpha(x)$ per-unit of mass at x (similarly to Peter, $\beta(y)$).
- For Peter and John to agree we must have

$$\alpha(x) + \beta(y) \leq c(x, y).$$

Kantorovich Relaxation: Primal Interpretation

- Vicky charges John $\alpha(x)$ per-unit of mass at x (similarly to Peter, $\beta(y)$).
- For Peter and John to agree we must have

$$\alpha(x) + \beta(y) \leq c(x, y).$$

- Vicky wishes to maximize her profit

$$\int \alpha(x)\,\mu(dx) + \int \beta(y)\,v(dy).$$

Kantorovich Relaxation: Primal Interpretation

- Vicky charges John $\alpha(x)$ per-unit of mass at x (similarly to Peter, $\beta(y)$).
- For Peter and John to agree we must have

$$\alpha(x) + \beta(y) \leq c(x, y).$$

- Vicky wishes to maximize her profit

$$\int \alpha(x) \mu(dx) + \int \beta(y) v(dy).$$

- Kantorovich duality says primal and dual optimal values coincide and

$$\alpha^*(x) + \beta^*(y) = c(x, y) \text{ - } \pi^* \text{ a.s. } <\!- \text{ complementary slackness}$$

Kantorovich Relaxation: Primal Interpretation

- Vicky charges John $\alpha(x)$ per-unit of mass at x (similarly to Peter, $\beta(y)$).
- For Peter and John to agree we must have

$$\alpha(x) + \beta(y) \leq c(x, y).$$

- Vicky wishes to maximize her profit

$$\int \alpha(x)\,\mu(dx) + \int \beta(y)\,v(dy).$$

- Kantorovich duality says primal and dual optimal values coincide and

$$\alpha^*(x) + \beta^*(y) = c(x, y) \text{ - } \pi^* \text{ a.s. } <\!\!- \text{ complementary slackness}$$

- Existence of dual optimizers: $c(x, y) \leq a(x) + b(y)$ so $E_\mu a(X) < \infty$, $E_\mu b(Y) < \infty$.

Proof Technique: Sketch of Strong Duality

- Suppose $\mathcal{X}$ and $\mathcal{Y}$ compact

$$
\begin{aligned}
&\inf_{\pi \geq 0} \sup_{\alpha,\beta} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c\left(x,y\right) \pi\left(dx,dy\right) \right. \\
&- \int_{\mathcal{X}\times\mathcal{Y}} \alpha\left(x\right) \pi\left(dx,dy\right) + \int_{\mathcal{X}} \alpha\left(x\right) \mu\left(dx\right) \\
&- \int_{\mathcal{X}\times\mathcal{Y}} \beta\left(y\right) \pi\left(dx,dy\right) + \int_{\mathcal{Y}} \beta\left(y\right) v\left(dy\right) \}
\end{aligned}
$$

Proof Technique: Sketch of Strong Duality

- Suppose $\mathcal{X}$ and $\mathcal{Y}$ compact

$$\inf_{\pi \geq 0} \sup_{\alpha,\beta} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c\left(x,y\right)\pi\left(dx,dy\right) \right.$$
$$- \int_{\mathcal{X}\times\mathcal{Y}} \alpha\left(x\right)\pi\left(dx,dy\right) + \int_{\mathcal{X}} \alpha\left(x\right)\mu\left(dx\right)$$
$$\left. - \int_{\mathcal{X}\times\mathcal{Y}} \beta\left(y\right)\pi\left(dx,dy\right) + \int_{\mathcal{Y}} \beta\left(y\right)v\left(dy\right) \right\}$$

- Swap sup and inf using **Sion's min-max theorem** by a compactness argument and conclude.

Proof Technique: Sketch of Strong Duality

- Suppose $\mathcal{X}$ and $\mathcal{Y}$ compact

$$
\begin{aligned}
&\inf_{\pi \geq 0} \sup_{\alpha,\beta} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c\left(x,y\right) \pi\left(dx,dy\right) \right. \\
&- \int_{\mathcal{X}\times\mathcal{Y}} \alpha\left(x\right) \pi\left(dx,dy\right) + \int_{\mathcal{X}} \alpha\left(x\right) \mu\left(dx\right) \\
&- \int_{\mathcal{X}\times\mathcal{Y}} \beta\left(y\right) \pi\left(dx,dy\right) + \int_{\mathcal{Y}} \beta\left(y\right) v\left(dy\right) \}
\end{aligned}
$$

- Swap sup and inf using **Sion's min-max theorem** by a compactness argument and conclude.
- *Some amount of work to extend to general Polish spaces.*

Economic Interpretations & Some Closed Form Solutions
(see e.g. A. Galichon's 2016 textbook & McCann 2013 notes).

Applications in Labor Markets

- Worker with skill x & company with technology y yield $\Psi(x, y)$ surplus.

Applications in Labor Markets

- Worker with skill x & company with technology y yield $\Psi(x, y)$ surplus.
- The population of workers is given by $\mu(x)$.

Applications in Labor Markets

- Worker with skill x & company with technology y yield $\Psi(x, y)$ surplus.
- The population of workers is given by $\mu(x)$.
- The population of companies is given by $v(y)$.

Applications in Labor Markets

- Worker with skill x & company with technology y yield $\Psi(x, y)$ surplus.
- The population of workers is given by $\mu(x)$.
- The population of companies is given by $v(y)$.
- The salary of worker x is $\alpha(x)$ & cost of technology y is $\beta(y)$

$$\alpha(x) + \beta(y) \geq \Psi(x, y).$$

Applications in Labor Markets

- Worker with skill x & company with technology y yield $\Psi(x, y)$ surplus.
- The population of workers is given by $\mu(x)$.
- The population of companies is given by $v(y)$.
- The salary of worker x is $\alpha(x)$ & cost of technology y is $\beta(y)$

$$\alpha(x) + \beta(y) \geq \Psi(x, y).$$

- Companies want to *minimize* total production cost

$$\int \alpha(x)\mu(x)\,dx + \int \beta(y)\,v(y)\,dy$$

Applications in Labor Markets

- Letting a central planner organize the Labor market.

Applications in Labor Markets

- Letting a central planner organize the Labor market.
- The planner wishes to maximize total surplus

$$\int \Psi(x, y)\, \pi(dx, dy)$$

Applications in Labor Markets

- Letting a central planner organize the Labor market.
- The planner wishes to maximize total surplus

$$\int \Psi(x, y)\, \pi(dx, dy)$$

- Over assignments $\pi(\cdot)$ which satisfy market clearing

$$\int_{\mathcal{Y}} \pi(dx, dy) = \mu(dx), \ \int_{\mathcal{X}} \pi(dx, dy) = v(dy).$$

Solving for Optimal Transport Coupling

- **Suppose that** $\Psi(x, y) = xy$, $\mu(x) = I(x \in [0, 1])$, $v(y) = e^{-y} I(y > 0)$.

Solving for Optimal Transport Coupling

- **Suppose that** $\Psi(x, y) = xy$**,** $\mu(x) = I(x \in [0, 1])$**,** $v(y) = e^{-y} I(y > 0)$**.**
- Solve primal by sampling: Let $\{X_i^n\}_{i=1}^n$ and $\{Y_i^n\}_{i=1}^n$ both i.i.d. from μ and v, respectively.

$$F_{\mu_n}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i^n \leq x), \ F_{v_n}(y) = \frac{1}{n} \sum_{j=1}^{n} I\left(Y_j^n \leq y\right)$$

Solving for Optimal Transport Coupling

- **Suppose that** $\Psi(x, y) = xy$, $\mu(x) = I(x \in [0, 1])$, $v(y) = e^{-y} I(y > 0)$.
- Solve primal by sampling: Let $\{X_i^n\}_{i=1}^n$ and $\{Y_i^n\}_{i=1}^n$ both i.i.d. from μ and v, respectively.

$$F_{\mu_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i^n \leq x), \; F_{v_n}(y) = \frac{1}{n}\sum_{j=1}^{n} I\left(Y_j^n \leq y\right)$$

- Consider

$$\max_{\pi\left(x_i^n, x_j^n\right)\geq 0} \sum_{i,j} \Psi\left(x_i^n, y_j^n\right)\pi\left(x_i^n, y_j^n\right)$$

$$\sum_j \pi\left(x_i^n, y_j^n\right) = \frac{1}{n}\;\forall x_i, \quad \sum_i \pi\left(x_i^n, y_j^n\right) = \frac{1}{n}\;\forall y_j.$$

Solving for Optimal Transport Coupling

- **Suppose that** $\Psi(x,y) = xy$**,** $\mu(x) = I(x \in [0,1])$**,** $v(y) = e^{-y} I(y > 0)$**.**
- Solve primal by sampling: Let $\{X_i^n\}_{i=1}^n$ and $\{Y_i^n\}_{i=1}^n$ both i.i.d. from μ and v, respectively.

$$F_{\mu_n}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i^n \leq x), \; F_{v_n}(y) = \frac{1}{n} \sum_{j=1}^{n} I\left(Y_j^n \leq y\right)$$

- Consider

$$\max_{\pi\left(x_i^n, x_j^n\right) \geq 0} \sum_{i,j} \Psi\left(x_i^n, y_j^n\right) \pi\left(x_i^n, y_j^n\right)$$

$$\sum_j \pi\left(x_i^n, y_j^n\right) = \frac{1}{n} \; \forall x_i, \quad \sum_i \pi\left(x_i^n, y_j^n\right) = \frac{1}{n} \; \forall y_j.$$

- **Clearly, simply sort and match is the solution!**

Solving for Optimal Transport Coupling

- Think of $Y_j^n = -\log\left(1 - U_j^n\right) = F_v^{-1}\left(U_j^n\right)$ for U_j^ns i.i.d. uniform$(0,1)$.

Solving for Optimal Transport Coupling

- Think of $Y_j^n = -\log\left(1 - U_j^n\right) = F_v^{-1}\left(U_j^n\right)$ for U_j^ns i.i.d. uniform$(0,1)$.
- The j-th order statistic $X_{(j)}^n$ is matched to $Y_{(j)}^n$.

Solving for Optimal Transport Coupling

- Think of $Y_j^n = -\log\left(1 - U_j^n\right) = F_v^{-1}\left(U_j^n\right)$ for U_j^ns i.i.d. uniform$(0,1)$.
- The j-th order statistic $X_{(j)}^n$ is matched to $Y_{(j)}^n$.
- As $n \to \infty$, $X_{(nt)}^n \to t$, so $Y_{(nt)}^n \to -\log(1-t)$.

Solving for Optimal Transport Coupling

- Think of $Y_j^n = -\log\left(1 - U_j^n\right) = F_v^{-1}\left(U_j^n\right)$ for U_j^ns i.i.d. uniform$(0,1)$.
- The j-th order statistic $X_{(j)}^n$ is matched to $Y_{(j)}^n$.
- As $n \to \infty$, $X_{(nt)}^n \to t$, so $Y_{(nt)}^n \to -\log(1-t)$.
- Thus, the optimal coupling as $n \to \infty$ is $X = U$ and $Y = -\log(1-U)$ (comonotonic coupling).

Solving for Optimal Transport Coupling

- Think of $Y_j^n = -\log\left(1 - U_j^n\right) = F_v^{-1}\left(U_j^n\right)$ for U_j^ns i.i.d. uniform$(0, 1)$.
- The j-th order statistic $X_{(j)}^n$ is matched to $Y_{(j)}^n$.
- As $n \to \infty$, $X_{(nt)}^n \to t$, so $Y_{(nt)}^n \to -\log(1 - t)$.
- Thus, the optimal coupling as $n \to \infty$ is $X = U$ and $Y = -\log(1 - U)$ (comonotonic coupling).
- In general, the optimal coupling is $X = F_\mu^{-1}(U)$ and $Y = F_v^{-1}(U)$.

Identities for Wasserstein Distances

- Comonotonic coupling is the solution if $\partial^2_{x,y}\Psi\left(x,y\right)\geq 0$ - supermodularity:

$$\Psi\left(x\vee x',y\vee y'\right)+\Psi\left(x\wedge x',y\wedge y'\right)\geq\Psi\left(x,y\right)+\Psi\left(x',y'\right)$$

- Comonotonic coupling is the solution if $\partial^2_{x,y}\Psi\left(x,y\right)\geq 0$ - supermodularity:

$$\Psi\left(x\vee x',y\vee y'\right)+\Psi\left(x\wedge x',y\wedge y'\right)\geq\Psi\left(x,y\right)+\Psi\left(x',y'\right)$$

- Or, for costs $c\left(x,y\right)=-\Psi\left(x,y\right)$, if $\partial^2_{x,y}c\left(x,y\right)\leq 0$ (submodularity).

Identities for Wasserstein Distances

- Comonotonic coupling is the solution if $\partial^2_{x,y}\Psi\left(x,y\right) \geq 0$ - supermodularity:

$$\Psi\left(x \vee x', y \vee y'\right) + \Psi\left(x \wedge x', y \wedge y'\right) \geq \Psi\left(x,y\right) + \Psi\left(x', y'\right)$$

- Or, for costs $c\left(x,y\right) = -\Psi\left(x,y\right)$, if $\partial^2_{x,y} c\left(x,y\right) \leq 0$ (submodularity).
- **Corollary:** Suppose $c\left(x,y\right) = \left|x-y\right|$ then $X = F_\mu^{-1}\left(U\right)$ and $Y = F_v^{-1}\left(U\right)$ thus

$$\begin{aligned} D_c\left(F_\mu, F_v\right) &= \int_0^1 \left|F_\mu^{-1}\left(u\right) - F_v^{-1}\left(u\right)\right| du \\ &= \int_{-\infty}^{\infty} \left|F_\mu\left(x\right) - F_v\left(x\right)\right| dx. \end{aligned}$$

Identities for Wasserstein Distances

- Comonotonic coupling is the solution if $\partial^2_{x,y}\Psi(x,y) \geq 0$ - supermodularity:

$$\Psi\left(x \vee x', y \vee y'\right) + \Psi\left(x \wedge x', y \wedge y'\right) \geq \Psi(x,y) + \Psi\left(x', y'\right)$$

- Or, for costs $c(x,y) = -\Psi(x,y)$, if $\partial^2_{x,y} c(x,y) \leq 0$ (submodularity).
- **Corollary:** Suppose $c(x,y) = |x-y|$ then $X = F_\mu^{-1}(U)$ and $Y = F_v^{-1}(U)$ thus

$$\begin{aligned} D_c\left(F_\mu, F_v\right) &= \int_0^1 \left|F_\mu^{-1}(u) - F_v^{-1}(u)\right| du \\ &= \int_{-\infty}^{\infty} \left|F_\mu(x) - F_v(x)\right| dx. \end{aligned}$$

- Similar identities are common for Wasserstein distances...

Interesting Insight on Salary Effects

- In equilibrium, by the envelope theorem

$$\dot{\beta}^*(y) = \frac{d}{dy}\sup_x\left[\Psi(x,y)-\alpha^*(x)\right] = \frac{\partial}{\partial y}\Psi(x_y,y) = x_y$$

$$\dot{\alpha}^*(x) = \frac{\partial}{\partial x}\Psi(x,y_x) = y_x = F_v^{-1}\left(F_\mu(x)\right).$$

Interesting Insight on Salary Effects

- In equilibrium, by the envelope theorem

$$\begin{aligned}\dot{\beta}^{*}(y) &= \frac{d}{dy}\sup_{x}\left[\Psi(x,y)-\alpha^{*}(x)\right]=\frac{\partial}{\partial y}\Psi(x_y,y)=x_y\\ \dot{\alpha}^{*}(x) &= \frac{\partial}{\partial x}\Psi(x,y_x)=y_x=F_{v}^{-1}\left(F_{\mu}(x)\right).\end{aligned}$$

- We also know that $y=-\log(1-x)$, or $x=1-\exp(-y)$

$$\begin{aligned}\beta^{*}(y) &= y+\exp(-y)-1+\beta^{*}(0).\\ \alpha^{*}(x)+\beta^{*}(-\log(1-x)) &= xy.\end{aligned}$$

Interesting Insight on Salary Effects

- In equilibrium, by the envelope theorem

$$\begin{aligned}\dot{\beta}^*(y) &= \frac{d}{dy}\sup_x\left[\Psi(x,y)-\alpha^*(x)\right] = \frac{\partial}{\partial y}\Psi(x_y,y) = x_y \\ \dot{\alpha}^*(x) &= \frac{\partial}{\partial x}\Psi(x,y_x) = y_x = F_v^{-1}\left(F_\mu(x)\right).\end{aligned}$$

- We also know that $y = -\log(1-x)$, or $x = 1-\exp(-y)$

$$\begin{aligned}\beta^*(y) &= y+\exp(-y)-1+\beta^*(0). \\ \alpha^*(x)+\beta^*(-\log(1-x)) &= xy.\end{aligned}$$

- What if $\Psi(x,y) \to \Psi(x,y)+f(x)$? (i.e. productivity changes).

Interesting Insight on Salary Effects

- In equilibrium, by the envelope theorem

$$
\begin{aligned}
\dot{\beta}^*(y) &= \frac{d}{dy}\sup_x\left[\Psi(x,y) - \alpha^*(x)\right] = \frac{\partial}{\partial y}\Psi(x_y, y) = x_y \\
\dot{\alpha}^*(x) &= \frac{\partial}{\partial x}\Psi(x, y_x) = y_x = F_v^{-1}\left(F_\mu(x)\right).
\end{aligned}
$$

- We also know that $y = -\log(1-x)$, or $x = 1 - \exp(-y)$

$$
\begin{aligned}
\beta^*(y) &= y + \exp(-y) - 1 + \beta^*(0). \\
\alpha^*(x) + \beta^*(-\log(1-x)) &= xy.
\end{aligned}
$$

- What if $\Psi(x,y) \to \Psi(x,y) + f(x)$? (i.e. productivity changes).
- *Answer: salaries increase if* $f(\cdot)$ is increasing.

Back to Wasserstein Distances

Additional properties of Optimal Transport Solutions: Kantorovich-Rubinstein Duality and Wasserstein GAN.

Back to Wasserstein Distances

- Consider the case $c(x, y) = d(x, y)$.

Back to Wasserstein Distances

- Consider the case $c(x,y) = d(x,y)$.
- Recall dual

$$\begin{aligned}&\max E_{\mu}\alpha(X) - E_{v}\beta(Y)\\ &s.t.\ \alpha(x) - \beta(y) \leq d(x,y)\ \ \forall\ \ x,y \in \mathcal{S}\ .\end{aligned}$$

Back to Wasserstein Distances

- Consider the case $c(x,y) = d(x,y)$.
- Recall dual

$$\begin{aligned} &\max E_{\mu}\alpha(X) - E_{v}\beta(Y) \\ &s.t.\ \alpha(x) - \beta(y) \leq d(x,y)\ \forall\ \ x,y \in \mathcal{S}\ . \end{aligned}$$

- Note that given β, we should pick

$$\alpha(x) = \beta^{d}(x) := \inf_{y}\{\beta(y) + d(x,y)\},$$

similarly once $\alpha(\cdot)$ is chosen, we could improve by picking

$$\beta^{dd}(y) = \sup_{x}\{\beta^{d}(x) - d(x,y)\}.$$

Transforms are Lipschitz

- Moreover, observe that $\beta^{d}\left(\cdot\right)$ is 1-Lipschitz

$$
\begin{aligned}
\beta^{d}\left(x\right) &= \inf_{y}\left\{\beta\left(y\right)+d\left(x,y\right)\right\} \text{ <- recall def} \\
\beta^{d}\left(x\right)-\beta^{d}\left(x'\right) &= \beta\left(y_{x}\right)+d\left(x,y_{x}\right) \\
&\quad -\beta\left(y_{x'}\right)-d\left(x,y_{x'}\right) \\
&\leq d\left(x,y_{x'}\right)-d\left(x,y_{x'}\right)\leq d\left(x,x'\right).
\end{aligned}
$$

Transforms are Lipschitz

- Moreover, observe that $\beta^d(\cdot)$ is 1-Lipschitz

$$
\begin{aligned}
\beta^d(x) &= \inf_y \{\beta(y) + d(x,y)\} \text{ <- recall def} \\
\beta^d(x) - \beta^d(x') &= \beta(y_x) + d(x, y_x) \\
&\quad -\beta(y_{x'}) - d(x, y_{x'}) \\
&\leq d(x, y_{x'}) - d(x, y_{x'}) \leq d(x, x').
\end{aligned}
$$

- Same argument is true for $\beta^{dd}(y)$.

The Transform of a Lipschitz Function is the Function Itself

- Moreover,

$$\beta^d(x) := \inf_y \{\beta(y) + d(x,y)\} \leq \beta(x)$$

and if β is 1-Lipschitz (meaning $|\beta(x) - \beta(y)| \leq d(x,y)$) then

$$\begin{aligned} \beta^d(x) - \beta(x) &= \inf_y \{d(x,y) + \beta(y) - \beta(x)\} \\ &\geq \inf_y \{d(x,y) - d(x,y)\} = 0. \end{aligned}$$

The Transform of a Lipschitz Function is the Function Itself

- Moreover,

$$\beta^{d}\left(x\right) := \inf_{y}\left\{\beta\left(y\right) + d\left(x,y\right)\right\} \leq \beta\left(x\right)$$

and if β is 1-Lipschitz (meaning $\left|\beta\left(x\right) - \beta\left(y\right)\right| \leq d\left(x,y\right)$) then

$$\begin{aligned}\beta^{d}\left(x\right) - \beta\left(x\right) &= \inf_{y}\left\{d\left(x,y\right) + \beta\left(y\right) - \beta\left(x\right)\right\} \\ &\geq \inf_{y}\left\{d\left(x,y\right) - d\left(x,y\right)\right\} = 0.\end{aligned}$$

- Consequently, if β is 1-Lipschitz $\beta = \beta^{d}$... So, the dual can be simplified.

Back to Wasserstein Distances

- Original Dual:

$$\max E_{\mu}\alpha\left(X\right)-E_{v}\beta\left(Y\right)$$
$$s.t.\ \alpha\left(x\right)-\beta\left(y\right)\leq d\left(x,y\right)\ \forall\ \ x,y\in\mathcal{S}\ .$$

Back to Wasserstein Distances

- Original Dual:

$$\max E_{\mu}\alpha\left(X\right)-E_{v}\beta\left(Y\right)$$
$$s.t.\ \alpha\left(x\right)-\beta\left(y\right)\leq d\left(x,y\right)\ \forall\ \ x,y\in\mathcal{S}\ .$$

- Simplified Dual (called Kantorovich duality result):

$$\max E_{\mu}\alpha\left(X\right)-E_{v}\alpha\left(Y\right)$$
$$s.t.\ \alpha \text{ is 1-Lipschitz }.$$

Back to Wasserstein Distances

- Original Dual:

$$\max E_{\mu}\alpha(X) - E_{v}\beta(Y)$$
$$s.t.\ \alpha(x) - \beta(y) \leq d(x,y)\ \ \forall\ \ x, y \in \mathcal{S}\ .$$

- Simplified Dual (called Kantorovich duality result):

$$\max E_{\mu}\alpha(X) - E_{v}\alpha(Y)$$
$$s.t.\ \alpha \text{ is 1-Lipschitz}\ .$$

- This is the basis for so-called Wasserstein GAN (Generative Adversarial Networks) – popular in artificial intelligence.

- Have you even thought about how to generate a "face" at random? (https://github.com/hindupuravinash/the-gan-zoo).

A Quick Discussion on Wasserstein GAN

- What's the formulation

$$\min_{\theta < \text{NN parameter}} D_d \left(v_\theta, \mu_n \right),$$

where μ_n represents the empirical measure of images.

A Quick Discussion on Wasserstein GAN

- What's the formulation

$$\min_{\theta < \text{NN parameter}} D_d \left(v_\theta, \mu_n \right),$$

where μ_n represents the empirical measure of images.

- $v_\theta \left(\cdot \right)$ is a probability measure generated by a Neural Network (NN), from initial random noise

A Quick Discussion on Wasserstein GAN

- What's the formulation

$$\min_{\theta < \text{NN parameter}} D_d\left(v_\theta, \mu_n\right),$$

  where μ_n represents the empirical measure of images.

- $v_\theta\left(\cdot\right)$ is a probability measure generated by a Neural Network (NN), from initial random noise
- θ represents the parameter of the network.

A Quick Discussion on Wasserstein GAN

- What's the formulation

$$\min_{\theta < \text{NN parameter}} D_d\left(v_\theta, \mu_n\right),$$

where μ_n represents the empirical measure of images.

- $v_\theta\left(\cdot\right)$ is a probability measure generated by a Neural Network (NN), from initial random noise
- θ represents the parameter of the network.
- By duality

$$\min_{\theta < \text{NN parameter}} \sup_{\alpha\text{-1-Lip}} \left\{ E_{v_\theta}\left(\alpha\left(X\right)\right) - E_{\mu_n}\left(\alpha\left(Y\right)\right) \right\}.$$

A Quick Discussion on Wasserstein GAN

- What's the formulation

$$\min_{\theta < \text{NN parameter}} D_d\left(v_\theta, \mu_n\right),$$

  where μ_n represents the empirical measure of images.
- $v_\theta\left(\cdot\right)$ is a probability measure generated by a Neural Network (NN), from initial random noise
- θ represents the parameter of the network.
- By duality

$$\min_{\theta < \text{NN parameter}} \sup_{\alpha\text{-1-Lip}} \left\{E_{v_\theta}\left(\alpha\left(X\right)\right) - E_{\mu_n}\left(\alpha\left(Y\right)\right)\right\}.$$

- Use another Neural Network to parameterize α (i.e. a 1-Lip function).

A Quick Discussion on Wasserstein GAN

- What's the formulation

$$\min_{\theta < \text{NN parameter}} D_d\left(v_\theta, \mu_n\right),$$

  where μ_n represents the empirical measure of images.
- $v_\theta\left(\cdot\right)$ is a probability measure generated by a Neural Network (NN), from initial random noise
- θ represents the parameter of the network.
- By duality

$$\min_{\theta < \text{NN parameter}} \sup_{\alpha\text{-1-Lip}} \left\{E_{v_\theta}\left(\alpha\left(X\right)\right) - E_{\mu_n}\left(\alpha\left(Y\right)\right)\right\}.$$

- Use another Neural Network to parameterize α (i.e. a 1-Lip function).
- Apply automatic differentiation to compute gradients & run stochastic gradient descent.

Optimal Transport with Quadratic Costs

- The case $c(x, y) = \|x - y\|_2^2 / 2$ is important because of its intuitive appeal and its theoretical properties.

Optimal Transport with Quadratic Costs

- The case $c(x, y) = \|x - y\|_2^2 / 2$ is important because of its intuitive appeal and its theoretical properties.
- We consider

$$D_c(\mu, v) = \min_{\pi}\{2^{-1} E_{\pi} \|X - Y\|_2^2 : \pi_X = \mu \text{ and } \pi_Y = v\}.$$

Optimal Transport with Quadratic Costs

- The case $c(x, y) = \|x - y\|_2^2 / 2$ is important because of its intuitive appeal and its theoretical properties.
- We consider

$$D_c(\mu, v) = \min_{\pi} \{2^{-1} E_{\pi} \|X - Y\|_2^2 : \pi_X = \mu \text{ and } \pi_Y = v\}.$$

- We assume that $E\|X\|_2^2 + E\|Y\|_2^2 < \infty$.

Optimal Transport with Quadratic Costs

- The case $c(x,y) = \|x-y\|_2^2/2$ is important because of its intuitive appeal and its theoretical properties.
- We consider

$$D_c(\mu, v) = \min_{\pi}\{2^{-1}E_\pi \|X - Y\|_2^2 : \pi_X = \mu \text{ and } \pi_Y = v\}.$$

- We assume that $E\|X\|_2^2 + E\|Y\|_2^2 < \infty$.
- So, the problem is equivalent to

$$\max_{\pi}\{E_\pi\left(X^T Y\right) : \pi_X = \mu \text{ and } \pi_Y = v\}.$$

Optimal Transport with Quadratic Costs

- The case $c\left(x,y\right)=\left\|x-y\right\|_2^2/2$ is important because of its intuitive appeal and its theoretical properties.
- We consider

$$D_c\left(\mu,v\right)=\min_{\pi}\{2^{-1}E_{\pi}\left\|X-Y\right\|_2^2:\pi_X=\mu \text{ and } \pi_Y=v\}.$$

- We assume that $E\left\|X\right\|_2^2+E\left\|Y\right\|_2^2<\infty$.
- So, the problem is equivalent to

$$\max_{\pi}\{E_{\pi}\left(X^TY\right):\pi_X=\mu \text{ and } \pi_Y=v\}.$$

- The dual is

$$\min\{E_{\mu}\alpha\left(X\right)+E_{v}\beta\left(Y\right):\alpha\left(x\right)+\beta\left(y\right)\geq x^Ty \text{ for } x,y\in S\}.$$

Optimal Transport with Quadratic Costs

- The dual is

$$\min\{E_{\mu}\alpha(X) + E_{v}\beta(Y) : \alpha(x) + \beta(y) \geq x^{T}y \text{ for } x, y \in S\}.$$

Optimal Transport with Quadratic Costs

- The dual is

$$\min\{E_\mu \alpha(X) + E_v \beta(Y) : \alpha(x) + \beta(y) \geq x^T y \text{ for } x, y \in S\}.$$

- Note now that given $\alpha(x)$ we improve the objective function choosing

$$\alpha^*(y) = \sup_x [x^T y - \alpha(x)],$$

which is convex.

Optimal Transport with Quadratic Costs

- The dual is

$$\min\{E_{\mu}\alpha(X)+E_{v}\beta(Y):\alpha(x)+\beta(y)\geq x^{T}y \text{ for } x,y\in S\}.$$

- Note now that given $\alpha(x)$ we improve the objective function choosing

$$\alpha^{*}(y)=\sup_{x}[x^{T}y-\alpha(x)],$$

which is convex.

- So, in the end the dual is simplified to

$$\min\{E_{\mu}\alpha(X)+E_{v}\alpha^{*}(Y):\alpha \text{ convex}\}.$$

Optimal Transport with Quadratic Costs

- Now, our goal is to characterize the optimal solution of the primal and dual problems.

Optimal Transport with Quadratic Costs

- Now, our goal is to characterize the optimal solution of the primal and dual problems.
- Suppose that μ has a density with respect to the Lebesgue measure.

Optimal Transport with Quadratic Costs

- Now, our goal is to characterize the optimal solution of the primal and dual problems.
- Suppose that μ has a density with respect to the Lebesgue measure.
- By complementary slackness

$$\alpha(x) + \alpha^*(y) = x^T y \quad \text{-}\pi^* \text{ a.s.}$$

Optimal Transport with Quadratic Costs

- Now, our goal is to characterize the optimal solution of the primal and dual problems.
- Suppose that μ has a density with respect to the Lebesgue measure.
- By complementary slackness

$$\alpha(x) + \alpha^*(y) = x^T y \ \text{-}\pi^* \text{ a.s.}$$

- But given x, equality holds if and only if $y \in \partial a(x)$ <– subdifferential (by convex analysis).

Optimal Transport with Quadratic Costs

- Now, our goal is to characterize the optimal solution of the primal and dual problems.
- Suppose that μ has a density with respect to the Lebesgue measure.
- By complementary slackness

$$\alpha(x) + \alpha^*(y) = x^T y \ \text{-}\pi^* \text{ a.s.}$$

- But given x, equality holds if and only if $y \in \partial a(x) <-$ subdifferential (by convex analysis).
- Similarly, given y, if and only if $x \in \partial \alpha^*(y)$.

Optimal Transport with Quadratic Costs

- Now, our goal is to characterize the optimal solution of the primal and dual problems.
- Suppose that μ has a density with respect to the Lebesgue measure.
- By complementary slackness

$$\alpha(x) + \alpha^*(y) = x^T y \ \ \text{-}\pi^* \text{ a.s.}$$

- But given x, equality holds if and only if $y \in \partial a(x)$ <– subdifferential (by convex analysis).
- Similarly, given y, if and only if $x \in \partial\alpha^*(y)$.
- But by Rademacher's theorem $\alpha(\cdot)$ is differentiable almost everywhere. So, given $X \sim \mu$, $Y = \nabla\alpha(X)$.

Optimal Transport with Quadratic Costs

- Consequently, this establishes Brennier's Theorem: If $c(x, y) = \|x - y\|_2^2 / 2$ then the optimal coupling

$$(X, Y) = (X, \nabla \alpha (X)),$$

where $\alpha(\cdot)$ is convex.

Optimal Transport with Quadratic Costs

- Consequently, this establishes Brennier's Theorem: If $c(x, y) = \|x - y\|_2^2 / 2$ then the optimal coupling

$$(X, Y) = (X, \nabla\alpha(X)),$$

where $\alpha(\cdot)$ is convex.

- The optimal $\nabla\alpha(\cdot)$ is unique almost surely: Suppose $\nabla\bar{\alpha}$ is another solution to the dual.

Optimal Transport with Quadratic Costs

- Consequently, this establishes Brennier's Theorem: If $c(x, y) = \|x - y\|_2^2 / 2$ then the optimal coupling

$$(X, Y) = (X, \nabla \alpha (X)),$$

  where $\alpha(\cdot)$ is convex.
- The optimal $\nabla \alpha (\cdot)$ is unique almost surely: Suppose $\nabla \bar{\alpha}$ is another solution to the dual.
- Then consider the couplings $(X, \nabla \alpha (X))$ and $(X, \nabla \bar{\alpha} (X))$ we have that for almost every x

$$\alpha (x) + \alpha^* (\nabla \bar{\alpha} (x)) = x^T \nabla \bar{\alpha} (x)$$

  (by complementary slackness).

Optimal Transport with Quadratic Costs

- Consequently, this establishes Brennier's Theorem: If $c(x, y) = \|x - y\|_2^2 / 2$ then the optimal coupling

$$(X, Y) = (X, \nabla\alpha(X)),$$

where $\alpha(\cdot)$ is convex.

- The optimal $\nabla\alpha(\cdot)$ is unique almost surely: Suppose $\nabla\bar{\alpha}$ is another solution to the dual.
- Then consider the couplings $(X, \nabla\alpha(X))$ and $(X, \nabla\bar{\alpha}(X))$ we have that for almost every x

$$\alpha(x) + \alpha^*(\nabla\bar{\alpha}(x)) = x^T \nabla\bar{\alpha}(x)$$

(by complementary slackness).

- Therefore $\nabla\bar{\alpha}(x) \in \partial\alpha(x)$ and by Rademacher $\nabla\bar{\alpha} = \nabla\alpha$ almost surely.

Optimal Transport with Quadratic Costs

- Example: Suppose that $X \sim N(0, I)$ and $Y \sim N(0, \Sigma)$ we want to transport X into Y optimally using the cost $c(x, y) = \|x - y\|_2^2 / 2$.

Optimal Transport with Quadratic Costs

- Example: Suppose that $X \sim N(0, I)$ and $Y \sim N(0, \Sigma)$ we want to transport X into Y optimally using the cost $c(x, y) = \|x - y\|_2^2 / 2$.
- We postulate that $\nabla \alpha(x) = Ax$ where A is positive definite.

Optimal Transport with Quadratic Costs

- Example: Suppose that $X \sim N(0, I)$ and $Y \sim N(0, \Sigma)$ we want to transport X into Y optimally using the cost $c(x, y) = \|x - y\|_2^2 / 2$.
- We postulate that $\nabla \alpha(x) = Ax$ where A is positive definite.
- **So, we must have that $A \cdot A = \Sigma$, the solution is that A is the polar factorization of Σ.**

Optimal Transport with Quadratic Costs

- Example: Suppose that $X \sim N(0, I)$ and $Y \sim N(0, \Sigma)$ we want to transport X into Y optimally using the cost $c(x, y) = \|x - y\|_2^2 / 2$.
- We postulate that $\nabla \alpha(x) = Ax$ where A is positive definite.
- **So, we must have that $A \cdot A = \Sigma$, the solution is that A is the polar factorization of Σ.**
- *From here it is easy to derive what the general optimal transport map is between two Gaussians (try this as an exercise).*

Illustration of Optimal Transport in Image Analysis

- Santambrogio (2010)'s illustration

The discussion is based on
B. & Murthy (2016)
https://arxiv.org/abs/1604.01446.
https://pubsonline.informs.org/doi/abs/10.1287/moor.2018.0936?journalCod

A Distributionally Robust Performance Analysis

- We are often interested in

$$E_{P_{true}}\left(f\left(X\right)\right)$$

for a complex model P_{true}.

A Distributionally Robust Performance Analysis

- We are often interested in

$$E_{P_{true}}\left(f\left(X\right)\right)$$

for a complex model P_{true}.

- Moreover, we wish to optimize, namely

$$\min_{\theta} E_{P_{true}}\left(h\left(X,\theta\right)\right).$$

A Distributionally Robust Performance Analysis

- We are often interested in

$$E_{P_{true}}\left(f\left(X\right)\right)$$

for a complex model P_{true}.

- Moreover, we wish to optimize, namely

$$\min_{\theta} E_{P_{true}}\left(h\left(X,\theta\right)\right).$$

- Model P_{true} might be unknown or too difficult to work with.

A Distributionally Robust Performance Analysis

- We are often interested in

$$E_{P_{true}}\left(f\left(X\right)\right)$$

for a complex model P_{true}.

- Moreover, we wish to optimize, namely

$$\min_{\theta} E_{P_{true}}\left(h\left(X,\theta\right)\right).$$

- Model P_{true} might be unknown or too difficult to work with.
- So, we introduce a proxy P_0 which provides a good trade-off between tractability and model fidelity (e.g. Brownian motion for random walk approximations).

A Distributionally Robust Performance Analysis

- For $f(\cdot)$ upper semicontinuous with $E_{P_0}|f(X)| < \infty$

$$\begin{aligned} &\sup E_P(f(Y)) \\ &D_c(P, P_0) \leq \delta\ , \end{aligned}$$

X takes values on a Polish space and $c(\cdot)$ is lower semi-continuous.

A Distributionally Robust Performance Analysis

- For $f(\cdot)$ upper semicontinuous with $E_{P_0}|f(X)| < \infty$

$$\begin{aligned} &\sup E_P(f(Y)) \\ &D_c(P, P_0) \leq \delta, \end{aligned}$$

X takes values on a Polish space and $c(\cdot)$ is lower semi-continuous.

- Also an infinite dimensional linear program

$$\begin{aligned} &\sup \int_{\mathcal{X}\times\mathcal{Y}} f(y)\,\pi(dx, dy) \\ &s.t. \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\,\pi(dx, dy) \leq \delta \\ &\int_{\mathcal{Y}} \pi(dx, dy) = P_0(dx). \end{aligned}$$

A Distributionally Robust Performance Analysis

- Formal duality:

$$
\begin{aligned}
Dual &= \inf_{\lambda \geq 0, \alpha} \left\{ \lambda\delta + \int \alpha(x) P_0(dx) \right\} \\
&\quad \lambda c(x, y) + \alpha(x) \geq f(y) .
\end{aligned}
$$

A Distributionally Robust Performance Analysis

- Formal duality:

$$Dual = \inf_{\lambda \geq 0, \alpha} \left\{ \lambda\delta + \int \alpha(x) P_0(dx) \right\}$$
$$\lambda c(x,y) + \alpha(x) \geq f(y).$$

- B. & Murthy (2016) - *No duality gap*:

$$Dual = \inf_{\lambda \geq 0} \left[\lambda\delta + E_0 \left(\sup_y \{ f(y) - \lambda c(X,y) \} \right) \right].$$

A Distributionally Robust Performance Analysis

- Formal duality:

$$Dual = \inf_{\lambda \geq 0, \alpha} \left\{ \lambda\delta + \int \alpha(x) P_0(dx) \right\}$$
$$\lambda c(x, y) + \alpha(x) \geq f(y) .$$

- B. & Murthy (2016) - *No duality gap*:

$$Dual = \inf_{\lambda \geq 0} \left[\lambda\delta + E_0 \left(\sup_y \{ f(y) - \lambda c(X, y) \} \right) \right].$$

- *We refer to this as RoPA Duality in this talk.*

A Distributionally Robust Performance Analysis

- Formal duality:

$$Dual = \inf_{\lambda \geq 0, \alpha} \left\{ \lambda\delta + \int \alpha(x) P_0(dx) \right\}$$
$$\lambda c(x, y) + \alpha(x) \geq f(y) \; .$$

- B. & Murthy (2016) - *No duality gap*:

$$Dual = \inf_{\lambda \geq 0} \left[\lambda\delta + E_0 \left(\sup_y \{ f(y) - \lambda c(X, y) \} \right) \right].$$

- *We refer to this as RoPA Duality in this talk.*
- Let us consider an important case first: $f(y) = I(y \in A)$ & $c(x, x) = 0$.

A Distributionally Robust Performance Analysis

- So, if $f(y) = I(y \in A)$ and $c_A(X) = \inf\{y \in A : c(x, y)\}$, then

$$Dual = \inf_{\lambda \geq 0} \left[\lambda\delta + E_0 \left(1 - \lambda c_A(X)\right)^+ \right] = P_0 \left(c_A(X) \leq 1/\lambda_*\right).$$

- So, if $f(y) = I(y \in A)$ and $c_A(X) = \inf\{y \in A : c(x, y)\}$, then

$$Dual = \inf_{\lambda \geq 0} \left[\lambda\delta + E_0 \left(1 - \lambda c_A(X)\right)^{+} \right] = P_0 \left(c_A(X) \leq 1/\lambda_*\right).$$

- If $c_A(X)$ is continuous under P_0 & $E_0(c_A(X)) \geq \delta$, then

$$\delta = E_0 \left[c_A(X) \, I\left(c_A(X) \leq 1/\lambda_*\right) \right].$$

Example: Model Uncertainty in Bankruptcy Calculations

- $R(t)$ = the reserve (perhaps multiple lines) at time t.

Example: Model Uncertainty in Bankruptcy Calculations

- $R(t)$ = the reserve (perhaps multiple lines) at time t.
- Bankruptcy probability (in finite time horizon T)

$$u_T = P_{true}\left(R(t) \in B \text{ for some } t \in [0, T]\right).$$

Example: Model Uncertainty in Bankruptcy Calculations

- $R(t)$ = the reserve (perhaps multiple lines) at time t.
- Bankruptcy probability (in finite time horizon T)

$$u_T = P_{true}\left(R(t) \in B \text{ for some } t \in [0, T]\right).$$

- B is a set which models bankruptcy.

Example: Model Uncertainty in Bankruptcy Calculations

- $R(t) =$ the reserve (perhaps multiple lines) at time t.
- Bankruptcy probability (in finite time horizon T)

$$u_T = P_{true}\left(R(t) \in B \text{ for some } t \in [0, T]\right).$$

- B is a set which models bankruptcy.
- **Problem:** Model (P_{true}) may be complex, intractable or simply unknown...

A Distributionally Robust Risk Analysis Formulation

- **Our solution:** Estimate u_T by solving

$$\sup_{D_c(P_0,P)\leq\delta} P_{true}\left(R\left(t\right) \in B \text{ for some } t \in \left[0, T\right]\right),$$

where P_0 is a *suitable* model.

A Distributionally Robust Risk Analysis Formulation

- **Our solution:** Estimate u_T by solving

$$\sup_{D_c(P_0,P)\leq\delta} P_{true}\left(R\left(t\right) \in B \text{ for some } t \in \left[0, T\right]\right),$$

where P_0 is a *suitable* model.

- P_0 = proxy for P_{true}.

A Distributionally Robust Risk Analysis Formulation

- **Our solution:** Estimate u_T by solving

$$\sup_{D_c(P_0,P)\leq\delta} P_{true}\left(R\left(t\right) \in B \text{ for some } t \in \left[0, T\right]\right),$$

where P_0 is a *suitable* model.

- P_0 = proxy for P_{true}.
- P_0 right trade-off between fidelity and tractability.

A Distributionally Robust Risk Analysis Formulation

- **Our solution:** Estimate u_T by solving

$$\sup_{D_c(P_0,P)\leq\delta} P_{true}\left(R\left(t\right)\in B \text{ for some } t\in\left[0,T\right]\right),$$

  where P_0 is a *suitable* model.
- P_0 = proxy for P_{true}.
- P_0 right trade-off between fidelity and tractability.
- δ is the distributional uncertainty size.

A Distributionally Robust Risk Analysis Formulation

- **Our solution:** Estimate u_T by solving

$$\sup_{D_c(P_0,P)\leq\delta} P_{true}\left(R\left(t\right)\in B \text{ for some } t\in\left[0,T\right]\right),$$

  where P_0 is a *suitable* model.
- P_0 = proxy for P_{true}.
- P_0 right trade-off between fidelity and tractability.
- δ is the distributional uncertainty size.
- $D_c\left(\cdot\right)$ is the distributional uncertainty region.

Desirable Elements of Distributionally Robust Formulation

- Would like $D_c(\cdot)$ to have wide flexibility (even non-parametric).

Desirable Elements of Distributionally Robust Formulation

- Would like $D_c(\cdot)$ to have wide flexibility (even non-parametric).
- Want optimization to be tractable.

Desirable Elements of Distributionally Robust Formulation

- Would like $D_c(\cdot)$ to have wide flexibility (even non-parametric).
- Want optimization to be tractable.
- *Want to preserve advantages of using P_0.*

Desirable Elements of Distributionally Robust Formulation

- Would like $D_c(\cdot)$ to have wide flexibility (even non-parametric).
- Want optimization to be tractable.
- *Want to preserve advantages of using P_0.*
- Want a way to estimate δ.

- Standard choices based on divergence (such as Kullback-Leibler) - Hansen & Sargent (2016)

$$D\left(v||\mu\right) = E_v\left(\log\left(\frac{dv}{d\mu}\right)\right).$$

Connections to Distributionally Robust Optimization

- Standard choices based on divergence (such as Kullback-Leibler) - Hansen & Sargent (2016)

$$D\left(v||\mu\right) = E_v\left(\log\left(\frac{dv}{d\mu}\right)\right).$$

- Robust Optimization: Ben-Tal, El Ghaoui, Nemirovski (2009).

Connections to Distributionally Robust Optimization

- Standard choices based on divergence (such as Kullback-Leibler) - Hansen & Sargent (2016)

$$D\left(v||\mu\right) = E_v\left(\log\left(\frac{dv}{d\mu}\right)\right).$$

- Robust Optimization: Ben-Tal, El Ghaoui, Nemirovski (2009).
- **Big problem: Absolute continuity may typically be violated...**

Connections to Distributionally Robust Optimization

- Standard choices based on divergence (such as Kullback-Leibler) - Hansen & Sargent (2016)

$$D\left(v||\mu\right) = E_v\left(\log\left(\frac{dv}{d\mu}\right)\right).$$

- Robust Optimization: Ben-Tal, El Ghaoui, Nemirovski (2009).
- **Big problem: Absolute continuity may typically be violated...**
- Think of using Brownian motion as a proxy model for $R\left(t\right)$...

Connections to Distributionally Robust Optimization

- Standard choices based on divergence (such as Kullback-Leibler) - Hansen & Sargent (2016)

$$D\left(v||\mu\right) = E_v\left(\log\left(\frac{dv}{d\mu}\right)\right).$$

- Robust Optimization: Ben-Tal, El Ghaoui, Nemirovski (2009).
- **Big problem: Absolute continuity may typically be violated...**
- Think of using Brownian motion as a proxy model for $R\left(t\right)$...
- **Optimal transport is a natural option!**

Application 1: Back to Classical Risk Problem

- Suppose that

$$
\begin{aligned}
c(x,y) &= d_J(x(\cdot), y(\cdot)) = \text{Skorokhod } J_1 \text{ metric.} \\
&= \inf_{\phi(\cdot) \text{ bijection}} \Big\{ \sup_{t \in [0,1]} |x(t) - y(\phi(t))|, \sup_{t \in [0,1]} |\phi(t) - t| \Big\}.
\end{aligned}
$$

Application 1: Back to Classical Risk Problem

- Suppose that

$$\begin{aligned} c(x,y) &= d_J(x(\cdot), y(\cdot)) = \text{Skorokhod } J_1 \text{ metric.} \\ &= \inf_{\phi(\cdot) \text{ bijection}} \{ \sup_{t\in[0,1]} |x(t) - y(\phi(t))|, \sup_{t\in[0,1]} |\phi(t) - t| \}. \end{aligned}$$

- If $R(t) = b - Z(t)$, then ruin during time interval $[0,1]$ is

$$B_b = \{R(\cdot) : 0 \geq \inf_{t\in[0,1]} R(t)\} = \{Z(\cdot) : b \leq \sup_{t\in[0,1]} Z(t)\}.$$

Application 1: Back to Classical Risk Problem

- Suppose that

$$
\begin{aligned}
c(x,y) &= d_J(x(\cdot), y(\cdot)) = \text{Skorokhod } J_1 \text{ metric.} \\
&= \inf_{\phi(\cdot) \text{ bijection}} \left\{ \sup_{t\in[0,1]} |x(t) - y(\phi(t))|, \sup_{t\in[0,1]} |\phi(t) - t| \right\}.
\end{aligned}
$$

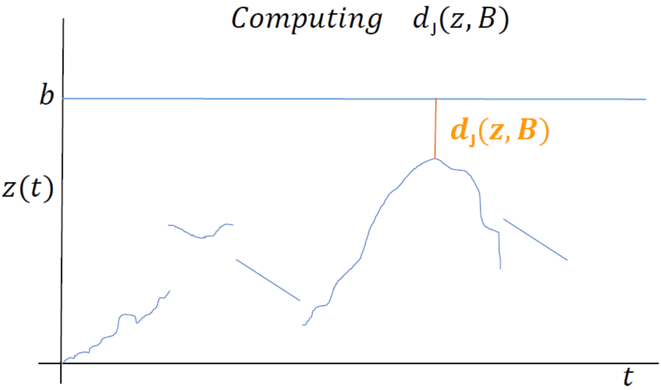
- If $R(t) = b - Z(t)$, then ruin during time interval $[0,1]$ is

$$
B_b = \{R(\cdot) : 0 \geq \inf_{t\in[0,1]} R(t)\} = \{Z(\cdot) : b \leq \sup_{t\in[0,1]} Z(t)\}.
$$

- **Let $P_0(\cdot)$ be the Wiener measure want to compute**

$$
\sup_{D_c(P_0,P)\leq\delta} P(Z \in B_b).
$$

Application 1: Computing Distance to Bankruptcy



- **So:** $\{c_{B_b}(Z) \leq 1/\lambda_*\} = \{\sup_{t\in[0,1]} Z(t) \geq b - 1/\lambda^*\}$, and

$$\sup_{D_c(P_0,P)\leq\delta} P(Z \in B_b) = P_0\left(\sup_{t\in[0,1]} Z(t) \geq b - 1/\lambda^*\right).$$

Application 1: Computing Uncertainty Size

- Note **any coupling** π so that $\pi_X = P_0$ and $\pi_Y = P$ satisfies

$$D_c\left(P_0, P\right) \leq E_\pi\left[c\left(X, Y\right)\right] \approx \delta.$$

Application 1: Computing Uncertainty Size

- Note **any coupling** π so that $\pi_X = P_0$ and $\pi_Y = P$ satisfies

$$D_c\left(P_0, P\right) \leq E_\pi\left[c\left(X, Y\right)\right] \approx \delta.$$

- So use any coupling between *evidence* and P_0 or expert knowledge.

Application 1: Computing Uncertainty Size

- Note **any coupling** π so that $\pi_X = P_0$ and $\pi_Y = P$ satisfies

$$D_c\left(P_0, P\right) \leq E_\pi\left[c\left(X, Y\right)\right] \approx \delta.$$

- So use any coupling between *evidence* and P_0 or expert knowledge.
- We discuss choosing δ non-parametrically momentarily.

Application 1: Illustration of Coupling

- Given arrivals and claim sizes let $Z(t) = m_2^{-1/2} \sum_{k=1}^{N(t)} (X_k - m_1)$

Algorithm 1 To embed the process $(Z(t) : t \geq 0)$ in Brownian motion $(B(t) : t \geq 0)$
Given: Brownian motion $B(t)$, moment m_1 and independent realizations of claim sizes $X_1, X_2, \ldots$

Initialize $\tau_0 := 0$ and $\Psi_0 := 0$. For $j \geq 1$, recursively define,

$$\tau_{j+1} := \inf\left\{ s \geq \tau_j : \sup_{\tau_j \leq r \leq s} B_r - B_s = X_{j+1} \right\}, \text{ and } \Psi_j := \Psi_{j-1} + X_j.$$

Define the auxiliary processes

$$\tilde{S}(t) := \sum_{j>0} \sup_{\tau_j \leq s \leq t} B(s) \mathbf{1}(\tau_j \leq t < \tau_{j+1}) \text{ and } \tilde{N}(t) := \sum_{j \geq 0} \Psi_j \mathbf{1}(\tau_j \leq t < \tau_{j+1}).$$

Let $A(t) := \tilde{N}(t) + \tilde{S}(t)$, and identify the time change $\sigma(t) := \inf\{s : A(s) = m_1 t\}$. Next, take the time changed version $Z(t) := \tilde{S}(\sigma(t))$.

Replace $Z(t)$ by $-Z(t)$ and $B(t)$ by $-B(t)$.

Application 1: Illustration of Coupling

- Given arrivals and claim sizes let $Z(t) = m_2^{-1/2} \sum_{k=1}^{N(t)} (X_k - m_1)$

Algorithm 1 To embed the process $(Z(t) : t \geq 0)$ in Brownian motion $(B(t) : t \geq 0)$
Given: Brownian motion $B(t)$, moment m_1 and independent realizations of claim sizes $X_1, X_2, \ldots$

Initialize $\tau_0 := 0$ and $\Psi_0 := 0$. For $j \geq 1$, recursively define,

$$\tau_{j+1} := \inf\left\{ s \geq \tau_j : \sup_{\tau_j \leq r \leq s} B_r - B_s = X_{j+1} \right\}, \text{ and } \Psi_j := \Psi_{j-1} + X_j.$$

Define the auxiliary processes

$$\tilde{S}(t) := \sum_{j>0} \sup_{\tau_j \leq s \leq t} B(s) \mathbf{1}(\tau_j \leq t < \tau_{j+1}) \text{ and } \tilde{N}(t) := \sum_{j \geq 0} \Psi_j \mathbf{1}(\tau_j \leq t < \tau_{j+1}).$$

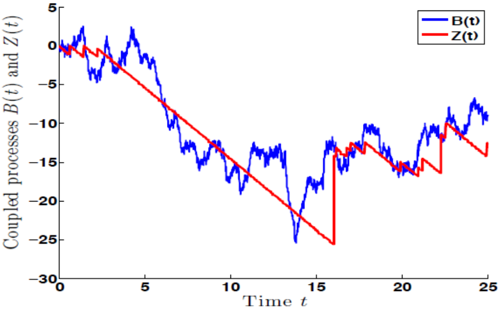
Let $A(t) := \tilde{N}(t) + \tilde{S}(t)$, and identify the time change $\sigma(t) := \inf\{s : A(s) = m_1 t\}$. Next, take the time changed version $Z(t) := \tilde{S}(\sigma(t))$.

Replace $Z(t)$ by $-Z(t)$ and $B(t)$ by $-B(t)$.

- See also Fomivoch, Gonzalez-Cazares, Ivanovs (2021).

Application 1: Coupling in Action

FIGURE 4. A coupled path output by Algorithm 1



Application 1: Numerical Example

- Assume Poisson arrivals.
- *Pareto claim sizes with index* **2.2** – $(P(V > t) = 1/(1+t)^{2.2})$.
- Cost $c(x, y) = d_J(x, y)^2$ <– note power of 2.
- Used Algorithm 1 to calibrate (estimating means and variances from data).

b	$\frac{P_0(\text{Ruin})}{P_{true}(\text{Ruin})}$	$\frac{P^*_{robust}(\text{Ruin})}{P_{true}(\text{Ruin})}$
100	1.07×10^{-1}	12.28
150	2.52×10^{-4}	10.65
200	5.35×10^{-8}	10.80
250	1.15×10^{-12}	10.98

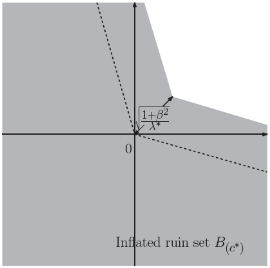
- See also Birghila, Aigner, Engelke (2021)

Additional Applications: Multidimensional Ruin Problems

- **https://arxiv.org/abs/1604.01446** contains more applications.

Additional Applications: Multidimensional Ruin Problems

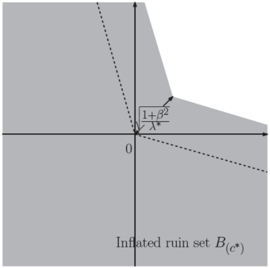
- **https://arxiv.org/abs/1604.01446** contains more applications.
- Control: $\min_\theta \sup_{P:D(P,P_0)\leq\delta} E[L(\theta, Z)]$ <– robust optimal reinsurance.



(b)Computation of worst-case ruin using the baseline measure

Additional Applications: Multidimensional Ruin Problems

- **https://arxiv.org/abs/1604.01446** contains more applications.
- Control: $\min_\theta \sup_{P:D(P,P_0)\leq\delta} E[L(\theta, Z)]$ <– robust optimal reinsurance.

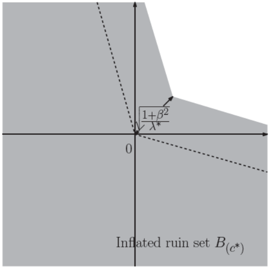


(b)Computation of worst-case ruin using the baseline measure

- Multidimensional risk processes (explicit evaluation of $c_B(x)$ for d_J metric).

Additional Applications: Multidimensional Ruin Problems

- **https://arxiv.org/abs/1604.01446** contains more applications.
- Control: $\min_\theta \sup_{P:D(P,P_0)\leq\delta} E[L(\theta, Z)]$ <– robust optimal reinsurance.

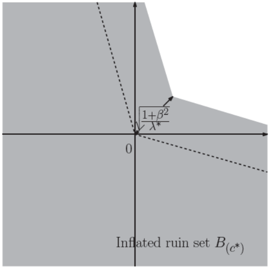


(b)Computation of worst-case ruin using the baseline measure

- Multidimensional risk processes (explicit evaluation of $c_B(x)$ for d_J metric).
- **Key insight: Geometry of target set often remains largely the same!**

Additional Applications: Multidimensional Ruin Problems

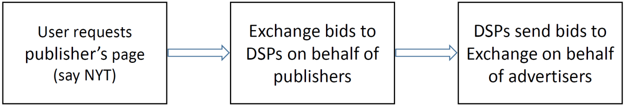
- **https://arxiv.org/abs/1604.01446** contains more applications.
- Control: $\min_\theta \sup_{P:D(P,P_0)\leq\delta} E[L(\theta, Z)]$ <– robust optimal reinsurance.



(b)Computation of worst-case ruin using the baseline measure

- Multidimensional risk processes (explicit evaluation of $c_B(x)$ for d_J metric).
- **Key insight: Geometry of target set often remains largely the same!**
- See also Engelke and Ivanovs (2017).

Background: (Very) Simplified version of Demand Side Platforms (DSPs)



Goal of DSP: Maximize revenue on behalf of advertisers

A Bit of Background on Online Advertising

- Until recently, most exchanges operated using second price auctions.

- Until recently, most exchanges operated using second price auctions.
- The optimal bidding policy in second price auctions is to bid truthfully.

A Bit of Background on Online Advertising

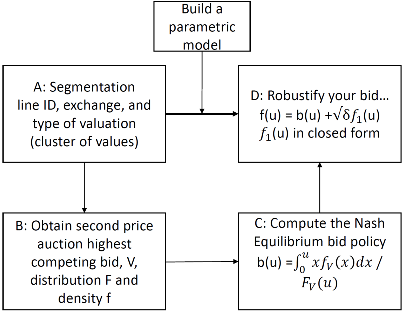
- Until recently, most exchanges operated using second price auctions.
- The optimal bidding policy in second price auctions is to bid truthfully.
- Now, first price auction exchanges have become popular.

A Bit of Background on Online Advertising

- Until recently, most exchanges operated using second price auctions.
- The optimal bidding policy in second price auctions is to bid truthfully.
- Now, first price auction exchanges have become popular.
- How to transfer information from second-price exchanges into first-price exchanges?

Summary of blue print

A –> B –> C –> D



Notations

- $U_i =$ (dlls/1000) value of the item in auction i *if we win*. We write $U_i = u_i$ when value is given.
- $b_i =$ (dlls/1000) is what we bid in the i-th auction (cost in 1st price auction).
- $V_i =$ (dlls/1000) is the highest competing bid in the i-th auction.
- $f_{V_i} =$ the probability density function of V_i.
- $F_{V_i} =$ the cumulative distribution function of V_i.

Model and Performance Measure

- A Simplified Model:

$$\max_{\{b_1,\ldots,b_n\}} \frac{1}{n} \sum_{i=1}^{n} \left(u_i - b_i\right) P\left(V_i \le b_i \middle| U_i = u_i\right),$$

where n is the number of auctions in a given time period, for instance, a day.

Model and Performance Measure

- A Simplified Model:

$$\max_{\{b_1,\ldots,b_n\}} \frac{1}{n}\sum_{i=1}^{n}(u_i - b_i)\, P\left(V_i \le b_i \middle| U_i = u_i\right),$$

  where n is the number of auctions in a given time period, for instance, a day.
- Assume auctions are split according to segments, such as line and exchange, to induce homogeneity.

Model and Performance Measure

- A Simplified Model:

$$\max_{\{b_1,\ldots,b_n\}} \frac{1}{n}\sum_{i=1}^{n}(u_i - b_i)\, P\left(V_i \le b_i \middle| U_i = u_i\right),$$

  where n is the number of auctions in a given time period, for instance, a day.
- Assume auctions are split according to segments, such as line and exchange, to induce homogeneity.
- *Homogeneity*: For each $i \neq j$

$$P\left(V_i \le b \middle| U_i = u\right) = P\left(V_j \le b \middle| U_j = u\right).$$

Model and Performance Measure

- A Simplified Model:

$$\max_{\{b_1,\ldots,b_n\}} \frac{1}{n}\sum_{i=1}^{n}(u_i - b_i)\,P\left(V_i \leq b_i \middle| U_i = u_i\right),$$

  where n is the number of auctions in a given time period, for instance, a day.

- Assume auctions are split according to segments, such as line and exchange, to induce homogeneity.
- *Homogeneity*: For each $i \neq j$

$$P\left(V_i \leq b \middle| U_i = u\right) = P\left(V_j \leq b \middle| U_j = u\right).$$

- Under homogeneity it suffices to solve

$$\max_{b}\,(u - b)\,P\left(V \leq b \middle| U = u\right).$$

Model and Performance Measure

- A Simplified Model:

$$\max_{\{b_1,\ldots,b_n\}} \frac{1}{n}\sum_{i=1}^{n}(u_i - b_i)\, P\left(V_i \le b_i \middle| U_i = u_i\right),$$

  where n is the number of auctions in a given time period, for instance, a day.
- Assume auctions are split according to segments, such as line and exchange, to induce homogeneity.
- *Homogeneity*: For each $i \neq j$

$$P\left(V_i \le b \middle| U_i = u\right) = P\left(V_j \le b \middle| U_j = u\right).$$

- Under homogeneity it suffices to solve

$$\max_b (u - b)\, P\left(V \le b \middle| U = u\right).$$

- *Also assume conditional independence.*

Dealing with Dependence

- Setting the derivative with respect to b equal to zero yields

$$b = u - F_{V|U=u}\left(b\right) / f_{V|U=u}\left(b\right).$$

Dealing with Dependence

- Setting the derivative with respect to b equal to zero yields

$$b = u - F_{V|U=u}(b) / f_{V|U=u}(b) .$$

- *Challenge:* The quantity

$$F_{V|U=u}(\cdot) \quad \text{and} \quad f_{V|U=u}(\cdot)$$

are virtually impossible to estimate in a first price auction setting.

Dealing with Dependence

- Setting the derivative with respect to b equal to zero yields

$$b = u - F_{V|U=u}(b) / f_{V|U=u}(b).$$

- *Challenge:* The quantity

$$F_{V|U=u}(\cdot) \quad \text{and} \quad f_{V|U=u}(\cdot)$$

  are virtually impossible to estimate in a first price auction setting.
- *Virtually ONLY solution:* Assume that V and U are conditionally independent given some other observable factor Θ.

Dealing with Dependence

- Setting the derivative with respect to b equal to zero yields

$$b = u - F_{V|U=u}(b) / f_{V|U=u}(b).$$

- *Challenge:* The quantity

$$F_{V|U=u}(\cdot) \quad \text{and} \quad f_{V|U=u}(\cdot)$$

  are virtually impossible to estimate in a first price auction setting.
- *Virtually ONLY solution:* Assume that V and U are conditionally independent given some other observable factor Θ.
- For example: Θ is a value type (i.e. $\Theta = k \Leftrightarrow U \in \mathcal{A}_k$) = segmentation across values (there are only a few segments).

Dealing with Dependence

- Setting the derivative with respect to b equal to zero yields

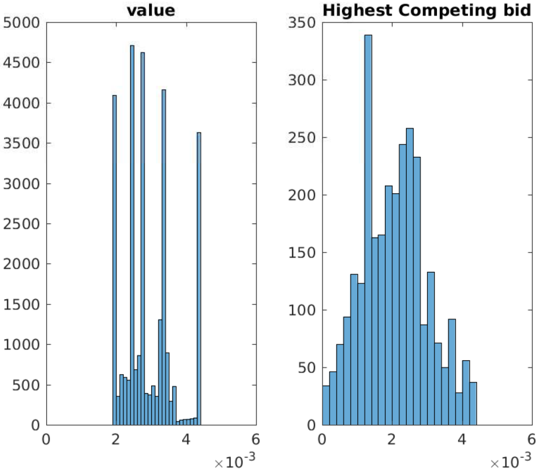
$$b = u - F_{V|U=u}(b) \, / \, f_{V|U=u}(b) \, .$$

- *Challenge:* The quantity

$$F_{V|U=u}(\cdot) \quad \text{and} \quad f_{V|U=u}(\cdot)$$

  are virtually impossible to estimate in a first price auction setting.
- *Virtually ONLY solution:* Assume that V and U are conditionally independent given some other observable factor Θ.
- For example: Θ is a value type (i.e. $\Theta = k \Leftrightarrow U \in \mathcal{A}_k) =$ segmentation across values (there are only a few segments).
- We go back to this in part II)...

Inducing Homogeneity and Conditional Independence



Quantifying Model Mispecifications

- Even if two exchanges run under second price auctions, their competitive landscapes may be different.

Quantifying Model Mispecifications

- Even if two exchanges run under second price auctions, their competitive landscapes may be different.
- So, if $\bar{V}$ is taken from exchange X, we need to recognize the possibility of model error.

Quantifying Model Mispecifications

- Even if two exchanges run under second price auctions, their competitive landscapes may be different.
- So, if $\bar{V}$ is taken from exchange X, we need to recognize the possibility of model error.
- We do this by introducing a metric to compare CDFs, say F and G

$$D\left(F, G\right) = \int_{-\infty}^{\infty} \left|F\left(x\right) - G\left(x\right)\right| dx.$$

Quantifying Model Mispecifications

- Even if two exchanges run under second price auctions, their competitive landscapes may be different.
- So, if $\bar{V}$ is taken from exchange X, we need to recognize the possibility of model error.
- We do this by introducing a metric to compare CDFs, say F and G

$$D(F,G) = \int_{-\infty}^{\infty} |F(x) - G(x)|\, dx.$$

- It turns out that

$$\begin{aligned} D(F,G) &= \min\{E(|X-Y|) \text{ over all joint distributions} \\ &\quad \text{such that } X \text{ has CDF } F \text{ and } Y \text{ has CDF } G\}. \end{aligned}$$

Quantifying Model Mispecifications

- We now want
$$\max_{b} \min_{D(F,F_V)\leq\delta} (u-b)\,F(b)\,.$$

- We now want
$$\max_b \min_{D(F,F_V)\leq\delta} (u-b)\, F(b).$$
- If we write $\bar{F}(x) = 1 - F(x) = P(V > x)$, then the inner minimization is equivalent to
$$\max_{D(F,F_V)\leq\delta} \bar{F}(b) = \max_{D(F,F_V)\leq\delta} P_F(V > b) = P_F(V > b - \lambda_b).$$

Quantifying Model Mispecifications

- We now want
$$\max_{b} \min_{D(F,F_V)\leq\delta} (u-b)\, F(b)\,.$$
- If we write $\bar{F}(x) = 1 - F(x) = P(V > x)$, then the inner minimization is equivalent to
$$\max_{D(F,F_V)\leq\delta} \bar{F}(b) = \max_{D(F,F_V)\leq\delta} P_F(V > b) = P_F(V > b - \lambda_b)\,.$$
- Let $\lambda = \lambda_b \geq 0$ be a Lagrange multiplier, the "worst case distribution" is
$$\begin{aligned} V^* &= V \cdot I(V > b) + b \cdot I(b - \lambda < V \leq b) \\ &\quad + V \cdot I(V \leq b - \lambda)\,. \end{aligned}$$

Quantifying Model Mispecifications

- We now want
$$\max_b \min_{D(F,F_V)\leq\delta} (u-b)\,F(b)\,.$$

- If we write $\bar{F}(x) = 1 - F(x) = P(V > x)$, then the inner minimization is equivalent to
$$\max_{D(F,F_V)\leq\delta} \bar{F}(b) = \max_{D(F,F_V)\leq\delta} P_F(V > b) = P_F(V > b - \lambda_b)\,.$$

- Let $\lambda = \lambda_b \geq 0$ be a Lagrange multiplier, the "worst case distribution" is
$$\begin{aligned} V^* &= V \cdot I(V > b) + b \cdot I(b - \lambda < V \leq b) \\ &\quad + V \cdot I(V \leq b - \lambda)\,. \end{aligned}$$

- Intuitively: re-arrange V as cheaply as possible to produce V^* so that $V^* > b$ happens (λ computed to satisfy cost constraint).

- Conclusion: We are trying to find the (Nash Equilibrium) policy $b^*(u) = f(u)$ so

$$\max_b \min_{D(F,F_{\bar{V}})\leq\delta} (u-b)\,F_{\bar{V}}\left(f^{-1}(b)\right)$$
$$= \max_b (u-b)\,F_{\bar{V}}\left(f^{-1}(b) - \lambda_{f^{-1}(b)}\right).$$

Quantifying Model Mispecifications

- Conclusion: We are trying to find the (Nash Equilibrium) policy $b^*(u) = f(u)$ so

$$\begin{aligned} & \max_b \min_{D(F, F_{\bar{V}}) \leq \delta} (u-b) F_{\bar{V}}\left(f^{-1}(b)\right) \\ = \quad & \max_b (u-b) F_{\bar{V}}\left(f^{-1}(b) - \lambda_{f^{-1}(b)}\right). \end{aligned}$$

- Optimizing over $b(\cdot)$ we obtain

$$b(u) = \frac{\int_0^u x f_{\bar{V}}(x - \lambda_x)\left(1 - \dot{\lambda}(x)\right) dx}{F_{\bar{V}}(u - \lambda_u)},$$

with

$$\int_{u-\lambda_u}^{u} (u-v) f_{\bar{V}}(v)\, dv = \delta.$$

Approximate Distributionally Robust Equilibrium Bidding Policies

- While the previous equations can be solved numerically, they may be a bit cumbersome to implement.

Approximate Distributionally Robust Equilibrium Bidding Policies

- While the previous equations can be solved numerically, they may be a bit cumbersome to implement.
- So, we provide an asymptotic expansion as $\delta \to 0$.

Approximate Distributionally Robust Equilibrium Bidding Policies

- While the previous equations can be solved numerically, they may be a bit cumbersome to implement.
- So, we provide an asymptotic expansion as $\delta \to 0$.
- This leads to a bidding strategy of the form

$$b_\delta(u) = b_0(u) + \delta^{1/2} b_1(u) + O(\delta),$$

where

$$b_0(u) = E(\bar{V} | \bar{V} \leq u) = \int_0^u x f_{\bar{V}}(x)\, dx / F_{\bar{V}}(x)$$

and

$$b_1(u) = \frac{\sqrt{2}}{F_{\bar{V}}(u)} \left(\int_0^u \sqrt{f_{\bar{V}}(x)} dx - \frac{f_{\bar{V}}(u)}{F_{\bar{V}}(u)} \int_0^u F_{\bar{V}}(x)\, dx \right).$$

Example

- Example 3: Back to logistic model

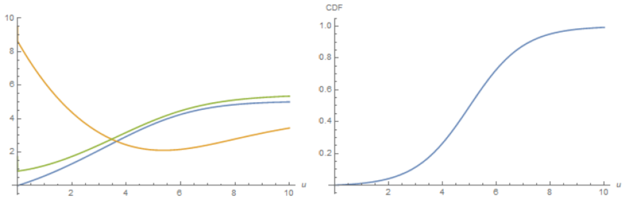
Example

- Example 3: Back to logistic model
- $P\left(\bar{V} \leq x\right) = \left(1 + \exp\left(-xc\right)\right) / \left(1 + \exp\left(a - xc\right)\right)$ for $a \in R$, $c > 0$.

Example

- Example 3: Back to logistic model
- $P\left(\bar{V} \leq x\right) = \left(1 + \exp\left(-xc\right)\right) / \left(1 + \exp\left(a - xc\right)\right)$ for $a \in R$, $c > 0$.
- $a = 5$, $c = 1$ and $\delta = .01$ (figures in \$/1000)

We show the bidding policy and CDF for $a = 5, c = 1, \delta = 0.01$ in the following plot.



(a) Bidding policy (b) CDF of V

Our Goal

So, now we want to add a player optimizing a decision and play the game:

$$\min_{\theta} \max_{D(P,P_n)\leq\delta} E\left(l\left(X,\theta\right)\right).$$

Based on: Robust Wasserstein Profile Inference (B., Murthy & Kang '16)
https://arxiv.org/abs/1610.05627
https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/robust-wasserstein-profile-inference-and-applications-to-machine-learning

- Consider estimating $\beta_* \in R^m$ in linear regression

$$Y_i = \beta X_i + e_i,$$

where $\{(Y_i, X_i)\}_{i=1}^n$ are data points.

- Consider estimating $\beta_* \in R^m$ in linear regression

$$Y_i = \beta X_i + e_i,$$

where $\{(Y_i, X_i)\}_{i=1}^n$ are data points.

- Optimal Least Squares approach consists in estimating β_* via

$$\min_\beta E_{P_n}\left[\left(Y - \beta^T X\right)^2\right] = \min_\beta \frac{1}{n}\sum_{i=1}^n \left(Y_i - \beta^T X_i\right)^2$$

Distributionally Robust Optimization in Machine Learning

- Consider estimating $\beta_* \in R^m$ in linear regression

$$Y_i = \beta X_i + e_i,$$

where $\{(Y_i, X_i)\}_{i=1}^n$ are data points.

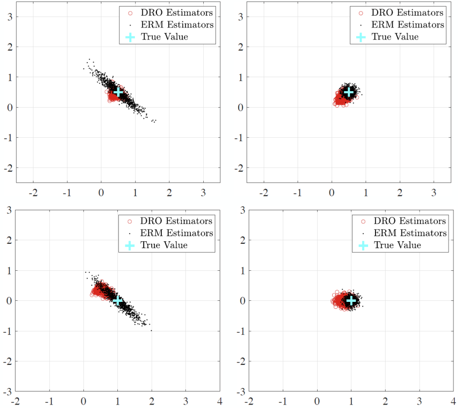
- Optimal Least Squares approach consists in estimating β_* via

$$\min_{\beta} E_{P_n}\left[\left(Y - \beta^T X\right)^2\right] = \min_{\beta} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \beta^T X_i\right)^2$$

- Apply the distributionally robust estimator based on optimal transport.

Applying Distributionally Robust Optimization in Linear Regression

Estimation of θ_* with DRO ($\circ$) and without DRO ($\circ$)



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

Connection to Sqrt-Lasso

Theorem (B., Kang, Murthy (2016)) Suppose that

$$c\left(\left(x,y\right),\left(x',y'\right)\right)=\begin{cases}\left\|x-x'\right\|_q^2 & \text{if} \quad y=y' \\ \infty & \text{if} \quad y\neq y'\end{cases}.$$

Then, if $1/p+1/q=1$

$$\max_{P:D_c(P,P_n)\leq\delta} E_P^{1/2}\left(\left(Y-\beta^T X\right)^2\right)=E_{P_n}^{1/2}\left[\left(Y-\beta^T X\right)^2\right]+\sqrt{\delta}\left\|\beta\right\|_p.$$

Remark 1: This is sqrt-Lasso (Belloni et al. (2011)).

Logistic Regression

- Classical classification model:

$$
\begin{aligned}
P(Y=1|X) &= \frac{\exp\left(\beta^T X\right)}{1+\exp\left(\beta^T X\right)} = \frac{1}{\exp\left(-\beta^T X\right)+1} \\
P(Y=-1|X) &= \frac{1}{1+\exp\left(\beta^T X\right)}
\end{aligned}
$$

Logistic Regression

- Classical classification model:

$$P(Y=1|X) = \frac{\exp\left(\beta^T X\right)}{1+\exp\left(\beta^T X\right)} = \frac{1}{\exp\left(-\beta^T X\right)+1}$$

$$P(Y=-1|X) = \frac{1}{1+\exp\left(\beta^T X\right)}$$

- The likelihood of (y, x) is:

$$-\log\left(1+\exp\left(-y\beta^T x\right)\right)$$

Logistic Regression

- Therefore, given $\{(y_i, x_i)\}_{i=1}^{n}$ maximum likelihood is equivalent to

$$\max_{\beta} - \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \beta^T x_i\right)\right).$$

Logistic Regression

- Therefore, given $\{(y_i, x_i)\}_{i=1}^{n}$ maximum likelihood is equivalent to

$$\max_{\beta} - \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \beta^T x_i\right)\right).$$

- Also equivalent to

$$\begin{aligned} & \min_{\beta} E_{P_n}\left[\log\left(1 + \exp\left(-Y\beta^T X\right)\right)\right] \\ = \quad & \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \beta^T x_i\right)\right). \end{aligned}$$

Regularized Logistic Regression

Theorem (B., Kang, Murthy (2016)) Suppose that

$$c\left((x,y),(x',y')\right) = \begin{cases} \|x-x'\|_q & \text{if } \quad y=y' \\ \infty & \text{if } \quad y\neq y' \end{cases}.$$

Then,

$$\sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\left[\log(1+e^{-Y\beta^T X})\right] = E_{P_n}\left[\log(1+e^{-Y\beta^T X})\right] + \delta\,\|\beta\|_p\,.$$

Remark 1: First studied via an approximation in Esfahani and Kuhn (2015).

Connection to Support Vector Machines

Theorem (B., Kang, Murthy (2016)) Suppose that

$$c\left(\left(x,y\right),\left(x',y'\right)\right)=\begin{cases}\left\|x-x'\right\|_q & \text{if} \quad y=y' \\ \infty & \text{if} \quad y\neq y'\end{cases}.$$

Then,

$$\sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P[\left(1-Y\beta^T X\right)^+] = E_{P_n}\left[\left(1-Y\beta^T X\right)^+\right]+\delta\left\|\beta\right\|_p.$$

Unification and Extensions of Regularized Estimators

- Distributionally Robust Optimization using Optimal Transport recovers many other estimators...

Unification and Extensions of Regularized Estimators

- Distributionally Robust Optimization using Optimal Transport recovers many other estimators...
- *Group Lasso*: B., & Kang (2016): **https://arxiv.org/abs/1705.04241**

Unification and Extensions of Regularized Estimators

- Distributionally Robust Optimization using Optimal Transport recovers many other estimators...
- *Group Lasso*: B., & Kang (2016): **https://arxiv.org/abs/1705.04241**
- *Generalized adaptive ridge*: B., Kang, Murthy, Zhang (2017): **https://arxiv.org/abs/1705.07152**

Unification and Extensions of Regularized Estimators

- Distributionally Robust Optimization using Optimal Transport recovers many other estimators...
- *Group Lasso*: B., & Kang (2016): **https://arxiv.org/abs/1705.04241**
- *Generalized adaptive ridge*: B., Kang, Murthy, Zhang (2017): **https://arxiv.org/abs/1705.07152**
- Semisupervised learning: B., and Kang (2016): **https://arxiv.org/abs/1702.08848**

Unification and Extensions of Regularized Estimators

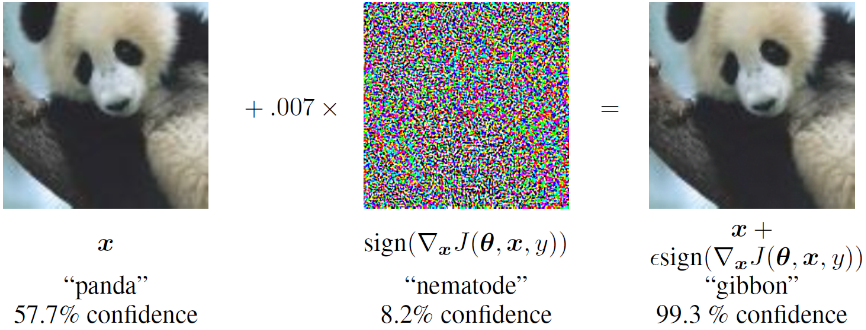
- Distributionally Robust Optimization using Optimal Transport recovers many other estimators...
- *Group Lasso*: B., & Kang (2016): **https://arxiv.org/abs/1705.04241**
- *Generalized adaptive ridge*: B., Kang, Murthy, Zhang (2017): **https://arxiv.org/abs/1705.07152**
- Semisupervised learning: B., and Kang (2016): **https://arxiv.org/abs/1702.08848**
- See the excellent tutorials by Kuhn et al (2019) and Rahimian & Mehrotra (2019).

Unification and Extensions of Regularized Estimators

- Distributionally Robust Optimization using Optimal Transport recovers many other estimators...
- *Group Lasso*: B., & Kang (2016): **https://arxiv.org/abs/1705.04241**
- *Generalized adaptive ridge*: B., Kang, Murthy, Zhang (2017): **https://arxiv.org/abs/1705.07152**
- Semisupervised learning: B., and Kang (2016): **https://arxiv.org/abs/1702.08848**
- See the excellent tutorials by Kuhn et al (2019) and Rahimian & Mehrotra (2019).
- Other areas in which optimal transport arises in machine learning

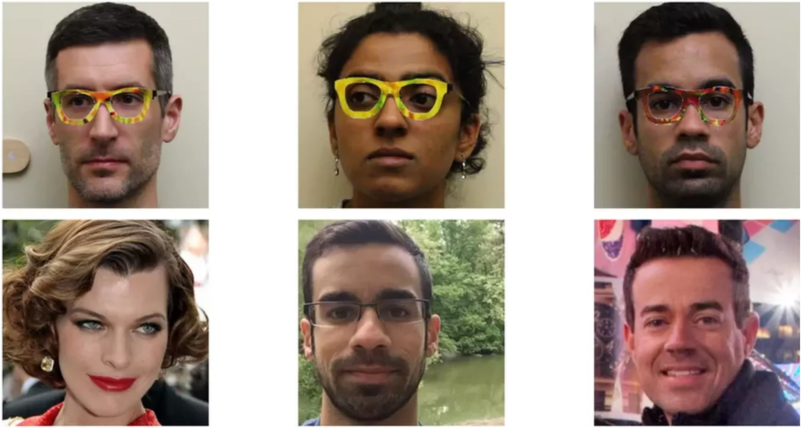
Deep Neural Networks: Adversarial Attacks

- Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, and Fergus (2014).



Deep Neural Networks: Adversarial Attacks

- Sharif, Bhagavatula, Bauer, and Reiter (2016)



Deep Neural Networks: Adversarial Attacks

- Picture from the BBC

Chinese man caught by facial recognition at pop concert

13 April 2018

Chinese police have used facial recognition technology to locate and arrest a man who was among a crowd of 60,000 concert goers.

How Regularization and Dual Norms Arise?

- Let us work out a simple example...

- Let us work out a simple example...
- Recall RoPA Duality: Pick $c\left((x,y),(x',y')\right) = \|(x,y)-(x',y')\|_q^2$

$$\max_{P:D_c(P,P_n)\le\delta} E_P\left(\left(\left(X,Y\right)\cdot(\beta,1)\right)^2\right)$$

$$= \min_{\lambda\ge 0}\left\{\lambda\delta + E_{P_n}\sup_{(x',y')}\left[\left(\left(x',y'\right)\cdot(\beta,1)\right)^2 - \lambda\left\|(X,Y)-(x',y')\right\|_q^2\right.\right.$$

- Let us work out a simple example...
- Recall RoPA Duality: Pick $c\left((x,y),(x',y')\right) = \left\|(x,y)-(x',y')\right\|_q^2$

$$\max_{P:D_c(P,P_n)\leq\delta} E_P\left(\left(\left(X,Y\right)\cdot\left(\beta,1\right)\right)^2\right)$$

$$= \min_{\lambda\geq 0}\left\{\lambda\delta + E_{P_n}\sup_{(x',y')}\left[\left(\left(x',y'\right)\cdot\left(\beta,1\right)\right)^2 - \lambda\left\|\left(X,Y\right)-\left(x',y'\right)\right\|_q^2\right.\right.$$

- Let's focus on the inside E_{P_n}...

How Regularization and Dual Norms Arise?

- Let $\Delta = (X, Y) - (x', y')$

$$
\begin{aligned}
&\sup_{(x',y')} \left[\left(\left(x', y' \right) \cdot (\beta, 1) \right)^2 - \lambda \left\| (X, Y) - \left(x', y' \right) \right\|_q^2 \right] \\
= &\sup_{\Delta} \left[\left((X, Y) \cdot (\beta, 1) - \Delta \cdot (\beta, 1) \right)^2 - \lambda \left\| \Delta \right\|_q^2 \right] \\
= &\sup_{\|\Delta\|_q} \left[\left(\left| (X, Y) \cdot (\beta, 1) \right| + \left\| \Delta \right\|_q \left\| (\beta, 1) \right\|_p \right)^2 - \lambda \left\| \Delta \right\|_q^2 \right]
\end{aligned}
$$

How Regularization and Dual Norms Arise?

- Let $\Delta = (X, Y) - (x', y')$

$$
\begin{aligned}
& \sup_{(x',y')} \left[\left(\left(x', y' \right) \cdot (\beta, 1) \right)^2 - \lambda \left\| (X, Y) - \left(x', y' \right) \right\|_q^2 \right] \\
= & \sup_{\Delta} \left[((X, Y) \cdot (\beta, 1) - \Delta \cdot (\beta, 1))^2 - \lambda \left\| \Delta \right\|_q^2 \right] \\
= & \sup_{\left\| \Delta \right\|_q} \left[(\left| (X, Y) \cdot (\beta, 1) \right| + \left\| \Delta \right\|_q \left\| (\beta, 1) \right\|_p)^2 - \lambda \left\| \Delta \right\|_q^2 \right]
\end{aligned}
$$

- Last equality uses $z \to z^2$ is symmetric around origin and $|a \cdot b| \leq \|a\|_p \|b\|_q$.

How Regularization and Dual Norms Arise?

- Let $\Delta = (X, Y) - (x', y')$

$$
\begin{aligned}
& \sup_{(x',y')} \left[\left(\left(x', y' \right) \cdot (\beta, 1) \right)^2 - \lambda \left\| (X, Y) - \left(x', y' \right) \right\|_q^2 \right] \\
= & \sup_{\Delta} \left[((X, Y) \cdot (\beta, 1) - \Delta \cdot (\beta, 1))^2 - \lambda \left\| \Delta \right\|_q^2 \right] \\
= & \sup_{\left\| \Delta \right\|_q} \left[(|(X, Y) \cdot (\beta, 1)| + \left\| \Delta \right\|_q \left\| (\beta, 1) \right\|_p)^2 - \lambda \left\| \Delta \right\|_q^2 \right]
\end{aligned}
$$

- Last equality uses $z \to z^2$ is symmetric around origin and $|a \cdot b| \leq \|a\|_p \|b\|_q$.
- Note problem is now one-dimensional (easily computable).

A Fully Worked Out Example: Support Vector Machines

- Use RoPA: with
$c\left(\left(x,y\right),\left(x',y'\right)\right) = \left\|x-x'\right\|_q I\left(y=y'\right) + \infty I\left(y\neq y'\right)$

$$
\begin{aligned}
&\sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\left[\left(1-Y\beta^T X\right)^+\right] \\
=\ & \min_{\lambda\geq 0}\left[\lambda\delta + E_{P_n}\left\{\max_x\left(\left(1-Y\beta^T x\right)^+ - \lambda\left\|x-X\right\|_q\right)\right\}\right] \\
=\ & \min_{\lambda\geq 0}\left[\lambda\delta + E_{P_n}\left\{\max_\Delta\left(\left(1-Y\beta^T X - Y\beta^T\Delta\right)^+ - \lambda\left\|\Delta\right\|_q\right)\right\}\right] \\
=\ & \min_{\lambda\geq 0}\left[\lambda\delta + E_{P_n}\left\{\max_\Delta\left(\left(1-Y\beta^T X + \left\|\beta\right\|_p\left\|\Delta\right\|_q\right)^+ - \lambda\left\|\Delta\right\|_q\right)\right\} \right. \\
=\ & \min_{\lambda\geq\left\|\beta\right\|_p}\left[\lambda\delta + E_{P_n}\left\{\max_{\left\|\Delta\right\|_q}\left(\left(1-Y\beta^T X + \left\|\beta\right\|_p\left\|\Delta\right\|_q\right)^+ - \lambda\left\|\Delta\right\|_q \right.\right.\right. \\
=\ & \min_{\lambda\geq\left\|\beta\right\|_p}\left[\lambda\delta + E_{P_n}\left(1-Y\beta^T X\right)^+\right] = \lambda\left\|\beta\right\|_p + E_{P_n}\left(1-Y\beta^T X\right)
\end{aligned}
$$

Explaining the Adversarial Attacks of Neural Networks

- So, in general

$$c\left(\left(x,y\right),\left(x',y'\right)\right) = \left\|x-x'\right\|_q I\left(y=y'\right) + \infty I\left(y\neq y'\right)$$

$$\begin{aligned}
&\sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\left[l\left(\theta, Y, X\right)\right] \\
&= \min_{\lambda\geq 0}\left[\lambda\delta + E_{P_n}\left\{\max_x\left(l\left(\theta, Y, x\right) - \lambda\left\|x-X\right\|_q\right)\right\}\right] \\
&= \min_{\lambda\geq 0}\left[\lambda\delta + E_{P_n}\left\{\max_\Delta\left(l\left(\theta, Y, X+\Delta\right) - \lambda\left\|\Delta\right\|_q\right)\right\}\right] \\
&= \min_{\lambda\geq 0}\left[\lambda\delta + E_{P_n}\left\{\max_\Delta\left(l\left(\theta, Y, X+\Delta/\lambda\right) - \left\|\Delta\right\|_q\right)\right\}\right].
\end{aligned}$$

Explaining the Adversarial Attacks of Neural Networks

- So, in general
$c\left(\left(x,y\right),\left(x',y'\right)\right) = \left\|x-x'\right\|_q I\left(y=y'\right) + \infty I\left(y \neq y'\right)$

$$
\begin{aligned}
& \sup_{P:\ \mathcal{D}_c(P,P_n)\leq\delta} E_P\left[l\left(\theta, Y, X\right)\right] \\
= \quad & \min_{\lambda\geq 0}\left[\lambda\delta + E_{P_n}\left\{\max_x\left(l\left(\theta, Y, x\right) - \lambda\left\|x-X\right\|_q\right)\right\}\right] \\
= \quad & \min_{\lambda\geq 0}\left[\lambda\delta + E_{P_n}\left\{\max_\Delta\left(l\left(\theta, Y, X+\Delta\right) - \lambda\left\|\Delta\right\|_q\right)\right\}\right] \\
= \quad & \min_{\lambda\geq 0}\left[\lambda\delta + E_{P_n}\left\{\max_\Delta\left(l\left(\theta, Y, X+\Delta/\lambda\right) - \left\|\Delta\right\|_q\right)\right\}\right].
\end{aligned}
$$

- If $\delta \approx 0$, then λ is large, so inner maximization

$$
\begin{aligned}
& \max_\Delta\left(l\left(\theta, Y, X+\Delta/\lambda\right) - \left\|\Delta\right\|_q\right) \\
\approx \quad & l\left(\theta, Y, X\right) + \left\|l_x\left(\theta, Y, X\right)\right\|_p \left\|\Delta\right\|_q/\lambda - \left\|\Delta\right\|_q
\end{aligned}
$$

Summary

- The worst case perturbation is given by Δ such that

$$l_x\left(\theta, Y, X\right) \cdot \Delta / \lambda = \left\| l_x\left(\theta, Y, X\right) \right\|_p \left\| \Delta \right\|_q / \lambda,$$

if $q = \infty$, then $\Delta = c \cdot sign\left(l_x\left(\theta, Y, X\right)\right)$.

Summary

- The worst case perturbation is given by Δ such that

$$l_x(\theta, Y, X) \cdot \Delta/\lambda = \|l_x(\theta, Y, X)\|_p \|\Delta\|_q /\lambda,$$

  if $q = \infty$, then $\Delta = c \cdot sign(l_x(\theta, Y, X))$.

- So, $\delta \approx 0$ means perturbing by

$$\epsilon \cdot sign(l_x(\theta, Y, X))$$

  for $\epsilon > 0$.

Summary

- The worst case perturbation is given by Δ such that

$$l_x(\theta, Y, X) \cdot \Delta / \lambda = \|l_x(\theta, Y, X)\|_p \|\Delta\|_q / \lambda,$$

  if $q = \infty$, then $\Delta = c \cdot sign(l_x(\theta, Y, X))$.
- So, $\delta \approx 0$ means perturbing by

$$\epsilon \cdot sign(l_x(\theta, Y, X))$$

  for $\epsilon > 0$.
- This explains the nature of the panda example given earlier.

Can We Defend Against Attacks?

- Naturally, it makes sense then to train networks using

$$
\begin{aligned}
& \min_{\theta} \max_{D(P,P_n)\leq\delta} E_P\left(l\left(\theta, Y, X\right)\right) \\
= \;& \min_{\theta}\{\lambda\delta + E_{P_n} \max_{x}[l\left(\theta, Y, x\right) - \lambda\left\|x - X\right\|_q].
\end{aligned}
$$

Can We Defend Against Attacks?

- Naturally, it makes sense then to train networks using

$$\begin{aligned} & \min_{\theta} \max_{D(P,P_n)\leq\delta} E_P\left(l\left(\theta, Y, X\right)\right) \\ = \quad & \min_{\theta}\{\lambda\delta + E_{P_n} \max_{x}[l\left(\theta, Y, x\right) - \lambda\left\|x - X\right\|_q]. \end{aligned}$$

- This will automatically protect against attacks.

Can We Defend Against Attacks?

- Naturally, it makes sense then to train networks using

$$\begin{aligned} & \min_{\theta} \max_{D(P,P_n)\leq\delta} E_P\left(l\left(\theta, Y, X\right)\right) \\ = \; & \min_{\theta}\{\lambda\delta + E_{P_n}\max_{x}[l\left(\theta, Y, x\right) - \lambda\left\|x - X\right\|_q]. \end{aligned}$$

- This will automatically protect against attacks.
- This is an active area of research currently.

Can We Defend Against Attacks?

- Naturally, it makes sense then to train networks using

$$\begin{aligned} & \min_{\theta} \max_{D(P,P_n)\leq\delta} E_P\left(l\left(\theta, Y, X\right)\right) \\ = \; & \min_{\theta}\{\lambda\delta + E_{P_n} \max_{x}[l\left(\theta, Y, x\right) - \lambda \left\|x - X\right\|_q]. \end{aligned}$$

- This will automatically protect against attacks.
- This is an active area of research currently.
- But there may be many possible attacks.

On Role of Transport Cost...

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**

On Role of Transport Cost...

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**
- Suppose that $\|x - x'\|_A^2 = (x - x')\, A\, (x - x)$ with A positive definite (Mahalanobis distance).

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**
- Suppose that $\|x - x'\|_A^2 = (x - x')\, A\, (x - x)$ with A positive definite (Mahalanobis distance).
- Then,

$$
\begin{aligned}
& \max_{P: D_c(P, P_n) \leq \delta} E_P^{1/2}\left(\left(Y - \beta^T X\right)^2\right) \\
= \quad & \min_{\beta} E_{P_n}^{1/2}\left[\left(Y - \beta^T X\right)^2\right] + \sqrt{\delta}\, \|\beta\|_{A^{-1}} .
\end{aligned}
$$

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**
- Suppose that $\|x-x'\|_A^2 = (x-x')\,A\,(x-x)$ with A positive definite (Mahalanobis distance).
- Then,

$$\begin{aligned} & \max_{P:D_c(P,P_n)\leq\delta} E_P^{1/2}\left(\left(Y-\beta^T X\right)^2\right) \\ = \; & \min_{\beta} E_{P_n}^{1/2}\left[\left(Y-\beta^T X\right)^2\right] + \sqrt{\delta}\,\|\beta\|_{A^{-1}}\,. \end{aligned}$$

- *Intuition: Think of A diagonal, encoding inverse variability of X_is...*

On Role of Transport Cost...

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**
- Suppose that $\|x - x'\|_A^2 = (x - x')\,A\,(x - x)$ with A positive definite (Mahalanobis distance).
- Then,

$$\begin{aligned} & \max_{P : D_c(P, P_n) \leq \delta} E_P^{1/2}\left(\left(Y - \beta^T X\right)^2\right) \\ = & \min_\beta E_{P_n}^{1/2}\left[\left(Y - \beta^T X\right)^2\right] + \sqrt{\delta}\,\|\beta\|_{A^{-1}}. \end{aligned}$$

- *Intuition: Think of A diagonal, encoding inverse variability of X_is...*
- **High variability —> cheap transportation —> high impact in risk estimation.**

On Role of Transport Cost...

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**

On Role of Transport Cost...

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**
- Suppose that $\|x - x'\|_{\Lambda}^{2} = (x - x')\,\Lambda\,(x - x)$ with Λ positive definite (Mahalanobis distance).

On Role of Transport Cost...

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**
- Suppose that $\|x - x'\|_{\Lambda}^{2} = (x - x') \Lambda (x - x)$ with Λ positive definite (Mahalanobis distance).
- Then,

$$
\begin{aligned}
& \max_{P: D_c(P, P_n) \leq \delta} E_P^{1/2}\left(\left(Y - \beta^T X\right)^2\right) \\
= \quad & \min_{\beta} E_{P_n}^{1/2}\left[\left(Y - \beta^T X\right)^2\right] + \sqrt{\delta} \, \|\beta\|_{\Lambda^{-1}} .
\end{aligned}
$$

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**
- Suppose that $\|x - x'\|_{\Lambda}^{2} = (x - x')\,\Lambda\,(x - x)$ with Λ positive definite (Mahalanobis distance).
- Then,

$$
\begin{aligned}
&\max_{P:D_c(P,P_n)\leq\delta} E_P^{1/2}\left(\left(Y-\beta^T X\right)^2\right) \\
=\;& \min_{\beta} E_{P_n}^{1/2}\left[\left(Y-\beta^T X\right)^2\right]+\sqrt{\delta}\,\|\beta\|_{\Lambda^{-1}}.
\end{aligned}
$$

- *Intuition: Think of Λ diagonal, encoding inverse variability of X_is...*

On Role of Transport Cost...

- **https://arxiv.org/abs/1705.07152: Data-driven chose of $c(\cdot)$.**
- Suppose that $\|x - x'\|_{\Lambda}^{2} = (x - x')\,\Lambda\,(x - x)$ with Λ positive definite (Mahalanobis distance).
- Then,

$$\max_{P:D_c(P,P_n)\leq\delta} E_P^{1/2}\left(\left(Y-\beta^T X\right)^2\right)$$

$$= \min_{\beta} E_{P_n}^{1/2}\left[\left(Y-\beta^T X\right)^2\right] + \sqrt{\delta}\,\|\beta\|_{\Lambda^{-1}}.$$

- *Intuition: Think of Λ diagonal, encoding inverse variability of X_is...*
- **High variability —> cheap transportation —> high impact in risk estimation.**

Connections to Statistical Analysis

https://arxiv.org/abs/1610.05627
Robust Wasserstein Profile Inference
B., Murthy & Kang '16

https://arxiv.org/abs/1906.01614
Confidence Regions in Wasserstein Distributionally Robust Estimation
B., Murthy & Si '19
Optimal size of uncertainty + Asymptotic Normality

Towards an Optimal Choice of Uncertainty Size

- How to choose uncertainty size in a data-driven way?

Towards an Optimal Choice of Uncertainty Size

- How to choose uncertainty size in a data-driven way?
- Once again, consider Lasso as example:

$$
\begin{aligned}
&\min_{\beta} \max_{P: D_c(P, P_n) \leq \delta} E_P^{1/2}\left(\left(Y - \beta^T X\right)^2\right) \\
= \quad &\min_{\beta} E_{P_n}^{1/2}\left[\left(Y - \beta^T X\right)^2\right] + \sqrt{\delta}\, \|\beta\|_p \,.
\end{aligned}
$$

Towards an Optimal Choice of Uncertainty Size

- How to choose uncertainty size in a data-driven way?
- Once again, consider Lasso as example:

$$
\begin{aligned}
& \min_{\beta} \max_{P: D_c(P, P_n) \le \delta} E_P^{1/2}\left(\left(Y - \beta^T X\right)^2\right) \\
= \quad & \min_{\beta} E_{P_n}^{1/2}\left[\left(Y - \beta^T X\right)^2\right] + \sqrt{\delta}\, \|\beta\|_p .
\end{aligned}
$$

- Use left hand side to define a statistical principle to choose δ.

Towards an Optimal Choice of Uncertainty Size

- How to choose uncertainty size in a data-driven way?
- Once again, consider Lasso as example:

$$\begin{aligned}
& \min_{\beta} \max_{P:D_c(P,P_n)\leq\delta} E_P^{1/2}\left(\left(Y-\beta^T X\right)^2\right) \\
= \; & \min_{\beta} E_{P_n}^{1/2}\left[\left(Y-\beta^T X\right)^2\right] + \sqrt{\delta}\,\|\beta\|_p .
\end{aligned}$$

- Use left hand side to define a statistical principle to choose δ.
- Important: Optimizing δ is equivalent to optimizing regularization.

Towards an Optimal Choice of Uncertainty Size

- One way to select δ: estimate $D\left(P_{true}, P_n\right)$?

Towards an Optimal Choice of Uncertainty Size

- One way to select δ: estimate $D\left(P_{true}, P_n\right)$?
- This was advocated and seems natural at first sight... but there is a big problem.

Towards an Optimal Choice of Uncertainty Size

- One way to select δ: estimate $D\left(P_{true}, P_n\right)$?
- This was advocated and seems natural at first sight... but there is a big problem.
- Consider the case $c\left(x, x'\right) = \left\|x - x'\right\|_\infty$ by Kantorovich-Rubinstein duality

$$\begin{aligned} D\left(P_{true}, P_n\right) &= \sup_{\alpha\, \in \, \mathrm{Lip}(1)} E_{P_{true}}\alpha\left(X\right) - E_{P_n}\alpha\left(X\right) \\ &= \sup_{\alpha\, \in \, \mathrm{Lip}(1)} \int \alpha\left(x\right)\left(dP_{true} - dP_n\right). \end{aligned}$$

Towards an Optimal Choice of Uncertainty Size

- One way to select δ: estimate $D\left(P_{true}, P_n\right)$?
- This was advocated and seems natural at first sight... but there is a big problem.
- Consider the case $c\left(x, x'\right) = \left\|x - x'\right\|_{\infty}$ by Kantorovich-Rubinstein duality

$$
\begin{aligned}
D\left(P_{true}, P_n\right) &= \sup_{\alpha\ \in\ \mathrm{Lip}(1)} E_{P_{true}}\alpha\left(X\right) - E_{P_n}\alpha\left(X\right) \\
&= \sup_{\alpha\ \in\ \mathrm{Lip}(1)} \int \alpha\left(x\right)\left(dP_{true} - dP_n\right).
\end{aligned}
$$

- Unfortunately, it turns out that typically $D\left(P_{true}, P_n\right) = O\left(n^{-1/d}\right)$ (Dudley '68) for $d > 2$.

Towards an Optimal Choice of Uncertainty Size

- So, even if statistics for $D\left(P_{true}, P_n\right) = O\left(n^{-1/d}\right)$ are known, this approach would suggest choosing $\delta = cn^{-1/d}$.

Towards an Optimal Choice of Uncertainty Size

- So, even if statistics for $D\left(P_{true}, P_n\right) = O\left(n^{-1/d}\right)$ are known, this approach would suggest choosing $\delta = cn^{-1/d}$.
- But this would imply solving (say for the logistic regression)

$$\min_{\beta}\left\{E_{P_n}\left[\log(1+e^{-Y\beta^T X})\right] + cn^{-1/d}\left\|\beta\right\|_1\right\}.$$

Towards an Optimal Choice of Uncertainty Size

- So, even if statistics for $D\left(P_{true}, P_n\right) = O\left(n^{-1/d}\right)$ are known, this approach would suggest choosing $\delta = cn^{-1/d}$.
- But this would imply solving (say for the logistic regression)

$$\min_{\beta}\left\{E_{P_n}\left[\log(1+e^{-Y\beta^T X})\right] + cn^{-1/d}\|\beta\|_1\right\}.$$

- But we know that letting $\delta = 0$ we typically obtain asymptotically normal estimators

$$\beta_n \approx \beta_{true} + n^{-1/2}N\left(0, \sigma^2\right).$$

Towards an Optimal Choice of Uncertainty Size

- So, even if statistics for $D\left(P_{true}, P_n\right) = O\left(n^{-1/d}\right)$ are known, this approach would suggest choosing $\delta = cn^{-1/d}$.
- But this would imply solving (say for the logistic regression)

$$\min_{\beta}\left\{E_{P_n}\left[\log(1+e^{-Y\beta^T X})\right] + cn^{-1/d}\left\|\beta\right\|_1\right\}.$$

- But we know that letting $\delta = 0$ we typically obtain asymptotically normal estimators

$$\beta_n \approx \beta_{true} + n^{-1/2}N\left(0,\sigma^2\right).$$

- So, using $\delta = cn^{-1/d}$ induces an error much bigger than $n^{-1/2}$ when $d > 2$.

Towards an Optimal Choice of Uncertainty Size

- Cross validation is typically the method of choice!

Towards an Optimal Choice of Uncertainty Size

- Cross validation is typically the method of choice!
- There is really nothing wrong with cross validation (especially if prediction is the goal).

Towards an Optimal Choice of Uncertainty Size

- Cross validation is typically the method of choice!
- There is really nothing wrong with cross validation (especially if prediction is the goal).
- Except that it could be quite data intensive + computationally heavy.

Towards an Optimal Choice of Uncertainty Size

- Cross validation is typically the method of choice!
- There is really nothing wrong with cross validation (especially if prediction is the goal).
- Except that it could be quite data intensive + computationally heavy.
- For k-fold cross validation to be consistent you need $k/n \to 1$ and $n - k \to \infty$ (Shao '93).

Towards an Optimal Choice of Uncertainty Size

- Cross validation is typically the method of choice!
- There is really nothing wrong with cross validation (especially if prediction is the goal).
- Except that it could be quite data intensive + computationally heavy.
- For k-fold cross validation to be consistent you need $k/n \longrightarrow 1$ and $n - k \longrightarrow \infty$ (Shao '93).
- So, for model selection you need k increasing.

Towards an Optimal Choice of Uncertainty Size

- Keep in mind linear regression problem

$$Y_i = \beta_*^T X_i + \epsilon_i.$$

Towards an Optimal Choice of Uncertainty Size

- Keep in mind linear regression problem

$$Y_i = \beta_*^T X_i + \epsilon_i.$$

- The *plausible model variations* of P_n are given by the set

$$\mathcal{U}_\delta(n) = \{P : D_c(P, P_n) \le \delta\}.$$

Towards an Optimal Choice of Uncertainty Size

- Keep in mind linear regression problem

$$Y_i = \beta_*^T X_i + \epsilon_i.$$

- The *plausible model variations* of P_n are given by the set

$$\mathcal{U}_\delta(n) = \{P : D_c(P, P_n) \leq \delta\}.$$

- Given $P \in \mathcal{U}_\delta(n)$, define $\bar{\beta}(P) = \arg\min E_P\left[\left(Y - \beta^T X\right)^2\right]$.

Towards an Optimal Choice of Uncertainty Size

- Keep in mind linear regression problem

$$Y_i = \beta_*^T X_i + \epsilon_i.$$

- The *plausible model variations* of P_n are given by the set

$$\mathcal{U}_\delta(n) = \{P : D_c(P, P_n) \leq \delta\}.$$

- Given $P \in \mathcal{U}_\delta(n)$, define $\bar{\beta}(P) = \arg\min E_P\left[\left(Y - \beta^T X\right)^2\right]$.
- It is natural to say that

$$\Lambda_\delta(n) = \{\bar{\beta}(P) : P \in \mathcal{U}_\delta(n)\}$$

are *plausible estimates* of β_*.

Optimal Choice of Uncertainty Size

- Given a confidence level $1-\alpha$ we advocate choosing δ via

$$\begin{aligned}&\min \delta\\&s.t.\quad P\left(\beta_{*}\in\Lambda_{\delta}\left(n\right)\right)\geq 1-\alpha\ .\end{aligned}$$

Optimal Choice of Uncertainty Size

- Given a confidence level $1-\alpha$ we advocate choosing δ via

$$\begin{aligned} &\min \delta \\ &s.t. \quad P\left(\beta_{*} \in \Lambda_{\delta}\left(n\right)\right) \geq 1-\alpha \ . \end{aligned}$$

- Equivalently: Find smallest confidence region $\Lambda_{\delta}\left(n\right)$ at level $1-\alpha$.

Optimal Choice of Uncertainty Size

- Given a confidence level $1-\alpha$ we advocate choosing δ via

$$\begin{aligned}&\min \delta\\ &s.t.\quad P\left(\beta_{*}\in\Lambda_{\delta}\left(n\right)\right)\geq 1-\alpha\ .\end{aligned}$$

- Equivalently: Find smallest confidence region $\Lambda_{\delta}\left(n\right)$ at level $1-\alpha$.
- In simple words: Find the smallest δ so that β_{*} is plausible with confidence level $1-\alpha$.

The Robust Wasserstein Profile Function

- The value $\bar{\beta}(P)$ is characterized by

$$E_P\left(\nabla_\beta\left(Y-\beta^T X\right)^2\right)=2E_P\left(\left(Y-\beta^T X\right)X\right)=0.$$

The Robust Wasserstein Profile Function

- The value $\bar{\beta}(P)$ is characterized by

$$E_P\left(\nabla_\beta\left(Y-\beta^T X\right)^2\right)=2E_P\left(\left(Y-\beta^T X\right)X\right)=0.$$

- Define the *Robust Wasserstein Profile (RWP) Function:*

$$R_n(\beta)=\min\{D_c(P,P_n):E_P\left(\left(Y-\beta^T X\right)X\right)=0\}.$$

The Robust Wasserstein Profile Function

- The value $\bar{\beta}(P)$ is characterized by

$$E_P\left(\nabla_\beta\left(Y-\beta^T X\right)^2\right)=2E_P\left(\left(Y-\beta^T X\right)X\right)=0.$$

- Define the *Robust Wasserstein Profile (RWP) Function:*

$$R_n(\beta)=\min\{D_c(P,P_n):E_P\left(\left(Y-\beta^T X\right)X\right)=0\}.$$

- Note that

$$R_n(\beta_*)\leq\delta\Longleftrightarrow\beta_*\in\Lambda_\delta(n)=\{\bar{\beta}(P):D(P,P_n)\leq\delta\}.$$

The Robust Wasserstein Profile Function

- The value $\bar{\beta}(P)$ is characterized by

$$E_P\left(\nabla_\beta\left(Y-\beta^T X\right)^2\right)=2E_P\left(\left(Y-\beta^T X\right)X\right)=0.$$

- Define the *Robust Wasserstein Profile (RWP) Function:*

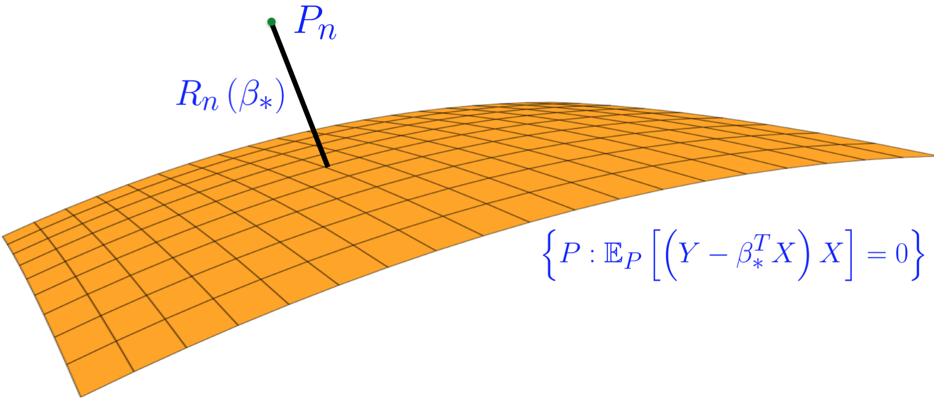
$$R_n(\beta)=\min\{D_c(P,P_n):E_P\left(\left(Y-\beta^T X\right)X\right)=0\}.$$

- Note that

$$R_n(\beta_*)\leq\delta\Longleftrightarrow\beta_*\in\Lambda_\delta(n)=\{\bar{\beta}(P):D(P,P_n)\leq\delta\}.$$

- **So δ is $1-\alpha$ quantile of $R_n(\beta_*)$!**

The Robust Wasserstein Profile Function



Computing Optimal Regularization Parameter

Theorem (B., Murthy, Kang (2016)) *Suppose that* $\{(Y_i, X_i)\}_{i=1}^n$ *is an i.i.d. sample with finite variance, with*

$$c\left((x,y),(x',y')\right) = \begin{cases} \|x - x'\|_q^2 & \text{if} \quad y = y' \\ \infty & \text{if} \quad y \neq y' \end{cases},$$

then

$$nR_n(\beta_*) \Rightarrow L_1,$$

where L_1 *is explicitly (to be computed in one moment)*

$$L_1 \overset{D}{\leq} L_2 := \frac{E[e^2]}{Var(e)} \|N(0, Cov(X))\|_q^2.$$

Remark: We recover same order of regularization (but L_1 gives the optimal constant!)

How to Use this Result?

- Compute η_α the quantile of L_1 (we'll see that L_1 is explicit) – say for $\alpha = .95$.

How to Use this Result?

- Compute η_α the quantile of L_1 (we'll see that L_1 is explicit) – say for $\alpha = .95$.
- The distribution of L_1 will depend on β_* but you can use any consistent plug-in estimator for β_* (same asymptotic convergence holds).

How to Use this Result?

- Compute η_α the quantile of L_1 (we'll see that L_1 is explicit) – say for $\alpha = .95$.
- The distribution of L_1 will depend on β_* but you can use any consistent plug-in estimator for β_* (same asymptotic convergence holds).
- The distribution of L_1 also depends on $Cov\left(X\right)$ but you again can use any consistent plug-in estimator.

How to Use this Result?

- Compute η_α the quantile of L_1 (we'll see that L_1 is explicit) – say for $\alpha = .95$.
- The distribution of L_1 will depend on β_* but you can use any consistent plug-in estimator for β_* (same asymptotic convergence holds).
- The distribution of L_1 also depends on $Cov(X)$ but you again can use any consistent plug-in estimator.
- So, using all of these estimators compute η_α and let $\delta = \eta_\alpha / n$.

Discussion on Optimal Uncertainty Size

- Optimal δ is of order $O\left(1/n\right)$ as opposed to $O\left(1/n^{1/d}\right)$ as advocated in the standard approach.

Discussion on Optimal Uncertainty Size

- Optimal δ is of order $O\left(1/n\right)$ as opposed to $O\left(1/n^{1/d}\right)$ as advocated in the standard approach.
- Note that $R_n\left(\beta_*\right)$ turns out to be parallel to Empirical Likelihood – Owen (1988).

Discussion on Optimal Uncertainty Size

- Optimal δ is of order $O\left(1/n\right)$ as opposed to $O\left(1/n^{1/d}\right)$ as advocated in the standard approach.
- Note that $R_n\left(\beta_*\right)$ turns out to be parallel to Empirical Likelihood – Owen (1988).
- So, although we are using $R_n\left(\beta_*\right)$ to compute optimal uncertainty sizes.

Discussion on Optimal Uncertainty Size

- Optimal δ is of order $O(1/n)$ as opposed to $O\left(1/n^{1/d}\right)$ as advocated in the standard approach.
- Note that $R_n(\beta_*)$ turns out to be parallel to Empirical Likelihood – Owen (1988).
- So, although we are using $R_n(\beta_*)$ to compute optimal uncertainty sizes.
- There is a broader connection to hypothesis testing (applications to *fairness* are explored in **https://arxiv.org/abs/2012.04800**)

Discussion on Optimal Uncertainty Size

- Optimal δ is of order $O(1/n)$ as opposed to $O\left(1/n^{1/d}\right)$ as advocated in the standard approach.
- Note that $R_n(\beta_*)$ turns out to be parallel to Empirical Likelihood – Owen (1988).
- So, although we are using $R_n(\beta_*)$ to compute optimal uncertainty sizes.
- There is a broader connection to hypothesis testing (applications to *fairness* are explored in **https://arxiv.org/abs/2012.04800**)
- Next, we'll see what is L_1 in the more general hypothesis testing setting.

More Generally Projections to Linear Manifolds

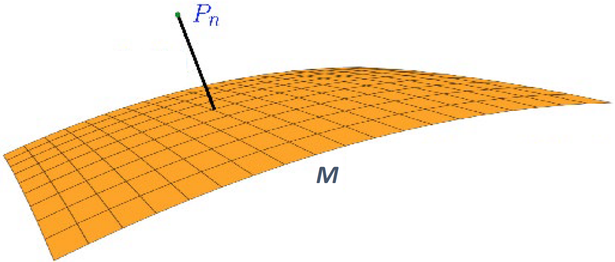
- Let

$$\mathcal{M} = \{P : E_P h_i(X) = 0 \text{ for } i = 1, ..., m\}$$

(i.e. distribution that are similar to P_* based on characteristics h_i)

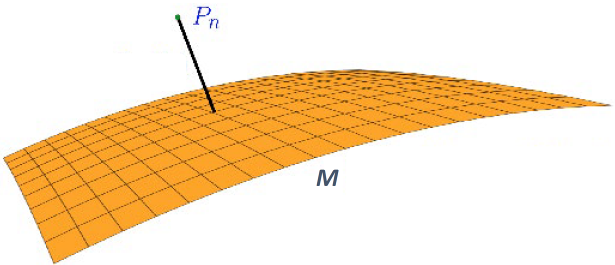
More Generally Projections to Linear Manifolds

- Let
$$\mathcal{M} = \{P : E_P h_i(X) = 0 \text{ for } i = 1, ..., m\}$$
(i.e. distribution that are similar to P_* based on characteristics h_i)
- **We have that** $R_n = D(P_n, \mathcal{M}) = \min\{D(P_n, P) : P \in \mathcal{M}\}$



More Generally Projections to Linear Manifolds

- Let
$$\mathcal{M} = \{P : E_P h_i(X) = 0 \text{ for } i = 1, ..., m\}$$
(i.e. distribution that are similar to P_* based on characteristics h_i)
- **We have that** $R_n = D(P_n, \mathcal{M}) = \min\{D(P_n, P) : P \in \mathcal{M}\}$



- P_n is the empirical measure on some data set.

Duality Results

Theorem (B., Kang, Murthy '19)

Suppose that $c(x,y) \geq 0$ is lower semicontinuous and define $H(x) = (h_1(x), ..., h_m(x))^T \in \mathbb{R}^m$ and suppose that $E_{P_*}(H(X))$ is in the interior of $\{H(x) : x \in \mathbb{R}^d\}$, then

$$R_n = \max_{\lambda \in R^m} \left\{ -E_{P_n}\left(\sup_y \{\lambda^T H(y) - c(X,y)\} \right) \right\}$$

Some Comments on Proof: Finite Support Essential

- Primal:

$$
\begin{aligned}
& \min \int \int c(x,y)\, \pi(dx,dy) \\
\int \int h_i(y)\, \pi(dx,dy) &= 0 \text{ for all } i = 1, \ldots, m. \\
\int \pi(dx,dy) &= P_n(dx)\,; \quad \pi(dx,dy) \geq 0.
\end{aligned}
$$

Some Comments on Proof: Finite Support Essential

- Primal:

$$
\begin{aligned}
& \min \int \int c(x,y)\,\pi(dx,dy) \\
\int \int h_i(y)\,\pi(dx,dy) &= 0 \text{ for all } i = 1, ..., m. \\
\int \pi(dx,dy) &= P_n(dx)\,; \quad \pi(dx,dy) \geq 0.
\end{aligned}
$$

- Dual:

$$
\begin{aligned}
& \sup_{\lambda \in \mathbb{R}^m} E_{P_n} \alpha(X) \\
& \lambda^T H(y) + \alpha(x) \leq c(x,y) \text{ for } x \in \{X_i\}_{i=1}^n,\ y \in \mathbb{R}^d\ .
\end{aligned}
$$

Some Comments on Proof: Finite Support Essential

- Primal:

$$
\begin{aligned}
& \min \int\int c(x,y)\,\pi(dx,dy) \\
\int\int h_i(y)\,\pi(dx,dy) &= 0 \text{ for all } i = 1,\ldots,m. \\
\int \pi(dx,dy) &= P_n(dx)\,; \quad \pi(dx,dy) \geq 0.
\end{aligned}
$$

- Dual:

$$
\begin{aligned}
& \sup_{\lambda \in \mathbb{R}^m} E_{P_n}\alpha(X) \\
& \lambda^T H(y) + \alpha(x) \leq c(x,y) \text{ for } x \in \{X_i\}_{i=1}^n,\ y \in \mathbb{R}^d\ .
\end{aligned}
$$

- Proof technique reduces to problem of moments (finitely many constraints in primal crucial).

Statistics: Limiting Distribution

Theorem (B., Kang, Murthy '19)

Suppose $c(x, y) = \|x - y\|^2$ *for* $r \geq 1$ *(and* $\|z\|_* = \sup_{\|x\| \leq 1} x^T z$ *is the dual norm of* $\|\cdot\|$*). Assume that duality holds and that* $Cov_{P_*}(H(X)) = G$ *exists. Then (under regularity assumptions to be discussed) if* $P_* \in \mathcal{M}$ *(recall* $P_* = P_\infty$ *the data generating distribution)*

$$nR_n \Rightarrow \psi^*(Z) = \sup_\theta [\theta \cdot Z - \psi(\theta)],$$

where $Z \sim N(0, G)$ *and*

$$\psi(\theta) = E_{P_*}\left[\left\|\theta^T DH(X)\right\|_*^2\right].$$

Remark: So, the solution is $\psi^*(Z)$ is a quadratic form of the Gaussian. *Let's study the structure of the projection.*

Intuition and Insights from the Proof

- By defining applying duality

$$R_n = \max_{\lambda}\{-E_{P_n} \max_{\Delta}\left[\lambda^T H\left(X+\Delta\right) - \left\|\Delta\right\|^2\right]\}.$$

Intuition and Insights from the Proof

- By defining applying duality

$$R_n = \max_{\lambda}\{-E_{P_n} \max_{\Delta} \left[\lambda^T H\left(X+\Delta\right) - \|\Delta\|^2\right]\}.$$

- Guessing scalings: $\Delta = O\left(n^{-1/2}\right)$ (since only $O\left(n^{-1/2}\right)$ transport will match constraints by the CLT).

Intuition and Insights from the Proof

- By defining applying duality

$$R_n = \max_{\lambda}\{-E_{P_n} \max_{\Delta} \left[\lambda^T H\left(X+\Delta\right) - \left\|\Delta\right\|^2\right]\}.$$

- Guessing scalings: $\Delta = O\left(n^{-1/2}\right)$ (since only $O\left(n^{-1/2}\right)$ transport will match constraints by the CLT).
- $R_n = O\left(n^{-1}\right)$ because $R_n^{1/2}$=distance to match constraints $=$ $O\left(n^{-1/2}\right)$.

Intuition and Insights from the Proof

- By defining applying duality

$$R_n = \max_{\lambda}\{-E_{P_n} \max_{\Delta} \left[\lambda^T H\left(X + \Delta\right) - \left\|\Delta\right\|^2\right]\}.$$

- Guessing scalings: $\Delta = O\left(n^{-1/2}\right)$ (since only $O\left(n^{-1/2}\right)$ transport will match constraints by the CLT).
- $R_n = O\left(n^{-1}\right)$ because $R_n^{1/2}$=distance to match constraints $=$ $O\left(n^{-1/2}\right)$.
- $\lambda =$ sensitivity with respect to change in constraints $=$ $O\left(n^{-1}/n^{-1/2}\right) = O\left(n^{-1/2}\right)$.

- Substitute $\Delta \leftarrow \Delta/n^{1/2}$:

$$
\begin{aligned}
R_n &= \max_{\lambda}\{-E_{P_n}\max_{\Delta}\left[\lambda^T H\left(X+\Delta/n^{1/2}\right)-\left\|\Delta/n^{1/2}\right\|^2\right]\} \\
&= \max_{\lambda}\{-\lambda^T E_{P_n} H\left(X\right) \\
&\quad -E_{P_n}\max_{\Delta}\left[\lambda^T\left(H\left(X+\Delta/n^{1/2}\right)-H\left(X\right)\right)-\left\|\Delta/n^{1/2}\right\|^2\right]\}.
\end{aligned}
$$

Intuition and Insights from the Proof

- Substitute $\lambda \leftarrow \lambda n^{-1/2}$ and use $H\left(X + \Delta/n^{1/2}\right) - H(X) \approx DH(X)\,\Delta/n^{1/2}$:

$$\begin{aligned} & \max_{\lambda}\{-n^{-1/2}\lambda^{T} E_{P_n}(H(X)) \\ & -E_{P_n}\max_{\Delta}\left[n^{-1}\lambda^{T} DH(X)\,\Delta - n^{-1}\|\Delta\|^{2}\right]\} \\ = \quad & n^{-1/2}\max_{\lambda}\{-n^{1/2}\lambda^{T} E_{P_n}(H(X)) \\ & -E_{P_n}\max_{\Delta}\left[\lambda^{T} DH(X)\,\Delta - \|\Delta\|^{2}\right]\}. \end{aligned}$$

Intuition and Insights from the Proof

- Substitute $\lambda \leftarrow \lambda n^{-1/2}$ and use $H\left(X + \Delta/n^{1/2}\right) - H\left(X\right) \approx DH\left(X\right)\Delta/n^{1/2}$:

$$
\begin{aligned}
& \max_{\lambda}\{-n^{-1/2}\lambda^{T}E_{P_n}\left(H\left(X\right)\right) \\
& -E_{P_n}\max_{\Delta}\left[n^{-1}\lambda^{T}DH\left(X\right)\Delta - n^{-1}\left\|\Delta\right\|^{2}\right]\} \\
= \quad & n^{-1/2}\max_{\lambda}\{-n^{1/2}\lambda^{T}E_{P_n}\left(H\left(X\right)\right) \\
& -E_{P_n}\max_{\Delta}\left[\lambda^{T}DH\left(X\right)\Delta - \left\|\Delta\right\|^{2}\right]\}.
\end{aligned}
$$

- Already can see all the elements in the result (at least formally) since $n^{1/2}\lambda^{T}E_{P_n}H\left(X\right) \Rightarrow \lambda^{T}Z$ (by the CLT).

Intuition and Insights from the Proof

- Conclude by noting

$$
\begin{aligned}
& E_{P_n} \max_{\Delta} \left[\lambda^T DH(X) \Delta - \|\Delta\|^2 \right] \\
= \quad & E_{P_n} \max_{\Delta} \left[\left\| \lambda^T DH(X) \right\|_* \|\Delta\| - \|\Delta\|^2 \right],
\end{aligned}
$$

with $\Delta_{opt}(X)$ dual ("parallel") to $\lambda^T D\bar{H}(X)$ and with $\|\Delta_{opt}(X)\| = 2^{-1} \left\| \lambda^T D\bar{H}(X) \right\|_*$.

Intuition and Insights from the Proof

- Conclude by noting

$$\begin{aligned} & E_{P_n} \max_{\Delta} \left[\lambda^T DH(X) \Delta - \|\Delta\|^2 \right] \\ = \ & E_{P_n} \max_{\Delta} \left[\left\| \lambda^T DH(X) \right\|_* \|\Delta\| - \|\Delta\|^2 \right], \end{aligned}$$

  with $\Delta_{opt}(X)$ dual ("parallel") to $\lambda^T D\bar{H}(X)$ and with $\|\Delta_{opt}(X)\| = 2^{-1} \left\| \lambda^T D\bar{H}(X) \right\|_*$.
- The map $X \rightarrow X + \Delta_{opt}(X) / n^{1/2}$ characterizes the optimal transport projection plan.

Intuition and Insights from the Proof

- Conclude by noting

$$\begin{aligned} & E_{P_n} \max_{\Delta} \left[\lambda^T DH(X) \Delta - \|\Delta\|^2 \right] \\ = \quad & E_{P_n} \max_{\Delta} \left[\left\| \lambda^T DH(X) \right\|_* \|\Delta\| - \|\Delta\|^2 \right], \end{aligned}$$

  with $\Delta_{opt}(X)$ dual ("parallel") to $\lambda^T D\bar{H}(X)$ and with $\|\Delta_{opt}(X)\| = 2^{-1} \left\| \lambda^T D\bar{H}(X) \right\|_*$.

- The map $X \rightarrow X + \Delta_{opt}(X) / n^{1/2}$ characterizes the optimal transport projection plan.
- This provides the elements and the intuition.

Intuition and Insights from the Proof

- Conclude by noting

$$
\begin{aligned}
& E_{P_n} \max_{\Delta} \left[\lambda^T DH(X) \Delta - \|\Delta\|^2 \right] \\
= \quad & E_{P_n} \max_{\Delta} \left[\left\| \lambda^T DH(X) \right\|_* \|\Delta\| - \|\Delta\|^2 \right],
\end{aligned}
$$

  with $\Delta_{opt}(X)$ dual ("parallel") to $\lambda^T D\bar{H}(X)$ and with $\|\Delta_{opt}(X)\| = 2^{-1} \left\| \lambda^T D\bar{H}(X) \right\|_*$.
- The map $X \to X + \Delta_{opt}(X) / n^{1/2}$ characterizes the optimal transport projection plan.
- This provides the elements and the intuition.
- Rigorous analysis requires compactifying over λ.

Infinite Dimensional Case

What about the infinite dimensional case?

Statistics: Limiting Distribution

Theorem (Si, B., Ghosh, Squillante '20)

Suppose $c(x, y) = \|x - y\|_2^2$ *and* $\mathcal{C} = \left\{ f\left(\theta^T x\right) : \theta \in \{\theta_1, ..., \theta_m\}, f \in \mathcal{F} \right\}$. *If domain is compact, under regularity conditions on* $\mathcal{F}$

$$nR_n \Rightarrow L = \sup_{f \in \mathcal{L}(\mathcal{C})} \left[-2Z(f) - E_{P_*}\left(\|Df(X)\|^2\right)\right],$$

where $Z(f)$ *is a Gaussian random field such that* $cov_{P_*}(Z(f), Z(g)) = cov_{P_*}(f(X), g(X))$.

Remark: Regularity condition, it is required that P_* has a density and that $\mathcal{F}$ satisfies

$$\sup_{f \in \mathcal{L}(\mathcal{F})} \frac{\sup_{x \in \Omega} \left| f''\left(\theta_i^T x\right) \right|^2}{\int_\Omega \left(f'\left(\theta_i^T z\right)\right)^2 dz} < \infty.$$

Comments

- Proof follows same elements as finite dimensional case (the compactification step is more involved).

Comments

- Proof follows same elements as finite dimensional case (the compactification step is more involved).
- Natural connection to a Poincaré inequality of the form

$$Var_{P_*}(f(X)) \leq cE_{P_*}\left(\|Df(X)\|^2\right)$$

arises naturally in the limit.

Asymptotic Normality

Once we know how to choose the size
of the uncertainty optimally
we can obtain asymptotically optimal estimators

Theorem (B., Murthy, Si (2019) https://arxiv.org/pdf/1906.01614.pdf)

Assume that $\{X_i : 1 \leq i \leq n\}$ *is an i.i.d. sample from* P_*. *Suppose* $l(\cdot)$ *is twice differentiable,* $l(x, \cdot)$ *convex,* $C = E\left(D_\beta^2 l(X, \beta_*)\right) \succ 0$ *(where* $\beta_* = \arg\min E_P(l(X, \beta))$*), then, with* $\delta_n^* = \eta / n$

$$
\begin{aligned}
n^{1/2}\left(\beta_n^{DRO}(0) - \beta_*\right) &\Rightarrow C^{-1} Z_0 \\
n^{1/2}\left(\beta_n^{DRO}(\delta_n^*) - \beta_n^{ERM}\right) &\Rightarrow \nabla v(\beta),
\end{aligned}
$$

Remark: *Recall* $Z_0 \sim N\left(0, Cov\left(D_\beta l(X, \beta_*)\right)\right)$ and $v(\beta) = \eta^{1/2} E_{P_n}^{1/2} \|D_x l(X, \beta)\|_q^2$

A Proof Sketch: Duality + Asymptotic Statistics

- Recall the duality result with $\delta_n = \eta / n$

$$\begin{aligned} & \max_{D(P,P_n)\leq\delta_n} E_P\left(l\left(X,\beta\right)\right) \\ = \quad & \max_{\lambda}\{\frac{\lambda\eta}{n} + E_{P_n} \max_{\Delta}\{l\left(X+\Delta,\beta\right) - \lambda\left\|\Delta\right\|_p^2\}. \end{aligned}$$

A Proof Sketch: Duality + Asymptotic Statistics

- Recall the duality result with $\delta_n = \eta/n$

$$\max_{D(P,P_n)\leq\delta_n} E_P\left(l\left(X,\beta\right)\right)$$
$$= \max_{\lambda}\{\frac{\lambda\eta}{n} + E_{P_n}\max_{\Delta}\{l\left(X+\Delta,\beta\right) - \lambda\left\|\Delta\right\|_p^2\}.$$

- Similar scaling as before: $\Delta \to \Delta/n^{1/2}$, $\lambda \to \lambda n^{1/2}$

$$\max_{\lambda}\left\{\frac{\lambda\eta}{n^{1/2}} + E_{P_n}\max_{\Delta}\left\{l\left(X+\frac{\Delta}{n^{1/2}},\beta\right) - \frac{\lambda}{n^{1/2}}\left\|\Delta\right\|_p^2\right\}\right\}$$
$$\approx E_{P_n} l\left(X,\beta\right) + n^{-1/2}\eta^{1/2}E_{P_n}^{1/2}\left\|D_x l\left(X,\beta\right)\right\|_q^2.$$

A Proof Sketch: Duality + Asymptotic Statistics

- Recall the duality result with $\delta_n = \eta/n$

$$\begin{aligned} & \max_{D(P,P_n)\leq\delta_n} E_P\left(l\left(X,\beta\right)\right) \\ = \quad & \max_{\lambda}\{\frac{\lambda\eta}{n} + E_{P_n}\max_{\Delta}\{l\left(X+\Delta,\beta\right) - \lambda\left\|\Delta\right\|_p^2\}. \end{aligned}$$

- Similar scaling as before: $\Delta \rightarrow \Delta/n^{1/2}$, $\lambda \rightarrow \lambda n^{1/2}$

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

- From this form, it is easy to guess the result...

A Proof Sketch: Duality + Asymptotic Statistics

- Recall the duality result with $\delta_n = \eta / n$

$$\begin{aligned} & \max_{D(P,P_n)\leq\delta_n} E_P\left(l\left(X,\beta\right)\right) \\ = \quad & \max_{\lambda}\{\frac{\lambda\eta}{n} + E_{P_n} \max_{\Delta}\{l\left(X+\Delta,\beta\right) - \lambda \left\|\Delta\right\|_p^2\}. \end{aligned}$$

- Similar scaling as before: $\Delta \rightarrow \Delta / n^{1/2}$, $\lambda \rightarrow \lambda n^{1/2}$

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

- From this form, it is easy to guess the result...
- Worst case adversary: $\Delta_{opt}\left(X_i\right)$ is parallel to $D_x l\left(X,\beta\right)$ & $\left\|\Delta_{opt}\left(X_i\right)\right\|_p = \left\|D_x l\left(X,\beta\right)\right\|_q / (2\lambda)$

Remember the Key Confidence Region?

- $\Lambda_{\delta_n^*}(n) = \{\bar{\beta}(P) = \arg\{\min E_P[l(X,\beta)] : D(P, P_n) \leq \delta_n^*\}$

Remember the Key Confidence Region?

- $\Lambda_{\delta_n^*}(n) = \{\bar{\beta}(P) = \arg\{\min E_P [l(X, \beta)] : D(P, P_n) \leq \delta_n^*\}$
- $\Lambda_{\delta_n^*}(n)$ **is the natural DRO confidence region & has desired coverage.**

Remember the Key Confidence Region?

- $\Lambda_{\delta_n^*}(n) = \{\bar{\beta}(P) = \arg\{\min E_P[l(X,\beta)] : D(P, P_n) \leq \delta_n^*\}$
- $\Lambda_{\delta_n^*}(n)$ **is the natural DRO confidence region & has desired coverage.**
- $\Lambda_{\delta_n^*}(n)$ **contains both the ERM solution (i.e. $\delta = 0$) and β_n^{DRO}.**

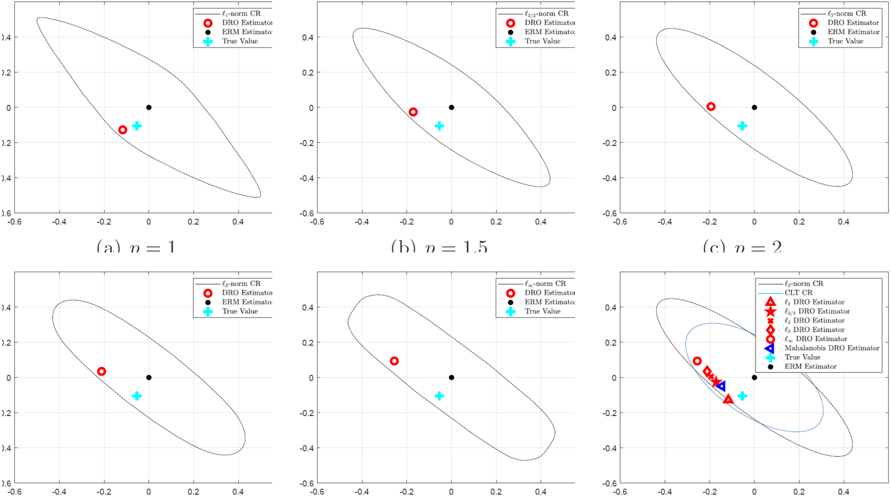
Remember the Key Confidence Region?

- $\Lambda_{\delta_n^*}(n) = \{\bar{\beta}(P) = \arg\{\min E_P[l(X,\beta)] : D(P, P_n) \leq \delta_n^*\}$
- $\Lambda_{\delta_n^*}(n)$ **is the natural DRO confidence region & has desired coverage.**
- $\Lambda_{\delta_n^*}(n)$ **contains both the ERM solution (i.e.** $\delta = 0$**) and** β_n^{DRO}**.**
- *Standard CLT confidence region does not necessarily contain* β_n^{DRO}*.*

Remember the Key Confidence Region?

- $\Lambda_{\delta_n^*}(n) = \{\bar{\beta}(P) = \arg\{\min E_P[l(X,\beta)] : D(P,P_n) \leq \delta_n^*\}$
- $\Lambda_{\delta_n^*}(n)$ **is the natural DRO confidence region & has desired coverage.**
- $\Lambda_{\delta_n^*}(n)$ **contains both the ERM solution (i.e.** $\delta = 0$**) and** β_n^{DRO}**.**
- *Standard CLT confidence region does not necessarily contain* β_n^{DRO}*.*
- $\Lambda_{\delta_n^*}(n) \approx C^{-1}Z_0 + \Lambda_\eta$ and $\Lambda_\eta = \{u : \psi^*(Cu) \leq \eta\}$

Geometry of Confidence Region?



(a) $p = 1$ (b) $p = 1.5$ (c) $p = 2$

(d) $p = 3$ (e) $p = \infty$ (f) CLT

Containment of the DRO Solution

- The fact that
$$\beta_n^{DRO} \in \Lambda_{\delta_n^*}(n).$$
is non-obvious.

Containment of the DRO Solution

- The fact that
$$\beta_n^{DRO} \in \Lambda_{\delta_n^*}(n).$$
is non-obvious.
- *It follows from the following duality result in* **B., Murthy and Si (2019) https://arxiv.org/pdf/1906.01614.pdf**
$$\inf_{\beta} \sup_{D(P,P_n)\leq\delta} E_P l(X,\beta) = \sup_{D(P,P_n)\leq\delta} \inf_{\beta} E_P l(X,\beta).$$

Standard CLT May Not Contain the DRO Solution

TABLE 1. Coverage Probability

β_0	ρ	ℓ_2 DRO confidence region		CLT confidence region	
		Coverage for β_n^{DRO}	Coverage for β_*	Coverage for β_n^{DRO}	Coverage for β_*
$\begin{bmatrix}0.5\\0.5\end{bmatrix}$	0.95	100.0%	94.5%	99.4%	94.6%
	0	100.0%	94.0%	97.1%	93.5%
	-0.95	100.0%	94.8%	75.8%	94.4%
$\begin{bmatrix}1.0\\0.0\end{bmatrix}$	0.95	100.0%	94.6%	93.7%	95.4%
	0	100.0%	94.6%	100%	94.1%
	-0.95	100.0%	95.3%	91.2%	94.9%

Summary Day 3

- Theory for optimal choice of uncertainty size in Wasserstein DRO.

Summary Day 3

- Theory for optimal choice of uncertainty size in Wasserstein DRO.
- Asymptotic normality of DRO results given optimal uncertainty size.

Summary Day 3

- Theory for optimal choice of uncertainty size in Wasserstein DRO.
- Asymptotic normality of DRO results given optimal uncertainty size.
- Existence of Nash equilibrium value in Wasserstein DRO.

Summary Day 3

- Theory for optimal choice of uncertainty size in Wasserstein DRO.
- Asymptotic normality of DRO results given optimal uncertainty size.
- Existence of Nash equilibrium value in Wasserstein DRO.
- Structure of the Nash equilibrium.

Summary Day 3

- Theory for optimal choice of uncertainty size in Wasserstein DRO.
- Asymptotic normality of DRO results given optimal uncertainty size.
- Existence of Nash equilibrium value in Wasserstein DRO.
- Structure of the Nash equilibrium.
- Connections to interesting projection problem $R_n = D(P_n, \mathcal{M})$:

$$nD(P_n, \mathcal{M}) \Rightarrow L.$$