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.

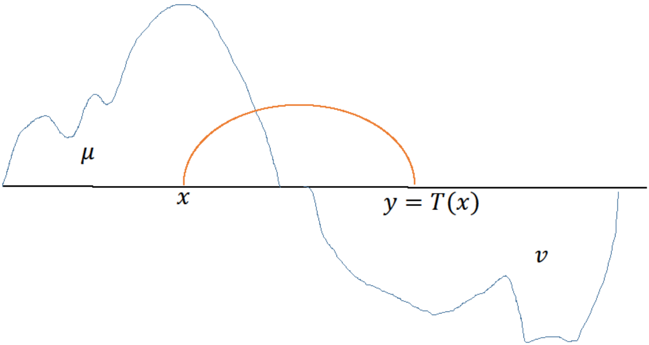
Agenda

- 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):

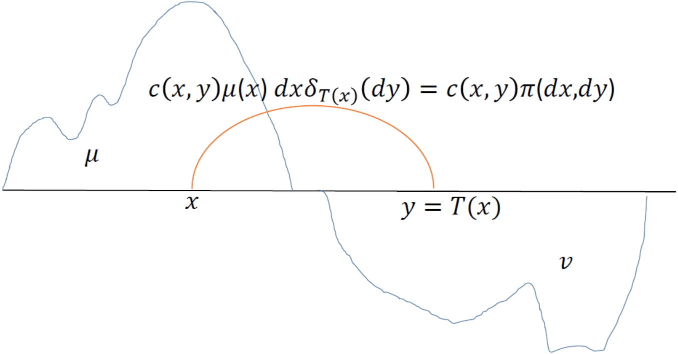
$$\begin{aligned} D_c(\mu, v) \ &: = \min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \pi(dx, dy) \\ & \int_{\mathcal{Y}} \pi(dx, dy) = \mu(dx), \int_{\mathcal{X}} \pi(dx, dy) = v(dy). \end{aligned}$$

- 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:

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

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

Kantorovich Relaxation: Primal Problem

- Keep in mind primal:

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

- 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:

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

- 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 $D_d(v, w)$.

- 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 $D_d(\mu, w)$.

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:

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

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

- 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

$\alpha^*(x) + \beta^*(y) = c(x, y)$ - π^* a.s. <– 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

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

Proof Technique: Sketch of Strong Duality

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

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

- 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

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

- 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.*

Application of Optimal Transport in Economics

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

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

- 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

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

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

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 $\partial^2_{x,y}\Psi(x,y) \geq 0$ - supermodularity:

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

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).

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

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

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

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

- 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 $\Psi(x,y)\rightarrow\Psi(x,y)+f(x)$? (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.

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 $\beta^d(\cdot)$ is 1-Lipschitz

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

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,
$$\beta^{d}(x) := \inf_{y}\{\beta(y) + d(x,y)\} \leq \beta(x)$$
and if β is 1-Lipschitz (meaning $|\beta(x) - \beta(y)| \leq d(x,y)$) then
$$\begin{aligned} \beta^{d}(x) - \beta(x) &= \inf_{y}\{d(x,y) + \beta(y) - \beta(x)\} \\ &\geq \inf_{y}\{d(x,y) - d(x,y)\} = 0. \end{aligned}$$

- 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:

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

- Simplified Dual (called Kantorovich duality result):

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

- 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 $c(x, y) = \|x - y\|_2^2 / 2$ 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

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

- The dual is

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

- 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

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

- 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

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

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

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

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

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

- But given x, equality holds if and only if $y\in\partial a\left(x\right)$ <– 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

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

- 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

$$E_{P_{true}}(f(X))$$

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 $E_{P_0}|f(X)| < \infty$

$$\sup E_P(f(Y))$$
$$D_c(P, P_0) \leq \delta ,$$

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

- Also an infinite dimensional linear program

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

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:

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

- 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:

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

- 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

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

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

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

- 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

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

  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.

Connections to Distributionally Robust Optimization

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

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

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)

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

- 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 $R(t)$...
- **Optimal transport is a natural option!**

Application 1: Back to Classical Risk Problem

- Suppose that

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

Application 1: Back to Classical Risk Problem

- Suppose that

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

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

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

Application 1: Back to Classical Risk Problem

- Suppose that

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

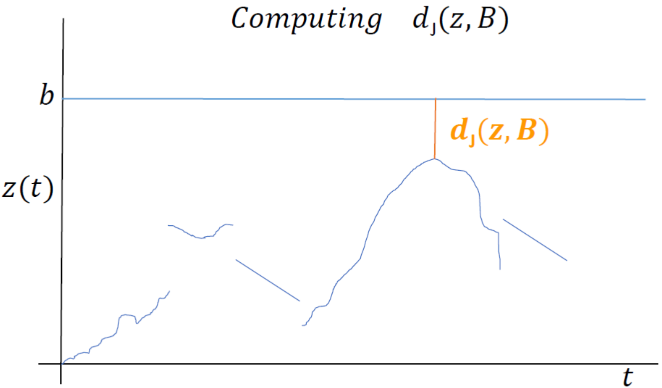
- If $R(t) = b - Z(t)$, then ruin during time interval $[0,1]$ is

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

- **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

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

Application 1: Computing Uncertainty Size

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

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

- 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

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

- 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

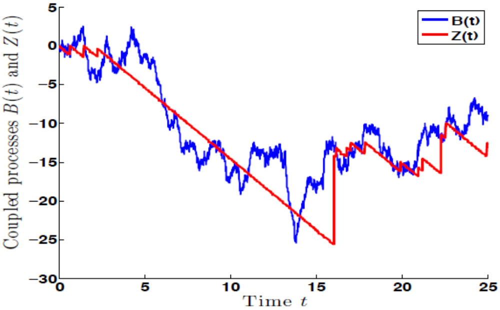
$$\tilde{S}(t) := \sum_{j>0} \sup_{\tau_j \leq s \leq t} B(s) \mathbf{1}(\tau_j \leq t < \tau_{j+1}) \text{ and } \tilde{N}(t) := \sum_{j \geq 0} \Psi_j \mathbf{1}(\tau_j \leq t < \tau_{j+1}).$$

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).

FIGURE 4. A coupled path output by Algorithm 1



Application 1: Numerical Example

- Assume Poisson arrivals.
- *Pareto claim sizes with index* **2.2** – $\left(P\left(V > t\right) = 1/\left(1+t\right)^{2.2}\right)$.
- Cost $c\left(x,y\right) = d_J\left(x,y\right)^2$ <– 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

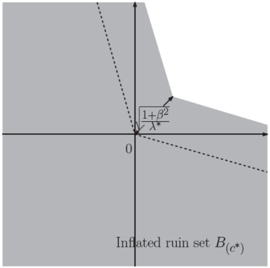
- See also Birghila, Aigner, Engelke (2021)

Additional Applications: Multidimensional Ruin Problems

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

Additional Applications: Multidimensional Ruin Problems

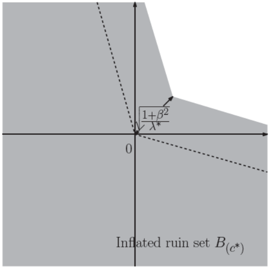
- **https://arxiv.org/abs/1604.01446** contains more applications.
- Control: $\min_\theta \sup_{P:D(P,P_0)\leq\delta} E[L(\theta, Z)]$ <– robust optimal reinsurance.



(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.

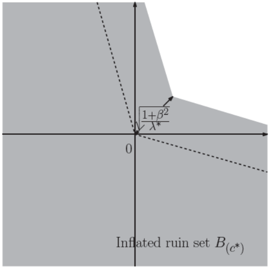


(b)Computation of worst-case ruin using the baseline measure

- 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.

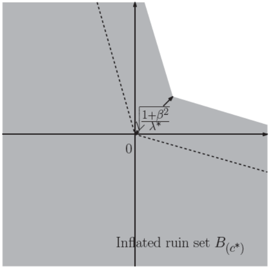


(b)Computation of worst-case ruin using the baseline measure

- 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).

A Bit of Background on Online Advertising

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

User requests publisher's page (say NYT) → Exchange bids to DSPs on behalf of publishers → DSPs send bids to Exchange on behalf of advertisers

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.

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.

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.

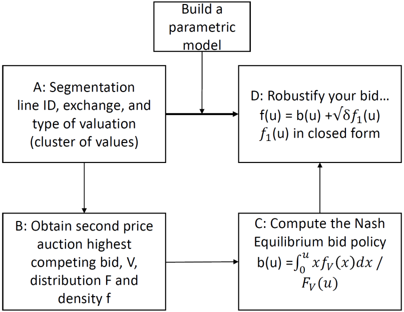
A Bit of Background on Online Advertising

- 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?

Transfer Information and Mitigation of Model Error

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:

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

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

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:

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

  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 $i \neq j$

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

Model and Performance Measure

- A Simplified Model:

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

  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 $i \neq j$

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

- Under homogeneity it suffices to solve

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

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

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

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

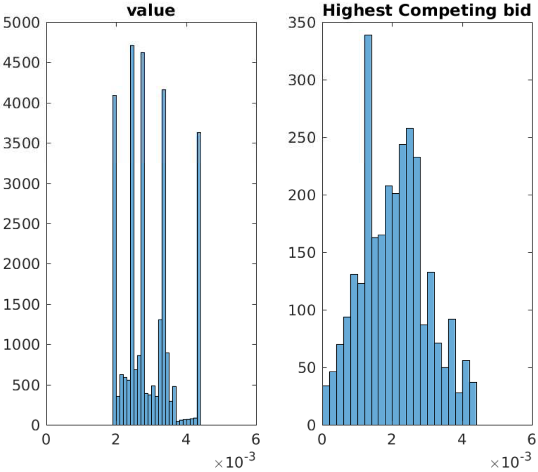
$$b = u - F_{V|U=u}(b) \,/\, f_{V|U=u}(b) \,.$$

- *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

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

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

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
$$\max_{D(F,F_V)\leq\delta} \bar{F}(b) = \max_{D(F,F_V)\leq\delta} P_F(V > b) = P_F(V > b - \lambda_b)\,.$$

Quantifying Model Mispecifications

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

Quantifying Model Mispecifications

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

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

- 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 $\delta \rightarrow 0$.
- 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

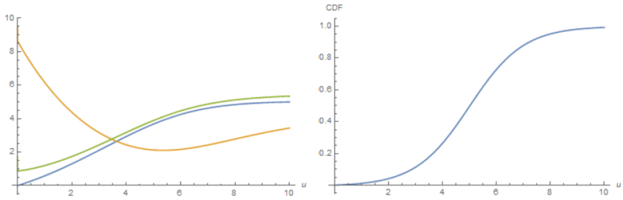
Example

- 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.

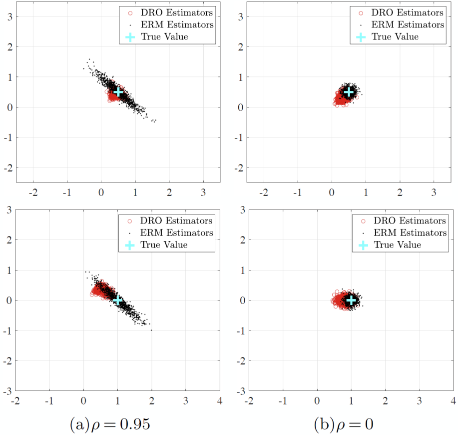
- Optimal Least Squares approach consists in estimating β_* via

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

- Apply the distributionally robust estimator based on optimal transport.

Applying Distributionally Robust Optimization in Linear Regression

Estimation of θ_* with DRO (○) and without DRO (∘)



Connection to Sqrt-Lasso

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

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

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

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

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

Logistic Regression

- Classical classification model:

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

Logistic Regression

- Classical classification model:

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

- 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).

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,

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

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**

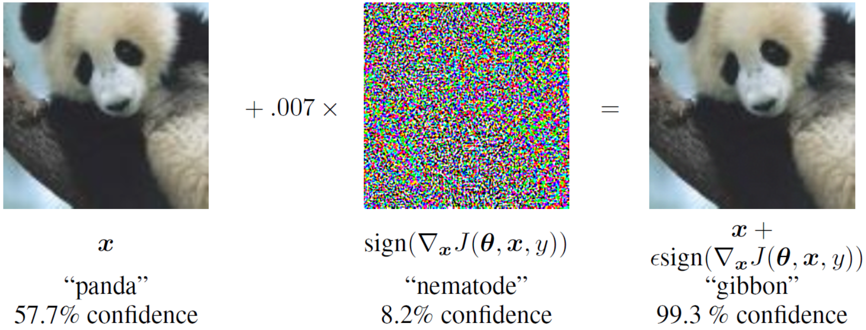
Unification and Extensions of Regularized Estimators

- 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).

- 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

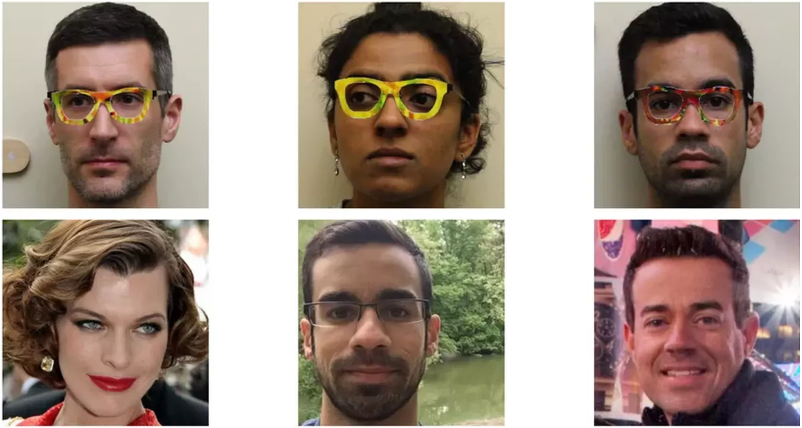
Deep Neural Networks: Adversarial Attacks

- 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


GETTY IMAGES

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...

How Regularization and Dual Norms Arise?

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

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

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

How Regularization and Dual Norms Arise?

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

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

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

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

How Regularization and Dual Norms Arise?

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

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

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

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

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

- 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
$c\left(\left(x,y\right),\left(x',y'\right)\right)=\left\|x-x'\right\|_{q}I\left(y=y'\right)+\infty I\left(y\neq y'\right)$

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

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

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

Summary

- The worst case perturbation is given by Δ such that

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

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

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

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

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

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

- This will automatically protect against attacks.
- 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,

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

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,

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

- *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,

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

- *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,

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

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,

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

- *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,

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

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

Connections to Statistical Analysis

Based on:

Robust Wasserstein Profile Inference (B., Murthy & Kang '16)

https://arxiv.org/abs/1610.05627

Highlight: How to choose size of uncertainty?

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:

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

Towards an Optimal Choice of Uncertainty Size

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

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

- 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:

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

- Use left hand side to define a statistical principle to choose δ.
- 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 $D(P_{true}, P_n)$?
- This was advocated and seems natural at first sight... but there is a big problem.
- Consider the case $c(x, x') = \|x - x'\|_\infty$ by Kantorovich-Rubinstein duality

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

Towards an Optimal Choice of Uncertainty Size

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

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

- The analysis of this object is extensively studied in the theory of Empirical Processes.

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

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

- The analysis of this object is extensively studied in the theory of Empirical Processes.
- 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)

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

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

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

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

- 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)

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

- 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 $d > 2$.
- So, instead, we'll focus on an optimal (in some sense to be explained) approach.

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

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

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 $\bar{\beta}(P) = \arg\min E_P\left(Y - \beta^T X\right)$.

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

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

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

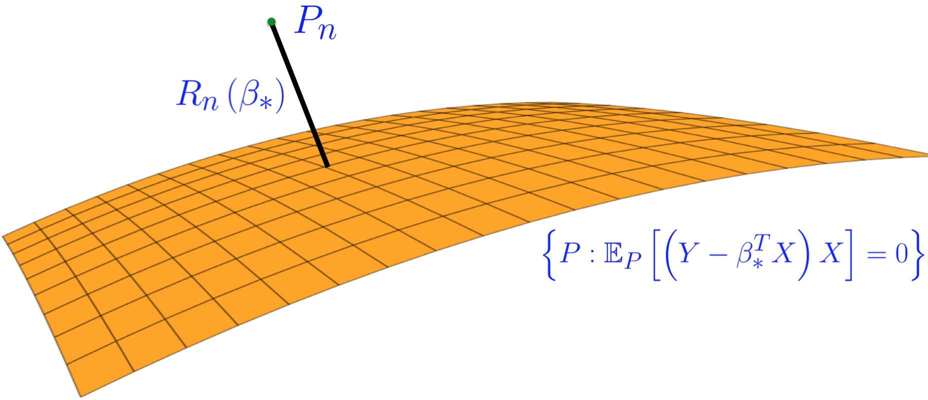
$$R_n(\beta)=\min\{D_c(P,P_n): E_P\left(\left(Y-\beta^T X\right)X\right)=0\}.$$

- 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 and*

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

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

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.
- We characterize the asymptotic constant (not only order) in optimal regularization:

$$P\left(L_1 \leq \eta_{1-\alpha}\right) = 1 - \alpha.$$

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.
- We characterize the asymptotic constant (not only order) in optimal regularization:
$$P\left(L_1 \leq \eta_{1-\alpha}\right) = 1 - \alpha.$$
- $R_n\left(\beta_*\right)$ is inspired by Empirical Likelihood – Owen (1988).

A Toy Example Illustrating Proof Techniques

- Consider

$$\min_{\beta} \max_{P:\mathcal{D}_c(P,P_n)\leq\delta} E\left[(Y-\beta)^2\right]$$

with $c(y,y') = (y-y')^{\rho}$ and define

$$\begin{aligned} R_n(\beta) &= \min_{\pi(dy,du)\geq 0} \int (y-u)^{\rho}\, \pi(dy,du) : \\ & \int_{u\in\mathbb{R}} \pi(dy,du) = \frac{1}{n}\delta_{\{Y_i\}}(dy) \;\; \forall i, \\ & 2\int\int (u-\beta)\, \pi(dy,du) = 0. \end{aligned}$$

A Toy Example Illustrating Proof Techniques

- Dual linear programming problem: Plug in $\beta = \beta_*$

$$
\begin{aligned}
R_n(\beta_*) &= \sup_{\lambda \in \mathbb{R}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sup_{u \in \mathbb{R}} \left\{ \lambda(u - \beta_*) - |Y_i - u|^{\rho} \right\} \right\} \\
&= \sup_{\lambda \in \mathbb{R}} \left\{ \begin{array}{c} -\frac{\lambda}{n} \sum_{i=1}^{n} (Y_i - \beta_*) \\ -\frac{1}{n} \sum_{i=1}^{n} \sup_{u \in \mathbb{R}} \left\{ \lambda(u - Y_i) - |Y_i - u|^{\rho} \right\} \end{array} \right\} \\
&= \sup_{\lambda} \left\{ -\frac{\lambda}{n} \sum_{i=1}^{n} (Y_i - \beta_*) - (\rho - 1) \left| \frac{\lambda}{\rho} \right|^{\frac{\rho}{\rho-1}} \right\} \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_*) \right|^{\rho} = \frac{1}{n^{1/2}} \left| N(0, \sigma^2) \right|^{\rho}.
\end{aligned}
$$

Computational Tools

- Fast computation of Optimal Transport Distances is an active topic of research currently.

Computational Tools

- Fast computation of Optimal Transport Distances is an active topic of research currently.
- Fastest algorithm for solving

$$\min_{\pi(i,j)\geq 0} \sum c(i,j)\,\pi(i,j)$$

$$\sum_{j=1}^{n} \pi(i,j) = \mu(i) \quad \forall i\ ,\ \sum_{i=1}^{n} \pi(i,j) = v(j) \quad \forall j$$

is in Quanrud '18 (arxiv/1810.05957B) and B., Kent, Jambulapati, Kent, Sidford '18, (arxiv/1810.07717) it runs in $O\left(n^2 \|c\|_\infty / \varepsilon\right)$ time for ε-additive error.

Computational Tools

- Fast computation of Optimal Transport Distances is an active topic of research currently.
- Fastest algorithm for solving

$$\min_{\pi(i,j)\geq 0} \sum c\left(i,j\right)\pi\left(i,j\right)$$
$$\sum_{j=1}^{n} \pi\left(i,j\right) = \mu\left(i\right) \quad \forall i\ , \sum_{i=1}^{n} \pi\left(i,j\right) = v\left(j\right) \quad \forall j$$

  is in Quanrud '18 (arxiv/1810.05957B) and B., Kent, Jambulapati, Kent, Sidford '18, (arxiv/1810.07717) it runs in $O\left(n^2 \left\|c\right\|_\infty / \varepsilon\right)$ time for ε-additive error.
- Optimal complexity algorithms for continuous problems is still an open problem.

Discussion: Some Open Problems

- Extensions: Optimal Transport with constrains, Optimal Martingale Transport.

Discussion: Some Open Problems

- Extensions: Optimal Transport with constrains, Optimal Martingale Transport.
- Computational methods: Typical approach is entropic regularization (new methods currently developed in the machine learning literature).

Conclusions

- Optimal transport (OT) is a powerful tool based on linear programming.

Conclusions

- Optimal transport (OT) is a powerful tool based on linear programming.
- Arises in Economics, Machine Learning, Operations Research, Statistics, etc.

Conclusions

- Optimal transport (OT) is a powerful tool based on linear programming.
- Arises in Economics, Machine Learning, Operations Research, Statistics, etc.
- OT can be used in path-space to quantify model error.

Conclusions

- Optimal transport (OT) is a powerful tool based on linear programming.
- Arises in Economics, Machine Learning, Operations Research, Statistics, etc.
- OT can be used in path-space to quantify model error.
- OT can be used for data-driven distributionally robust optimization.

Conclusions

- Optimal transport (OT) is a powerful tool based on linear programming.
- Arises in Economics, Machine Learning, Operations Research, Statistics, etc.
- OT can be used in path-space to quantify model error.
- OT can be used for data-driven distributionally robust optimization.
- Cost function in OT can be used to improve out-of-sample performance.

Conclusions

- Optimal transport (OT) is a powerful tool based on linear programming.
- Arises in Economics, Machine Learning, Operations Research, Statistics, etc.
- OT can be used in path-space to quantify model error.
- OT can be used for data-driven distributionally robust optimization.
- Cost function in OT can be used to improve out-of-sample performance.
- OT can be used for statistical inference using RWP function.