Optimal Transport Methods in Operations Research and Statistics

Jose Blanchet (based on work with F. He, Y. Kang, K. Murthy, F. Zhang).

Stanford University (Management Science and Engineering), and Columbia University (Department of Statistics and Department of IEOR).

Goal:

Goal: Introduce optimal transport techniques and applications in OR & Statistics

Optimal transport is useful tool in model robustness, equilibrium, and machine learning!

Agenda

- Introduction to Optimal Transport

Agenda

- Introduction to Optimal Transport
- Economic Interpretations and Wasserstein Distances

Agenda

- Introduction to Optimal Transport
- Economic Interpretations and Wasserstein Distances
- Applications in Stochastic Operations Research

Agenda

- Introduction to Optimal Transport
- Economic Interpretations and Wasserstein Distances
- Applications in Stochastic Operations Research
- Applications in Distributionally Robust Optimization

Agenda

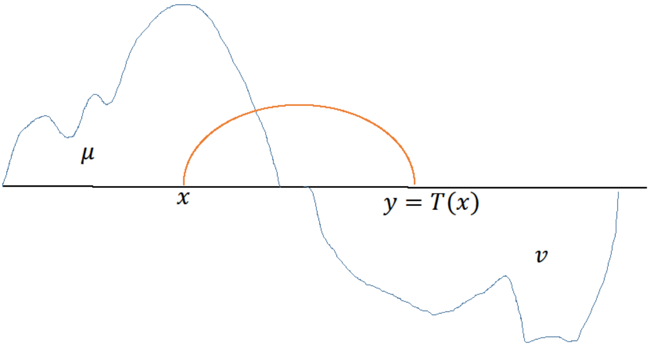
- Introduction to Optimal Transport
- Economic Interpretations and Wasserstein Distances
- Applications in Stochastic Operations Research
- Applications in Distributionally Robust Optimization
- Applications in Statistics

Introduction to Optimal Transport

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):

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

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

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):

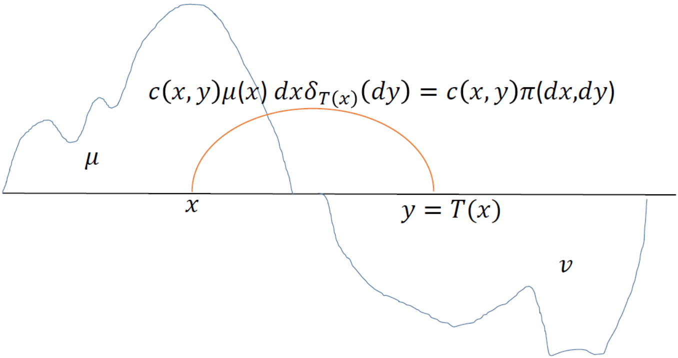
$$\begin{aligned} D_c(\mu, v) \quad : \quad &= \min_{\pi(dx,dy)\geq 0} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\, \pi(dx, dy) \\ &\int_{\mathcal{Y}} \pi(dx, dy) = \mu(dx), \int_{\mathcal{X}} \pi(dx, dy) = v(dy). \end{aligned}$$

- If $c(x, y) = d^p(x, y)$ (d-metric) then $D_c^{1/p}(\mu, v)$ is a p-Wasserstein metric.

Illustration of Optimal Transport Costs

- Monge's solution would take the form

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



Kantorovich Relaxation: Dual Problem

- Primal has always a solution for $c(\cdot) \geq 0$ lower semicontinuous.

Kantorovich Relaxation: Dual Problem

- Primal has always a solution for $c(\cdot) \geq 0$ lower semicontinuous.
- Linear programming (Dual):

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

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

Kantorovich Relaxation: Dual Problem

- Primal has always a solution for $c(\cdot) \geq 0$ lower semicontinuous.
- Linear programming (Dual):

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

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

- Dual α and β can be taken over continuous functions.

Kantorovich Relaxation: Dual Problem

- Primal has always a solution for $c(\cdot) \geq 0$ lower semicontinuous.
- Linear programming (Dual):

$$\sup_{\alpha,\beta} \int_{\mathcal{X}} \alpha(x)\,\mu(dx) + \int_{\mathcal{Y}} \beta(y)\,v(dy)$$
$$\alpha(x) + \beta(y) \leq c(x,y) \;\; \forall (x,y) \in \mathcal{X} \times \mathcal{Y}\,.$$

- Dual α and β can be taken over continuous functions.
- Complementary slackness: Equality holds on the support of π^* (primal optimizer).

Kantorovich Relaxation: Primal Interpretation

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

Kantorovich Relaxation: Primal Interpretation

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

Kantorovich Relaxation: Primal Interpretation

- John wants to remove of a pile of sand, $\mu(\cdot)$.
- Peter wants to cover a sinkhole, $v(\cdot)$.
- Cost for John and Peter 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

- John wants to remove of a pile of sand, $\mu(\cdot)$.
- Peter wants to cover a sinkhole, $v(\cdot)$.
- Cost for John and Peter 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 Maria, who has a business...

Kantorovich Relaxation: Primal Interpretation

- John wants to remove of a pile of sand, $\mu(\cdot)$.
- Peter wants to cover a sinkhole, $v(\cdot)$.
- Cost for John and Peter 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 Maria, who has a business...
- Maria promises to transport on behalf of John and Peter the whole amount.

Kantorovich Relaxation: Primal Interpretation

- Maria charges John $\alpha(x)$ per-unit of mass at x (similarly to Peter).

Kantorovich Relaxation: Primal Interpretation

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

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

Kantorovich Relaxation: Primal Interpretation

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

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

- Maria wishes to maximize her profit

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

Kantorovich Relaxation: Primal Interpretation

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

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

- Maria 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 (under mild regularity)

$$\begin{aligned} \alpha^*(x) &= \inf_y \{c(x, y) - \beta^*(y)\} \\ \beta^*(y) &= \inf_x \{c(x, y) - \alpha^*(x)\}. \end{aligned}$$

Proof Techniques

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

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

Proof Techniques

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

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

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

Proof Techniques

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

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

- Swap sup and inf using Sion's min-max theorem by a compactness argument and conclude.
- *Significant amount of work needed to extend to general Polish spaces and construct the dual optimizers (primal a bit easier).*

Optimal Transport Applications

Optimal Transport has gained popularity in many areas including: image analysis, economics, statistics, machine learning...

The rest of the talk mostly concerns applications to OR and Statistics but we'll briefly touch upon others, including economics...

Illustration of Optimal Transport in Image Analysis

- Santambrogio (2010)'s illustration

Application of Optimal Transport in Economics

Economic Interpretations
(see e.g. A. Galichon's 2016 textbook & McCaan 2013 notes).

Applications in Labor Markets

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

Applications in Labor Markets

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

Applications in Labor Markets

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

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

- Companies want to *minimize* total production cost

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

Applications in Labor Markets

- Letting a central planner organize the Labor market

Applications in Labor Markets

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

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

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.

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

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.

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

- Consider

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

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

Solving for Optimal Transport Coupling

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

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

- Consider

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

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

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

Solving for Optimal Transport Coupling

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

Solving for Optimal Transport Coupling

- Think of $Y_j^n = -\log\left(1 - U_j^n\right)$ 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 $Y_j^n = -\log\left(1 - U_j^n\right)$ 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 $n \longrightarrow \infty$, $X_{(nt)}^n \longrightarrow t$, so $Y_{(nt)}^n \longrightarrow -\log(1-t)$.

Solving for Optimal Transport Coupling

- Think of $Y_j^n = -\log\left(1 - U_j^n\right)$ 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 $n \longrightarrow \infty$, $X_{(nt)}^n \longrightarrow t$, so $Y_{(nt)}^n \longrightarrow -\log(1-t)$.
- Thus, the optimal coupling as $n \longrightarrow \infty$ is $X = U$ and $Y = -\log(1-U)$ (comonotonic coupling).

Identities for Wasserstein Distances

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

Identities for Wasserstein Distances

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

$$D_c(F_\mu, F_v) = \int_0^1 \left| F_\mu^{-1}(u) - F_v^{-1}(u) \right| du.$$

Identities for Wasserstein Distances

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

$$D_c(F_\mu, F_v) = \int_0^1 \left| F_\mu^{-1}(u) - F_v^{-1}(u) \right| du.$$

- Similar identities are common for Wasserstein distances...

Interesting Insight on Salary Effects

- In equilibrium, by the envelope theorem

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

Interesting Insight on Salary Effects

- In equilibrium, by the envelope theorem

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

- 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

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

- 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 grows).

Interesting Insight on Salary Effects

- In equilibrium, by the envelope theorem

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

- 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 grows).
- *Answer: salaries grows if* $f(\cdot)$ *is increasing.*

Application of Optimal Transport in Stochastic OR
Blanchet and Murthy (2016)
https://arxiv.org/abs/1604.01446.

Insight: Diffusion approximations and optimal transport

A Distributionally Robust Performance Analysis

- In Stochastic OR we are often interested in evaluating

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

for a complex model P_{true}

A Distributionally Robust Performance Analysis

- In Stochastic OR we are often interested in evaluating

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

  for a complex model P_{true}
- Moreover, we wish to control / optimize it

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

A Distributionally Robust Performance Analysis

- In Stochastic OR we are often interested in evaluating

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

  for a complex model P_{true}
- Moreover, we wish to control / optimize it

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

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

A Distributionally Robust Performance Analysis

- In Stochastic OR we are often interested in evaluating

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

  for a complex model P_{true}

- Moreover, we wish to control / optimize it

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

- 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 heavy-traffic approximations).

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.

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:

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

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

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

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

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)

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

Example: Model Uncertainty in Bankruptcy Calculations

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

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

- 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)

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

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

A Distributionally Robust Risk Analysis Formulation

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

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

where P_0 is a *suitable* model.

A Distributionally Robust Risk Analysis Formulation

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

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

  where P_0 is a *suitable* model.

- P_0 = proxy for P_{true}.

A Distributionally Robust Risk Analysis Formulation

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

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

where P_0 is a *suitable* model.

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

A Distributionally Robust Risk Analysis Formulation

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

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

where P_0 is a *suitable* model.

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

A Distributionally Robust Risk Analysis Formulation

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

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

where P_0 is a *suitable* model.

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

Desirable Elements of Distributionally Robust Formulation

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

Desirable Elements of Distributionally Robust Formulation

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

Desirable Elements of Distributionally Robust Formulation

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

Desirable Elements of Distributionally Robust Formulation

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

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)

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

- 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 $R(t)$...

Connections to Distributionally Robust Optimization

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

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

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

Application 1: Back to Classical Risk Problem

- Suppose that

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

Application 1: Back to Classical Risk Problem

- Suppose that

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

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

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

Application 1: Back to Classical Risk Problem

- Suppose that

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

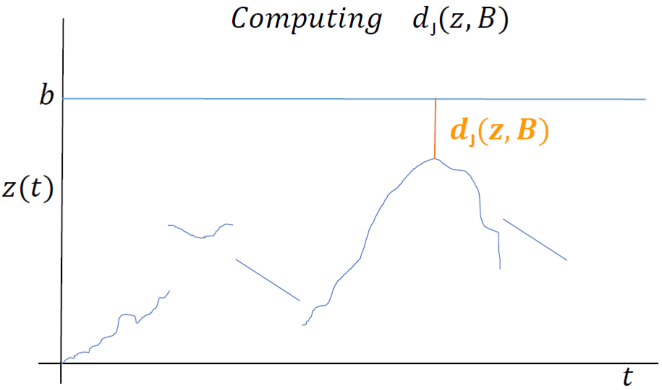
- If $R(t) = b - Z(t)$, then ruin during time interval $[0,1]$ is

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

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

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

Application 1: Computing Distance to Bankruptcy



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

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

Application 1: Computing Uncertainty Size

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

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

Application 1: Computing Uncertainty Size

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

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

- 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

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

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

Application 1: Illustration of Coupling

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

Algorithm 1 To embed the process $(Z(t) : t \geq 0)$ in Brownian motion $(B(t) : t \geq 0)$

Given: Brownian motion $B(t)$, moment m_1 and independent realizations of claim sizes $X_1, X_2, \ldots$

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

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

Define the auxiliary processes

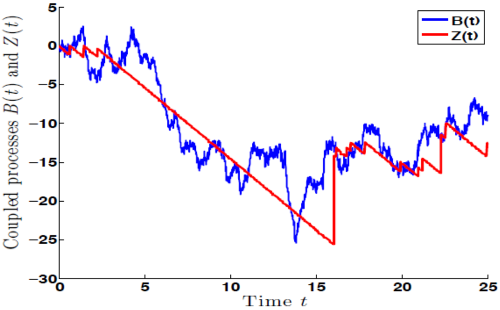
$$\tilde{S}(t) := \sum_{j>0} \sup_{\tau_j \leq s \leq t} B(s) \mathbf{1}(\tau_j \leq t < \tau_{j+1}) \text{ and } \tilde{N}(t) := \sum_{j \geq 0} \Psi_j \mathbf{1}(\tau_j \leq t < \tau_{j+1}).$$

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

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

Application 1: Coupling in Action

FIGURE 4. A coupled path output by Algorithm 1



Application 1: Numerical Example

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

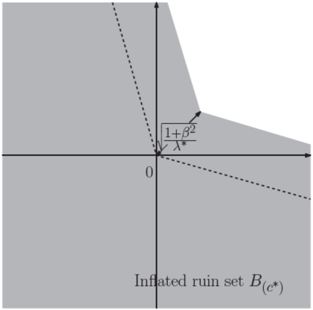
b	$\frac{P_0(\text{Ruin})}{P_{true}(\text{Ruin})}$	$\frac{P^*_{robust}(\text{Ruin})}{P_{true}(\text{Ruin})}$
100	1.07×10^{-1}	12.28
150	2.52×10^{-4}	10.65
200	5.35×10^{-8}	10.80
250	1.15×10^{-12}	10.98

Additional Applications: Multidimensional Ruin Problems

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

Additional Applications: Multidimensional Ruin Problems

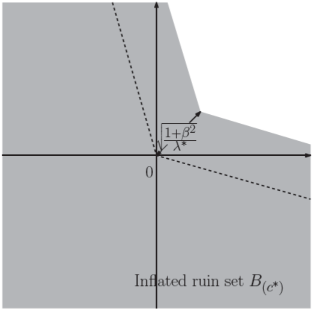
- **https://arxiv.org/abs/1604.01446** contains more applications.
- Control: $\min_\theta \sup_{P:D(P,P_0)\leq\delta} E[L(\theta, Z)]$ <– robust optimal reinsurance.



(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.

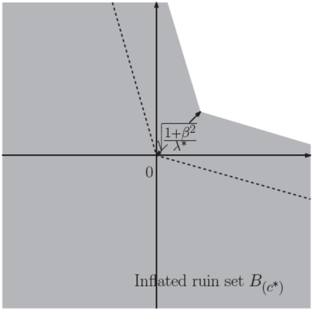


(b)Computation of worst-case ruin using the baseline measure

- 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**

Connections to Distributionally Robust Optimization

Based on:

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

https://arxiv.org/abs/1610.05627

Highlight: Additional insights into why optimal transport...

- 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.

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

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

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.

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)).
Remark 2: Uses RoPA duality theorem & **"judicious choice of** $c(\cdot)$"

Connection to 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: *Approximate* connection studied in Esfahani and Kuhn (2015).

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...
- *Support Vector Machines*: B., Kang, Murthy (2016) - **https://arxiv.org/abs/1610.05627**

Unification and Extensions of Regularized Estimators

- Distributionally Robust Optimization using Optimal Transport recovers many other estimators...
- *Support Vector Machines*: B., Kang, Murthy (2016) - **https://arxiv.org/abs/1610.05627**
- *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...
- *Support Vector Machines*: B., Kang, Murthy (2016) - **https://arxiv.org/abs/1610.05627**
- *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...
- *Support Vector Machines*: B., Kang, Murthy (2016) - **https://arxiv.org/abs/1610.05627**
- *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**

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 $c\left((x,y),(x',y')\right) = \|(x,y)-(x',y')\|_q^2$

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

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

How Regularization and Dual Norms Arise?

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

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

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

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

How Regularization and Dual Norms Arise?

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

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

How Regularization and Dual Norms Arise?

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

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

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

How Regularization and Dual Norms Arise?

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

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

- 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).

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).

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,

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

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,

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

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

- *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,

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

- *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,

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

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

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...*

On Role of Transport Cost...

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

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

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

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

On Role of Transport Cost...

- **Comparing L_1 regularization vs data-driven cost regularization: real data**

		BC	BN	QSAR	Magic
3*LRL1	Train	$.185 \pm .123$	$.080 \pm .030$	$.614 \pm .038$	$.548 \pm .087$
	Test	$.428 \pm .338$	$.340 \pm .228$	$.755 \pm .019$	$.610 \pm .050$
	Accur	$.929 \pm .023$	$.930 \pm .042$	$.646 \pm .036$	$.665 \pm .045$
3*DRO-NL	Train	$.032 \pm .015$	$.113 \pm .035$	$.339 \pm .044$	$.381 \pm .084$
	Test	$.119 \pm .044$	$.194 \pm .067$	$.554 \pm .032$	$.576 \pm .049$
	Accur	$.955 \pm .016$	$.931 \pm .036$	$.736 \pm .027$	$.730 \pm .043$
Num Predictors		30	4	30	10
Train Size		40	20	80	30
Test Size		329	752	475	9990

Table: Numerical results for real data sets.

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:

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

Towards an Optimal Choice of Uncertainty Size

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

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

- 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:

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

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

- 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

- "Standard" way to pick δ (Esfahani and Kuhn (2015)).

Towards an Optimal Choice of Uncertainty Size

- "Standard" way to pick δ (Esfahani and Kuhn (2015)).
- Estimate $D\left(P_{true}, P_n\right)$ using concentration of measure results.

Towards an Optimal Choice of Uncertainty Size

- "Standard" way to pick δ (Esfahani and Kuhn (2015)).
- Estimate $D\left(P_{true}, P_n\right)$ using concentration of measure results.
- Not a good idea: rate of convergence of the form $O\left(1/n^{1/d}\right)$ (d is the data dimension).

Towards an Optimal Choice of Uncertainty Size

- "Standard" way to pick δ (Esfahani and Kuhn (2015)).
- Estimate $D\left(P_{true}, P_n\right)$ using concentration of measure results.
- Not a good idea: rate of convergence of the form $O\left(1/n^{1/d}\right)$ (d is the data dimension).
- Instead we seek an optimal 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

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

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

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

- 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

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

- 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:*

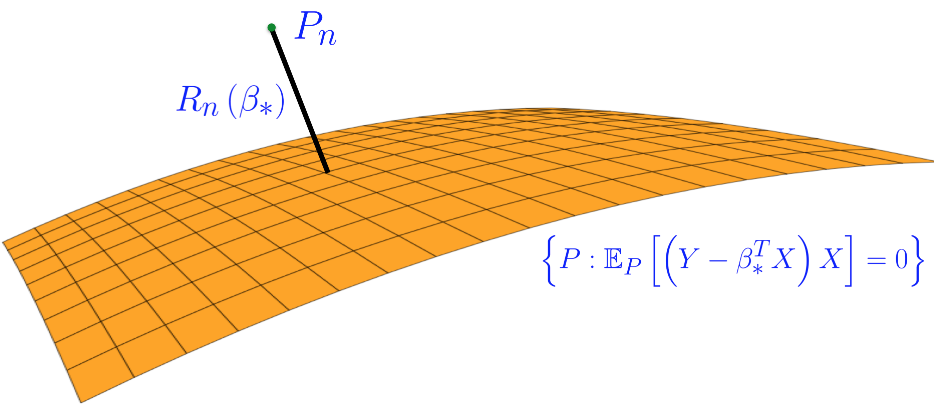
$$R_n(\beta)=\min\{D_c(P,P_n):E_P\left(\left(Y-\beta^T X\right)X\right)=0\}.$$

- Note that

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

- **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).

Discussion on Optimal Uncertainty Size

- Optimal δ is of order $O(1/n)$ as opposed to $O\left(1/n^{1/d}\right)$ as advocated in the standard approach.
- We characterize the asymptotic constant (not only order) in optimal regularization:

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

- $R_n(\beta_*)$ is inspired by Empirical Likelihood – Owen (1988).
- Lam & Zhou (2015) use Empirical Likelihood in DRO, but focus on divergence.

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 $\beta = \beta_*$

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

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.
- OT costs are natural for computing model uncertainty.

Conclusions

- Optimal transport (OT) is a powerful tool based on linear programming.
- OT costs are natural for computing model uncertainty.
- OT can be used in path-space to quantify error in diffusion approximations.

Conclusions

- Optimal transport (OT) is a powerful tool based on linear programming.
- OT costs are natural for computing model uncertainty.
- OT can be used in path-space to quantify error in diffusion approximations.
- OT can be used for data-driven distributionally robust optimization.

Conclusions

- Optimal transport (OT) is a powerful tool based on linear programming.
- OT costs are natural for computing model uncertainty.
- OT can be used in path-space to quantify error in diffusion approximations.
- 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.
- OT costs are natural for computing model uncertainty.
- OT can be used in path-space to quantify error in diffusion approximations.
- 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.