Stanford

Unifying Distributionally Robust Optimization

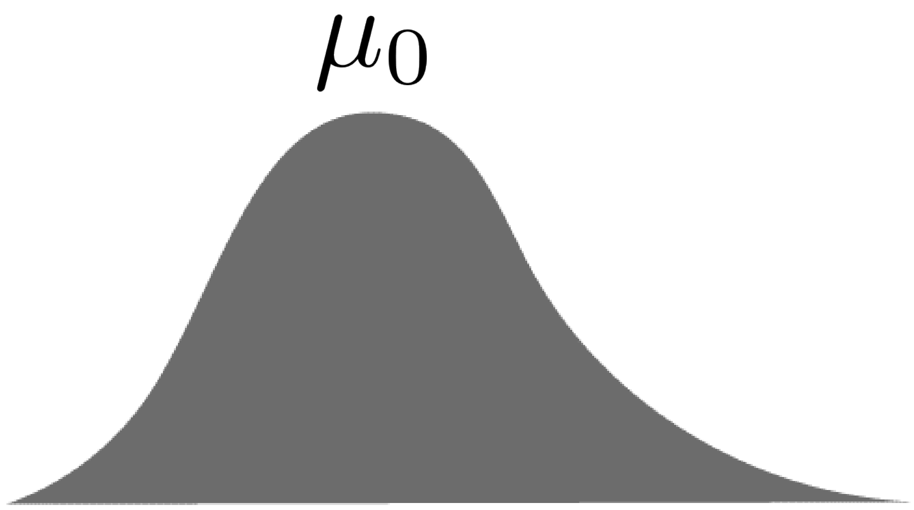
Optimal transport Approach

Jiajin Li

Joint work with Jose Blanchet, Daniel Kuhn and Bahar Tahksen

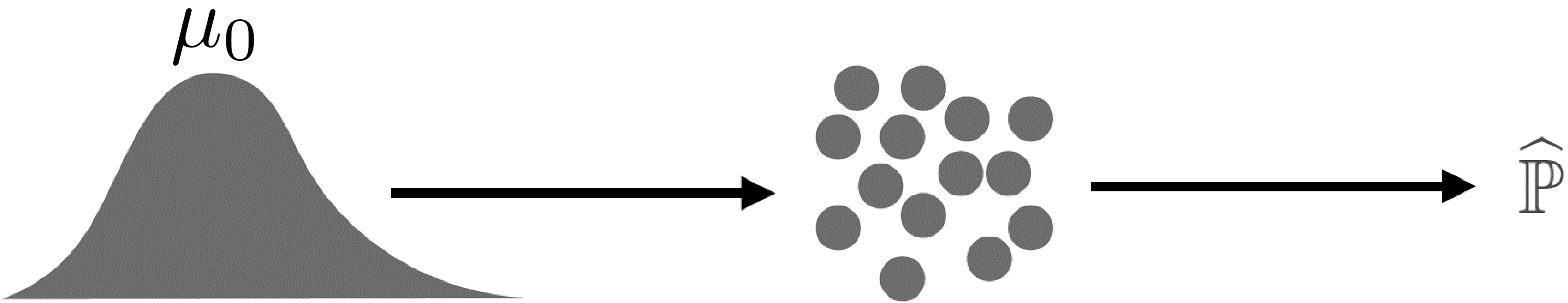
Optimization under Uncertainty

$$\boxed{\inf_{\ell \in \mathcal{L}} \mathbb{E}_{\mu_0}[\ell(Z)]}$$

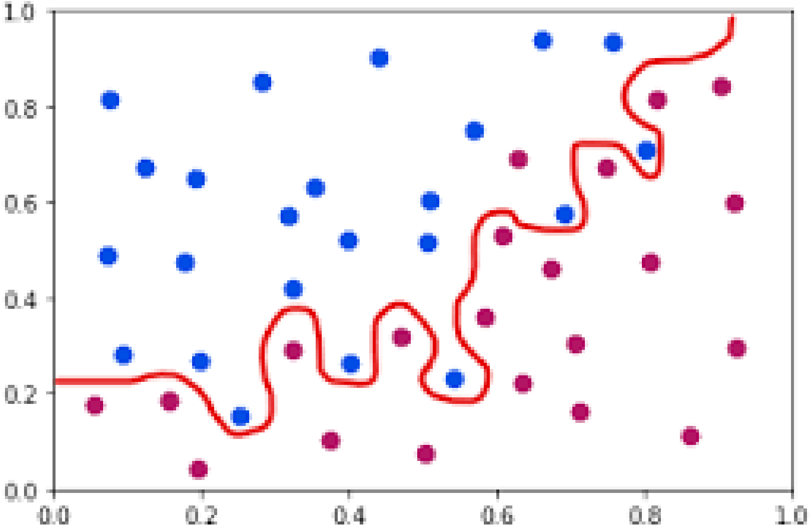


Optimization under Uncertainty

$$\inf_{\ell \in \mathcal{L}} \mathbb{E}_{\hat{\mu}}[\ell(Z)]$$

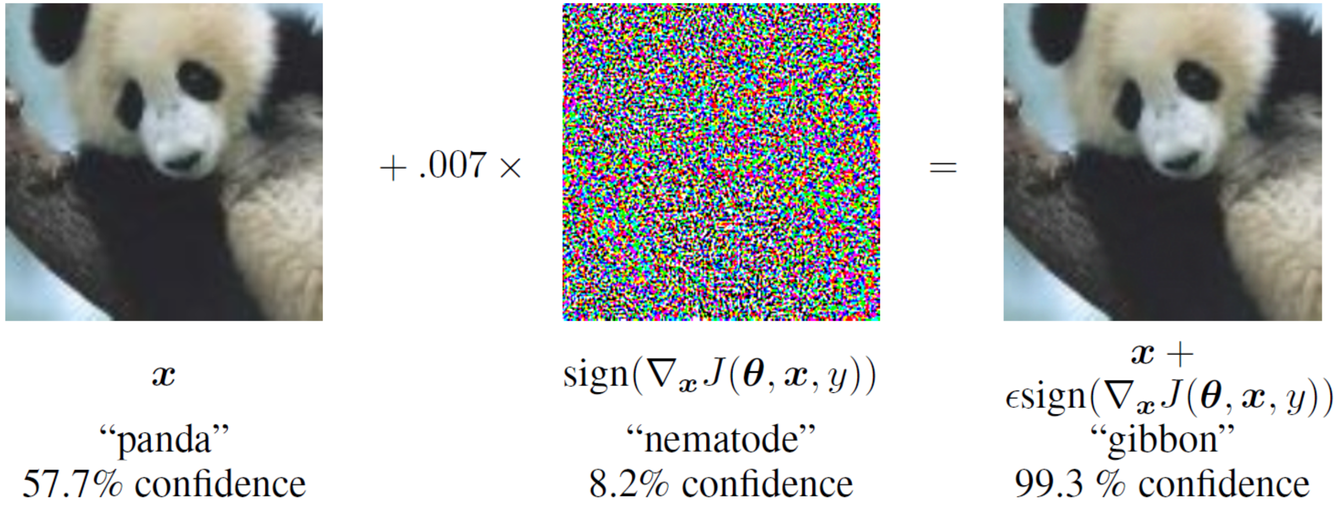


SAA Often Fails



- ➢ **Overfitting**

- ➢ **Adversarial Attack**

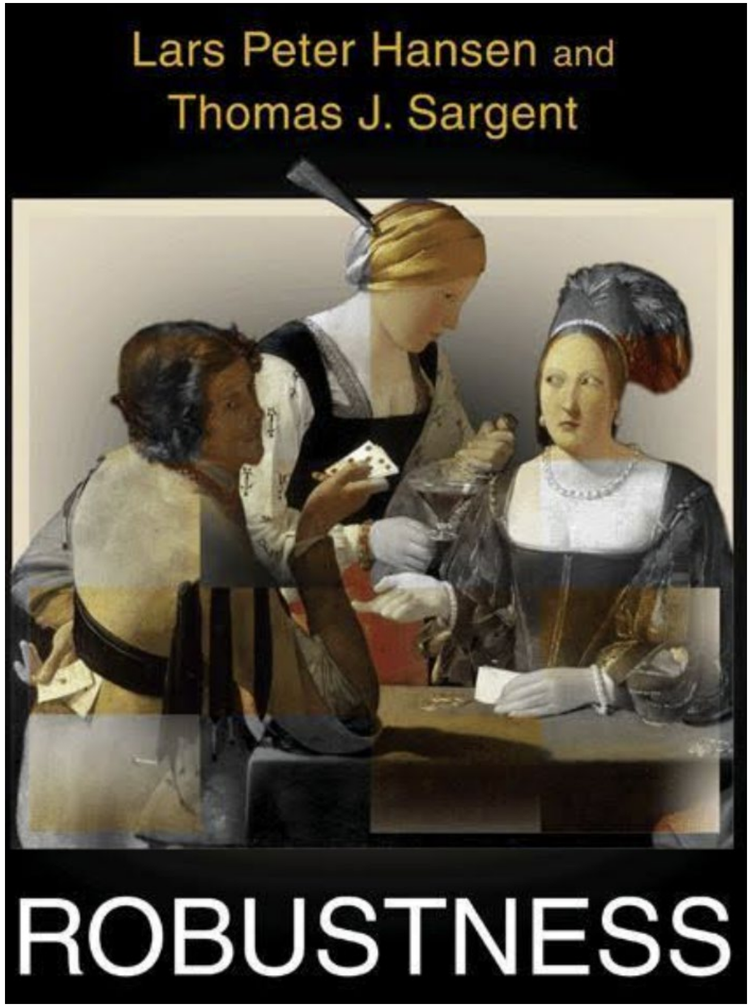


Model Misspecification

$$\inf_{\ell\in\mathcal{L}} \sup_{\mu\in\mathcal{B}} \mathbb{E}_{\mu}[\ell(Z)]$$

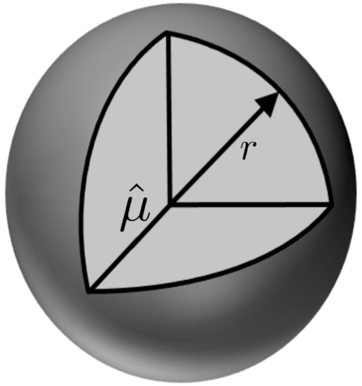
What makes sense for choosing the set $\mathcal{B}$?

Optimal in Gilboa-Schmeidler (2008) decision theoretic (multiple prior) sense.



Ambiguity Set

- A set of distribution around the baseline distribution $\hat{\mu}$:



$$\mathcal{B} := \{\mu \in \mathcal{P}(\mathcal{Z}) : \mathbb{D}(\mu, \hat{\mu}) \leq r\}$$

Desirable Properties:

- Non-parametric
- Tractable
- Explainable

How to choose the “discrepancy” $D(\cdot,\cdot)$?

How to Choose Probability Metric?

<u>**Two**</u> natural ways to model changes in distributions.

A) Our model gets the likelihood of outcomes wrong.

B) Our model gets the outcomes wrong.

Traditionally A) and B) are seen as separate mechanisms.

Approach A) <u>Divergence</u>: Dupuis, James & Peterson '00; Hansen & Sargent '01, '08; Nilim & El Ghaoui '02, '03; Iyengar '05; A. Ben-Tal, L. El Ghaoui, & A. Nemirovski '09; Bertsimas & Sim '04; Bertsimas, Brown, Caramanis '13; Lim & Shanthikumar '04; Lam '13, '17; Csiszár & Breuer '13; Jiang & Guan '12; Hu & Hong '13; Wang, Glynn & Ye '14; Bayrakskan & Love '15; Duchi, Glynn & Namkoong '16; Bertsimas, Gupta & Kallus ’13, (LD-) Van Bary et al. ‘17

Approach B) <u>Wasserstein</u>: Scarf '58; Hampel '73; Huber '81; Pflug & Wozabal '07; Mehrotra & Zhang '14; Esfahani & Kuhn '15; Blanchet & Murthy '16; Gao & Kleywegt '16; Duchi & Namkoong 17’, (Sinkhorn-) Wang et al. ’21.

How to Choose Probability Metric?

Two natural ways to model changes in distributions.

A) C
outc

B) C
wro

Traditionally A) and B) are seen as separate mechanisms.

Approach A) Divergence: Dupuis, James & Peterson '00; Hansen & Sargent '01, '08; Nilim & El Ghaoui '02, '03; Iyengar '05; A. Ben-Tal, L. El Ghaoui, & A. Nemirovski '09;
nis '13; Lim
er '13;
Ye '14;
g '16;
al. '17

Huber
Esfahani
egt '16;
21.

Can we address model misspecification in terms of both **likelihoods** and **actual outcomes?**

Why is this important?

Why is this important?

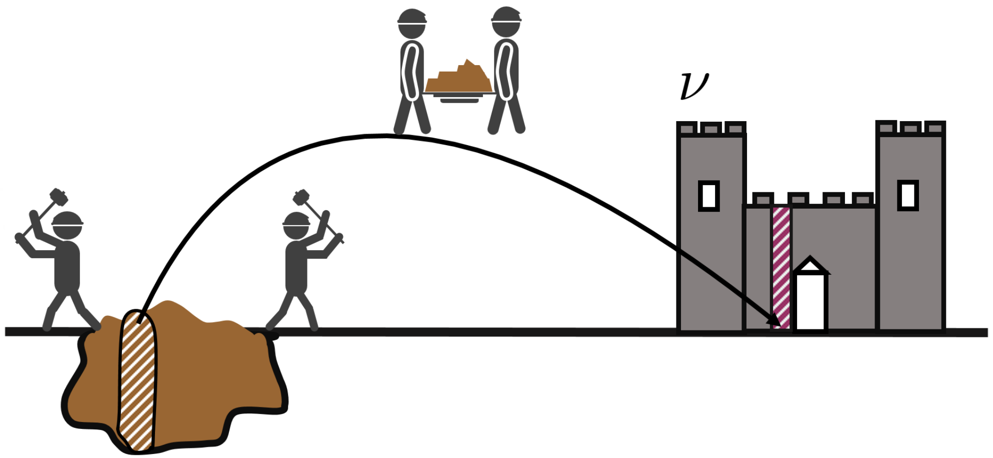
A unified theory of DRO will position the area well to address a key question…

How to *practically* choose the distributional uncertainty set?

Even experts see these DRO models as fundamentally different – in some sense, they are not…

A Unified View of DRO via Optimal Transport Approach with Conditional Moment Constraints

What is Optimal Transport?



Transportation cost function

$$\mathbb{D}_c(\mu, \hat{\mu}) := \begin{cases} \min\limits_{\pi(dZ, d\hat{Z}) \geq 0} \int_{Z \in \mathcal{Z}} \int_{\hat{Z} \in \mathcal{Z}} c(Z, \hat{Z}) \pi(dZ, d\hat{Z}) \\ \text{s.t.} \quad \int_{Z \in \mathcal{Z}} \pi(dZ, d\hat{Z}) = \hat{\mu}(d\hat{Z}) \\ \qquad \int_{Z \in \mathcal{Z}} \pi(dZ, d\hat{Z}) = \mu(dZ) \end{cases}$$

- Linear programming problem ("Monge-Kantorovich")

[Villani (2003)]

Unifying Formulation

- OT-DRO with conditional moment constraints

$$\sup_{\mu} \{\mathbb{E}_\mu[\ell(Z)] : \mathbb{D}(\mu, \hat{\mu}) \le r\}$$

$$Z \to (V, W)$$

$$\begin{aligned} \sup \quad & \mathbb{E}_\pi[W \cdot \ell(Z)] \\ \text{s.t.} \quad & \pi \in \mathcal{P}((\mathcal{V} \times \mathcal{W}) \times (\mathcal{V} \times \mathcal{W})) \\ & \pi_{(\hat{V}, \hat{W})} = \hat{\nu} \\ & \mathbb{E}_\pi[W | \mathcal{G}] = 1 \quad \pi\text{-a.s.} \\ & \mathbb{E}_\pi[c((V, W), (\hat{V}, \hat{W}))] \le r. \end{aligned}$$

Lifting Technique!

Unifying Formulation

- OT-DRO with conditional moment constraints

$$\sup_{\mu} \{\mathbb{E}_{\mu}[\ell(Z)] : \mathbb{D}(\mu, \hat{\mu}) \leq r\}$$

$$Z \to (V, W)$$

$$\begin{aligned} \sup \quad & \mathbb{E}_{\pi}[W \cdot \ell(Z)] \\ \text{s.t.} \quad & \pi \in \mathcal{P}((\mathcal{V} \times \mathcal{W}) \times (\mathcal{V} \times \mathcal{W})) \\ & \pi_{(\hat{V}, \hat{W})} = \hat{\nu} \\ & \boxed{\mathbb{E}_{\pi}[W|\mathcal{G}] = 1 \quad \pi\text{-a.s}} \\ & \mathbb{E}_{\pi}[c((V, W), (\hat{V}, \hat{W}))] \leq r. \end{aligned}$$

Lifting Technique!

If G is a trivial sigma field, then

$$\mathbb{E}_{\pi}[W|\mathcal{G}] = \mathbb{E}_{\pi}[W] = 1.$$

Unifying Formulation

- OT-DRO with conditional moment constraints

$$\sup_{\mu} \{\mathbb{E}_{\mu}[\ell(Z)] : \mathbb{D}(\mu, \hat{\mu}) \le r\}$$

$$Z \to (V, W)$$

$$\begin{aligned} \sup \quad & \mathbb{E}_{\pi}[W \cdot \ell(Z)] \\ \text{s.t.} \quad & \pi \in \mathcal{P}((\mathcal{V} \times \mathcal{W}) \times (\mathcal{V} \times \mathcal{W})) \\ & \pi_{(\hat{V}, \hat{W})} = \hat{\nu} \\ & \boxed{\mathbb{E}_{\pi}[W | \mathcal{G}] = 1 \quad \pi\text{-a.s}} \\ & \mathbb{E}_{\pi}[c((V, W), (\hat{V}, \hat{W}))] \le r. \end{aligned}$$

Lifting Technique!

If G is the smallest sigma field generated by hat V, then

$$\mathbb{E}_{\pi}[W | \mathcal{G}] = \mathbb{E}_{\pi}[W | \hat{V}] = 1, \pi\text{-a.s.}$$

Unifying Formulation

- OT-DRO with conditional moment constraints

$$\sup_{\mu} \{\mathbb{E}_{\mu}[\ell(Z)] : \mathbb{D}(\mu, \hat{\mu}) \leq r\} \longrightarrow$$

$$Z \to (V, W)$$

$$\begin{aligned}
\sup \quad & \mathbb{E}_{\pi}[W \cdot \ell(Z)] \\
\text{s.t.} \quad & \pi \in \mathcal{P}((\mathcal{V} \times \mathcal{W}) \times (\mathcal{V} \times \mathcal{W})) \\
& \pi_{(\hat{V}, \hat{W})} = \hat{\nu} \\
& \mathbb{E}_{\pi}[W | \mathcal{G}] = 1 \quad \pi\text{-a.s} \\
& \mathbb{E}_{\pi}[c((V, W), (\hat{V}, \hat{W}))] \leq r.
\end{aligned}$$

$$(\ell, \mathcal{Z}, \mathbb{D}, \hat{\mu}, r) \to (\mathcal{V}, \mathcal{W}, \mathcal{G}, \hat{\nu}, c, r)$$

Unifying Formulation

➢ OT-DRO with "Martingale" Constraint

$$\begin{aligned} \sup_{\pi \in \Pi((\Omega\times\mathbb{R}_+)\times(\Omega\times\mathbb{R}_+))} \quad & \mathbb{E}_\pi[\ell(\theta,\xi)\cdot\zeta] \\ \text{s.t.} \quad & \mathbb{E}_\pi\left[c_M((\xi,\zeta),(\xi',\zeta'))\right] \le \delta, \\ & \mathbb{E}_\pi\left[\zeta \mid \zeta'\right] = \zeta', \\ & \boxed{\Pi_{(\xi',\zeta')}\pi = \hat{\nu}.} \end{aligned}$$

This is the "**baseline model**" which is constrained to be $\hat{\nu}$.

We Recover Most DRO Formulations

Pick cost functions & reference measure

Recover → Most DRO Formulations!

Wasserstein DRO

$$\sup_{\pi \in \mathcal{P}(\Omega\times\Omega)} \left\{ \mathbb{E}_\pi[\ell(\theta,\xi)] : \mathbb{E}_\pi\left[c(\xi,\xi')\right] \le \delta, \Pi_{\xi'}\pi = \widehat{\mathbb{P}}_n \right\}$$

$$\begin{aligned} \sup_{\pi\in\Pi((\Omega\times\mathbb{R}_+)\times(\Omega\times\mathbb{R}_+))} \quad & \mathbb{E}_\pi[\ell(\theta,\xi)\cdot\zeta] \\ \text{s.t.} \quad & \mathbb{E}_\pi\left[c(\xi,\xi') + \boxed{\infty\cdot\mathbb{I}_{\zeta\neq\zeta'}}\right] \le \delta, \\ & \cancel{\mathbb{E}_\pi\left[\zeta \mid \zeta'\right] = \zeta',} \xrightarrow{} \text{Automatically satisfied} \\ & \Pi_{(\xi',\zeta')}\pi = \widehat{\mathbb{P}}_n \times \delta_1. \end{aligned}$$

$\zeta = 1, \pi$-a.s!

ϕ-divergence [Csiszar, 1963,1967]

For two probability measures $\mathbb{Q}\,\text{and}\,\widehat{\mathbb{P}}_n \in \mathcal{P}(\Omega)$, we let ρ be a dominating measure of $\mathbb{Q}\,\text{and}\,\widehat{\mathbb{P}}_n$(i.e., $\mathbb{Q} \ll \rho\,\text{and}\,\widehat{\mathbb{P}}_n \ll \rho$). Then, the ϕ divergence between $\mathbb{Q}\,\text{and}\,\widehat{\mathbb{P}}_n$ is defined, independently of ρ, by

$$D_\phi(\mathbb{Q}, \widehat{\mathbb{P}}_n) = \int_\Omega \frac{d\widehat{\mathbb{P}}_n}{d\rho}(\xi)\phi\left(\frac{d\mathbb{Q}}{d\rho}(\xi)/\frac{d\widehat{\mathbb{P}}_n}{d\rho}(\xi)\right) d\rho(\xi),$$

where $0 \cdot \phi\left(\frac{0}{0}\right) := 0,\ \text{and}\, 0 \cdot \phi\left(\frac{a}{0}\right) := \boxed{a \lim_{t \to 0} \frac{\phi(t)}{t} := \alpha\phi'_\infty}, \forall \alpha > 0.$

The speed of growth of ϕ at infty.

ϕ-divergence [Csiszar, 1963,1967]

For two probability measures $\mathbb{Q}\,\text{and}\,\widehat{\mathbb{P}}_n \in \mathcal{P}(\Omega)$, we let ρ be a dominating measure of $\mathbb{Q}\,\text{and}\,\widehat{\mathbb{P}}_n$(i.e., $\mathbb{Q} \ll \rho\,\text{and}\,\widehat{\mathbb{P}}_n \ll \rho$). Then, the ϕ divergence between

Fact 1 (Decomposition)

$$D_\phi(\mathbb{Q},\widehat{\mathbb{P}}_n) = \int_\Omega \phi\left(\frac{d\mathbb{Q}}{d\rho}(\xi)\Big/\frac{d\widehat{\mathbb{P}}_n}{d\rho}(\xi)\right) d\widehat{\mathbb{P}}_n(\xi), +\phi'_\infty \mathbb{Q}\left(\frac{d\widehat{\mathbb{P}}_n}{d\rho}(\xi)=0\right).$$

where $0\cdot\phi\left(\frac{0}{0}\right) := 0,\ \text{and}\ 0\cdot\phi\left(\frac{a}{0}\right) := \boxed{a\lim_{t\to 0}\frac{\phi(t)}{t} := a\phi'_\infty}, \forall a>0.$

The speed of growth of ϕ at infty.

ϕ-divergence [Csiszar, 1963,1967]

For two probability measures $\mathbb{Q}$ and $\widehat{\mathbb{P}}_n \in \mathcal{P}(\Omega)$, we let ρ be a dominating measure of $\mathbb{Q}$ and $\widehat{\mathbb{P}}_n$(i.e., $\mathbb{Q} \ll \rho$ and $\widehat{\mathbb{P}}_n \ll \rho$). Then, the ϕ divergence between

Example 1 (Kullback–Leibler divergence)

$$\phi'_\infty = \lim_{t\to\infty} \frac{t \log t}{t} = +\infty$$

$$D_{\mathbf{KL}}(\mathbb{Q}, \widehat{\mathbb{P}}_n) = \begin{cases} \int_\Omega \phi\left(\frac{d\mathbb{Q}}{d\widehat{\mathbb{P}}_n}\right) d\widehat{\mathbb{P}}_n, & \mathbb{Q} \ll \widehat{\mathbb{P}}_n \\ \infty, & \text{Otherwise} \end{cases}$$

The speed of growth of ϕ at infty.

ϕ-divergence [Csiszar, 1963,1967]

For two probability measures $\mathbb{Q}$ and $\widehat{\mathbb{P}}_n \in \mathcal{P}(\Omega)$, we let ρ be a dominating measure of $\mathbb{Q}$ and $\widehat{\mathbb{P}}_n$(i.e., $\mathbb{Q} \ll \rho$ and $\widehat{\mathbb{P}}_n \ll \rho$). Then, the ϕ divergence between

Fact 2 (Asymmetry)

$$D_\phi(\mathbb{Q}, \widehat{\mathbb{P}}_n) = D_\psi(\widehat{\mathbb{P}}_n, \mathbb{Q}),$$

where $\psi(t) = t\phi\left(\frac{1}{t}\right)$ represents the **Cisizar dual** of $\phi(t)$.

The speed of growth of ϕ at infty.

ϕ-divergence DRO [$\phi'_\infty = \infty$]

Likelihood ratio ζ

$$\sup_{\mathbb{Q}\in\mathcal{P}(\Omega)} \left\{ \mathbb{E}_{\mathbb{Q}}[\ell(\theta,\xi)] : \mathbb{E}_{\widehat{\mathbb{P}}_n}\left[\phi\left(\frac{d\mathbb{Q}}{d\widehat{\mathbb{P}}_n}\right)\right] \le \delta \right\}$$

$$\begin{aligned} \sup_{\pi\in\Pi((\Omega\times\mathbb{R}_+)\times(\Omega\times\mathbb{R}_+))} \quad & \mathbb{E}_\pi[\ell(\theta,\xi)\cdot\zeta] \\ \text{s.t.} \quad & \mathbb{E}_\pi\left[\infty\cdot\mathbb{I}_{\xi\neq\xi'} + (\phi(\zeta)-\phi(\zeta'))^+\right] \le \delta, \\ & \mathbb{E}_\pi[\zeta] = 1, \\ & \Pi_{(\xi',\zeta')}\pi = \widehat{\mathbb{P}}_n \times \delta_1. \end{aligned}$$

$\phi(1) = 0$!

ϕ-divergence DRO [$\phi'_\infty < \infty$]

$$\sup_{\mathbb{Q}\in\mathcal{P}(\Omega)} \left\{ \mathbb{E}_{\mathbb{Q}}[\ell(\theta,\xi)] : \mathbb{E}_\rho\left[\frac{d\widehat{\mathbb{P}}_n}{d\rho}\phi\left(\frac{d\mathbb{Q}}{d\rho}\Big/\frac{d\widehat{\mathbb{P}}_n}{d\rho}\right)\right] \le \delta \right\}$$

Fact : Suppose that Ω is compact, the worst-case distribution will be supported on **N+1 point**, i.e., $\xi'_{n+1} \in \arg\max_{\xi\in\Omega} \ell(\theta,\xi)$

$$\hat{\nu}(d\xi', d\zeta') = \frac{1-\epsilon}{n}\sum_i \delta_{(\xi'_i, 1-\epsilon)} + \epsilon\delta_{(\xi'_{n+1}, 0)}$$

ϕ-divergence DRO [$\phi'_\infty < \infty$]

$$
\begin{aligned}
\sup_{\pi \in \Pi((\Omega \times \mathbb{R}_+) \times (\Omega \times \mathbb{R}_+))} \quad & \mathbb{E}_\pi [\ell(\theta, \xi) \cdot \zeta] \\
\text{s.t.} \quad & \mathbb{E}_\pi \left[\infty \cdot \mathbb{I}_{\xi \neq \xi'} + \frac{d\hat{\mu}}{d\rho}(\xi) \phi \left(\zeta / \frac{d\hat{\mu}}{d\rho}(\xi) \right) \right] \leq \delta, \\
& \mathbb{E}_\pi [\zeta \mid \zeta'] = \zeta', \\
& \Pi_{(\xi', \zeta')} \pi = \hat{\nu}.
\end{aligned}
$$

$$\rho = (1 - \epsilon) \sum_{i=1}^{n} \delta_{\xi'_i} + \epsilon \delta_{\xi'_{n+1}}, \text{ for } \epsilon \in (0,1)$$

$$\hat{\nu}(d\xi', d\zeta') = \frac{1-\epsilon}{n} \sum_i \delta_{(\xi'_i, 1-\epsilon)} + \epsilon \delta_{(\xi'_{n+1}, 0)}$$

ϕ-divergence DRO

Divergence	$\phi(t),\ t \geqslant 0$
Kullback-Leibler	$t \cdot \log(t)$
Burg Entropy	$-\log(t) + t - 1$
J-divergence	$(t-1)\log(t)$
χ^2-distance	$\frac{1}{t}(t-1)^2$
Modified χ^2-distance	$(t-1)^2$
Hellinger distance	$(\sqrt{t}-1)^2$
χ-divergence of order $n > 1$	$\|t-1\|^n$
Variation distance	$\|t-1\|$
Cressie-Read	$\frac{1-\theta+\theta t-t^\theta}{\theta(1-\theta)},\ \theta \neq 0, 1$

Sinkhorn DRO

- **Entropic regularization** (Sinkhorn Distance) is popular in Optimal Transport applications in AI [Cuturi. (2013), Peyré & Cuturi. 2017]

- This motivated Wang et al. (2021) to consider the formulation:

Normal distribution

$$
\begin{aligned}
&\sup_{\pi \in \mathcal{P}(\Omega \times \Omega)} \quad \mathbb{E}_\pi[\ell(\theta, \xi)] \\
&\text{s.t., } \mathbb{E}_\pi\left[c(\xi, \xi') + \epsilon \log\left(\frac{d\pi(\xi, \xi')}{\boxed{d\eta(\xi)}\, d\widehat{\mathbb{P}}_n(\xi')}\right)\right] \leq \delta \\
&\Pi_{\xi'} \pi = \widehat{\mathbb{P}}_n
\end{aligned}
$$

How to recover this formulation?

Sinkhorn DRO

Normal distribution

$$\sup_{\pi \in \mathcal{P}(\Omega \times \Omega)} \mathbb{E}_{\pi}[\ell(\theta, \xi)]$$

$$\text{s.t.,}\ \mathbb{E}_{\pi}\left[c(\xi, \xi') + \epsilon \log\left(\frac{d\pi(\xi, \xi')}{\boxed{d\eta(\xi)}\, d\widehat{\mathbb{P}}_n(\xi')}\right)\right] \leq \delta$$

$$\Pi_{\xi'}\pi = \widehat{\mathbb{P}}_n$$

Again lift the outcome space to $\Omega \times \Omega \times R^{+}$!

Sinkhorn DRO -> KL-DRO

Following Wang et al. (2021) , we define the kernel distribution as

$$d\nu_{\xi',\epsilon}(\xi) := \frac{\exp(-\frac{c(\xi,\xi')}{\epsilon})}{\int_\Omega \exp(-\frac{c(x,\xi')}{\epsilon})d\eta(x)} d\eta(\xi),$$

and the new reference measure as:

$$d\gamma_0(\xi,\xi') = d\nu_{\xi',\epsilon}(\xi) \times d\widehat{\mathbb{P}}_n(\xi').$$

$$\boxed{\sup_{\pi \in \mathcal{P}(\Omega\times\Omega)} \left\{ \mathbb{E}_\pi[\ell(\theta,\xi)] : \epsilon\mathbb{E}_{\gamma_0}\left[\log\left(\frac{d\pi}{d\gamma_0}\right)\right] \le \bar{\delta}, \pi_{\xi'} = \widehat{\mathbb{P}}_n \right\}}$$

What about new – more powerful formulations?

New DRO Model

- The adversary has the ability to modify both the **actual outcomes** and the **associated probability**.

$$
\begin{aligned}
\sup_{\pi}\ & \mathbb{E}_{\pi}[\ell(\theta,\xi)\cdot\zeta] \\
\text{s.t.}\ & \mathbb{E}_{\pi}\left[\boxed{\gamma_1\zeta\cdot c(\xi,\xi') + \gamma_2\left(\phi(\zeta)-\phi(\zeta')\right)^{+}}\right] \le \delta, \\
& \mathbb{E}_{\pi}[\zeta] = 1, \\
& \Pi_{(\xi',\zeta')}\pi = \widehat{\mathbb{P}}_n \times \delta_1 .
\end{aligned}
\tag{P}
$$

$\gamma_1 = \infty$, KL-DRO

New DRO Model

- The adversary has the ability to modify both the **actual outcomes** and the **associated probability**.

$$
\begin{aligned}
\sup_{\pi}\ & \mathbb{E}_{\pi}\left[\ell(\theta,\xi)\cdot\zeta\right] \\
\text{s.t.}\ & \mathbb{E}_{\pi}\left[\boxed{\gamma_1\zeta\cdot c(\xi,\xi') + \gamma_2\left(\phi(\zeta)-\phi(\zeta')\right)^{+}}\right] \le \delta, \\
& \mathbb{E}_{\pi}\left[\zeta\right] = 1, \\
& \Pi_{(\xi',\zeta')}\pi = \widehat{\mathbb{P}}_n \times \delta_1.
\end{aligned}
\qquad (\mathrm{P})
$$

$\gamma_2 = \infty$, Wasserstein DRO

Reformulation Result

Theorem. (P) is equivalent to

$$\min_{\lambda \geq 0} \lambda\delta + \lambda\gamma_2 \log\left(\mathbb{E}_{\widehat{\mathbb{P}}_n}\left[\exp\left(\frac{\ell^c_{\lambda\gamma_1}(\xi')}{\lambda\gamma_2}\right)\right]\right)$$

where $\phi(\zeta) = \zeta\log(\zeta) - \zeta + 1$ and the c-transform of $\ell(\cdot)$ with penalty $\lambda\gamma_1$ is defined as

$$\ell^c_{\lambda\gamma_1}(\xi') = \sup_{\xi\in\Omega}\{\ell(\theta,\xi) - \lambda\gamma_1 c(\xi,\xi')\}$$

Reformulation Result

Theorem. (P) is equivalent to

$$\min_{\lambda \geq 0} \lambda\delta + \lambda\gamma_2 \log\left(\mathbb{E}_{\widehat{\mathbb{P}}_n}\left[\exp\left(\frac{\ell^c_{\lambda\gamma_1}(\xi')}{\lambda\gamma_2}\right)\right]\right)$$

where $\phi(\zeta) = \zeta\log(\zeta) - \zeta + 1$ and the c-transform of $\ell(\cdot)$ with penalty $\lambda\gamma_1$ is defined as

$$\ell^c_{\lambda\gamma_1}(\xi') = \sup_{\xi\in\Omega}\{\ell(\theta,\xi) - \lambda\gamma_1 c(\xi,\xi')\}$$

γ_1 and γ_2 play a critical role in controlling the LIKELIHOOD ERROR hedge vs OUTCOME ERROR hedge ← AND STILL TRACTABLE!

Optimal Transport Plan

Structure of the worst case $\pi^\star$: It must be concentrated on:

Perturbation on actual outcomes!

$$\left\{ (\xi, \zeta) \in \Omega \times \mathbb{R}^+ : \begin{bmatrix} \boxed{\xi} \in \arg\max\limits_{\xi \in \Omega} \left[\ell(\theta, \xi) - \lambda^\star \gamma_1 c(\xi, \xi_i') \right] \\ \boxed{\zeta} = \exp\left(\frac{\ell(\theta,\xi) - \alpha^\star}{\lambda^\star \gamma_2} - \frac{\gamma_1 c(\xi, \xi_i')}{\gamma_2} \right) \end{bmatrix}, \forall i \in [N] \right\},$$

Perturbation on data weights!

Where $\alpha^\star$ is the dual variable of $\mathbb{E}_\pi[\zeta] = 1$.

Tractability

Theorem 2. Suppose that the loss function $\ell(\theta, \cdot)$ is a pointwise maximum of concave functions and $c(\xi, \xi') = \|\xi - \xi'\|_p$, (P) can be reformulated as a finite convex program.

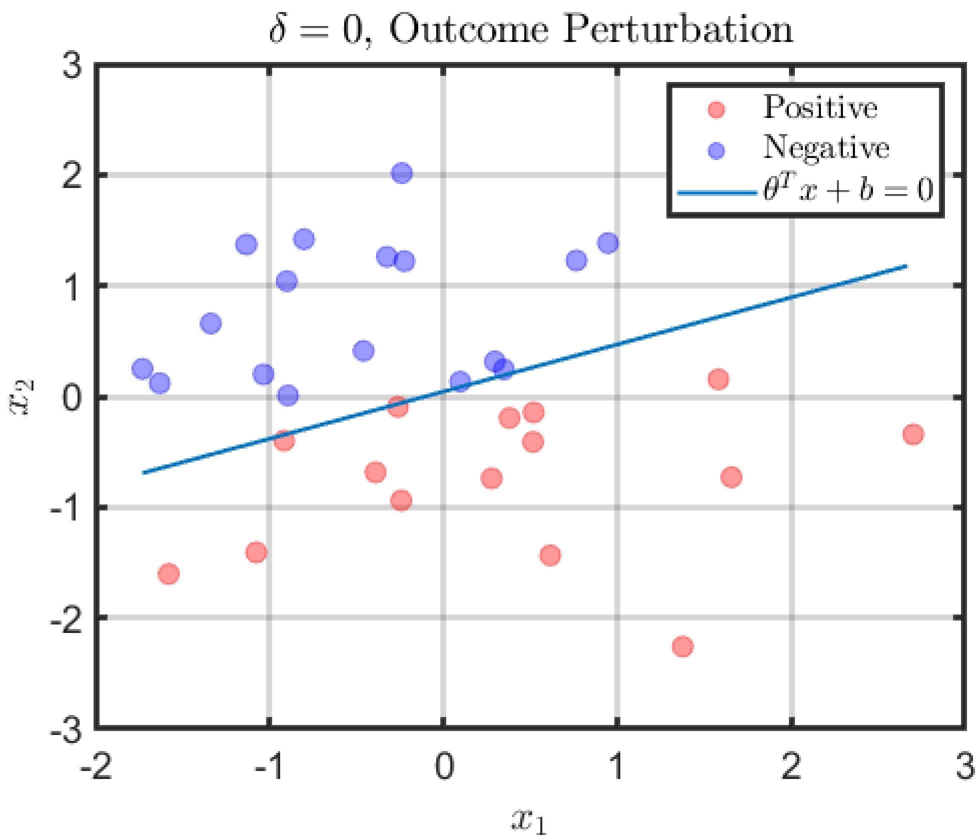
Can approximate any convex function as the maxima of affine functions…

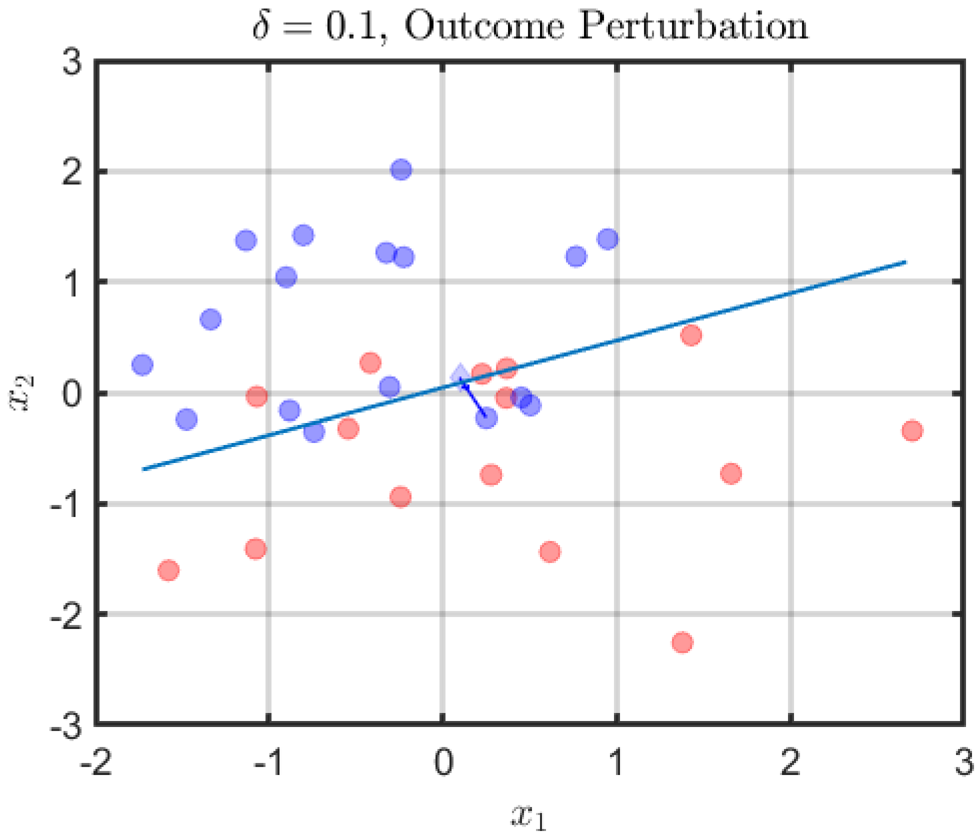
Mosek & Gurobi help!

Support Vector Machine

- Binary Classification Problem $\xi = (x, y), y \in \{+1, -1\}$
- Cost function $c(\xi, \xi') = \|x - x'\|_2^2 + \infty|y - y'|$
- Loss function $\ell(\theta, \xi) = \max(1 - y \cdot (\theta^T x + b), 0)$
- Hyperparameter: $\gamma_1 = \gamma_2 = 1$
- Dimension = 2

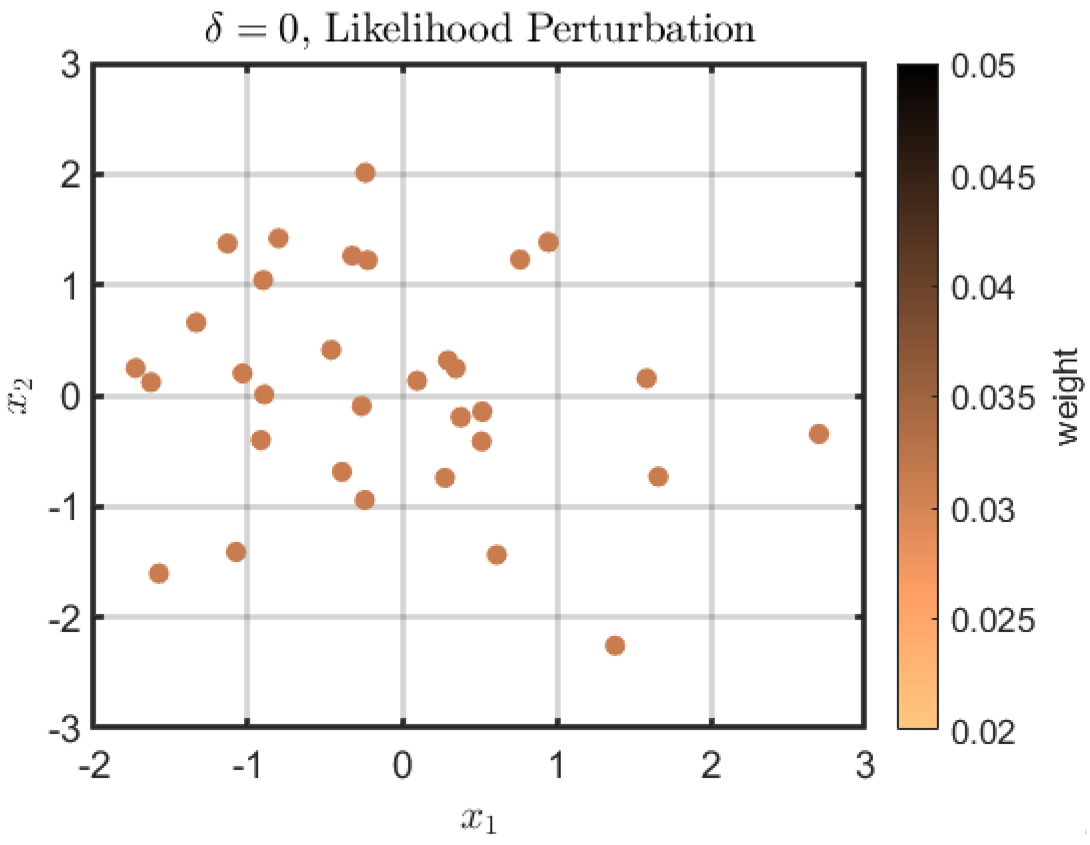
Worst-Case Distribution Visualization

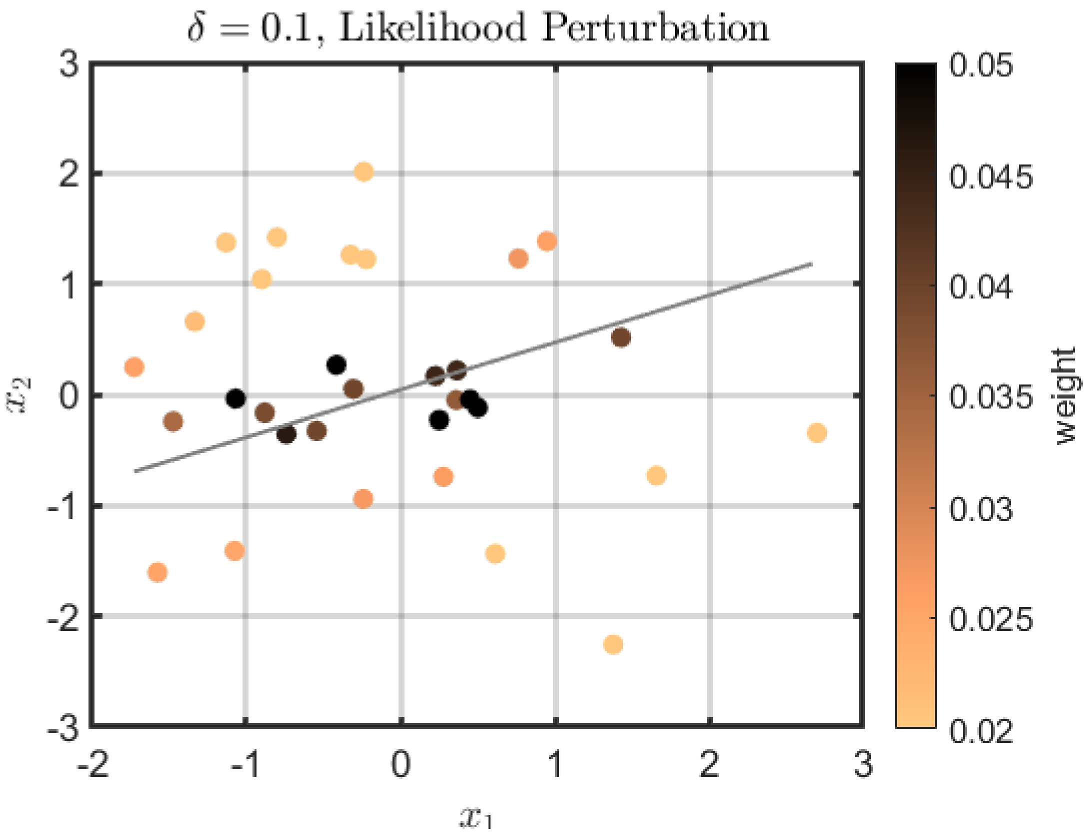




Outcome Perturbation!

Worst-Case Distribution Visualization





Likelihood Perturbation!

Reference:

1. Jose Blanchet*, Daniel Kuhn*, **Jiajin Li***, Bahar Tahksen* (Alphabetical order). Unifying Distributionally Robust Optimization via Optimal Transport Theory. **Working Paper**.

2. **Jiajin Li**, Sirui Lin, Jose Blanchet, Viet Anh Nguyen. Tikhonov Regularization is Optimal Transport Robust under Martingale Constraints, Neural Information Processing Systems (**NeurIPS**), 2022.

Thank you! Q&A?

Strong Duality Theorem

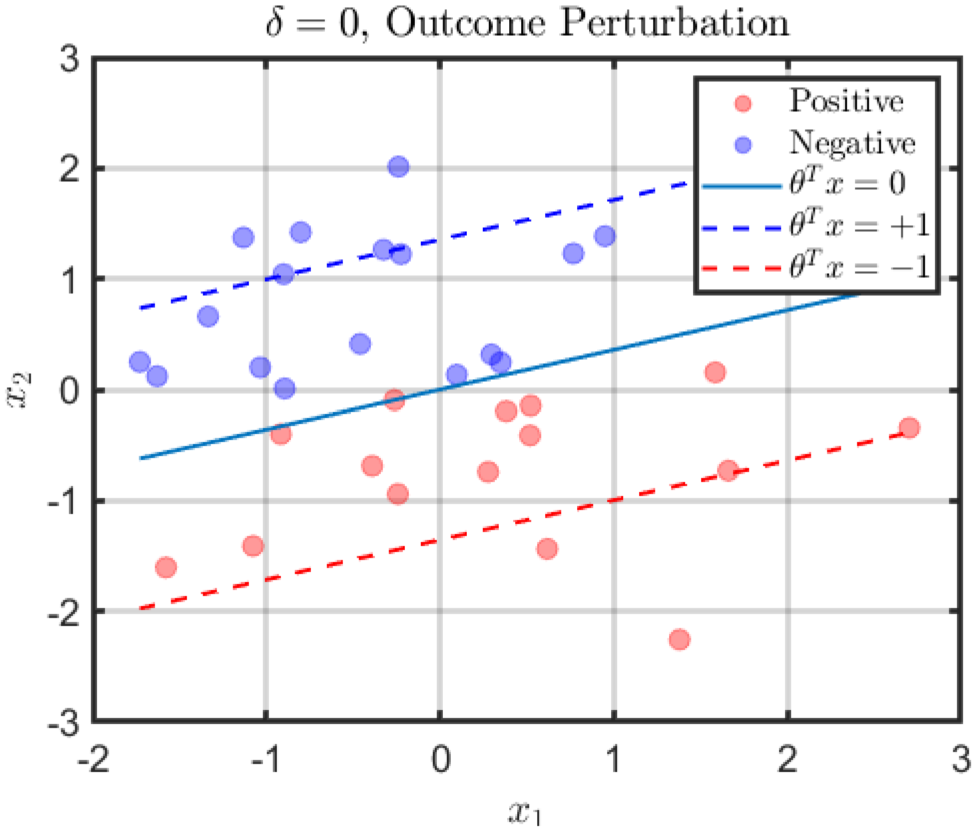
If the reference measure $\hat{\nu}$ is discrete, we have

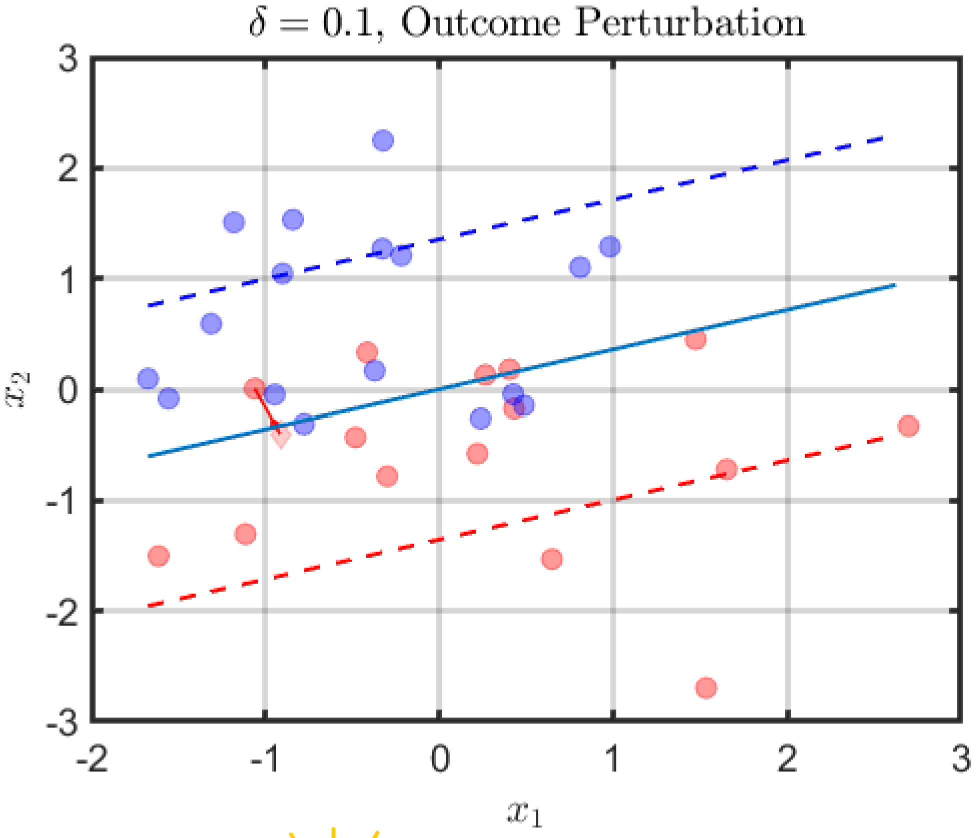
$$\begin{aligned} \sup_{\pi \in \Pi((\Omega\times\mathbb{R}_+)\times(\Omega\times\mathbb{R}_+))} \quad & \mathbb{E}_\pi[\ell(\theta,\xi)\cdot\zeta] \\ \text{s.t.} \quad & \mathbb{E}_\pi\left[c_M((\xi,\zeta),(\xi',\zeta'))\right] \le \delta, \\ & \mathbb{E}_\pi\left[\zeta \mid \zeta'\right] = \zeta', \\ & \Pi_{(\xi',\zeta')}\pi = \hat{\nu} \end{aligned}$$

$$= \inf_{\lambda\ge 0, \alpha\in\mathbb{R}^N} \lambda\delta + \mathbb{E}_{\hat{\nu}}\left[\sup_{(\xi,\zeta)} \ell(\theta,\xi)\cdot\zeta - \lambda c_M((\xi,\zeta),(\xi',\zeta')) + \alpha(\zeta')(\zeta-\zeta')\right]$$

Jiajin Li, Sirui Lin, Jose Blanchet, Viet Anh Nguyen. Tikhonov Regularization is Optimal Transport Robust under Martingale Constraints, Neural Information Processing Systems (**NeurIPS**), 2022.

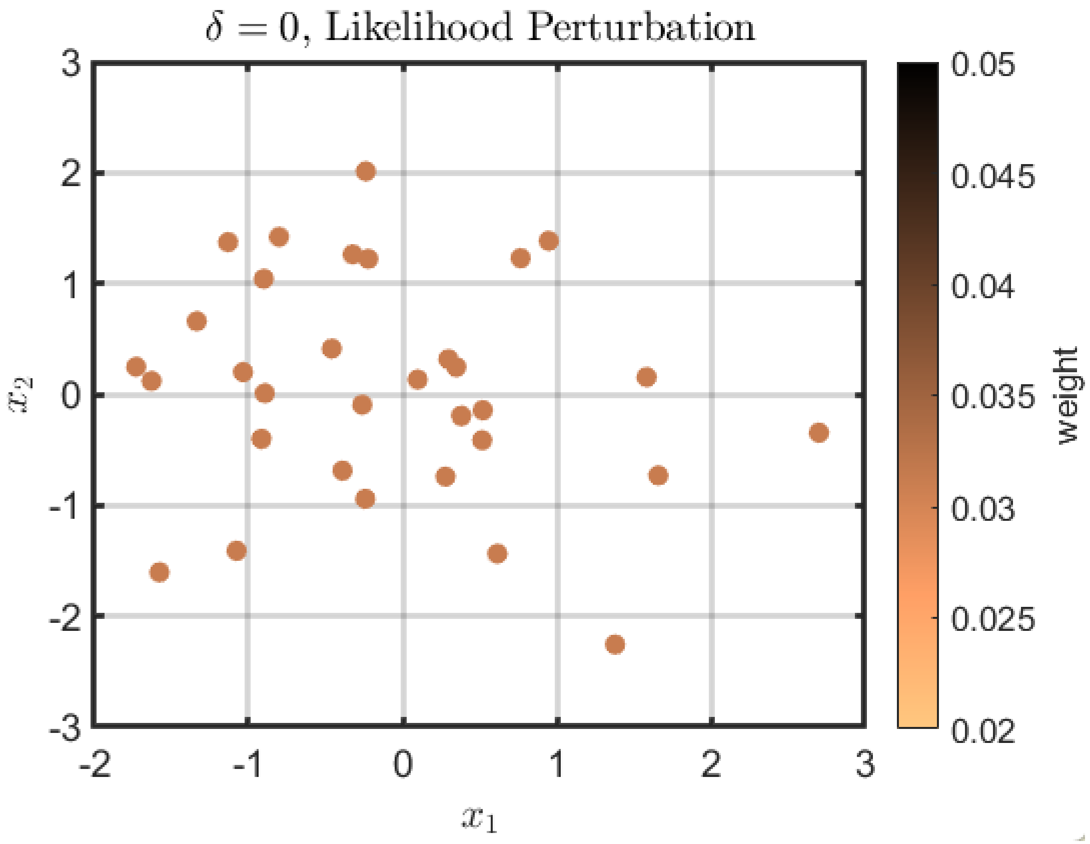
Worst-Case Distribution Visualization

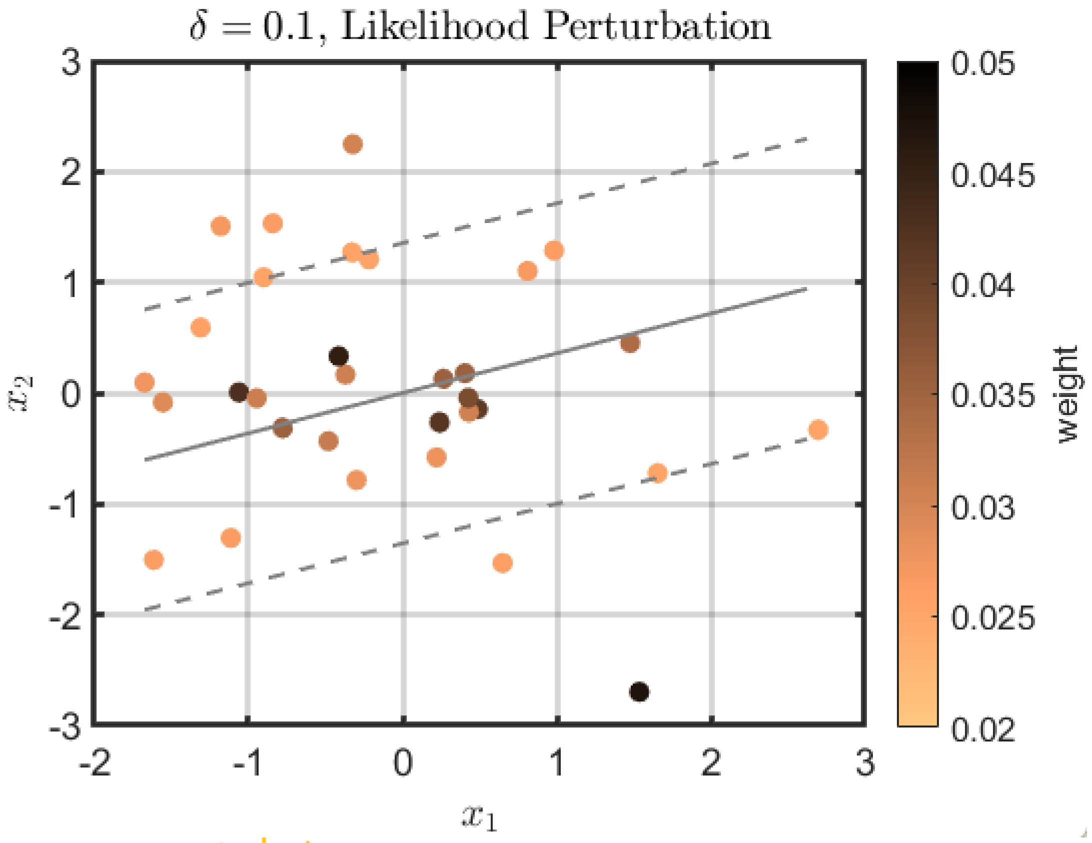




Outcome Perturbation!

Worst-Case Distribution Visualization





Likelihood Perturbation!