Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning

Daniel Kuhn, Peyman Mohajerin Esfahani,
Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh

Risk Analytics and Optimization Chair
École Polytechnique Fédérale de Lausanne
rao.epfl.ch

FNSNF
Swiss National Science Foundation

EPFL

Decision-Making under Uncertainty

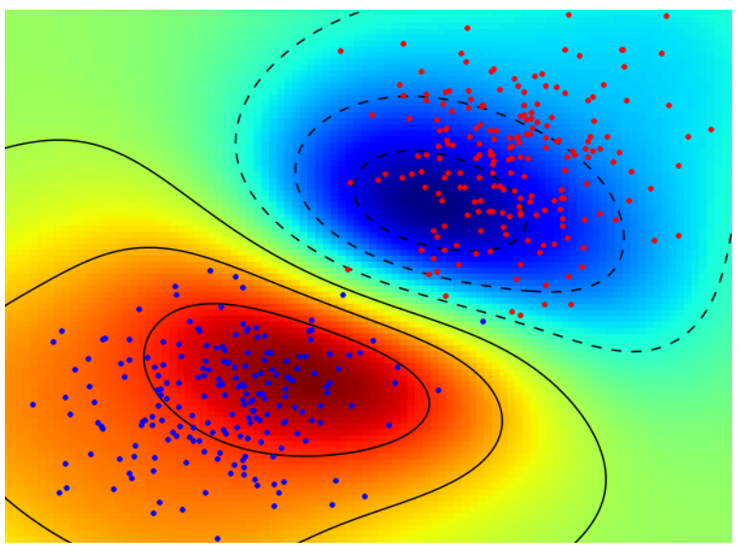
Risk: $\mathcal{R}(\mathbb{P}, \ell) = \mathbb{E}^{\mathbb{P}}[\ell(\boldsymbol{\xi})]$

Optimal risk: $\mathcal{R}(\mathbb{P}, \mathcal{L}) = \inf_{\ell \in \mathcal{L}} \mathbb{E}^{\mathbb{P}}[\ell(\boldsymbol{\xi})]$

Applications:

Supply Chain Mgmt.

Portfolio Mgmt.



Machine Learning

Data-Driven Decision-Making

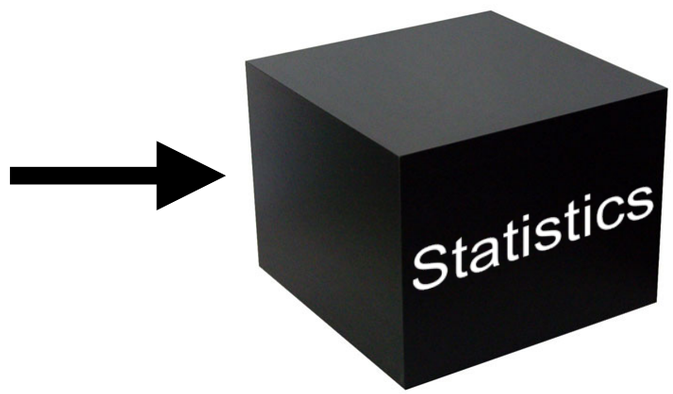
Available information:

Structural: $\mathbb{P}$ supported on $\Xi \subseteq \mathbb{R}^m$

Statistical: $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Nominal problem:

training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N$



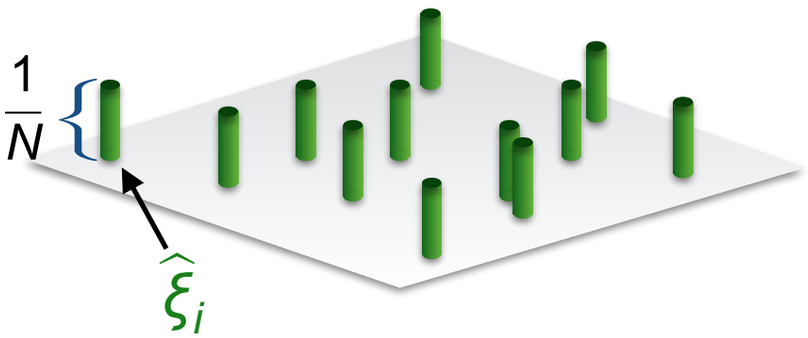
nominal distribution $\widehat{\mathbb{P}}_N$

Nominal risk: $\mathcal{R}(\widehat{\mathbb{P}}_N, \ell)$

Optimal nominal risk: $\mathcal{R}(\widehat{\mathbb{P}}_N, \mathcal{L})$

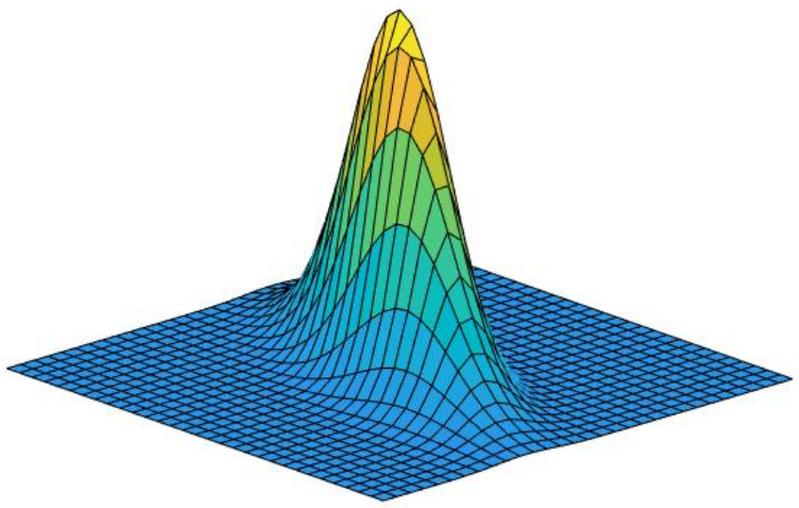
Nominal Distribution

Non-parametric estimators:



Empirical distribution: $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$

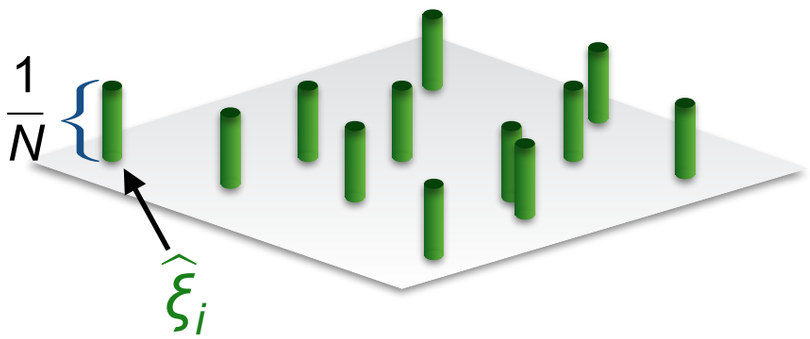
Parametric estimators:



Elliptical distribution: $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}_N, \widehat{\Sigma}_N)$

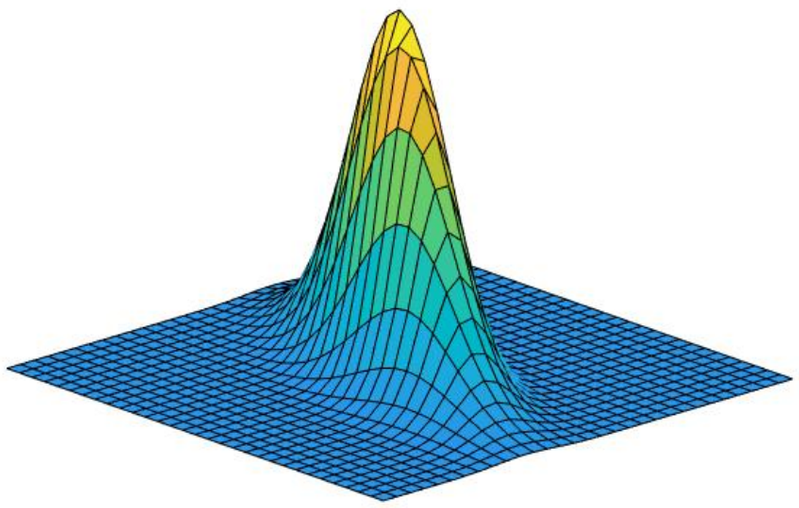
Nominal Distribution

Non-parametric estimators:



Empirical distribution: $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$

Parametric estimators:



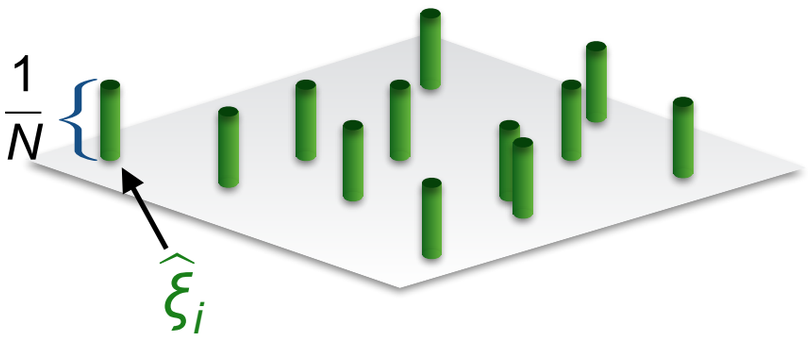
Elliptical distribution: $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}_N, \widehat{\Sigma}_N)$

density generator

Density function: $f(\xi) = C \det(\widehat{\Sigma}_N)^{-1} g((\xi - \widehat{\mu}_N)\widehat{\Sigma}_N^{-1}(\xi - \widehat{\mu}_N))$

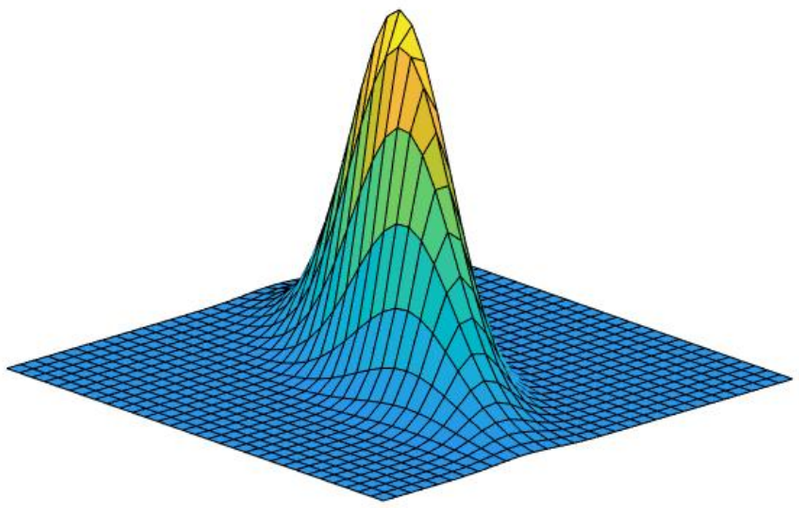
Nominal Distribution

Non-parametric estimators:



Empirical distribution: $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$

Parametric estimators:

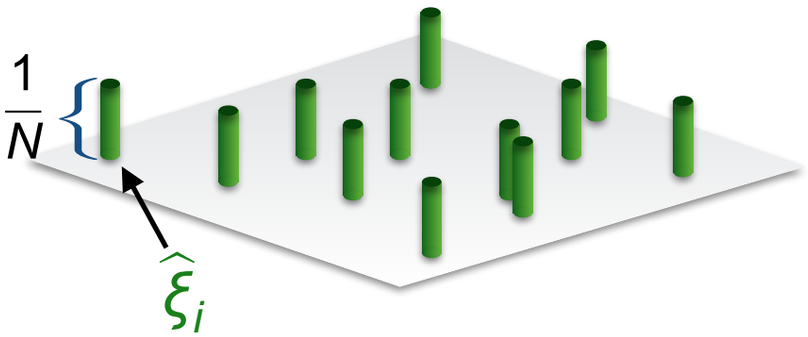


Elliptical distribution: $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}_N, \widehat{\Sigma}_N)$

mean

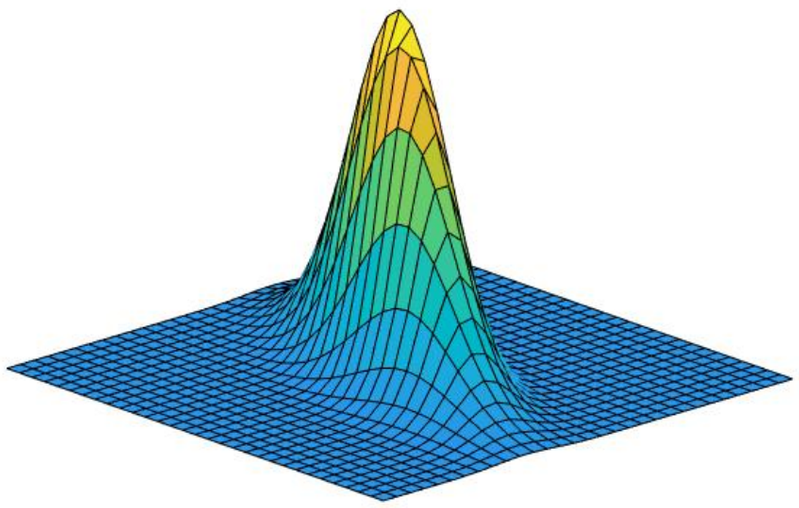
Nominal Distribution

Non-parametric estimators:



Empirical distribution: $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$

Parametric estimators:



Elliptical distribution: $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}_N, \widehat{\Sigma}_N)$

covariance matrix

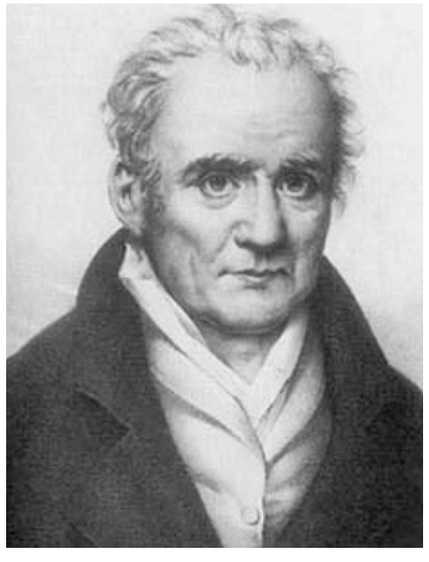
Estimation Errors

The nominal distribution $\widehat{\mathbb{P}}_N$

- depends on $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N$ and is thus random;
- differs from the data-generating distribution $\mathbb{P}$.

Q: How to measure estimation errors?

A: Use the Monge / Kantorovich / Wasserstein distance!



Gaspard Monge
(1746 – 1818)

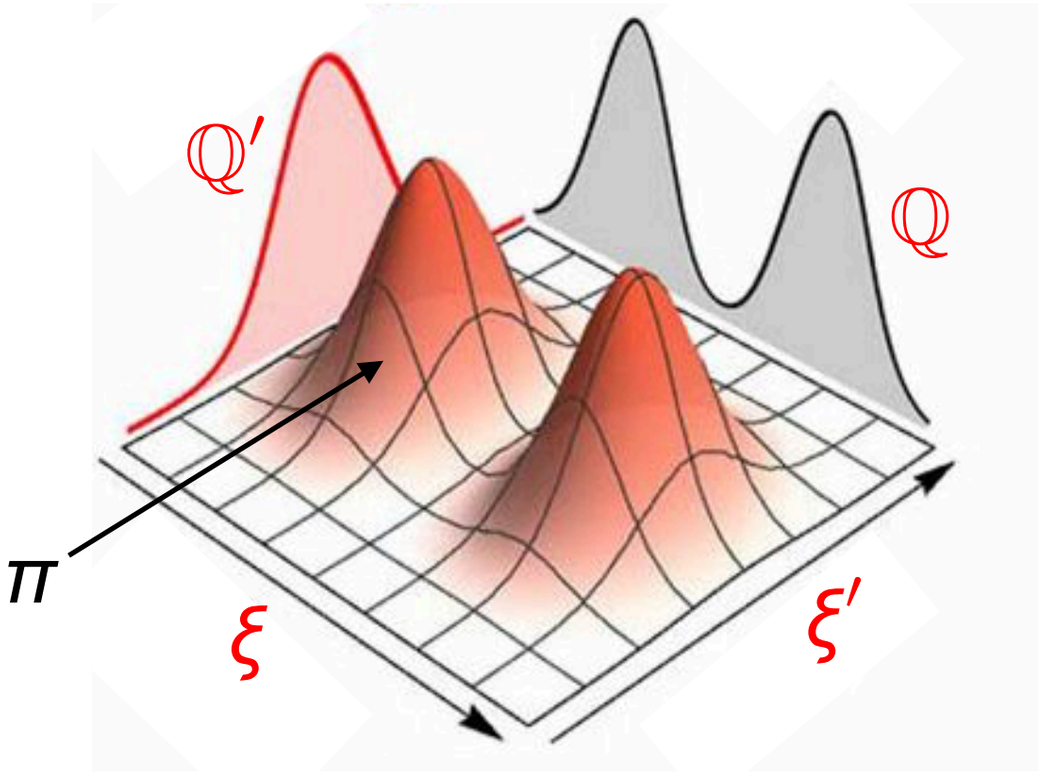
Leonid Kantorovich
(1912 – 1986)

Leonid Vaserštein
(*1944)

Wasserstein Distance

Definition:

$$W_p(\mathbb{Q}, \mathbb{Q}') = \left(\inf_{\pi \in \Pi(\mathbb{Q}, \mathbb{Q}')} \int_{\mathbb{R}^m \times \mathbb{R}^m} \|\xi - \xi'\|^p \, \pi(\mathrm{d}\xi, \mathrm{d}\xi') \right)^{\frac{1}{p}}$$

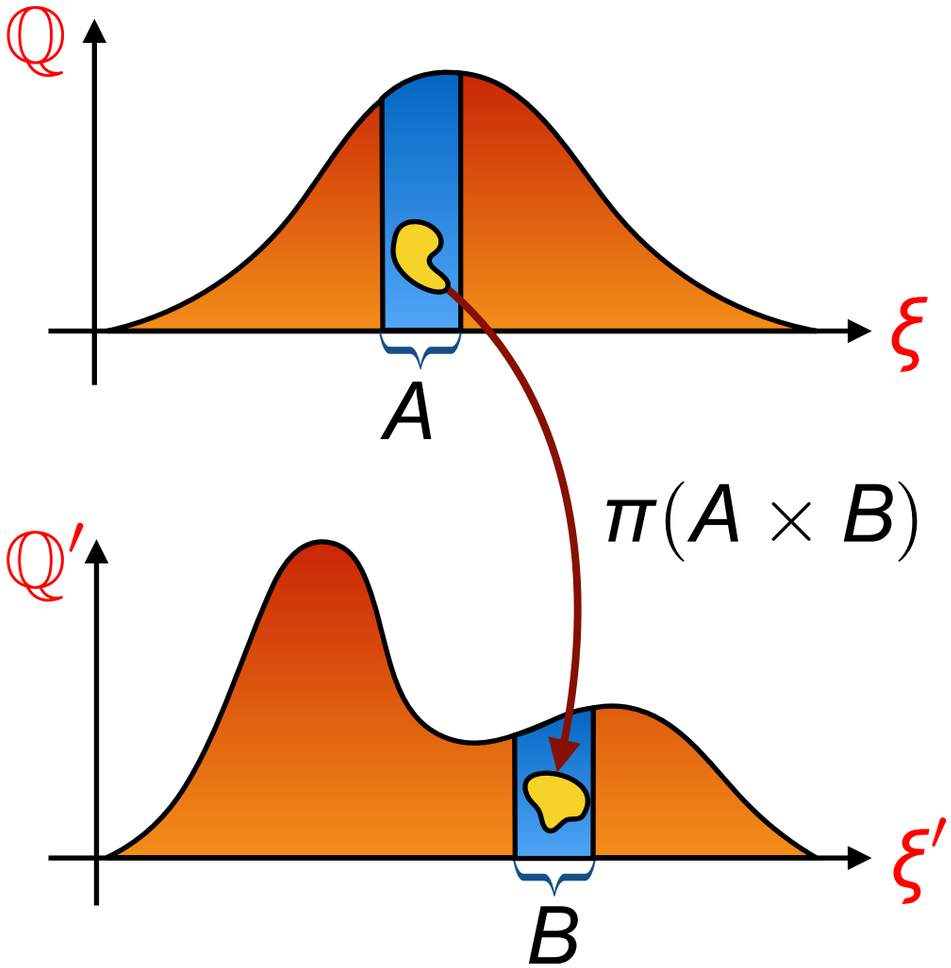


$\Pi(\mathbb{Q}, \mathbb{Q}') =$ set of couplings with marginals $\mathbb{Q}$ and $\mathbb{Q}'$

Wasserstein Distance

Definition:

$$W_p(\mathbb{Q}, \mathbb{Q}') = \left(\inf_{\pi \in \Pi(\mathbb{Q}, \mathbb{Q}')} \int_{\mathbb{R}^m \times \mathbb{R}^m} \| \xi - \xi' \|^p \, \pi(\mathrm{d}\xi, \mathrm{d}\xi') \right)^{\frac{1}{p}}$$



$$\pi(A \times B) = \left\{ \begin{array}{l} \text{mass moved from} \\ \text{source region } A \text{ to} \\ \text{target region } B \end{array} \right.$$

$$\| \xi - \xi' \|^p = \left\{ \begin{array}{l} \text{price paid for moving} \\ \text{mass from } \xi \text{ to } \xi' \end{array} \right.$$

Stability Theory

Assume:

- $\ell(\xi)$ is Lipschitz continuous with $\mathrm{Lip}(\ell) = L$;
- the estimation error in $\widehat{\mathbb{P}}_N$ is small, $W_1(\widehat{\mathbb{P}}_N, \mathbb{P}) \leq \varepsilon$.

Question: How big is the estimation error in $\mathcal{R}(\widehat{\mathbb{P}}_N, \ell)$?

Dupačová, *Ann. Oper. Res.*, 1990; Römisch & Schultz, *Ann. Oper. Res.*, 1991.

Stability Theory

Assume:

- $\ell(\xi)$ is Lipschitz continuous with $\mathrm{Lip}(\ell) = L$;
- the estimation error in $\widehat{\mathbb{P}}_N$ is small, $W_1(\widehat{\mathbb{P}}_N, \mathbb{P}) \leq \varepsilon$.

Question: How big is the estimation error in $\mathcal{R}(\widehat{\mathbb{P}}_N, \ell)$?

$$\left|\mathcal{R}(\widehat{\mathbb{P}}_N, \ell) - \mathcal{R}(\mathbb{P}, \ell)\right| = L \cdot \left|\mathbb{E}^{\widehat{\mathbb{P}}_N}[\ell(\xi)/L] - \mathbb{E}^{\mathbb{P}}[\ell(\xi)/L]\right|$$

$$\leq L \cdot W_1(\widehat{\mathbb{P}}_N, \mathbb{P}) \leq L \cdot \varepsilon$$

↑ K-R Theorem

$\Longrightarrow$ Estimation error at most amplified by L

Dupačová, *Ann. Oper. Res.*, 1990; Römisch & Schultz, *Ann. Oper. Res.*, 1991.

Stability Theory

Assume:

- $\ell(\xi)$ is Lipschitz continuous with $\mathrm{Lip}(\ell) \leq L \;\forall \ell \in \mathcal{L}$;
- the estimation error in $\widehat{\mathbb{P}}_N$ is small, $W_1(\widehat{\mathbb{P}}_N, \mathbb{P}) \leq \varepsilon$.

Question: How big is the estimation error in $\mathcal{R}(\widehat{\mathbb{P}}_N, \mathcal{L})$?

Dupačová, *Ann. Oper. Res.*, 1990; Römisch & Schultz, *Ann. Oper. Res.*, 1991.

Stability Theory

Assume:

- $\ell(\xi)$ is Lipschitz continuous with $\mathrm{Lip}(\ell) \leq L \; \forall \ell \in \mathcal{L}$;
- the estimation error in $\widehat{\mathbb{P}}_N$ is small, $W_1(\widehat{\mathbb{P}}_N, \mathbb{P}) \leq \varepsilon$.

Question: How big is the estimation error in $\mathcal{R}(\widehat{\mathbb{P}}_N, \mathcal{L})$?

$$\left| \mathcal{R}(\widehat{\mathbb{P}}_N, \mathcal{L}) - \mathcal{R}(\mathbb{P}, \mathcal{L}) \right| \leq L \cdot \varepsilon$$

$\Longrightarrow$ Estimation error at most amplified by L

Dupačová, *Ann. Oper. Res.*, 1990; Römisch & Schultz, *Ann. Oper. Res.*, 1991.

Distributionally Robust Optimization (DRO)

Nominal risk: $\mathcal{R}(\widehat{\mathbb{P}}_N, \ell)$

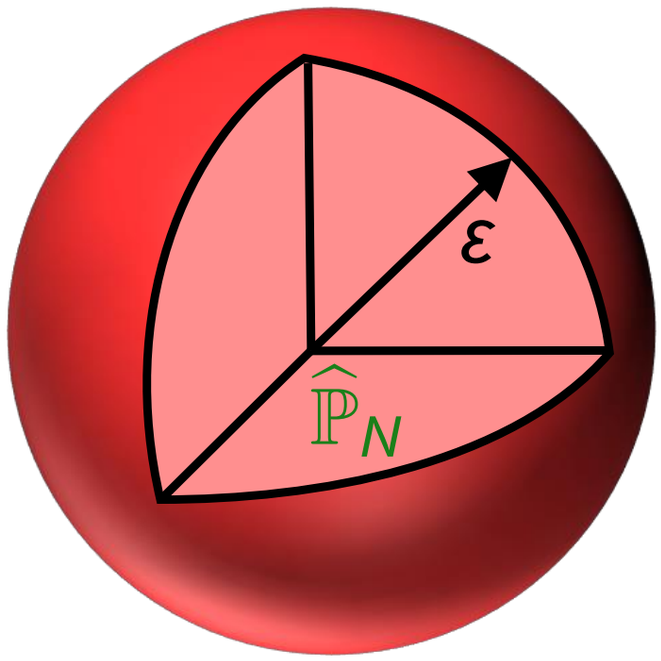
Optimal nominal risk: $\mathcal{R}(\widehat{\mathbb{P}}_N, \mathcal{L})$

Q: If we have found a $\widehat{\mathbb{P}}_N$ that approximates $\mathbb{P}$ best in Wasserstein distance, how can we further reduce the out-of-sample risk?

A: Robustify the risk w.r.t. all distributions $\mathbb{Q}$ with $W_p(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon$

Wasserstein DRO

Definition: $\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N) = \left\{ \mathbb{Q} \in \mathcal{P}(\Xi) : W_p(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \varepsilon \right\}$



Contains every $\mathbb{Q}$ obtainable by reshaping $\widehat{\mathbb{P}}_N$ at a cost of at most ε

Worst-case risk: $\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) = \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)]$

Worst-case optimal risk: $\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \mathcal{L}) = \inf_{\ell \in \mathcal{L}} \mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell)$

Pflug & Wozabal, *Quant. Finance*, 2007; Pflug & Pichler, *Springer*, 2014.

Three Simple Bounds on the Worst-Case Expectation

Robust Lower Bound (any p)

Theorem: If $\widehat{\mathbb{P}}_N$ is the empirical distribution, then

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \geq \left\{ \begin{array}{lll} \sup & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \ell(\widehat{\xi}_i + \theta_i) & \\ \text{s.t.} & \theta_i \in \mathbb{R}^m & \forall i \in [N] \\ & \widehat{\xi}_i + \theta_i \in \Xi & \forall i \in [N] \\ & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \|\theta_i\|^p \leq \varepsilon^p. & \end{array} \right.$$

Lipschitz Regularization (any p)

Theorem: The worst-case risk is bounded above by the Lipschitz-regularized nominal risk, i.e.,

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \leq \mathcal{R}(\widehat{\mathbb{P}}_N, \ell) + \varepsilon \cdot \mathsf{Lip}(\ell).$$

Gelbrich Bound ($p = 2$)

Theorem: If $\mathbb{Q} \sim (\mu, \Sigma)$ and $\mathbb{Q}' \sim (\mu', \Sigma')$, then[1)]

$$W_2(\mathbb{Q}, \mathbb{Q}') \geq \sqrt{\|\mu - \mu'\|_2^2 + \mathsf{Tr}\left(\Sigma + \Sigma' - 2\left(\Sigma^{\frac{1}{2}}\Sigma'\Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)}.$$

[1)] Gelbrich, *Mathematische Nachrichten*, 1990.

Gelbrich Bound ($p = 2$)

Theorem: If $\mathbb{Q} \sim (\mu, \Sigma)$ and $\mathbb{Q}' \sim (\mu', \Sigma')$, then[1)]

$$W_2(\mathbb{Q}, \mathbb{Q}') \geq \underbrace{\sqrt{\|\mu - \mu'\|_2^2 + \mathsf{Tr}\left(\Sigma + \Sigma' - 2\left(\Sigma^{\frac{1}{2}}\Sigma'\Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)}}_{= d_{\mathsf{G}}\left((\mu, \Sigma), (\mu', \Sigma')\right)}.$$

[1)] Gelbrich, *Mathematische Nachrichten*, 1990.

Gelbrich Bound ($p = 2$)

Theorem: If $\mathbb{Q} \sim (\mu, \Sigma)$ and $\mathbb{Q}' \sim (\mu', \Sigma')$, then[1)]

$$W_2(\mathbb{Q}, \mathbb{Q}') \geq \underbrace{\sqrt{\|\mu - \mu'\|_2^2 + \mathsf{Tr}\left(\Sigma + \Sigma' - 2\left(\Sigma^{\frac{1}{2}}\Sigma'\Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)}}_{= d_{\mathsf{G}}\left((\mu, \Sigma), (\mu', \Sigma')\right)}.$$

The bound is exact if $\mathbb{Q}, \mathbb{Q}'$ are elliptical with the same generator.

[1)] Gelbrich, *Mathematische Nachrichten*, 1990.

Gelbrich Bound ($p = 2$)

Uncertainty set for mean vectors and covariance matrices:

$$\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = \left\{ (\mu, \Sigma) \in \mathbb{R}^m \times \mathbb{S}_+^m : d_{\mathsf{G}} \left((\widehat{\mu}, \widehat{\Sigma}), (\mu, \Sigma) \right) \leq \varepsilon^2 \right\}$$

Gelbrich Bound ($p = 2$)

Uncertainty set for mean vectors and covariance matrices:

$$\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = \left\{ (\mu, \Sigma) \in \mathbb{R}^m \times \mathbb{S}_+^m : d_{\mathsf{G}} \left((\widehat{\mu}, \widehat{\Sigma}), (\mu, \Sigma) \right) \leq \varepsilon^2 \right\}$$

Mean-covariance “projection”:

$$T : \mathbb{Q} \mapsto \left(\mathbb{E}^{\mathbb{Q}}[\xi], \mathbb{E}^{\mathbb{Q}}[\xi\, \xi^\top] - \mathbb{E}^{\mathbb{Q}}[\xi]\, \mathbb{E}^{\mathbb{Q}}[\xi]^\top \right)$$

Gelbrich Bound ($p = 2$)

Uncertainty set for mean vectors and covariance matrices:

$$\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = \left\{ (\mu, \Sigma) \in \mathbb{R}^m \times \mathbb{S}_+^m : d_{\mathsf{G}} \left((\widehat{\mu}, \widehat{\Sigma}), (\mu, \Sigma) \right) \leq \varepsilon^2 \right\}$$

Mean-covariance “projection”:

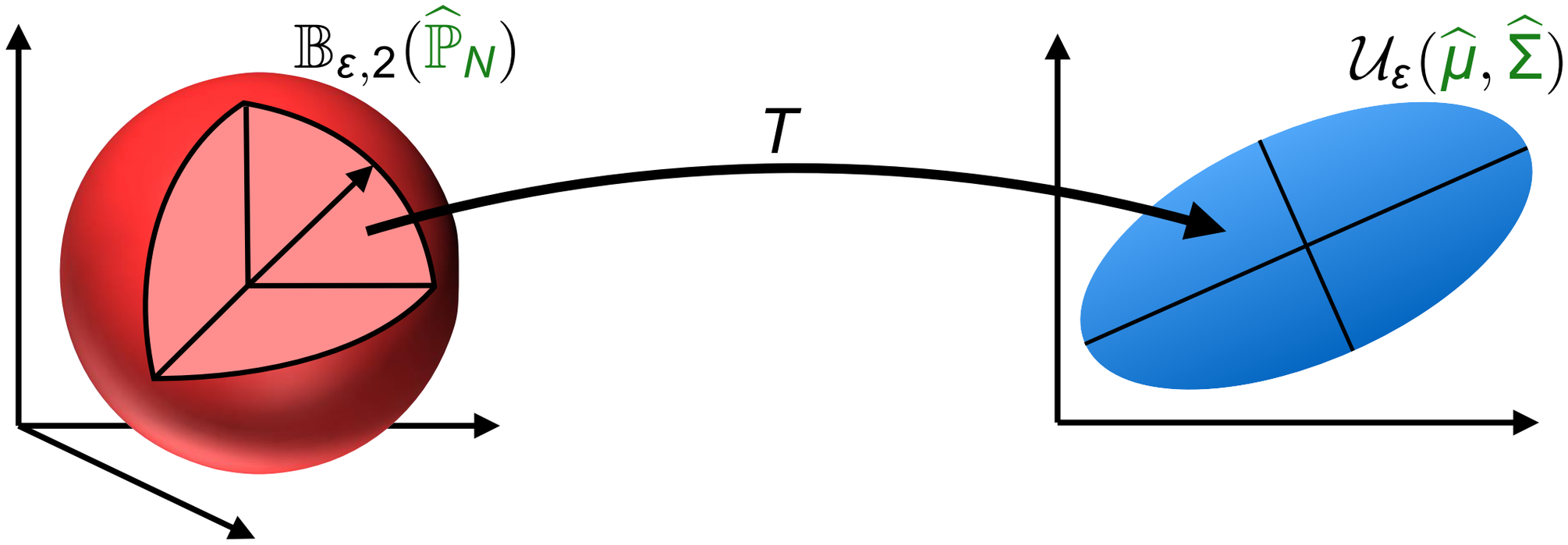
$$T : \mathbb{Q} \mapsto \big(\underbrace{\mathbb{E}^{\mathbb{Q}}[\xi]}_{\mu}, \underbrace{\mathbb{E}^{\mathbb{Q}}[\xi\, \xi^\top] - \mathbb{E}^{\mathbb{Q}}[\xi]\, \mathbb{E}^{\mathbb{Q}}[\xi]^\top}_{\Sigma} \big)$$

Gelbrich Bound ($p = 2$)

Uncertainty set for mean vectors and covariance matrices:

$$\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = \left\{ (\mu, \Sigma) \in \mathbb{R}^m \times \mathbb{S}_+^m : d_{\mathsf{G}} \left((\widehat{\mu}, \widehat{\Sigma}), (\mu, \Sigma) \right) \leq \varepsilon^2 \right\}$$

Mean-covariance “projection”:

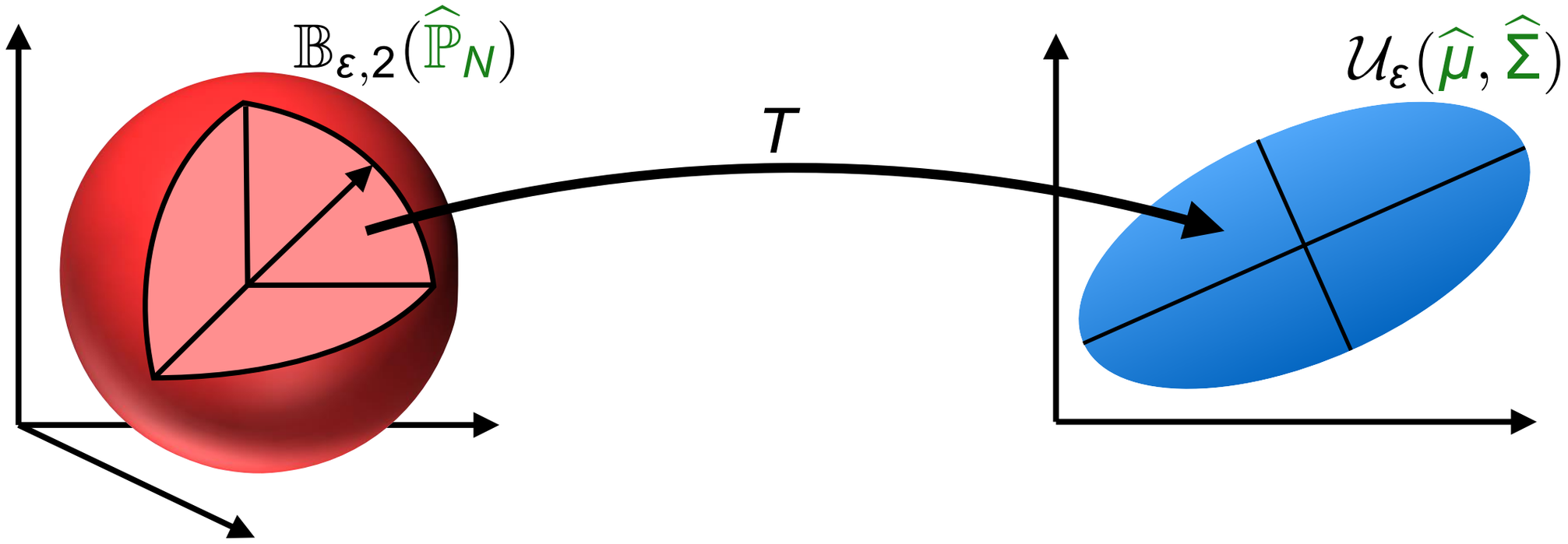


Gelbrich Bound ($p = 2$)

Uncertainty set for mean vectors and covariance matrices:

$$\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = \left\{ (\mu, \Sigma) \in \mathbb{R}^m \times \mathbb{S}_+^m : d_{\mathsf{G}} \left((\widehat{\mu}, \widehat{\Sigma}), (\mu, \Sigma) \right) \leq \varepsilon^2 \right\}$$

Mean-covariance “projection”:



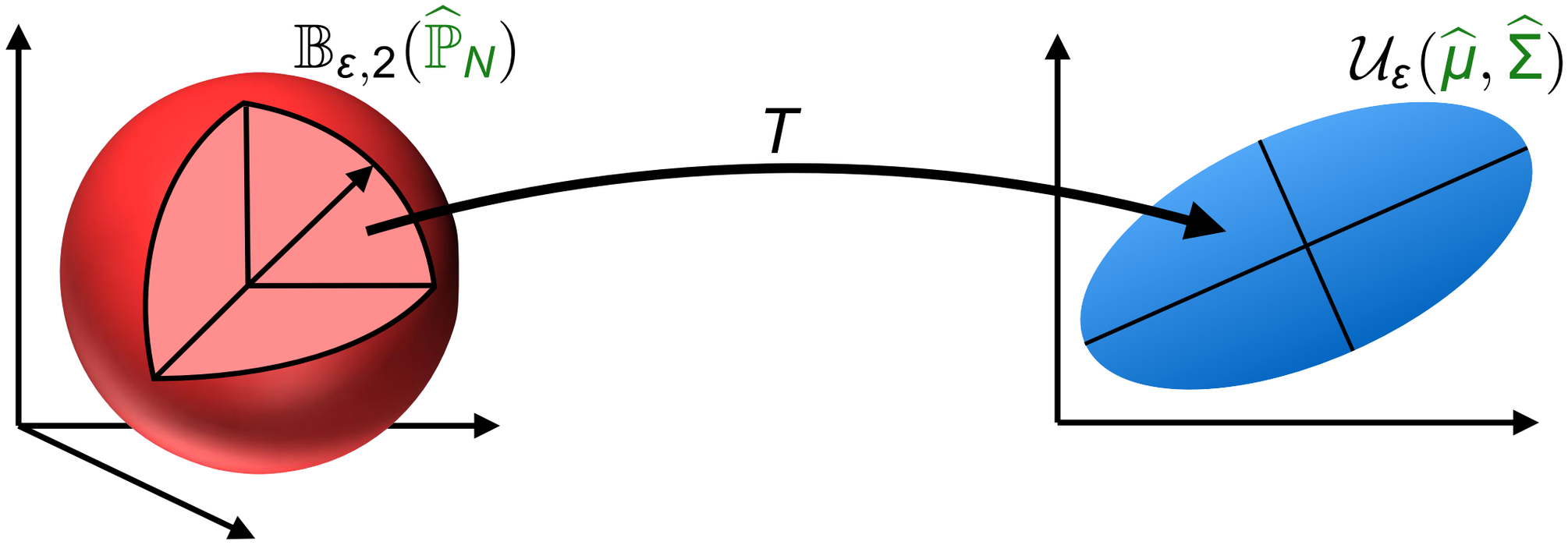
Note: $T(\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)) \subseteq \mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$

Gelbrich Bound ($p = 2$)

Uncertainty set for mean vectors and covariance matrices:

$$\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = \left\{ (\mu, \Sigma) \in \mathbb{R}^m \times \mathbb{S}_+^m : d_{\mathsf{G}}\left((\widehat{\mu}, \widehat{\Sigma}), (\mu, \Sigma) \right) \leq \varepsilon^2 \right\}$$

Mean-covariance “projection”:



Note: $T(\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)) = \mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ if $\widehat{\mathbb{P}}_N$ is elliptical!

Gelbrich Bound ($p = 2$)

The Gelbrich hull:

$$\mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = T^{-1}(\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}))$$

Gelbrich Bound ($p = 2$)

The Gelbrich hull:

$$\mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = T^{-1}(\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}))$$

Note: $\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N) \subseteq \mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$

Gelbrich Bound ($p = 2$)

The Gelbrich hull:

$$\mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = T^{-1}(\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}))$$

Note: $\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N) \subseteq \mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$

Theorem: If $\widehat{\mathbb{P}}_N$ has mean $\widehat{\mu}$ and covariance matrix $\widehat{\Sigma}$, then

$$\mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N, \ell) \leq \sup_{\mathbb{Q} \in \mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})} \mathbb{E}^{\mathbb{Q}}[\ell(\boldsymbol{\xi})].$$

Gelbrich Bound ($p = 2$)

The Gelbrich hull:

$$\mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}) = T^{-1}(\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}))$$

Note: $\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N) \subseteq \mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$

Theorem: If $\widehat{\mathbb{P}}_N$ has mean $\widehat{\mu}$ and covariance matrix $\widehat{\Sigma}$, then

$$\mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N, \ell) \leq \underbrace{\sup_{\mathbb{Q} \in \mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})} \mathbb{E}^{\mathbb{Q}}[\ell(\boldsymbol{\xi})]}_{= \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell)}.$$

Efficient Computation of the Worst-Case Risk

Strong Duality

Theorem: If ℓ is u.s.c. and integrable under $\widehat{\mathbb{P}}_N$, then[1)]

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) = \inf_{\gamma \geq 0} \; \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\ell_\gamma(\xi) \right] + \gamma \varepsilon^p,$$

where $\ell_\gamma(\xi) = \sup_{z \in \Xi} \ell(z) - \gamma \|z - \xi\|^p$.

[1)] Mohajerin Esfahani & Kuhn, *Math. Program.*, 2018; Guan & Zhao, *Oper. Res. Lett.*, 2018; Blanchet & Murthy, *Math. Oper. Res.*, 2019; Gao & Kleywegt, *arxiv*, 2017.

Strong Duality

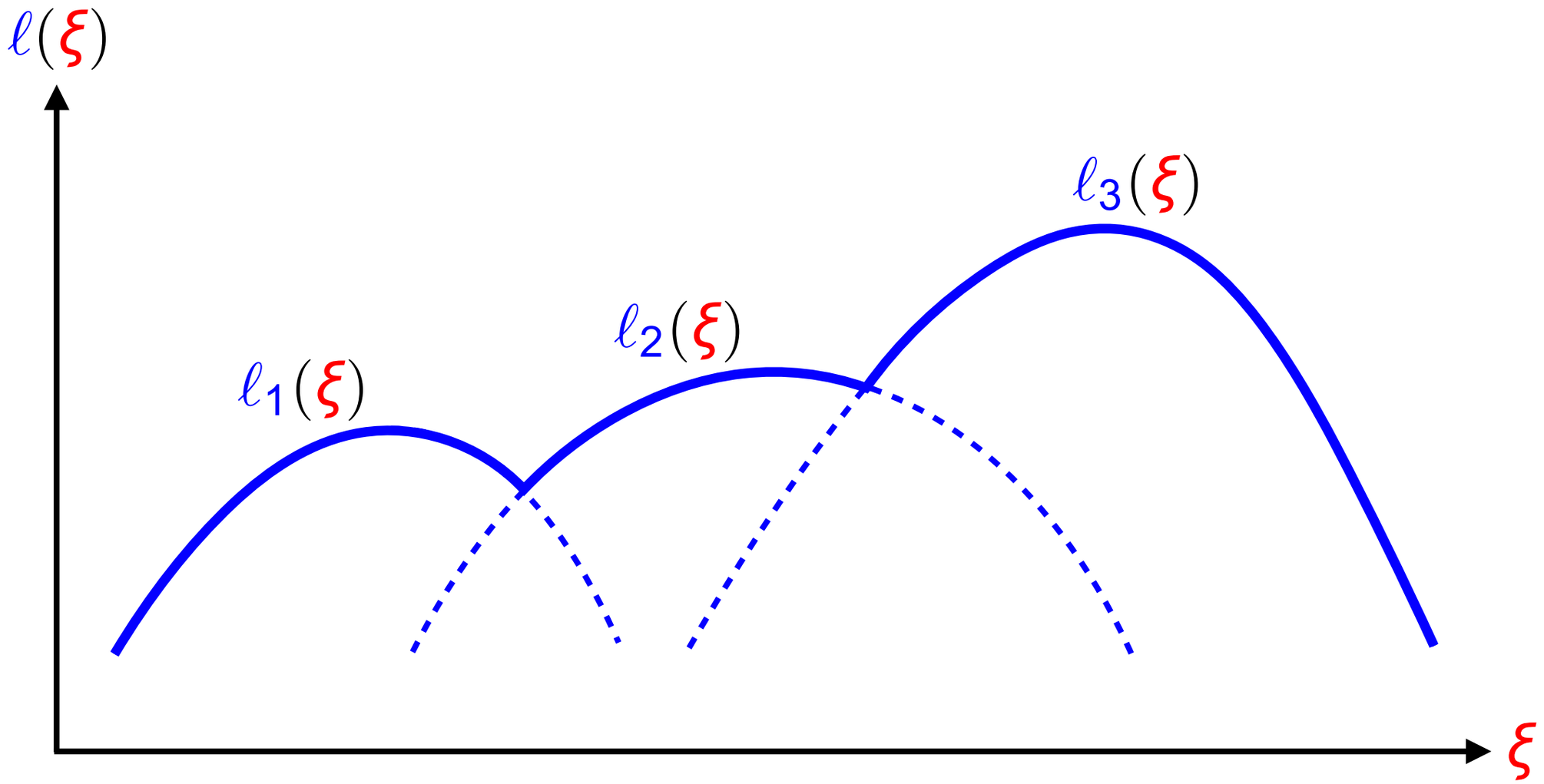
Theorem: If ℓ is u.s.c. and integrable under $\widehat{\mathbb{P}}_N$, then[1)]

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) = \inf_{\gamma \geq 0} \; \mathbb{E}^{\widehat{\mathbb{P}}_N} \left[\ell_\gamma(\xi) \right] + \gamma \varepsilon^p,$$

where $\ell_\gamma(\xi) = \underbrace{\sup_{z \in \Xi} \ell(z) - \gamma \|z - \xi\|^p}_{\text{Moreau envelope}}$.

[1)] Mohajerin Esfahani & Kuhn, *Math. Program.*, 2018; Guan & Zhao, *Oper. Res. Lett.*, 2018; Blanchet & Murthy, *Math. Oper. Res.*, 2019; Gao & Kleywegt, *arxiv*, 2017.

Piecewise Concave Loss



$\ell(\xi) = \max_j \ell_j(\xi)$, where each ℓ_j is proper, concave and u.s.c.

Piecewise Concave Loss

Theorem: If Ξ is convex, $\widehat{\mathbb{P}}_N$ is the empirical distribution and $\ell(\xi)$ is piecewise concave, then

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \;=$$

$$\begin{array}{lll} \inf & \gamma\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{s.t.} & \gamma \in \mathbb{R}_+,\ s_i \in \mathbb{R},\ u_{ij} \in \mathbb{R}^m,\ v_{ij} \in \mathbb{R}^m & \forall i,j \\ & [-\ell_j]^*(u_{ij} - v_{ij}) + \sigma_\Xi(v_{ij}) - u_{ij}{}^\top \widehat{\xi}_i + \varphi(q)\,\gamma \left\| \dfrac{u_{ij}}{\gamma} \right\|_*^q \le s_i & \forall i,j, \end{array}$$

where $\frac{1}{p} + \frac{1}{q} = 1$ and $\varphi(q) = (q-1)^{q-1}/q^q$.

Zhen, Kuhn & Wiesemann, *Working Paper*, 2019.

Piecewise Concave Loss

Finite convex optimization problem!
Size polynomial in # samples and # pieces!

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \;=$$

$$\begin{array}{ll} \inf & \gamma\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{s.t.} & \gamma \in \mathbb{R}_+,\ s_i \in \mathbb{R},\ u_{ij} \in \mathbb{R}^m,\ v_{ij} \in \mathbb{R}^m \qquad \forall i,j \\ & \underbrace{[-\ell_j]^*(u_{ij} - v_{ij})}_{\text{convex conjugate of } -\ell_j} + \underbrace{\sigma_\Xi(v_{ij})}_{\text{support function of } \Xi} - u_{ij}^\top \widehat{\xi}_i + \varphi(q)\,\underbrace{\gamma \left\| \frac{u_{ij}}{\gamma} \right\|_*^q}_{\text{perspective of } \|\cdot\|_*^q} \le s_i \quad \forall i,j \end{array}$$

Zhen, Kuhn & Wiesemann, *Working Paper*, 2019.

Piecewise Concave Loss

Theorem: If Ξ is convex, $\widehat{\mathbb{P}}_N$ is the empirical distribution and $\ell(\xi)$ is piecewise concave, then

$$\mathcal{R}_{\varepsilon,p}(\widehat{\mathbb{P}}_N, \ell) \;=\; \max \quad \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ij}\ell_j\left(\widehat{\xi}_i + \frac{\theta_{ij}}{\alpha_{ij}}\right)$$

$$\begin{array}{lll}
\text{s.t.} & \alpha_{ij} \in \mathbb{R}_+,\ \theta_{ij} \in \mathbb{R}^m & \forall i \in [N],\ \forall j \in [J] \\
& \widehat{\xi}_i + \dfrac{\theta_{ij}}{\alpha_{ij}} \in \Xi & \forall i \in [N],\ \forall j \in [J] \\
& \displaystyle\sum_{j=1}^{J} \alpha_{ij} = 1 & \forall i \in [N] \\
& \displaystyle\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} \alpha_{ij}\left\|\frac{\theta_{ij}}{\alpha_{ij}}\right\|^p \le \varepsilon^p. &
\end{array}$$

Zhen, Kuhn & Wiesemann, *Working Paper*, 2019.

Main Takeaways

- if ℓ is piecewise concave and $\widehat{\mathbb{P}}_N$ empirical, the worst-case risk can be computed by convex optimization for *any* $p \in [1, \infty]$;

  $\Longrightarrow$ generalizes earlier results for $p = 1$;[1)]

- if ℓ is concave and $\widehat{\mathbb{P}}_N$ empirical, the robust lower bound is exact;

- the dual convex program
  - provides a finite reduction of the worst-case risk problem;
  - is always solvable;
  - can be used to find a worst-case distribution (may not exist for $p = 1$, always exists for $p > 1$).

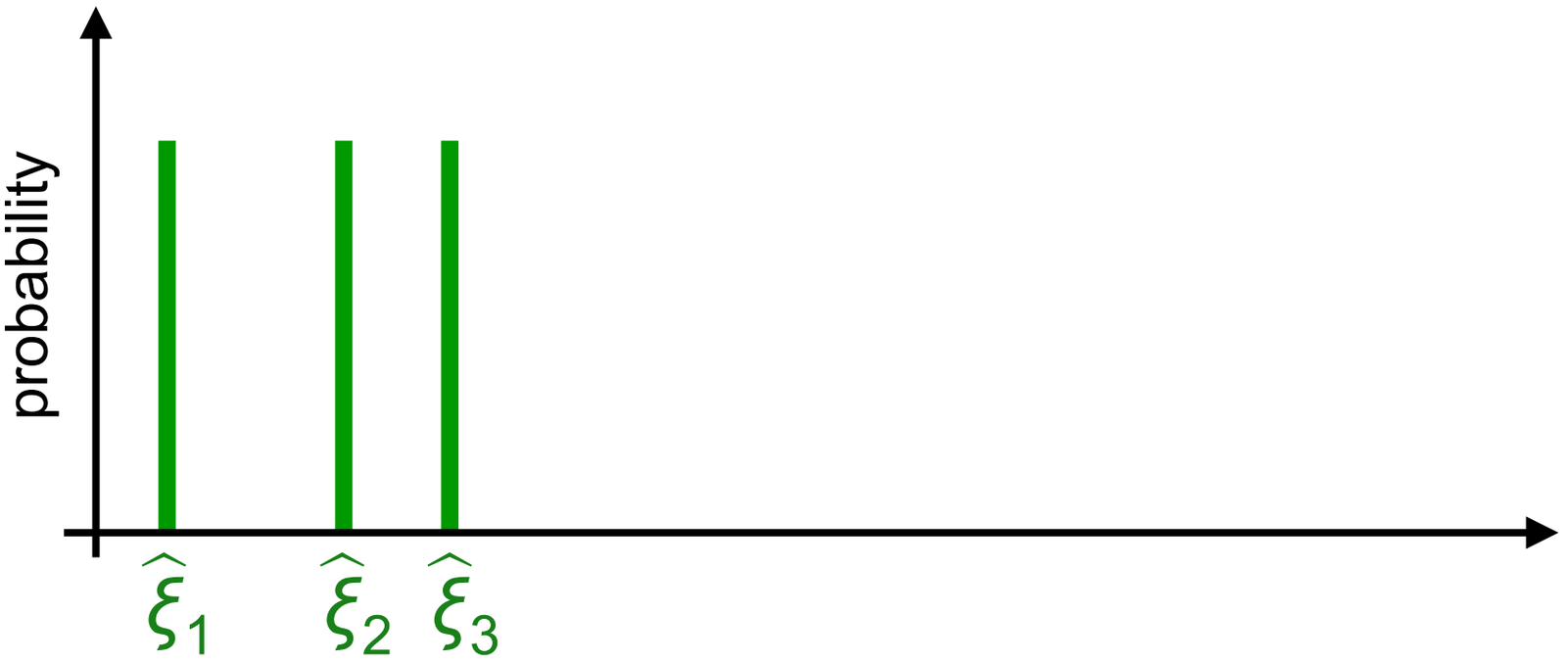
[1)] Mohajerin Esfahani & Kuhn, *Math. Program.*, 2018.

Convex Loss

Theorem: If $\Xi = \mathbb{R}^m$ and $\ell(\xi)$ is convex, then

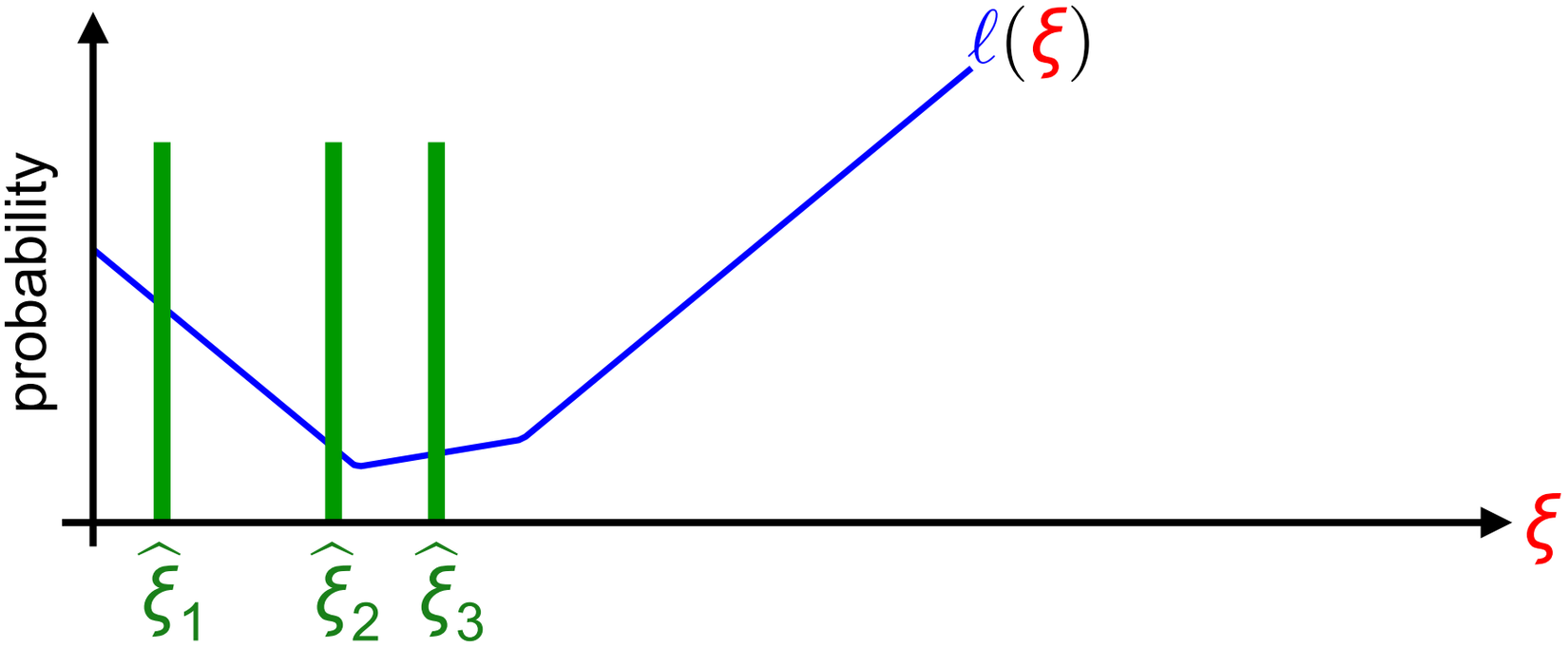
$$\mathcal{R}_{\varepsilon,1}(\widehat{\mathbb{P}}_N, \ell) = \mathcal{R}(\widehat{\mathbb{P}}_N, \ell) + \varepsilon \cdot \mathrm{Lip}(\ell).$$

Worst-Case Risk for $p = 1$



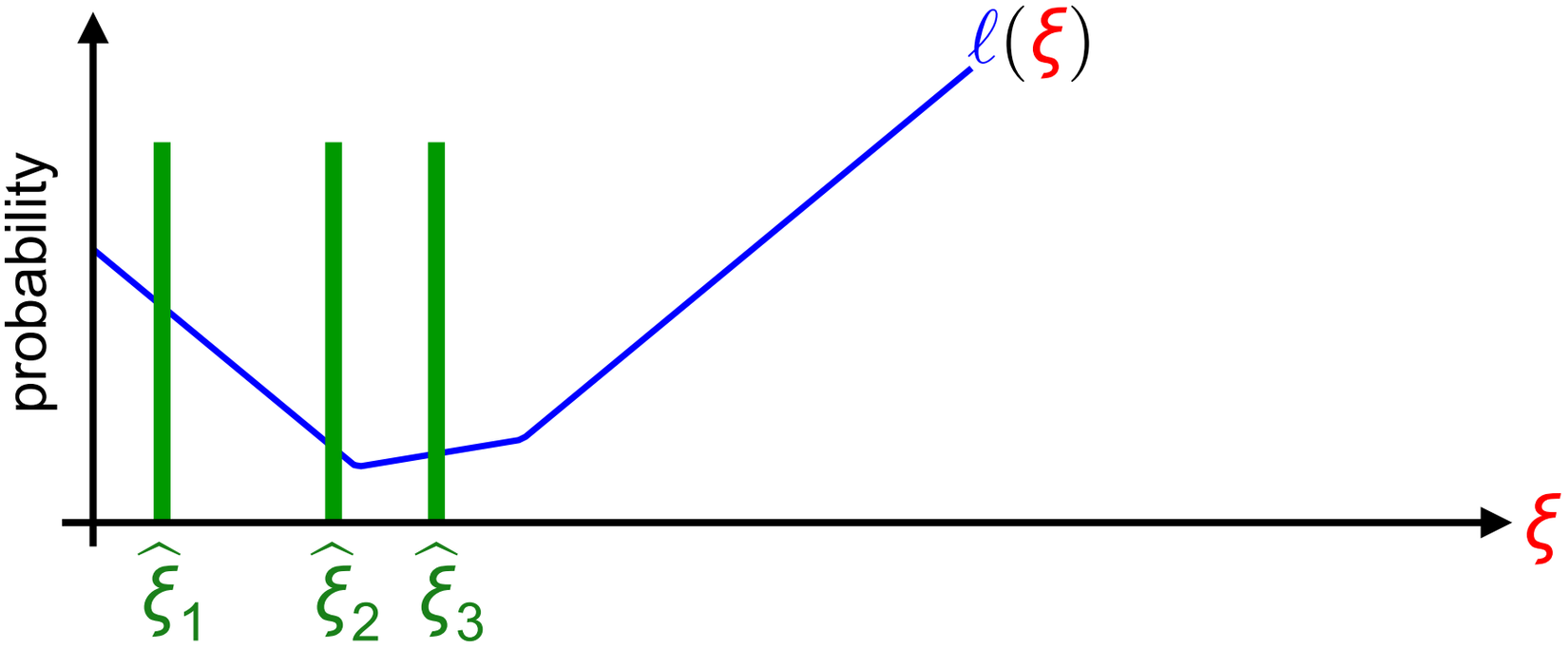
nominal distribution = empirical distribution
support set $\Xi = \mathbb{R}$
Wasserstein radius $\varepsilon = 0.7$

Worst-Case Risk for $p = 1$



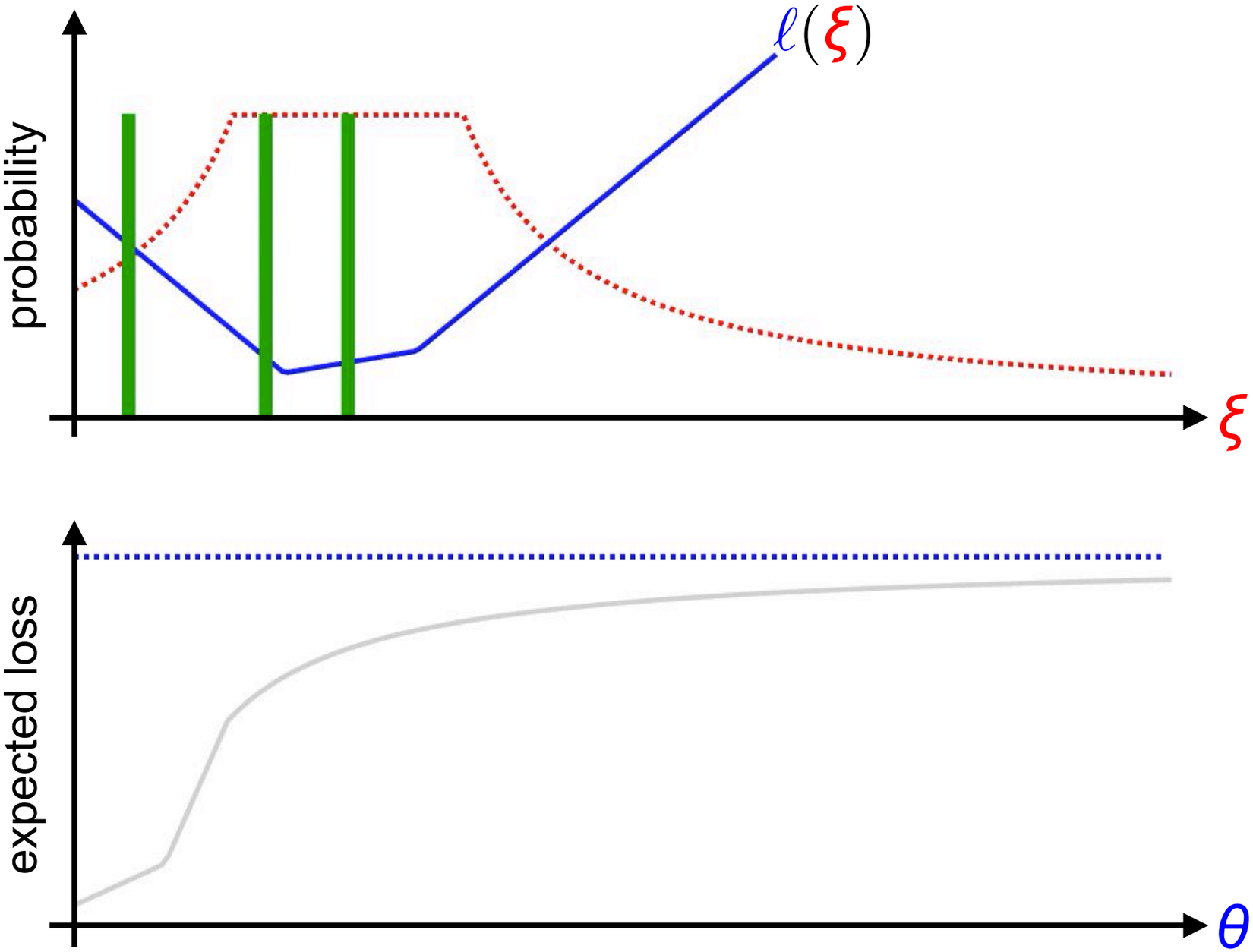
Convex and Lipschitz continuous loss

Worst-Case Risk for $p = 1$

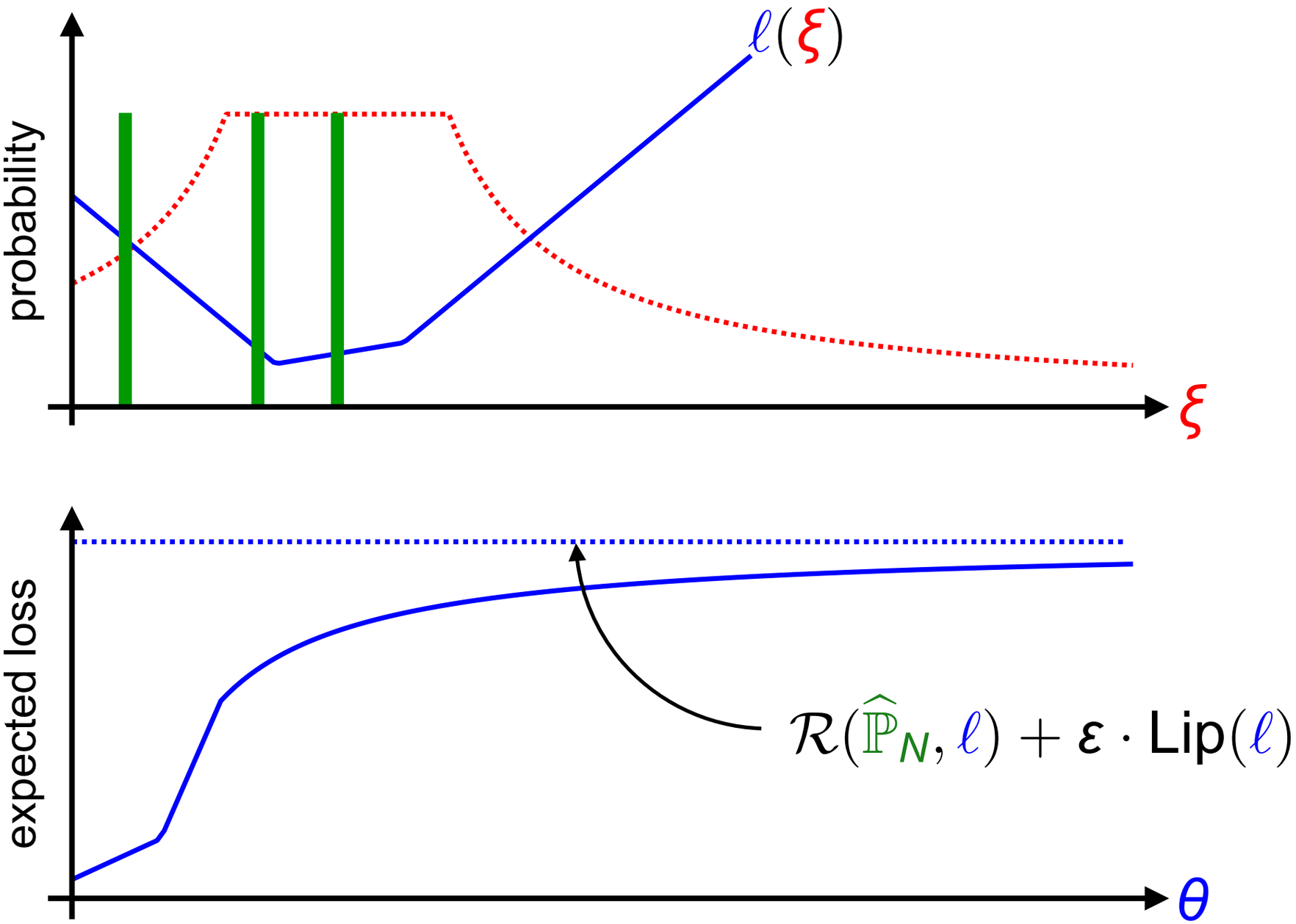


Increase $\widehat{\xi}_3$ by θ (and reduce its probability if necessary)

Worst-Case Risk for $p = 1$



Worst-Case Risk for $p = 1$



Main Takeaways

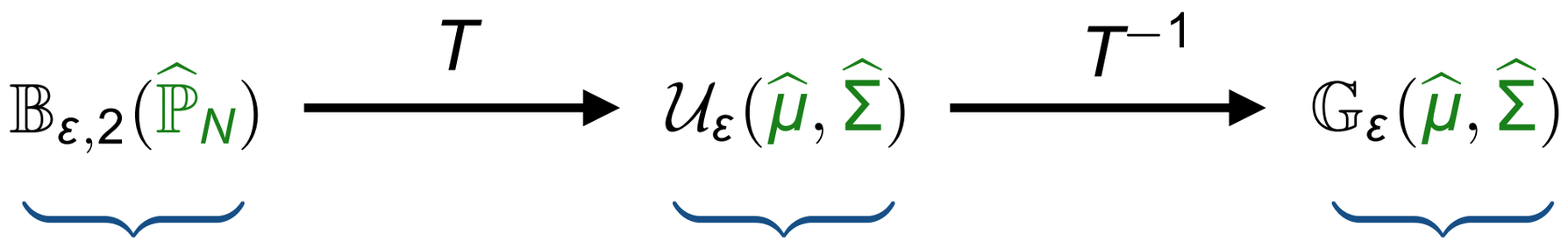
- if ℓ is convex, $\Xi = \mathbb{R}^m$ and $p = 1$, the worst case risk coincides with the Lipschitz-regularized nominal risk;
- no distribution attains the worst-case risk;
- the worst-case risk is attained asymptotically: push a linearly decreasing mass to infinity in the direction of steepest ascent;
- computing the worst-case risk is generally hard but tractable in special cases.

Computing the Gelbrich Bound

Recall: $\mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N, \ell) \leq \overline{\mathcal{R}}_{\varepsilon}(\widehat{\mu}, \widehat{\Sigma}, \ell)$

Computing the Gelbrich Bound

Recall: $\mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N, \ell) \leq \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell)$



Computing the Gelbrich Bound

Recall: $\mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N, \ell) \leq \overline{\mathcal{R}}_{\varepsilon}(\widehat{\mu}, \widehat{\Sigma}, \ell)$

$$\mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N) \xrightarrow{T} \mathcal{U}_{\varepsilon}(\widehat{\mu}, \widehat{\Sigma}) \xrightarrow{T^{-1}} \mathbb{G}_{\varepsilon}(\widehat{\mu}, \widehat{\Sigma})$$

Wasserstein ball — mean-covariance uncertainty set — Gelbrich hull

$$\mathbb{G}_{\varepsilon}(\widehat{\mu}, \widehat{\Sigma}) = \bigcup_{(\mu,\Sigma)\in\mathcal{U}_{\varepsilon}(\widehat{\mu},\widehat{\Sigma})} \mathcal{P}(\mu, \Sigma)$$

Chebyshev ambiguity set

Computing the Gelbrich Bound

Recall: $\mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N, \ell) \leq \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell)$

$$\overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) = \sup_{\mathbb{Q} \in \mathbb{G}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)]$$

Computing the Gelbrich Bound

Recall: $\mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N, \ell) \leq \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell)$

$$\overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) = \sup_{(\mu,\Sigma)\in\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathcal{P}(\mu,\Sigma)} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)]$$

Computing the Gelbrich Bound

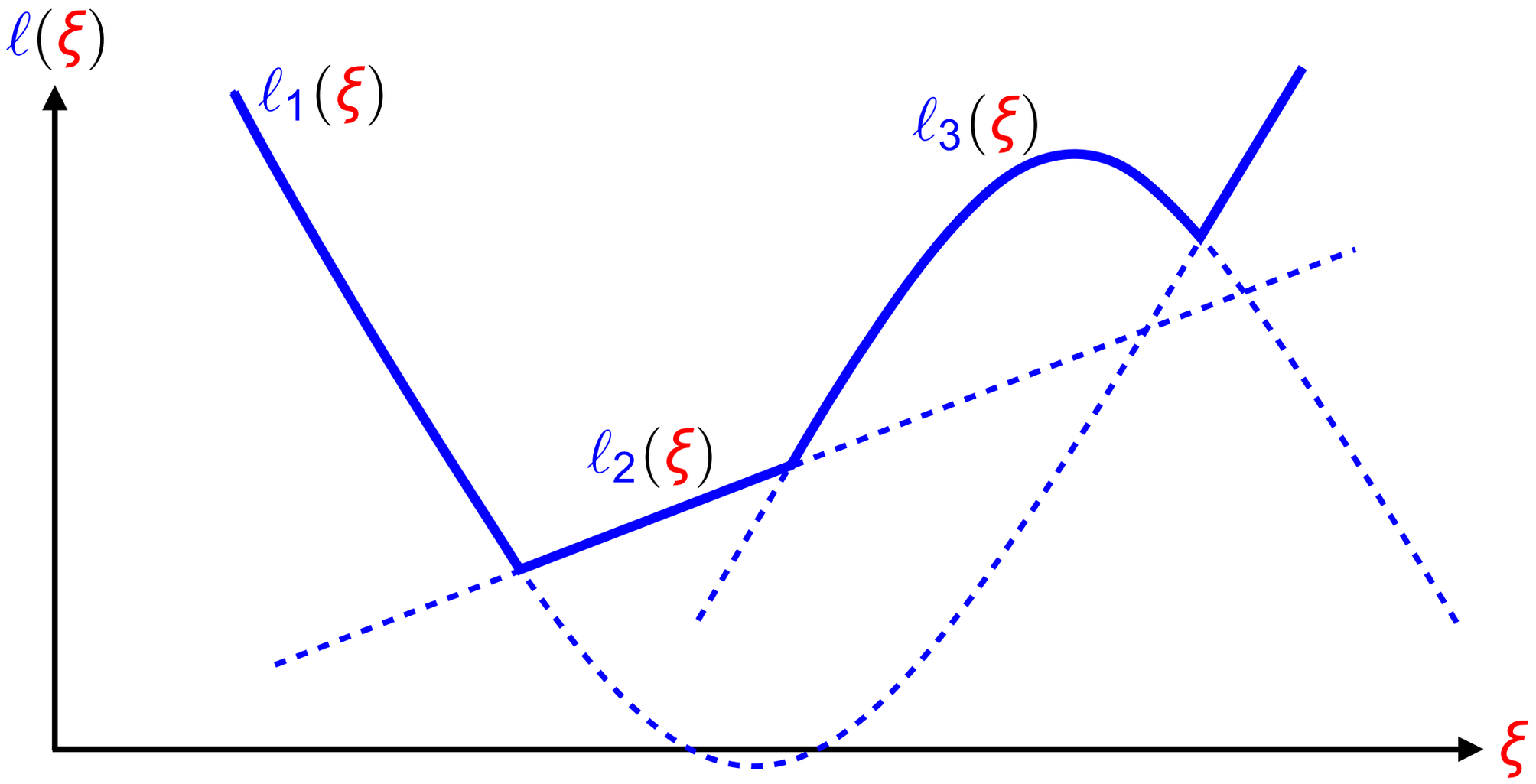
Recall: $\mathcal{R}_{\varepsilon,2}(\widehat{\mathbb{P}}_N, \ell) \leq \overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell)$

$$\overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) = \sup_{(\mu,\Sigma)\in\mathcal{U}_\varepsilon(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathcal{P}(\mu,\Sigma)} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)]$$

2^{nd} layer of robustness

Chebyshev risk

Piecewise Quadratic Loss



$$\ell(\xi) = \max_j \ell_j(\xi), \text{ where } \ell_j(\xi) = \xi^\top Q_j \xi + 2q_j^\top \xi + q_j^0$$

Piecewise Quadratic Loss

Theorem: If $\Xi = \mathbb{R}^m$, and $\ell(\xi)$ is piecewise quadratic, then the Gelbrich risk satisfies

$$\overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) \;=\; \inf \;\; y_0 + \gamma\left(\varepsilon^2 - \|\widehat{\mu}\|_2^2 - \mathsf{Tr}[\widehat{\Sigma}]\right) + z + \mathsf{Tr}[Z]$$

$$\text{s.t.} \;\; \gamma \in \mathbb{R}_+,\; y_0 \in \mathbb{R},\; y \in \mathbb{R}^m,\; Y \in \mathbb{S}^m,\; x \in \mathbb{R}_+,\; Z \in \mathbb{S}_+^m$$

$$\begin{bmatrix} \gamma I - Y & y + \gamma\widehat{\mu} \\ y^\top + \gamma\widehat{\mu}^\top & z \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \gamma I - Y & \gamma\widehat{\Sigma}^{\frac{1}{2}} \\ \gamma\widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0$$

$$\begin{bmatrix} Y - Q_j & y - q_j \\ y^\top - q_j^\top & y_0 - q_j^0 \end{bmatrix} \succeq 0 \quad \forall j \in [J].$$

Piecewise Quadratic Loss

Theorem: If $\Xi = \mathbb{R}^m$, and $\ell(\xi)$ is piecewise quadratic, then the Gelbrich risk satisfies

$$\overline{\mathcal{R}}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}, \ell) \;=$$

$$\begin{array}{ll} \max & \displaystyle\sum_{j=1}^{J} \mathrm{Tr}[Q_j \Theta_j] + 2 q_j^\top \theta_j + q_j^0 \alpha_j \\ \text{s.t.} & \mu \in \mathbb{R}^m,\ \Sigma \in \mathbb{S}_+^m,\ \alpha_j \in \mathbb{R}_+,\ \theta_j \in \mathbb{R}^m,\ \Theta_j \in \mathbb{S}_+^m \quad \forall j \in [J] \\ & \begin{bmatrix} \Theta_j & \theta_j \\ \theta_j^\top & \alpha_j \end{bmatrix} \succeq 0 \quad \forall j \in [J] \\ & \displaystyle\sum_{j=1}^{J} \alpha_j = 1,\ \sum_{j=1}^{J} \theta_j = \mu,\ \sum_{j=1}^{J} \Theta_j = \Sigma + \mu\mu^\top,\ (\mu, \Sigma) \in \mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma}). \end{array}$$

Properties of the Gelbrich Bound

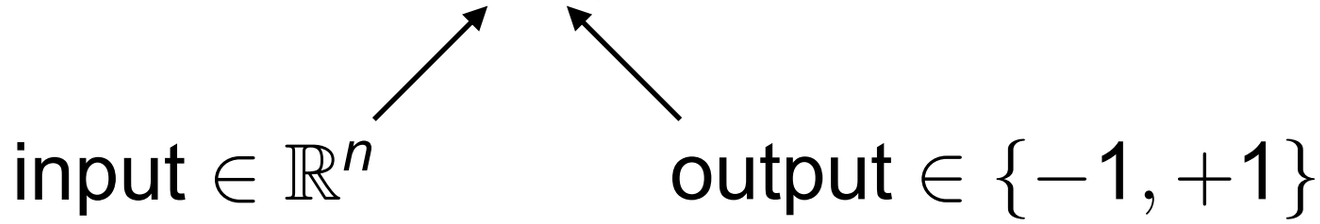
- bounds the Wasserstein risk and inherits its statistical guarantees;
- can be computed via convex optimization;
- induces the uncertainty set $\mathcal{U}_\varepsilon(\widehat{\mu}, \widehat{\Sigma})$ for μ and Σ

  $\Longrightarrow$ outer robust program *convex* for *any* loss function;
- exact if ℓ is quadratic and $\mathbb{P}_N$ elliptical;
- worst-case distribution available in closed form if ℓ is quadratic and $\widehat{\mathbb{P}}_N$ elliptical; hard to compute if ℓ is piecewise quadratic.

Applications in Machine Learning

Classification

Random vector: $\xi = (x, y) \sim \mathbb{P}$



Classification

Random vector: $\xi = (x, y) \sim \mathbb{P}$

Goal: predict output from input

Classification

Random vector: $\xi = (x, y) \sim \mathbb{P}$

Goal: predict output from input

Ideal prediction: $\begin{cases} +1 & \text{if } \mathbb{P}[y = +1 | x] \geq 50\% \\ -1 & \text{if } \mathbb{P}[y = -1 | x] > 50\% \end{cases}$

Classification

Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Classification

Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $\operatorname{sign}(w^\top x)$

$\Longrightarrow$ prediction correct if $y \cdot w^\top x > 0$

Classification

Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $\operatorname{sign}(w^\top x)$

$\Longrightarrow$ prediction error $L(y \cdot w^\top x)$

Classification

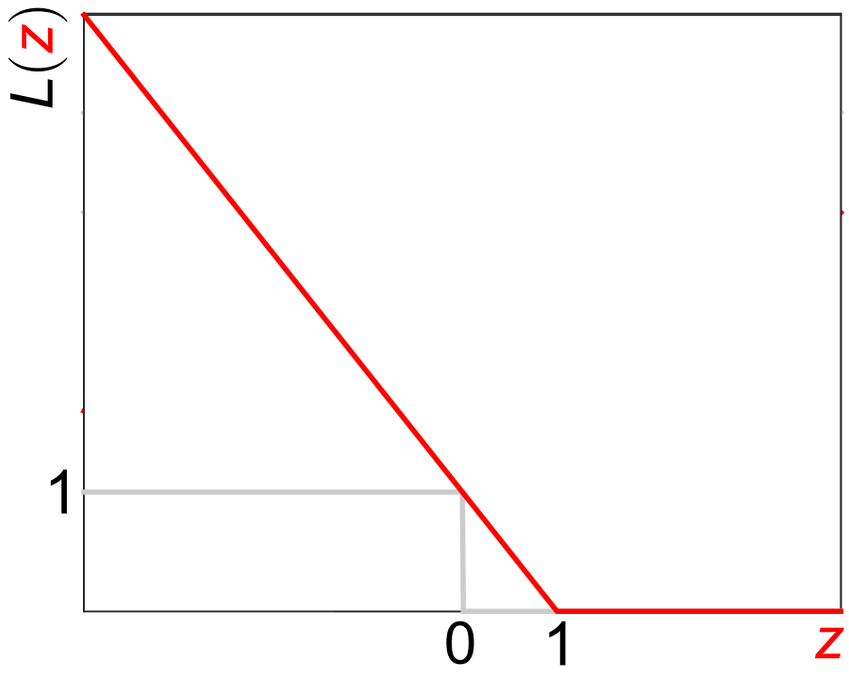
Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $\operatorname{sign}(w^\top x)$

$\Longrightarrow$ prediction error $L(y \cdot w^\top x)$



Support vector machine:

Hinge loss:

$$L(z) = \max\{0, 1 - z\}$$

Classification

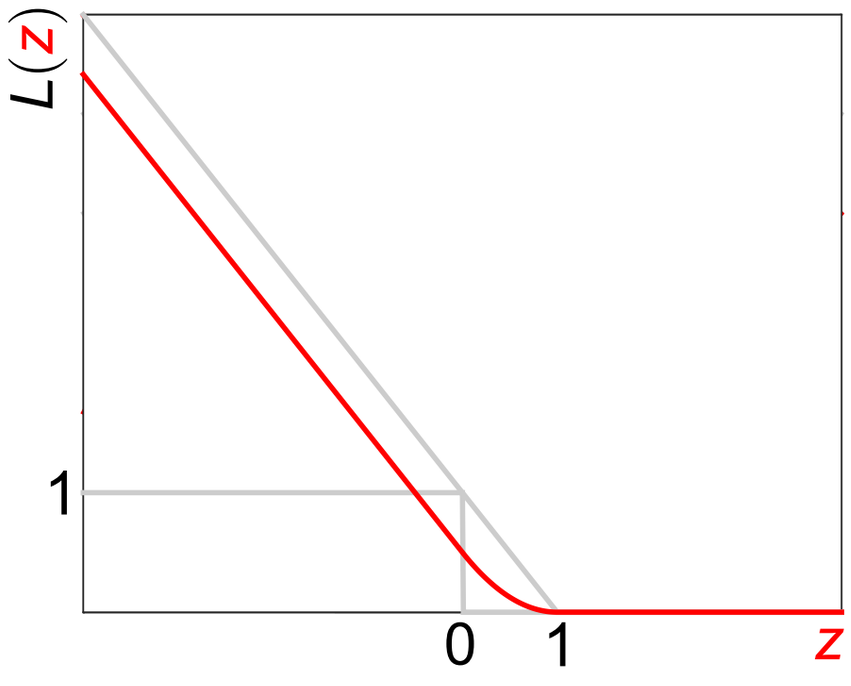
Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $\operatorname{sign}(w^\top x)$

$\Longrightarrow$ prediction error $L(y \cdot w^\top x)$



Smooth support vector machine:

Smooth hinge loss:

$$L(z) = \begin{cases} \frac{1}{2} - z & \text{if } z \leq 0 \\ \frac{1}{2}(1-z)^2 & \text{if } 0 < z < 1 \\ 0 & \text{else} \end{cases}$$

Classification

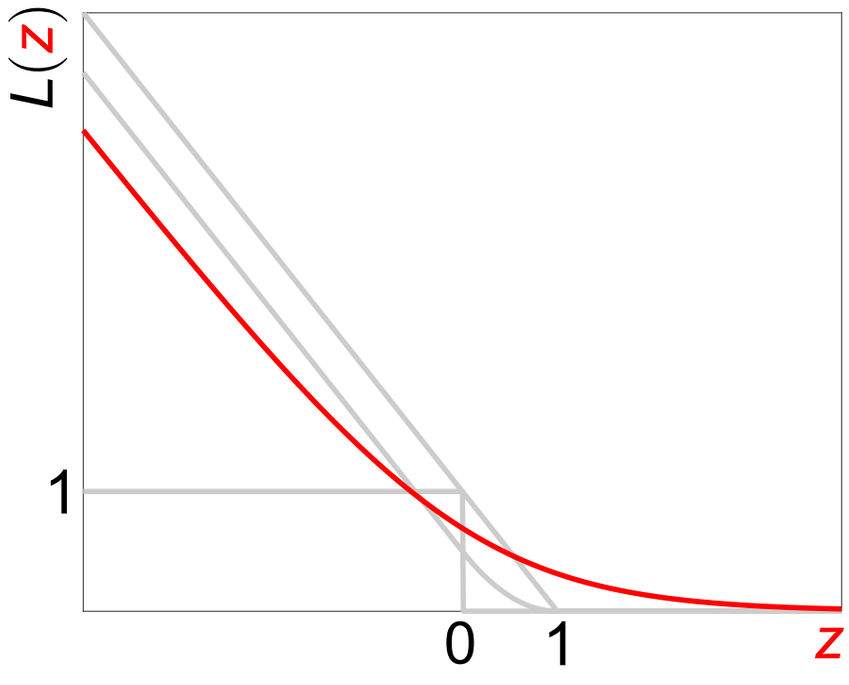
Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $\operatorname{sign}(w^\top x)$

$\Longrightarrow$ prediction error $L(y \cdot w^\top x)$



Logistic regression:

Logloss:

$$L(z) = \log(1 + e^{-z})$$

Classification

Empirical risk minimization:

$$\inf_{w \in \mathbb{R}^n} \mathbb{E}^{\widehat{\mathbb{P}}_N}[L(y \cdot w^\top x)]$$

Classification

Empirical risk minimization:

$$\inf_{w \in \mathbb{R}^n} \underbrace{\mathbb{E}^{\widehat{\mathbb{P}}_N}[L(y \cdot w^\top x)]}_{\text{nominal risk}}$$

- replace true risk with nominal risk
- $\widehat{\mathbb{P}}_N =$ empirical distribution
- susceptible to overfitting

Classification

Distributionally robust classification:

$$\inf_{w \in \mathbb{R}^n} \quad \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_N)} \quad \mathbb{E}^{\mathbb{Q}}[L(y \cdot w^\top x)]$$

Classification

Distributionally robust classification:

$$\inf_{w \in \mathbb{R}^n} \quad \underbrace{\sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[L(y \cdot w^\top x)]}_{\text{worst-case risk}}$$

- replace true risk with worst-case risk
- norm on input-output space: $\|\xi\| = \|x\| + \frac{\kappa}{2}|y|$
- $\kappa =$ cost of flipping a label

Classification

Distributionally robust classification:

$$\inf_{w \in \mathbb{R}^n} \quad \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_N)} \quad \mathbb{E}^{\mathbb{Q}}[L(y \cdot w^\top x)] \qquad (*)$$

Theorem: If $\kappa = \infty$ and $p = 1$ then[1)]

$$(*) \;\equiv\; \inf_{w \in \mathbb{R}^n} \; \mathbb{E}^{\widehat{\mathbb{P}}_N}[L(y \cdot w^\top x)] + \varepsilon \cdot \mathrm{Lip}(L) \cdot \|w\|_*.$$

[1)] Shafieezadeh-Abadeh, Kuhn & Mohajerin Esfahani, *J. Mach. Learn. Res.*, 2019; Blanchet, Kang & Murthy, *arXiv*, 2016; Gao, Chen & Kleywegt, *arXiv*, 2017.

Classification

Distributionally robust classification:

$$\inf_{w \in \mathbb{R}^n} \quad \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,1}(\widehat{\mathbb{P}}_N)} \quad \mathbb{E}^{\mathbb{Q}}[L(y \cdot w^\top x)] \qquad (*)$$

Theorem: If $\kappa = \infty$ and $L(z)$ is a standard loss function, then

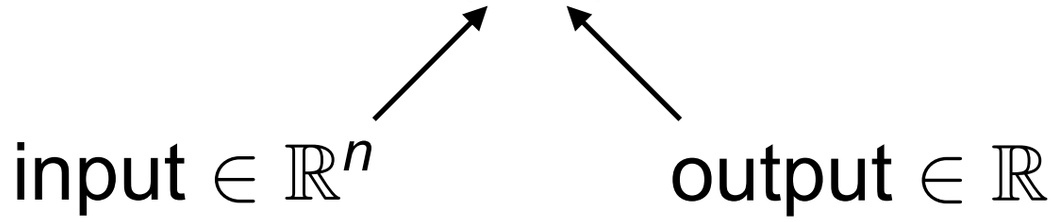
$$(*) \;\equiv\; \inf_{w \in \mathbb{R}^n} \; \underbrace{\mathbb{E}^{\widehat{\mathbb{P}}_N}[L(y \cdot w^\top x)]}_{\text{empirical risk}} + \underbrace{\varepsilon \cdot \mathrm{Lip}(L) \cdot \|w\|_*}_{\substack{\text{regularization} \\ \text{term}}}.$$

---

[1] Shafieezadeh-Abadeh, Kuhn & Mohajerin Esfahani, *J. Mach. Learn. Res.*, 2019; Blanchet, Kang & Murthy, *arXiv*, 2016; Gao, Chen & Kleywegt, *arXiv*, 2017.

Regression

Random vector: $\xi = (x, y) \sim \mathbb{P}$



Regression

Random vector: $\xi = (x, y) \sim \mathbb{P}$

Goal: predict output from input

Regression

Random vector: $\xi = (x, y) \sim \mathbb{P}$

Goal: predict output from input

Ideal prediction: $\mathbb{E}^{\mathbb{P}}[\, y \,|\, x \,]$

Regression

Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Regression

Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $w^\top x$

$\Longrightarrow$ prediction error $L(w^\top x - y)$

Regression

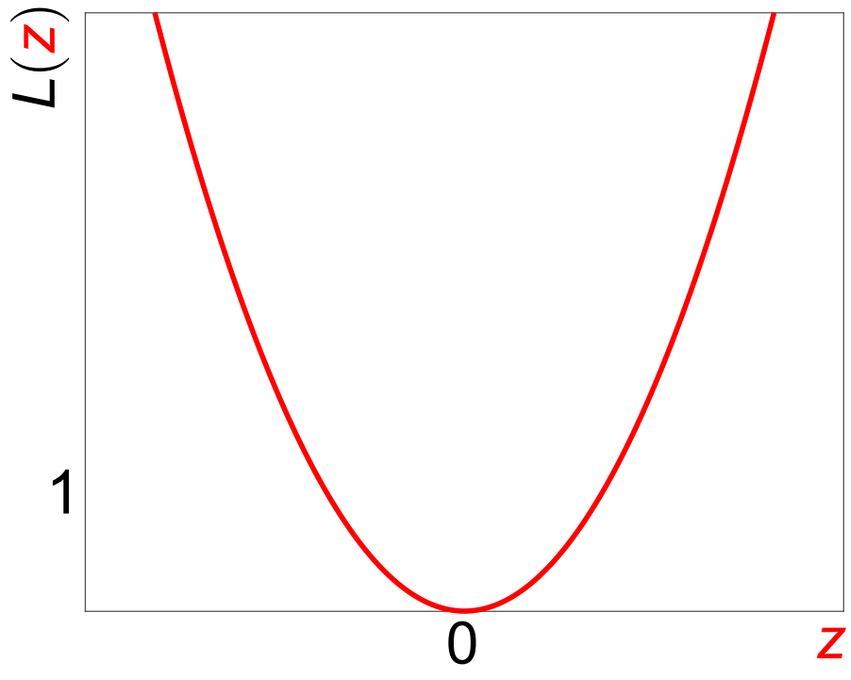
Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $w^\top x$

$\Longrightarrow$ prediction error $L(w^\top x - y)$



Ordinary least squares:

Square loss:

$L(z) = \frac{1}{2} z^2$

Regression

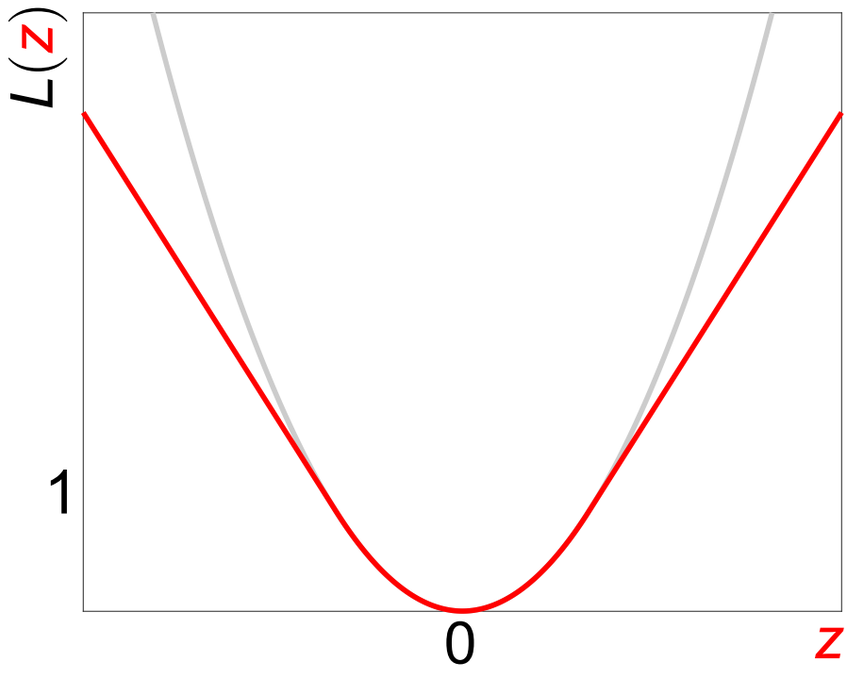
Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $w^\top x$

$\Longrightarrow$ prediction error $L(w^\top x - y)$



Robust regression:

Huber loss:

$$L(z) = \begin{cases} \frac{1}{2} z^2 & \text{if } |z| \leq \delta \\ \delta(|z| - \frac{1}{2}\delta) & \text{else} \end{cases}$$

Regression

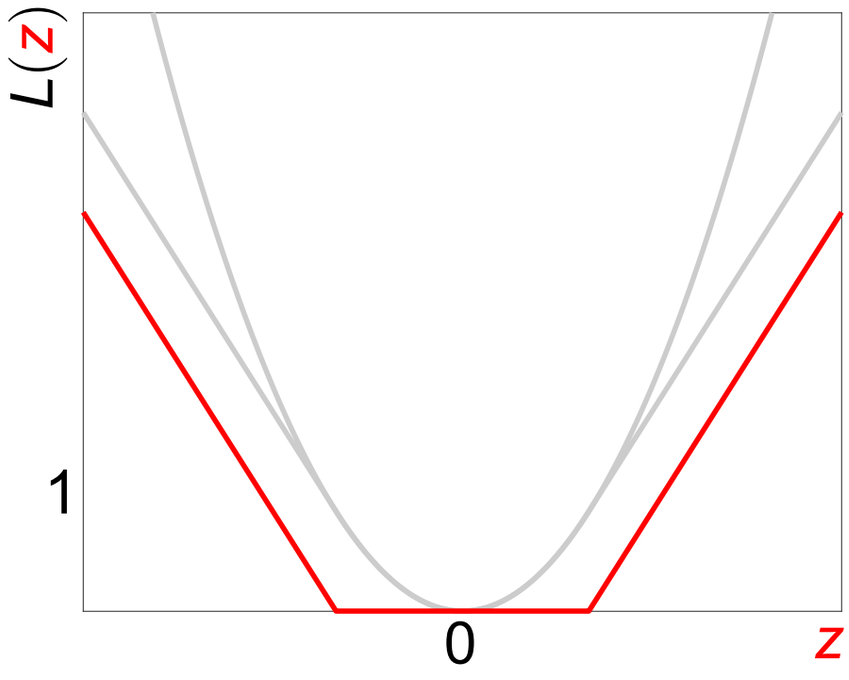
Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $w^\top x$

$\Longrightarrow$ prediction error $L(w^\top x - y)$



Support vector regression:

δ-insensitive loss:

$$L(z) = \max\{0, |z| - \varepsilon\}$$

Regression

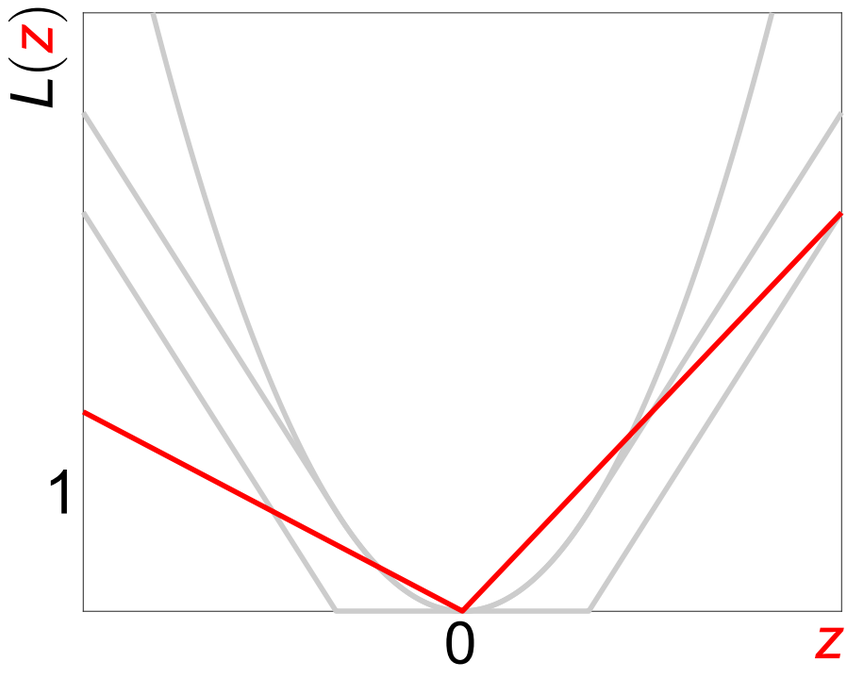
Random vector: $\xi = (x, y) \sim \mathbb{P}$

Reality:

- $\mathbb{P}$ unknown
- we are given training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \sim \mathbb{P}^N$

Practical approach: predict output as $w^\top x$

$\Longrightarrow$ prediction error $L(w^\top x - y)$



Quantile regression:

Pinball loss:

$$L(z) = \max\{-\tau z, (1-\tau) z\}$$

Regression

Empirical risk minimization:

$$\inf_{w \in \mathbb{R}^n} \mathbb{E}^{\widehat{\mathbb{P}}_N}[L(w^\top x - y)]$$

Regression

Distributionally robust regression:

$$\inf_{w \in \mathbb{R}^n} \quad \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \quad \mathbb{E}^{\mathbb{Q}}[L(w^\top x - y)]$$

Regression

Distributionally robust regression:

$$\inf_{w\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[L(w^\top x - y)]$$

- norm on input-output space: $\|\xi\| = \|x\| + \frac{\kappa}{2}|y|$

Regression

Distributionally robust regression:

$$\inf_{w\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[L(w^\top x - y)] \qquad (*)$$

Theorem: If $\kappa = \infty$ and $p = 1$, then[1)]

$$(*) \;\equiv\; \inf_{w\in\mathbb{R}^n} \mathbb{E}^{\widehat{\mathbb{P}}_N}[L(w^\top x - y)] + \varepsilon \cdot \mathrm{Lip}(L) \cdot \|w\|_*.$$

[1)] Shafieezadeh-Abadeh, Kuhn & Mohajerin Esfahani, *J. Mach. Learn. Res.*, 2019; Blanchet, Kang & Murthy, *arXiv*, 2016; Gao, Chen & Kleywegt, *arXiv*, 2017.

Regression

Distributionally robust regression:

$$\inf_{w\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[L(w^\top x - y)] \qquad (*)$$

Theorem: If $\kappa = \infty$ and $p = 1$, then[1)]

$$(*) \;\equiv\; \inf_{w\in\mathbb{R}^n} \underbrace{\mathbb{E}^{\widehat{\mathbb{P}}_N}[L(w^\top x - y)]}_{\text{empirical risk}} + \underbrace{\varepsilon \cdot \mathrm{Lip}(L) \cdot \|w\|_*}_{\substack{\text{regularization}\\ \text{term}}}.$$

[1)] Shafieezadeh-Abadeh, Kuhn & Mohajerin Esfahani, *J. Mach. Learn. Res.*, 2019; Blanchet, Kang & Murthy, *arXiv*, 2016; Gao, Chen & Kleywegt, *arXiv*, 2017.

Regression

Distributionally robust regression:

$$\inf_{w\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[L(w^\top x - y)] \qquad (*)$$

Theorem: If $\kappa = \infty$, $p = 2$ and $L(z)$ is the square error, then[1)]

$$(*) \;\equiv\; \inf_{w\in\mathbb{R}^n} \left[\left(\mathbb{E}^{\widehat{\mathbb{P}}_N}[L(w^\top x - y)]\right)^{\frac{1}{2}} + \varepsilon\cdot\|w\|_*\right]^2 .$$

[1)] Blanchet, Kang & Murthy, *arXiv*, 2016.

Regression

Distributionally robust regression:

$$\inf_{w\in\mathbb{R}^n} \sup_{\mathbb{Q}\in\mathbb{B}_{\varepsilon,p}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}}[L(w^\top x - y)] \qquad (*)$$

Theorem: If $\kappa = \infty$, $p = 2$ and $L(z)$ is the square error, then[1)]

$$(*) \equiv \inf_{w\in\mathbb{R}^n} \left[\underbrace{\left(\mathbb{E}^{\widehat{\mathbb{P}}_N}[L(w^\top x - y)]\right)^{\frac{1}{2}}}_{\text{square root of empirical risk}} + \underbrace{\varepsilon \cdot \|w\|_*}_{\text{regularization term}} \right]^2 .$$

$$\Longrightarrow \quad \text{generalized LASSO estimation problem}$$

[1)] Blanchet, Kang & Murthy, *arXiv*, 2016.

Maximum Likelihood Estimation

Given: Training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \in \mathbb{R}^m$ i.i.d. from $\mathcal{N}(\mu, \Sigma)$

Maximum Likelihood Estimation

Given: Training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \in \mathbb{R}^m$ i.i.d. from $\mathcal{N}(\mu, \Sigma)$

Input: Covariance matrix Σ

Output: Precision matrix $X = \Sigma^{-1}$

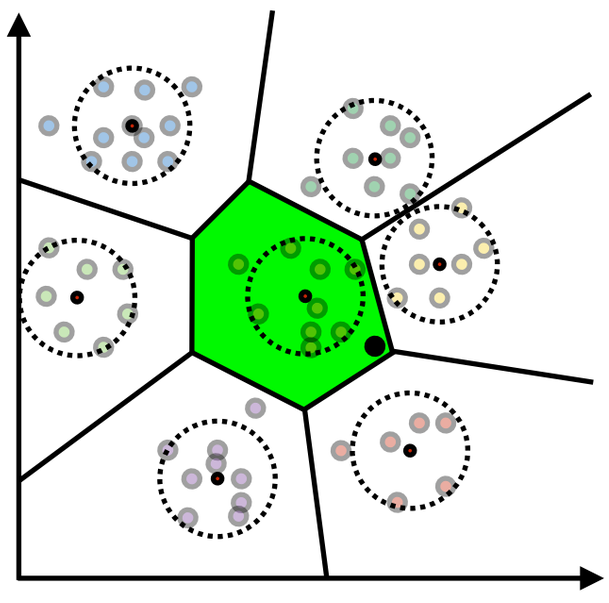
Maximum Likelihood Estimation

Given: Training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \in \mathbb{R}^m$ i.i.d. from $\mathcal{N}(\mu, \Sigma)$

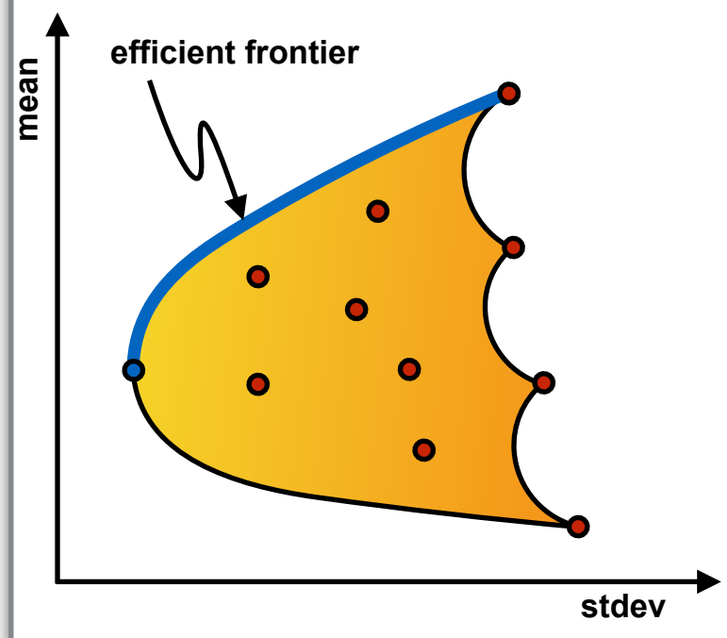
Input: Covariance matrix Σ

Output: Precision matrix $X = \Sigma^{-1}$

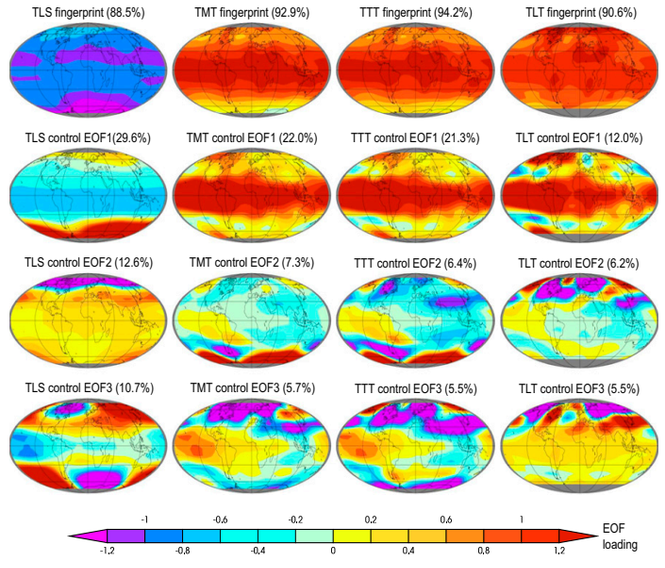
Linear discriminant analysis[1)]



Markowitz portfolio analysis[2)]



Optimal fingerprint method[3)]



1) Fisher, *Ann. Eugen.*, 1936.
2) Markowitz, *J. Financ.*, 1952.
3) Ribes et al., *Clim. Dynam.*, 2009.

Maximum Likelihood Estimation

Given: Training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \in \mathbb{R}^m$ i.i.d. from $\mathcal{N}(\mu, \Sigma)$

Empirical estimators:

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\xi}_i \quad \text{and} \quad \widehat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\widehat{\xi}_i - \widehat{\mu}) (\widehat{\xi}_i - \widehat{\mu})^\top$$

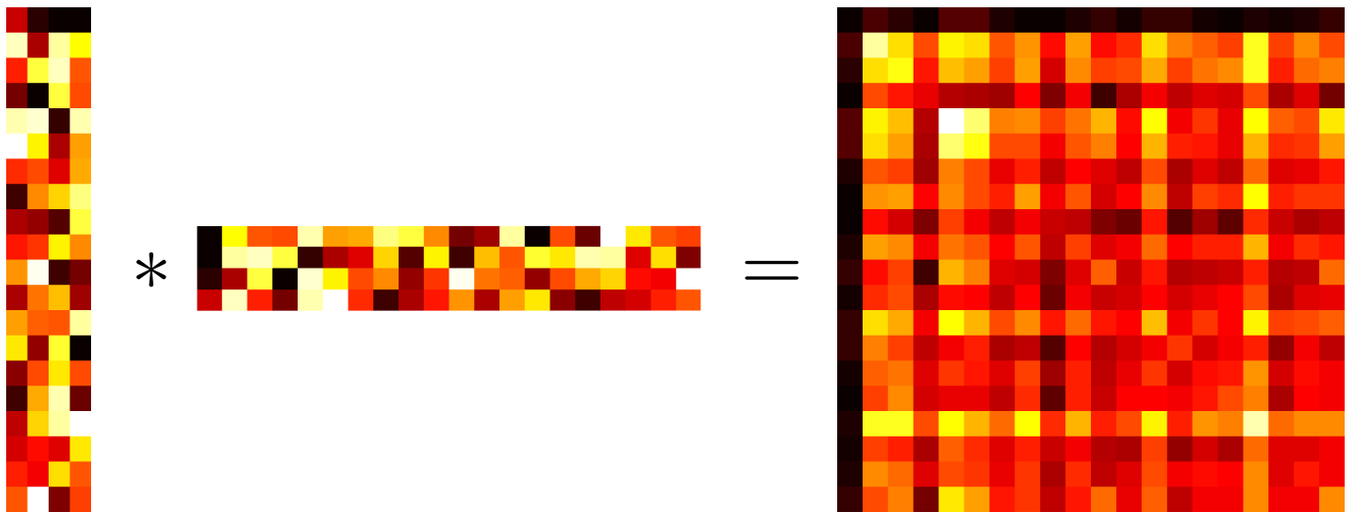
Maximum Likelihood Estimation

Given: Training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \in \mathbb{R}^m$ i.i.d. from $\mathcal{N}(\mu, \Sigma)$

Empirical estimators:

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\xi}_i \quad \text{and} \quad \widehat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\widehat{\xi}_i - \widehat{\mu}) (\widehat{\xi}_i - \widehat{\mu})^\top$$

- $m > N \implies \widehat{\Sigma}$ not invertible



Maximum Likelihood Estimation

Given: Training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \in \mathbb{R}^m$ i.i.d. from $\mathcal{N}(\mu, \Sigma)$

Maximum likelihood estimators:

$$\inf_{\mu \in \mathbb{R}^m, X \in \mathbb{S}^m_+} \left\{ -\log\det X + \frac{1}{N} \sum_{i=1}^{N} (\widehat{\xi}_i - \mu) X (\widehat{\xi}_i - \mu)^\top \right\}$$

Maximum Likelihood Estimation

Given: Training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \in \mathbb{R}^m$ i.i.d. from $\mathcal{N}(\mu, \Sigma)$

Maximum likelihood estimators:

$$\inf_{\mu \in \mathbb{R}^m, X \in \mathbb{S}^m_+} \left\{ -\log\det X + \frac{1}{N} \sum_{i=1}^{N} (\widehat{\xi}_i - \mu) X (\widehat{\xi}_i - \mu)^\top \right\}$$

- $m > N \implies$ problem unbounded!
- $m \leq N \implies \mu^\star = \widehat{\mu},\ X^\star = \widehat{\Sigma}^{-1}$

Maximum Likelihood Estimation

Given: Training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \in \mathbb{R}^m$ i.i.d. from $\mathcal{N}(\mu, \Sigma)$

Robust maximum likelihood estimators:[1)]

$$\inf_{\mu \in \mathbb{R}^m, X \in \mathbb{S}_+^m} \left\{ -\log\det X + \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}} \left[(\xi - \mu) X (\xi - \mu)^\top \right] \right\}$$

[1)] Nguyen, Kuhn & Mohajerin Esfahani, *arxiv*, 2018.

Maximum Likelihood Estimation

Given: Training samples $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N \in \mathbb{R}^m$ i.i.d. from $\mathcal{N}(\mu, \Sigma)$

Robust maximum likelihood estimators:[1)]

$$\inf_{\mu \in \mathbb{R}^m, X \in \mathbb{S}_+^m} \left\{ -\log\det X + \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}} \left[(\xi - \mu) X (\xi - \mu)^\top \right] \right\}$$

- normal nominal distribution $\widehat{\mathbb{P}}_N = \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$
- recover ML problem for $\varepsilon = 0$
- regularization by robustification

[1)] Nguyen, Kuhn & Mohajerin Esfahani, *arxiv*, 2018.

Wasserstein Shrinkage Estimator[1)]

Theorem: If $\varepsilon > 0$ and $\widehat{\Sigma} = \sum_{i=1}^{m} \lambda_i \cdot v_i v_i^\top$, then

$$\mu^\star = \widehat{\mu} \quad \text{and} \quad X^\star = \sum_{i=1}^{m} x_i^\star \cdot v_i v_i^\top,$$

where

$$x_i^\star = \gamma^\star \left[1 - \tfrac{1}{2} \left(\sqrt{\lambda_i^2 (\gamma^\star)^2 + 4 \lambda_i \gamma^\star} - \lambda_i \gamma^\star \right) \right]$$

and $\gamma^\star$ is the unique positive solution of

$$\left(\varepsilon^2 - \tfrac{1}{2} \sum_{i=1}^{m} \lambda_i \right) \gamma - m + \tfrac{1}{2} \sum_{i=1}^{m} \sqrt{\lambda_i^2 \gamma^2 + 4 \lambda_i \gamma} = 0.$$

[1)] Nguyen, Kuhn & Mohajerin Esfahani, *arxiv*, 2018.

Wasserstein Shrinkage Estimator[1)]

Theorem: If $\varepsilon > 0$ and $\widehat{\Sigma} = \sum_{i=1}^m \lambda_i \cdot v_i v_i^\top$, then

$$X^\star = \sum_{i=1}^m x_i^\star \cdot v_i v_i^\top$$

where

$$x_i^\star = \gamma^\star \left[1 - \tfrac{1}{2}\left(\sqrt{\lambda_i^2(\gamma^\star)^2 + 4\lambda_i\gamma^\star} - \lambda_i\gamma^\star\right)\right]$$

and $\gamma^\star$ is the unique positive solution of

$$\left(\varepsilon^2 - \tfrac{1}{2}\sum_{i=1}^m \lambda_i\right)\gamma - m + \tfrac{1}{2}\sum_{i=1}^m \sqrt{\lambda_i^2\gamma^2 + 4\lambda_i\gamma} = 0.$$

- $X^\star$ and $\widehat{\Sigma}$ commute $\implies$ $X^\star$ is **rotation-equivariant**
- $x_i^\star > 0$ for all i $\implies$ $X^\star$ is **invertible**
- $X^\star$ is **easy to compute** (spectral decomposition + bisection)

[1)] Nguyen, Kuhn & Mohajerin Esfahani, *arxiv*, 2018.

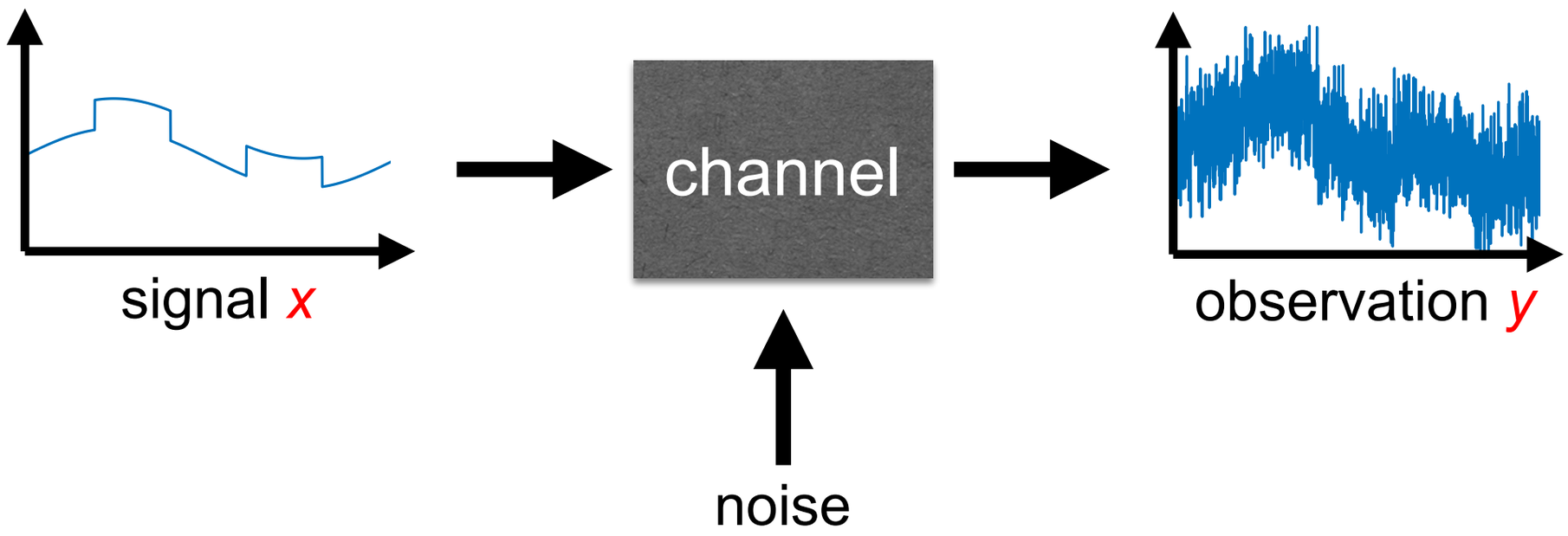
Wasserstein Shrinkage Estimator

Properties of the Wasserstein shrinkage estimator:

- exists for $m > N$ if $\varepsilon > 0$
- rotation-equivariant
- preserves the order of the eigenvalues of $\widehat{\Sigma}$
- condition number improves as $\varepsilon \to \infty$
- same complexity as spectral decomposition of $\widehat{\Sigma}$
- offers out-of-sample guarantee on Stein's loss
- follows from robustifying the ML estimation problem
- similar performance as graphical lasso at computational cost of linear shrinkage estimator

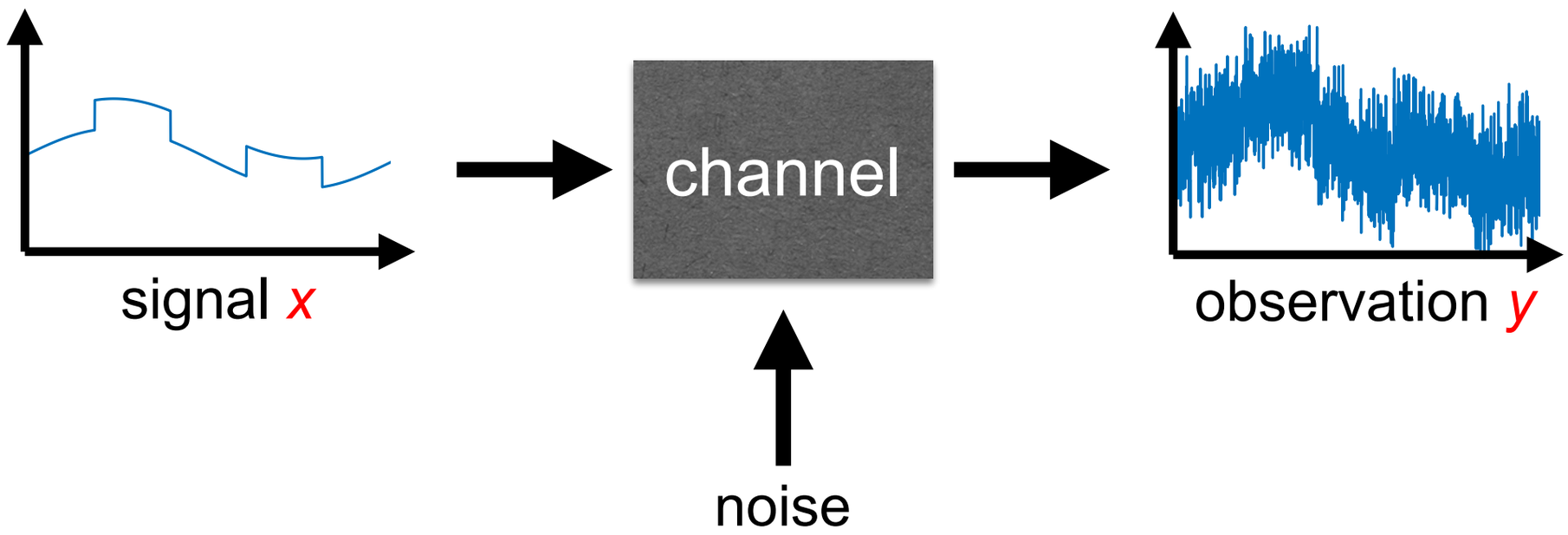
Minimum Mean Square Error Estimation

Problem: Infer signal $x \in \mathbb{R}^{m_x}$ from noisy observation $y \in \mathbb{R}^{m_y}$



Minimum Mean Square Error Estimation

Problem: Infer signal $x \in \mathbb{R}^{m_x}$ from noisy observation $y \in \mathbb{R}^{m_y}$



The distribution $\mathbb{P}$ of $\xi = (x, y) \in \mathbb{R}^m$, $m = m_x + m_y$, is unknown.

Elliptical nominal distribution: $\widehat{\mathbb{P}}_N = \mathcal{E}_g(\widehat{\mu}_N, \widehat{\Sigma}_N)$

Minimum Mean Square Error Estimation

Problem: Infer signal $x \in \mathbb{R}^{m_x}$ from noisy observation $y \in \mathbb{R}^{m_y}$

Robust MMSE estimation problem:[1)]

$$\inf_{\psi(\cdot)} \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)} \mathbb{E}^{\mathbb{Q}} \left[\| x - \psi(y) \|_2^2 \right]$$

[1)] Shafieezadeh-Abadeh, Nguyen, Kuhn & Mohajerin Esfahani, *NeuRIPS*, 2018.

Minimum Mean Square Error Estimation

Problem: Infer signal $x \in \mathbb{R}^{m_x}$ from noisy observation $y \in \mathbb{R}^{m_y}$

Robust MMSE estimation problem:[1)]

$$\inf_{\psi(\cdot)} \quad \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)} \quad \mathbb{E}^{\mathbb{Q}} \left[\| x - \psi(y) \|_2^2 \right]$$

estimator
(any function of y)

[1)] Shafieezadeh-Abadeh, Nguyen, Kuhn & Mohajerin Esfahani, *NeuRIPS*, 2018.

Minimum Mean Square Error Estimation

Problem: Infer signal $x \in \mathbb{R}^{m_x}$ from noisy observation $y \in \mathbb{R}^{m_y}$

Robust MMSE estimation problem:[1)]

$$\inf_{\psi(\cdot)} \sup_{\mathbb{Q} \in \mathbb{B}_{\varepsilon,2}(\widehat{\mathbb{P}}_N)} \underbrace{\mathbb{E}^{\mathbb{Q}}\left[\|x - \psi(y)\|_2^2\right]}_{\text{MSE}}$$

[1)] Shafieezadeh-Abadeh, Nguyen, Kuhn & Mohajerin Esfahani, *NeuRIPS*, 2018.

Minimum Mean Square Error Estimation

Theorem:[1] If $\widehat{\Sigma} \succ 0$, then the estimation problem is equivalent to

$$\begin{aligned} \max \quad & \operatorname{Tr}\left[\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right] \\ \text{s.t.} \quad & \Sigma_{xx} \in \mathbb{S}_+^{m_x}, \quad \Sigma_{yy} \in \mathbb{S}_+^{m_y}, \quad \Sigma_{xy} = \Sigma_{yx}^\top \in \mathbb{R}^{m_x \times m_y} \\ & \Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}, \quad \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\left(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \leq \varepsilon^2. \end{aligned}$$

If $\Sigma_{xx}^\star$, $\Sigma_{yy}^\star$ and $\Sigma_{xy}^\star$ are optimal, then

$$\psi^\star(y) = \Sigma_{xy}^\star(\Sigma_{yy}^\star)^{-1}(y - \widehat{\mu}_y) + \widehat{\mu}_x$$

is a robust MMSE estimator.

---

[1] Shafieezadeh-Abadeh, Nguyen, Kuhn & Mohajerin Esfahani, *NeuRIPS*, 2018.

Minimum Mean Square Error Estimation

Theorem:[1] If $\widehat{\Sigma} \succ 0$, then the estimation problem is equivalent to

$$\begin{array}{ll} \max & \operatorname{Tr}\left[\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\right] \\ \text{s.t.} & \Sigma_{xx} \in \mathbb{S}_+^{m_x}, \quad \Sigma_{yy} \in \mathbb{S}_+^{m_y}, \quad \Sigma_{xy} = \Sigma_{yx}^\top \in \mathbb{R}^{m_x \times m_y} \\ & \Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}, \quad \operatorname{Tr}\left[\Sigma + \widehat{\Sigma} - 2\left(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \leq \varepsilon^2. \end{array}$$

If $\Sigma_{xx}^\star$, $\Sigma_{yy}^\star$ and $\Sigma_{xy}^\star$ are optimal, then

$$\psi^\star(y) = \Sigma_{xy}^\star(\Sigma_{yy}^\star)^{-1}(y - \widehat{\mu}_y) + \widehat{\mu}_x$$

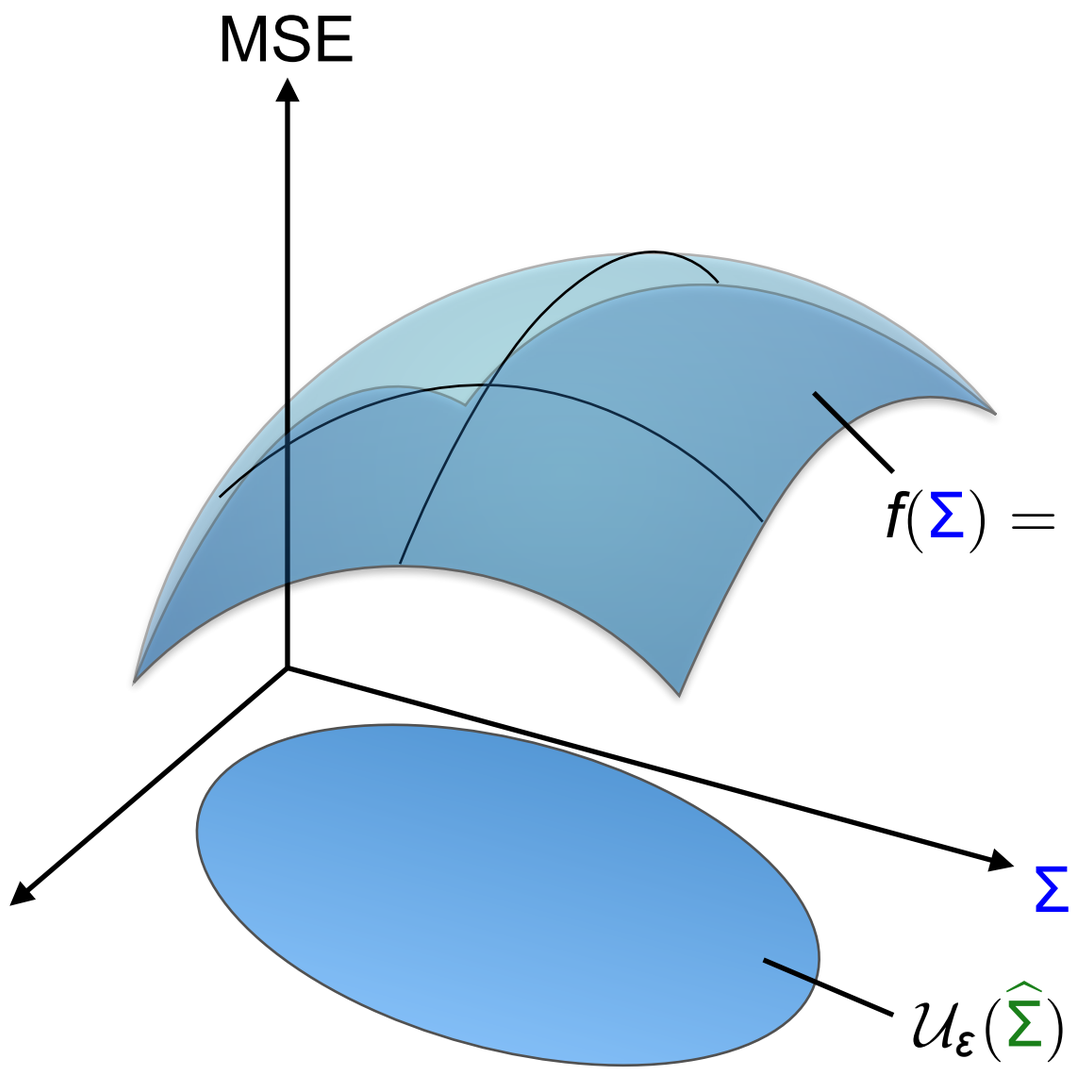
is a robust MMSE estimator.

- $\psi^\star$ is an **affine decision rule**
- $\psi^\star$ and $\mathbb{Q}^\star = \mathcal{E}_g(\widehat{\mu}, \Sigma^\star)$ form a **Nash equilibrium**

---

[1] Shafieezadeh-Abadeh, Nguyen, Kuhn & Mohajerin Esfahani, *NeuRIPS*, 2018.

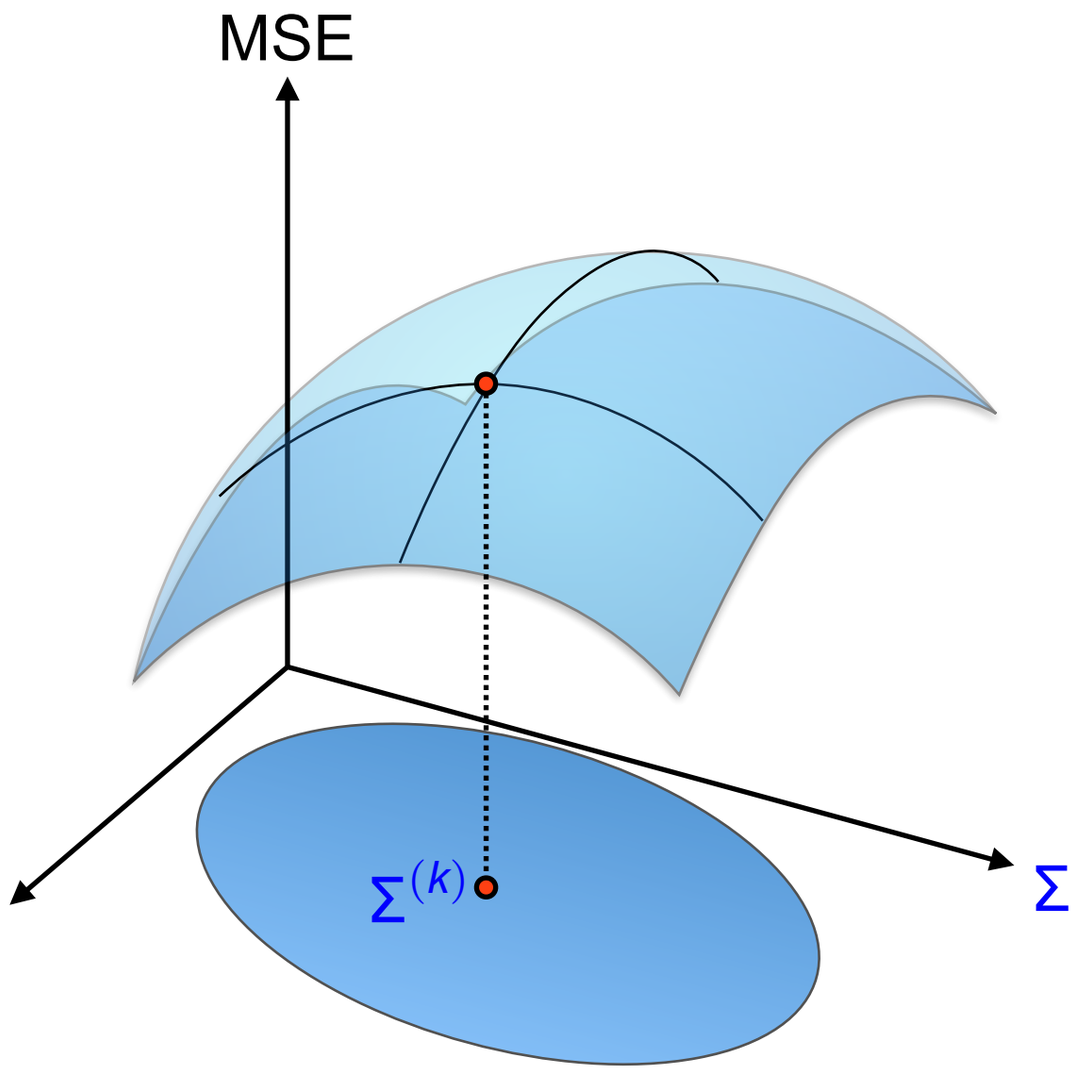
Frank-Wolfe Algorithm



Task: Solve the nonlinear SDP

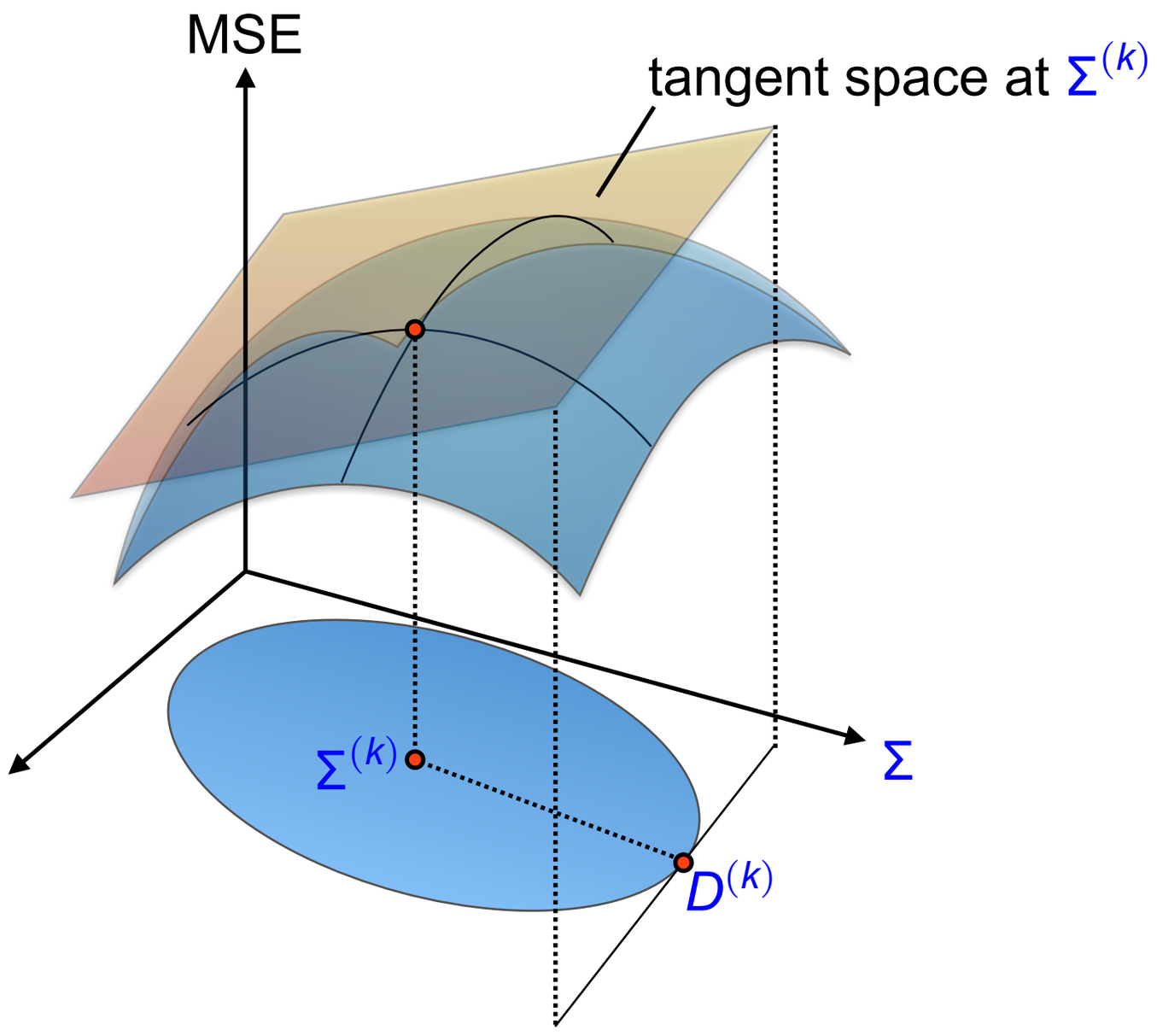
$$\max\left\{f(\Sigma) : \Sigma \in \mathcal{U}_\varepsilon(\widehat{\Sigma}),\ \Sigma \succeq 0\right\}$$

Frank-Wolfe Algorithm



Step 0: Pick initial iterate $\Sigma^{(k)}$

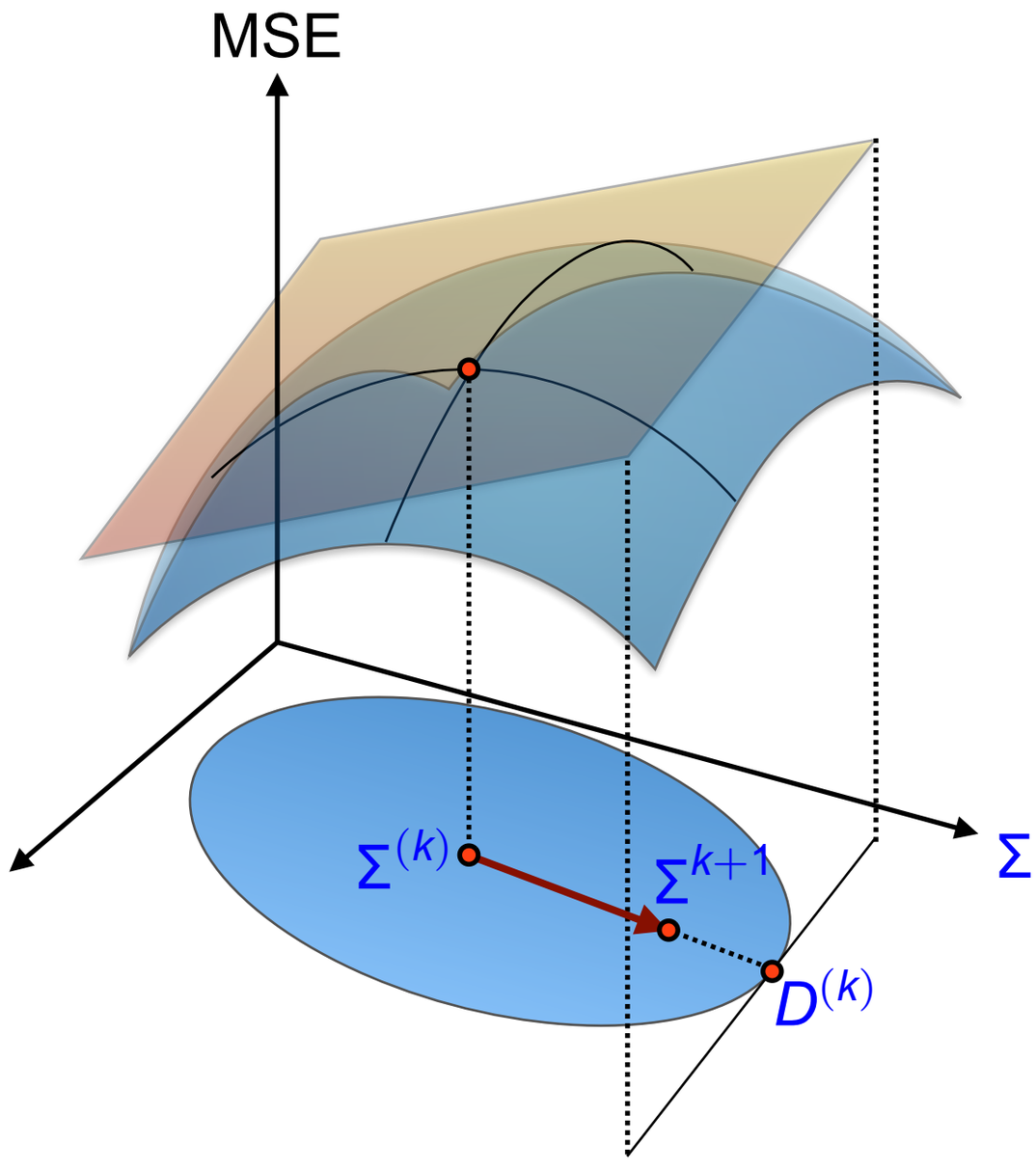
Frank-Wolfe Algorithm



Step 1: Solve direction-finding subproblem

$$D^{(k)} \in \arg\max \left\{ \mathrm{Tr}\left[D \nabla f(\Sigma^{(k)}) \right] : D \in \mathcal{U}_\varepsilon(\widehat{\Sigma}),\ D \succeq 0 \right\}$$

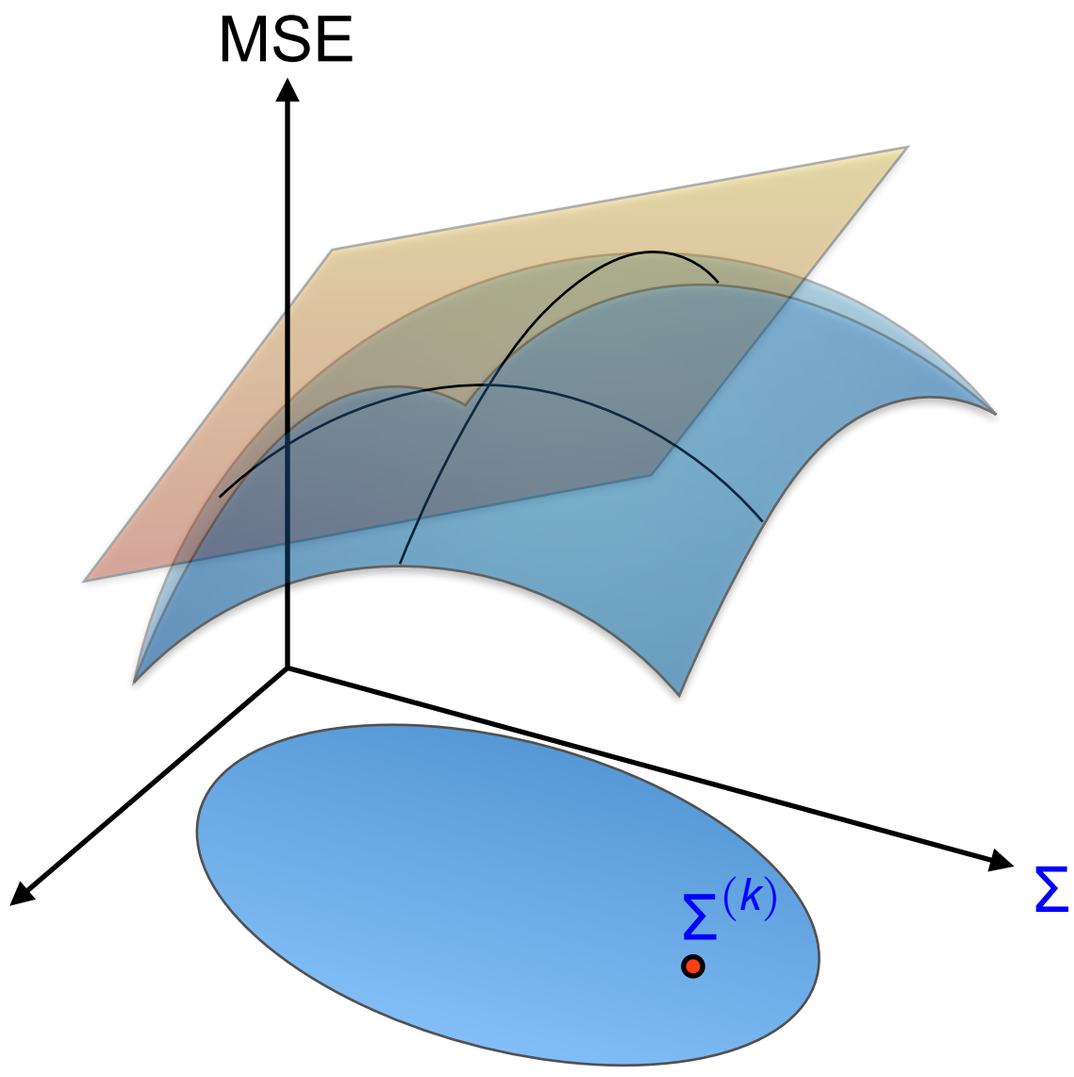
Frank-Wolfe Algorithm



Step 2: Construct new iterate

$$\Sigma^{(k+1)} = \alpha_k \cdot D^{(k)} + (1 - \alpha_k) \cdot \Sigma^{(k)}$$

Frank-Wolfe Algorithm



Step 3: Set $k \leftarrow k + 1$ and go to Step 1.

Frank-Wolfe Algorithm

Key benefits:[1)]

1. Nonlinear SDP is “nice”:
   - f is concave and smooth, ∇f is bounded away from 0
   - $\mathcal{U}_\varepsilon(\widehat{\Sigma})$ is strongly convex

[1)] Shafieezadeh-Abadeh, Nguyen, Kuhn & Mohajerin Esfahani, *NeuRIPS*, 2018.

Frank-Wolfe Algorithm

Key benefits:[1)]

1. Nonlinear SDP is “nice”:
   - f is concave and smooth, ∇f is bounded away from 0
   - $\mathcal{U}_\varepsilon(\widehat{\Sigma})$ is strongly convex

$\Longrightarrow$ **convergence of FW is linear!**

[1)] Shafieezadeh-Abadeh, Nguyen, Kuhn & Mohajerin Esfahani, *NeuRIPS*, 2018.

Frank-Wolfe Algorithm

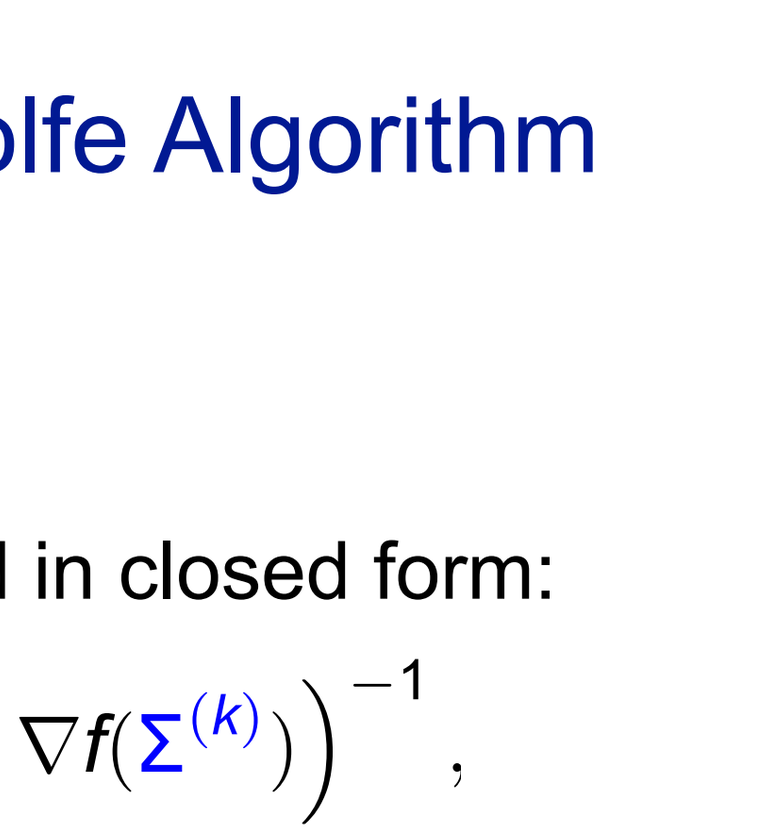
Key benefits:[1)]

2. Direction-finding subproblem can be solved in closed form:

$$D^{(k)} = (\gamma^\star)^2 \left(\gamma^\star I - \nabla f(\Sigma^{(k)})\right)^{-1} \widehat{\Sigma} \left(\gamma^\star I - \nabla f(\Sigma^{(k)})\right)^{-1},$$

where $\gamma^\star$ solves

$$\mathrm{Tr}\left[\widehat{\Sigma}\left(I - \gamma\left[\gamma I - \nabla f(\Sigma^{(k)})\right]^{-1}\right)^2\right] = \varepsilon^2.$$



[1)] Shafieezadeh-Abadeh, Nguyen, Kuhn & Mohajerin Esfahani, *NeuRIPS*, 2018.

Frank-Wolfe Algorithm

Key benefits:[1)]

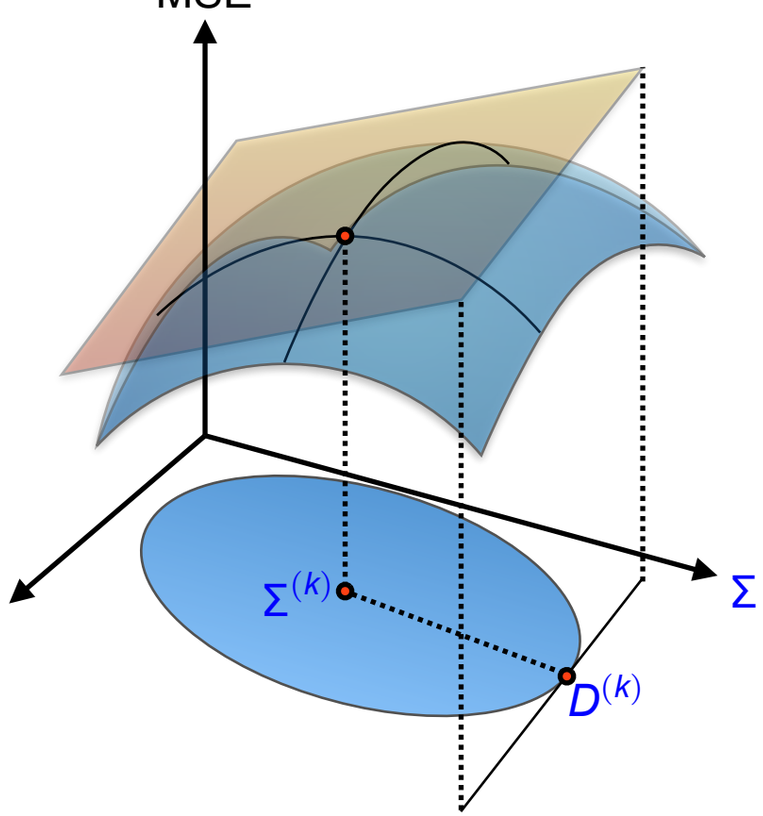
2. Direction-finding subproblem can be solved in closed form:

$$D^{(k)} = (\gamma^\star)^2 \left(\gamma^\star I - \nabla f(\Sigma^{(k)})\right)^{-1} \widehat{\Sigma} \left(\gamma^\star I - \nabla f(\Sigma^{(k)})\right)^{-1},$$

where $\gamma^\star$ solves

$$\mathrm{Tr}\left[\widehat{\Sigma}\left(I - \gamma\left[\gamma I - \nabla f(\Sigma^{(k)})\right]^{-1}\right)^2\right] = \varepsilon^2.$$

$\Longrightarrow$ **use bisection; extremely fast!**



[1)] Shafieezadeh-Abadeh, Nguyen, Kuhn & Mohajerin Esfahani, *NeuRIPS*, 2018.

Summary

Wasserstein DRO machine learning models:

- Explain commonly used regularization terms
- Explain the benefits of shrinkage estimators
- Motivate new regularization schemes
- Equivalent to efficiently solvable convex programs
- Sometimes solvable in quasi-closed form (bisection!)
- Sometimes solvable via fast FW algorithms

This Talk is Based on…

[1] D. Kuhn, P. Mohajerin Esfahani, V. Nguyen and S. Shafieezadeh Abadeh. **Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning**. INFORMS TutORials in Operations Research. 2019.

[2] P. Mohajerin Esfahani and D. Kuhn. **Data-Driven Distributionally Robust Optimization using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations**. *Mathematical Programming* 171(1–2), 115–166, 2018.

[3] V. Nguyen, D. Kuhn & P. Mohajerin Esfahani. **Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator**. Working Paper, 2018.

[4] S. Shafieezadeh Abadeh, P. Mohajerin Esfahani and D. Kuhn. **Distributionally Robust Logistic Regression**. *Neural Information Processing Systems*, 2015.

[5] S. Shafieezadeh Abadeh, P. Mohajerin Esfahani and D. Kuhn. **Regularization via Mass Transportation**. *Journal of Machine Learning Research* 20(103), 1–68, 2019.

[6] S. Shafieezadeh Abadeh, P. Mohajerin Esfahani, V. Nguyen and D. Kuhn. **Wasserstein Distributionally Robust Kalman Filtering**. *Neural Information Processing Systems*, 2018.

[7] J. (T.) Zhen, D. Kuhn & W. Wiesemann. **Nonlinear Distributionally Robust Optimization**. Working Paper, 2019.

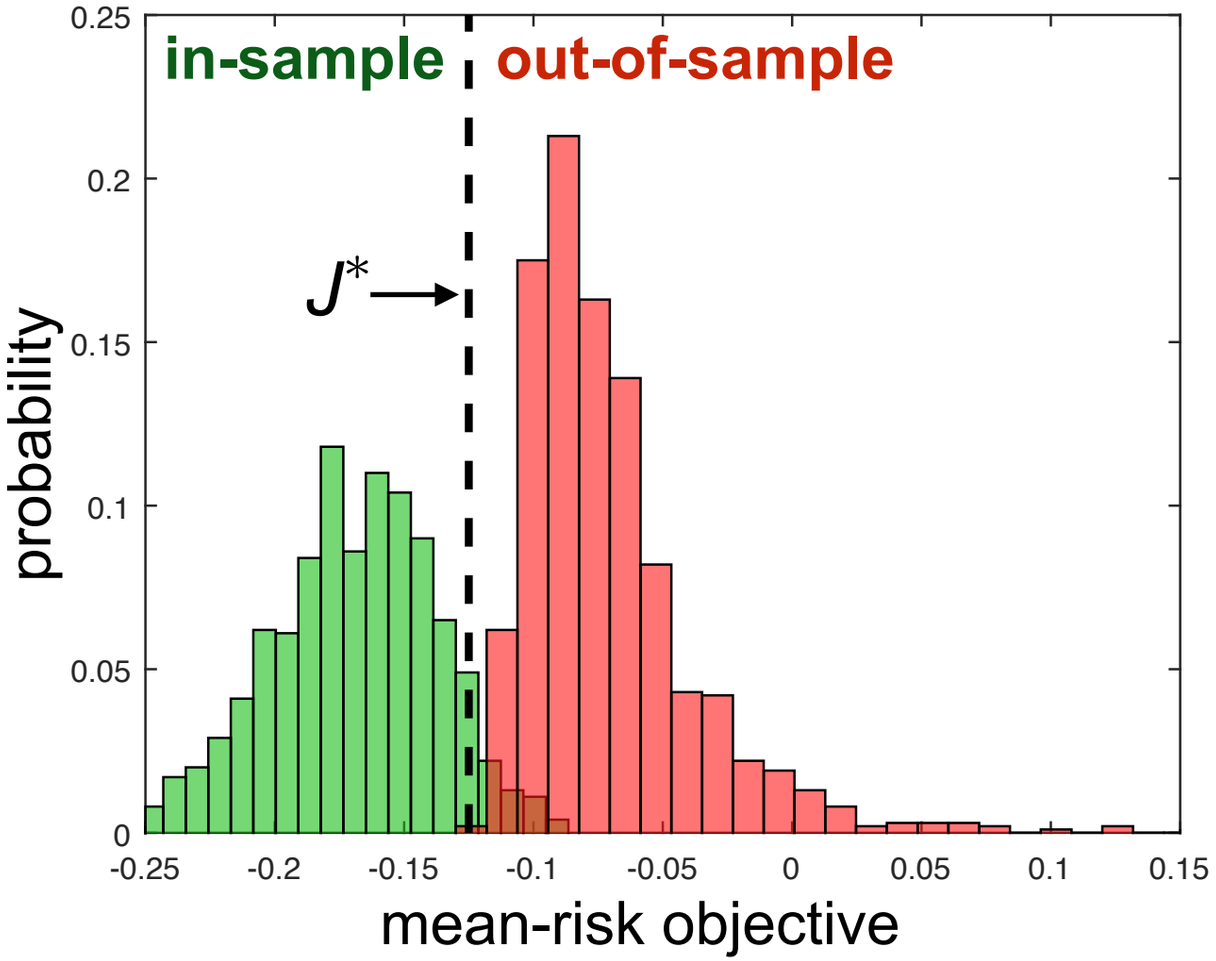
SAA with Scarce Data

Mean-risk portfolio problem

$$\min_{x \in \mathcal{X}} \left\{ \mathbb{E}^{\mathbb{P}}\left[-x^\top \xi \right] + \rho\, \mathbb{P}\text{-CVaR}_\alpha(-x^\top \xi) \right\}$$

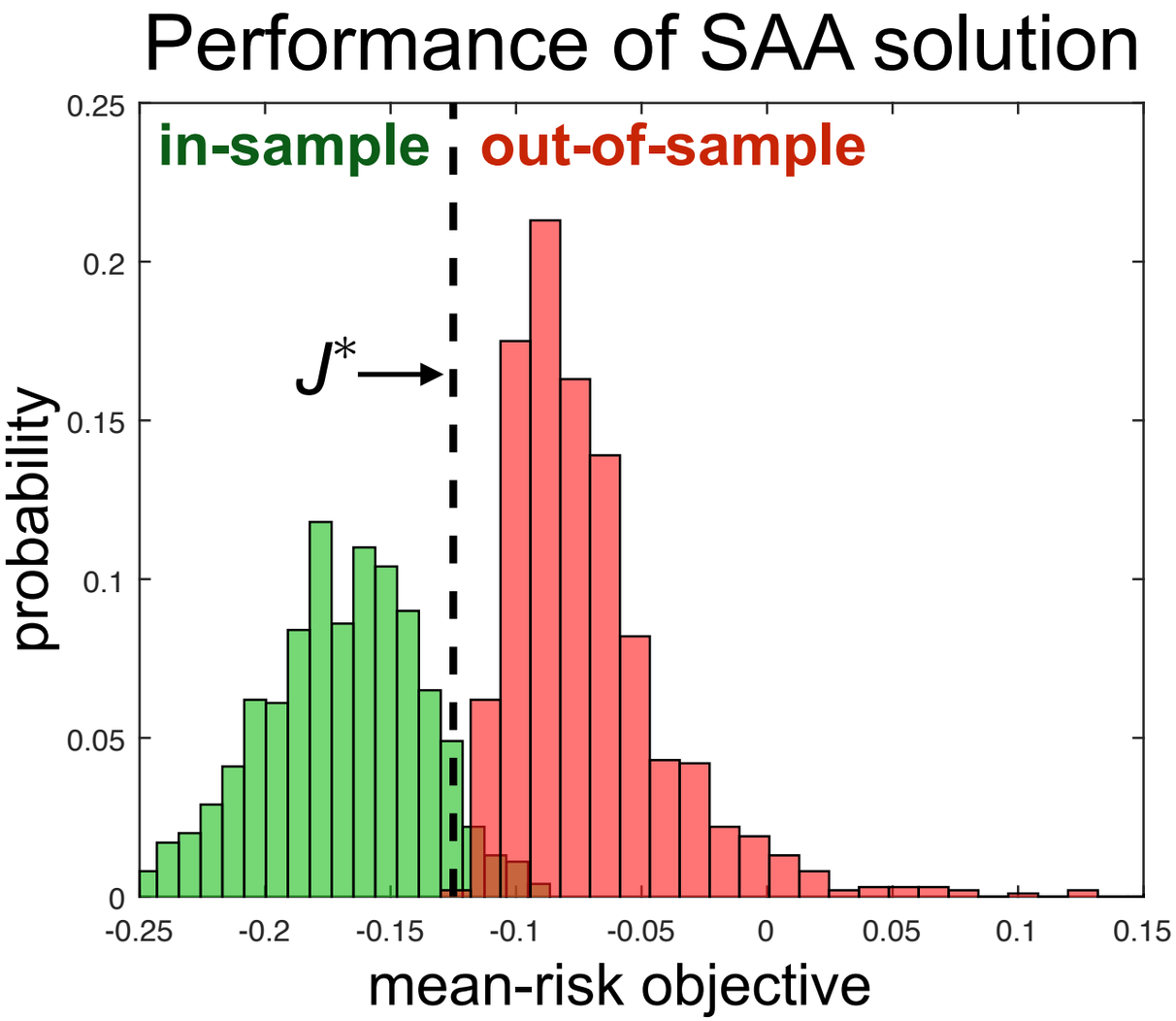
- 10 assets
- $\rho = 10$
- $\alpha = 20\%$
- $\xi_i = \psi + \zeta_i$ where $\psi \sim \mathcal{N}(0, 2\%)$ and $\zeta_i \sim \mathcal{N}(i \times 3\%, i \times 2.5\%)$

Performance of SAA solution



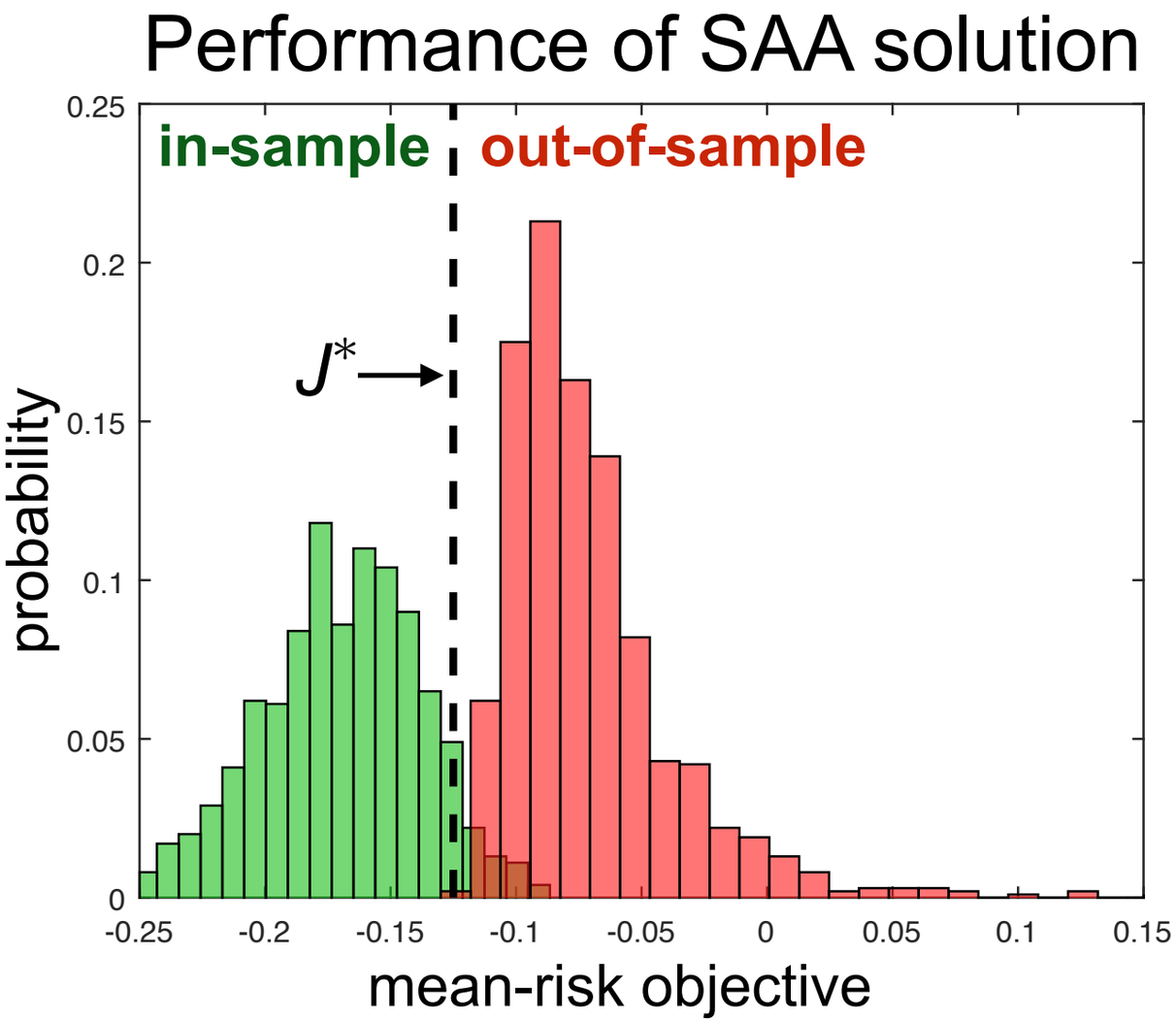
- 30 training samples
- in-sample: optimistic bias
- out-of-sample: pessimistic bias

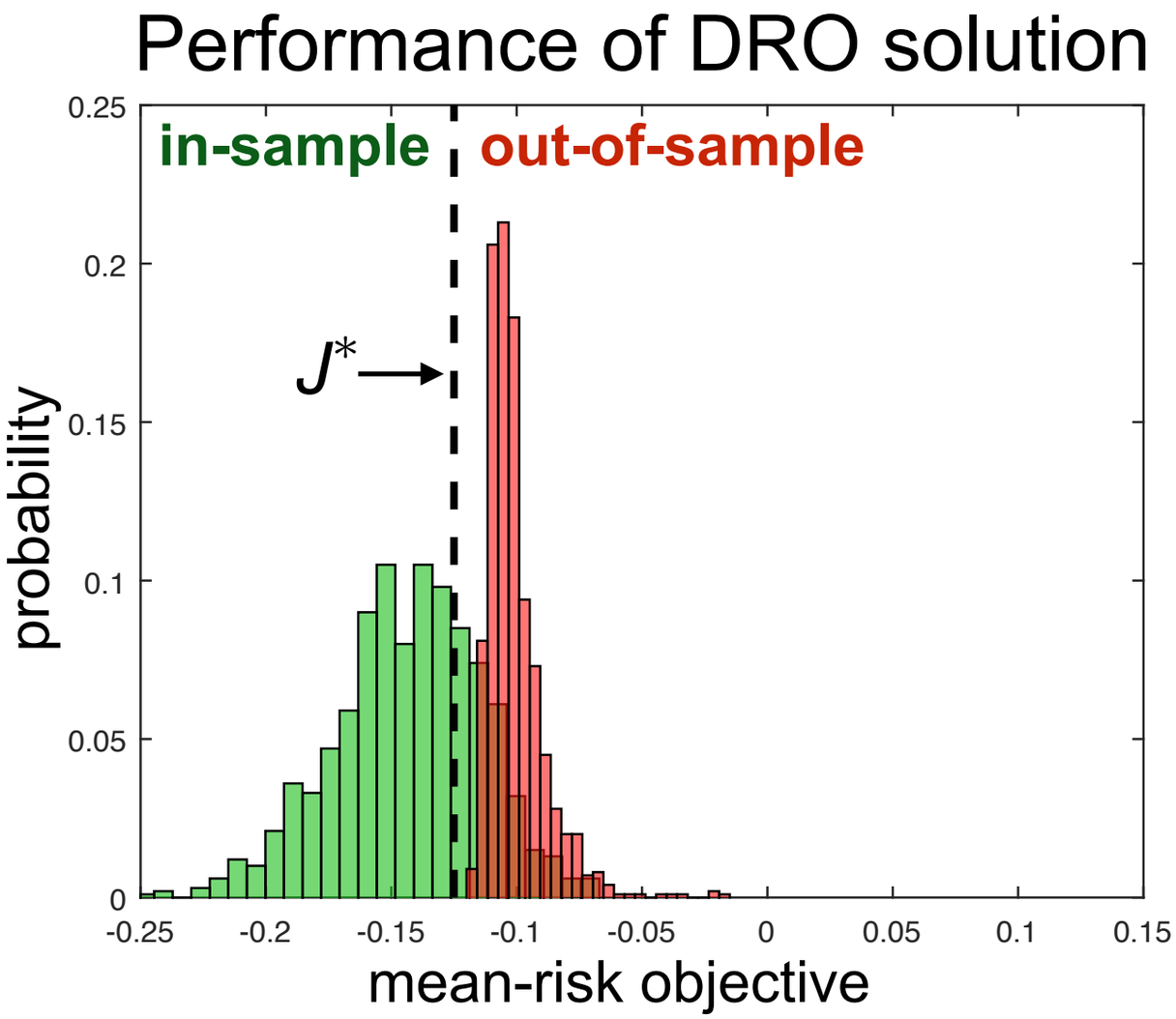
DRO with Scarce Data



- in-sample: optimistic bias
- out-of-sample: pessimistic bias

DRO with Scarce Data



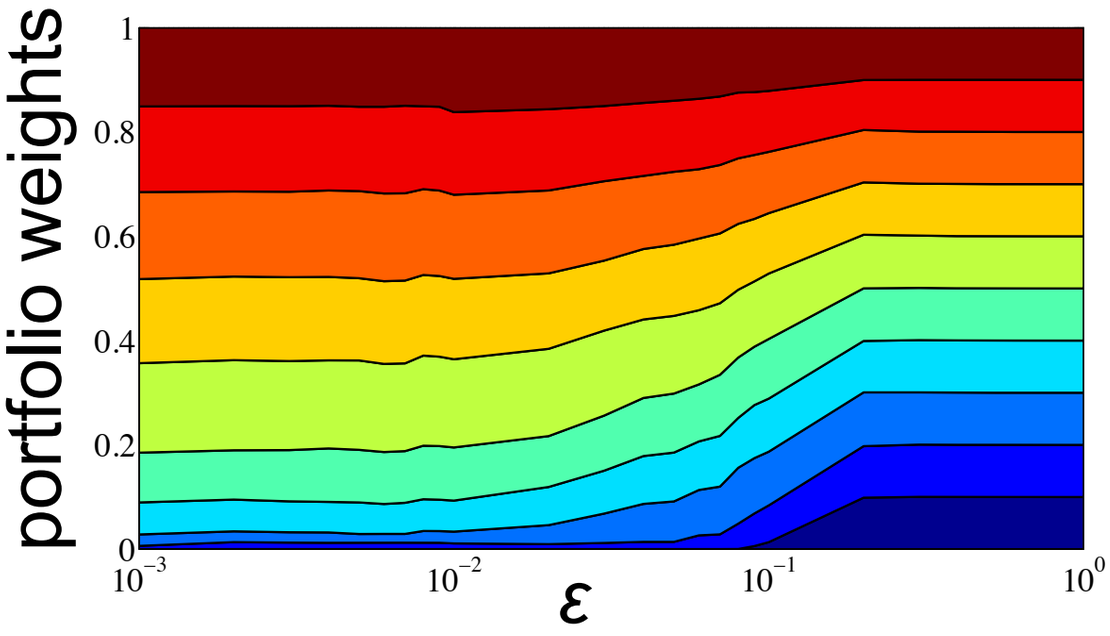


- in-sample: optimistic bias
- out-of-sample: pessimistic bias
- **DRO reduces bias & post-decision disappointment**

Application 1: Portfolio Selection

$$\min_{x \in \mathcal{X}} \left\{ \mathbb{E}^{\mathbb{P}}\left[-x^\top \xi \right] + \rho\, \mathbb{P}\text{-CVaR}_\alpha(-x^\top \xi) \right\}$$

- 10 assets
- $\rho = 10$
- $\alpha = 20\%$
- $\xi_i = \psi + \zeta_i$ where $\psi \sim \mathcal{N}(0, 2\%)$ and $\zeta_i \sim \mathcal{N}(i \times 3\%, i \times 2.5\%)$



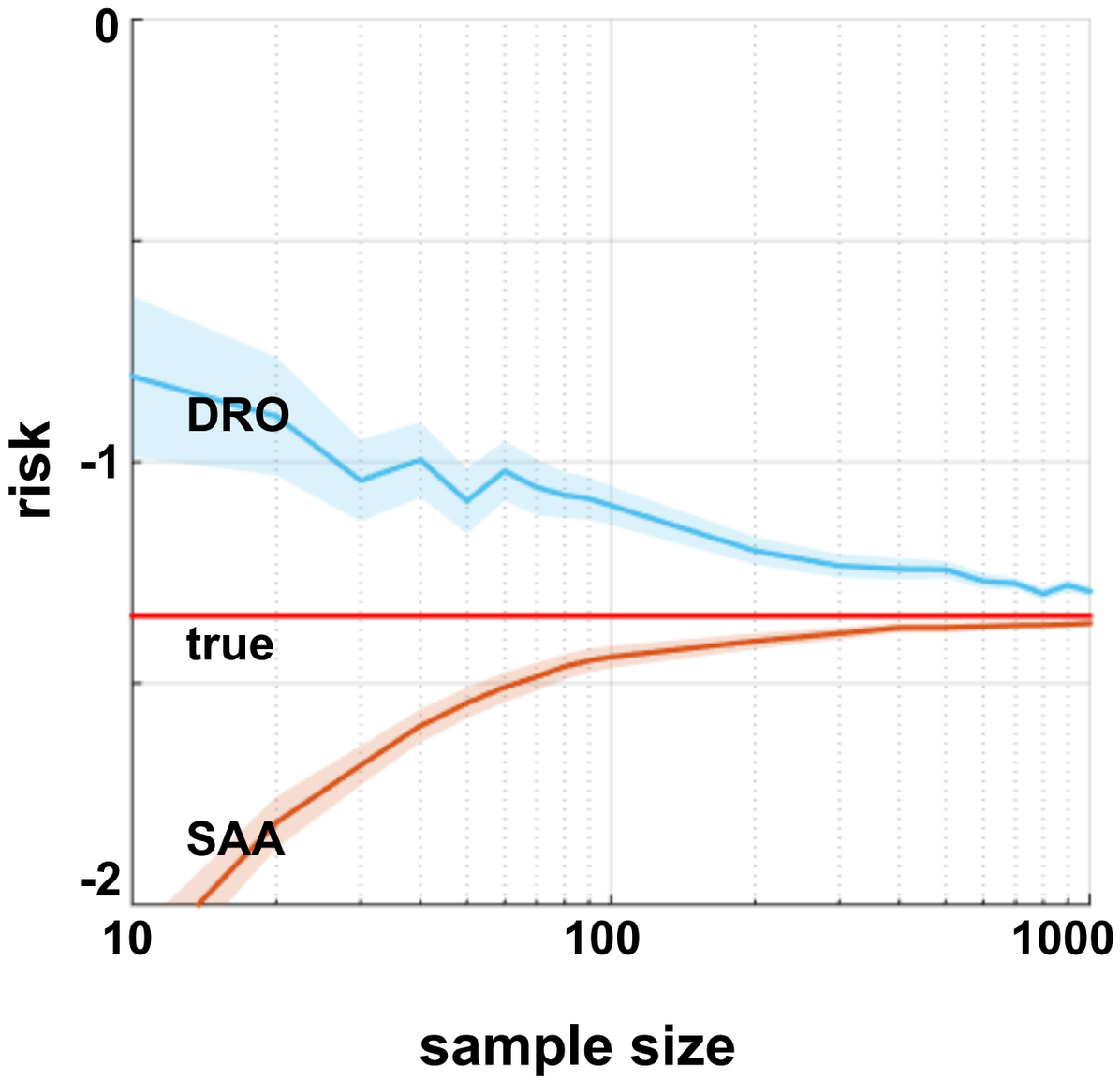
Fact: The 1/n portfolio is hard to beat out of sample.[1)]

Possible Explanation: It is optimal for $\varepsilon \to \infty$.[2)]

[1)] DeMiguel, Garlappi & Uppal, *Rev. Financ. Stud.*, 2009;
[2)] Pflug, Pichler & Wozabal, *J. Bank. Financ.*, 2012.

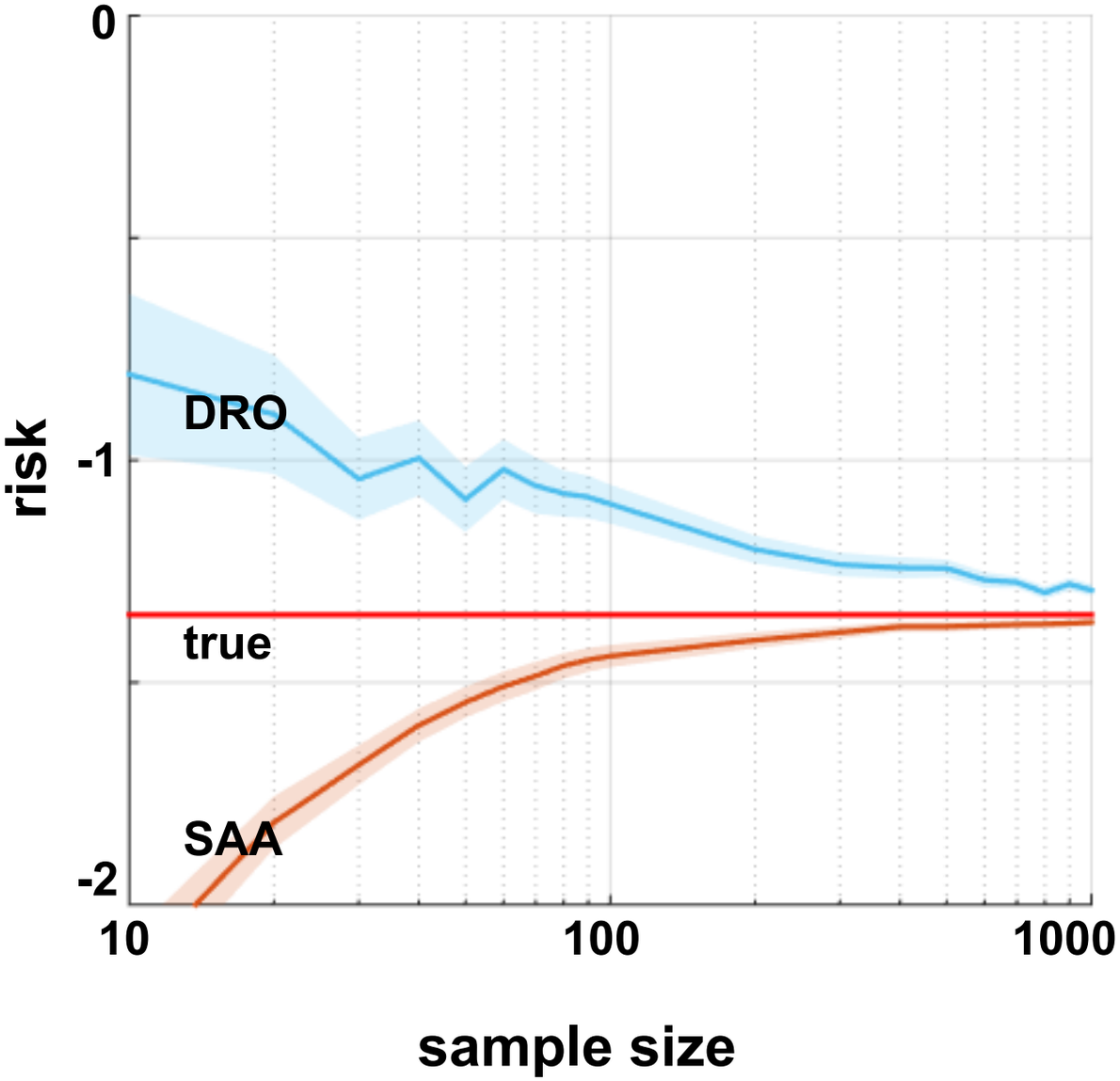
Learning Curves

what we **think** to get …

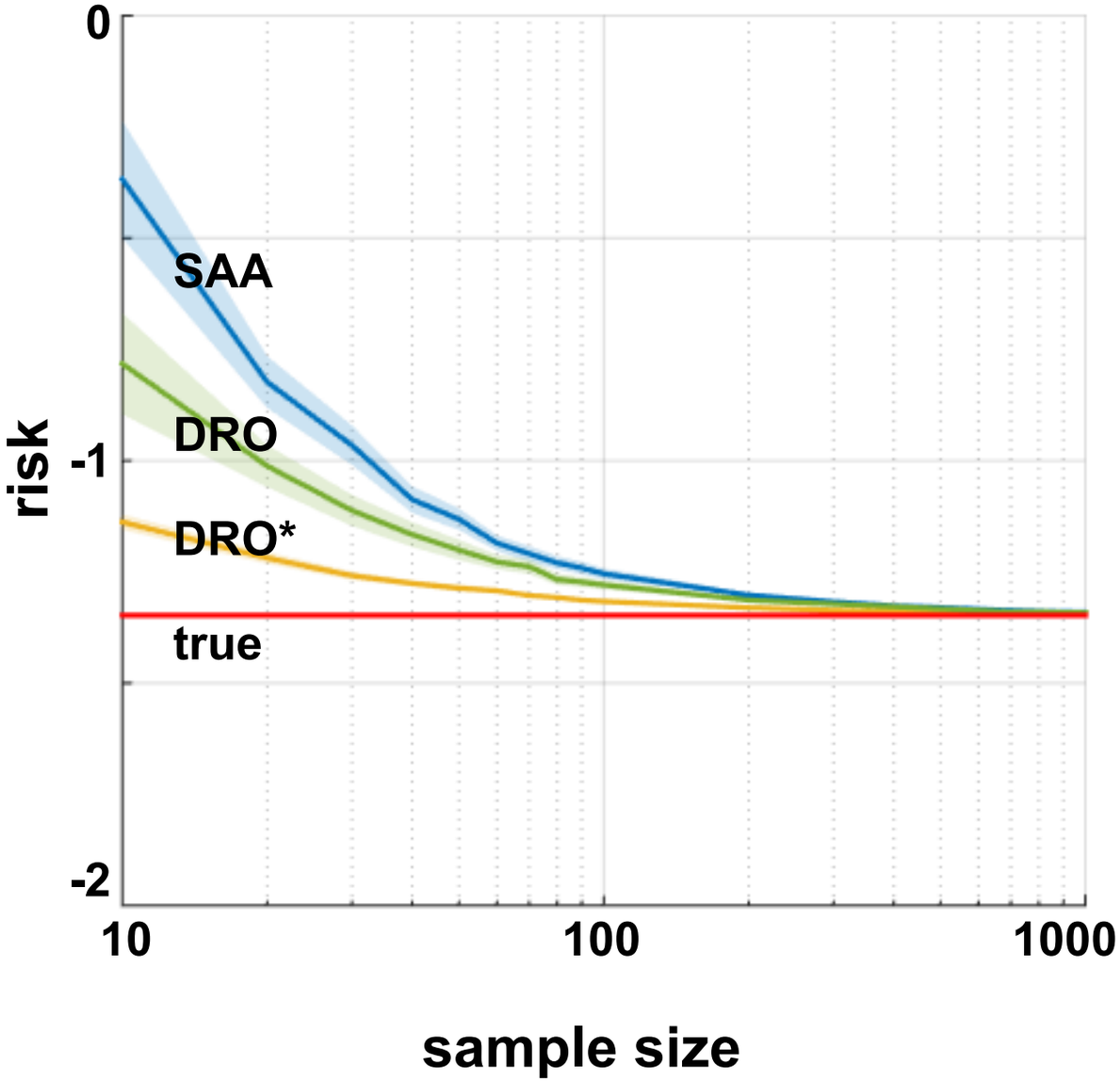


Learning Curves

what we **think** to get …



what we **actually** get …



Tractability

Worst-case expectation problem:

$$\sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}\left[\ell(\boldsymbol{\xi})\right]$$

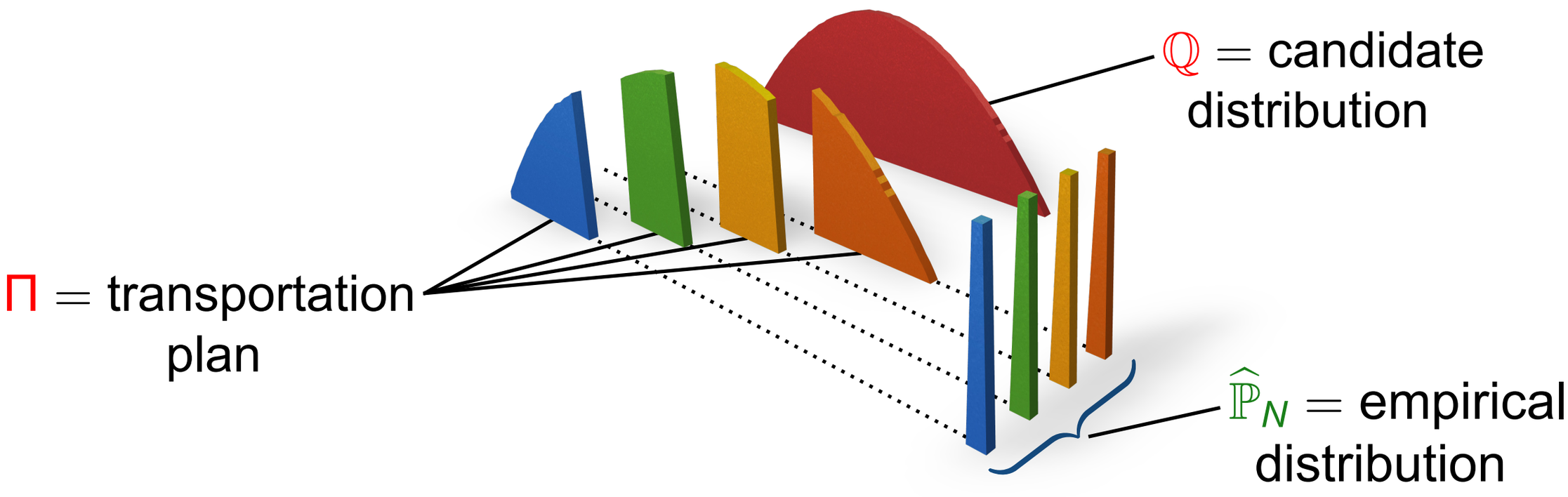
Tractability

Use the Kantorovich-Rubinstein theorem:

$$\sup_{\Pi,\mathbb{Q}} \quad \int_{\Xi} \ell(\xi)\, \mathbb{Q}(\mathrm{d}\xi)$$

$$\text{s.t.} \quad \int_{\Xi^2} \|\xi - \xi'\|\, \Pi(\mathrm{d}\xi, \mathrm{d}\xi') \leq \varepsilon$$

$$\left\{ \begin{array}{l} \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\ \text{with marginals } \mathbb{Q} \text{ and } \widehat{\mathbb{P}}_N, \text{ respectively} \end{array} \right.$$

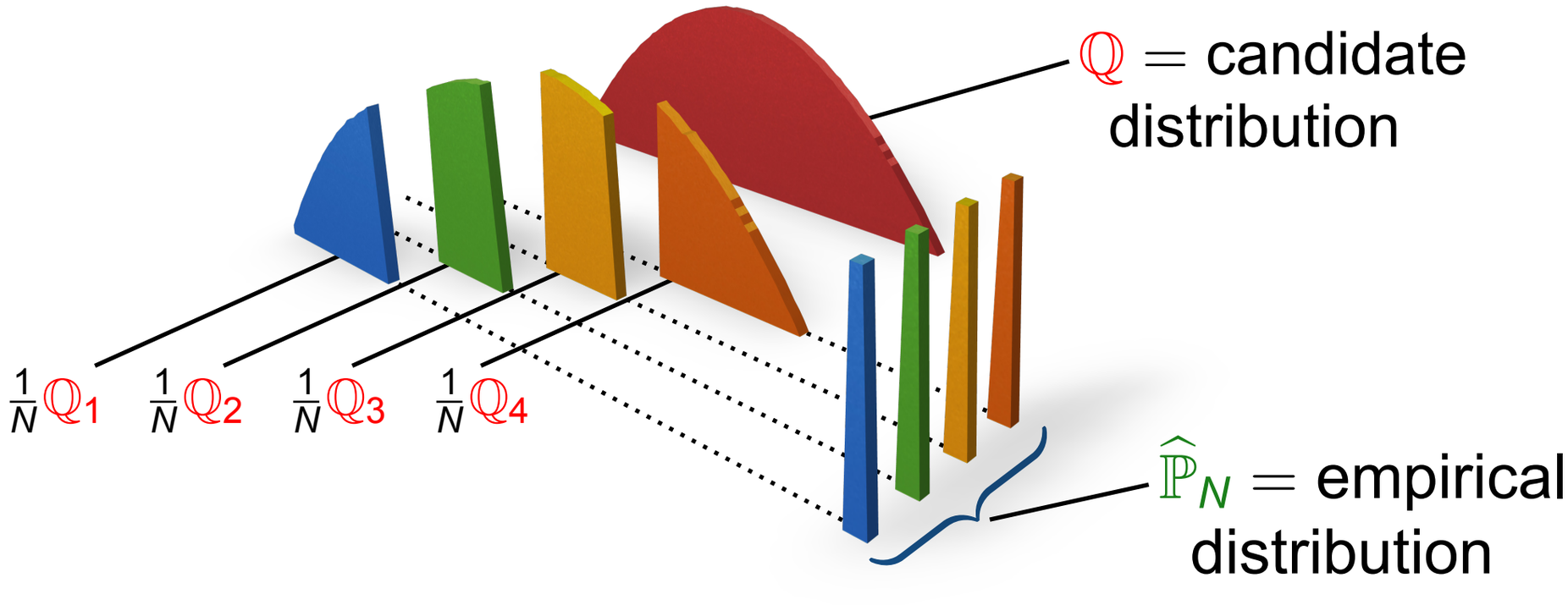


Tractability

Decompose Π into $\mathbb{Q}_1, \ldots, \mathbb{Q}_N$:

$$\sup_{\mathbb{Q}_i} \quad \frac{1}{N} \sum_{i=1}^{N} \int_{\Xi} \ell(\boldsymbol{\xi})\, \mathbb{Q}_i(\mathrm{d}\boldsymbol{\xi})$$

$$\text{s.t.} \quad \frac{1}{N} \sum_{i=1}^{N} \int_{\Xi} \|\boldsymbol{\xi} - \widehat{\boldsymbol{\xi}}_i\|\, \mathbb{Q}_i(\mathrm{d}\boldsymbol{\xi}) \leq \varepsilon$$



Tractability

Dual of the moment problem is a robust program:

$$\begin{array}{ll} \inf\limits_{\lambda, s_i} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{s.t.} & \ell(\xi) - \lambda\|\xi - \widehat{\xi}_i\| \leq s_i \quad \forall \xi \in \Xi, \quad \forall i \leq N \\ & \lambda > 0 \end{array}$$

Tractability

Introduce the indicator function of Ξ:

$$\inf_{\lambda, s_i} \quad \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i$$

$$\text{s.t.} \quad \ell(\xi) - \lambda\|\xi - \widehat{\xi}_i\| - \delta_{\Xi}(\xi) \leq s_i \quad \forall \xi \in \mathbb{R}^m, \quad \forall i \leq N$$

$$\lambda > 0$$

Tractability

Reformulate robust program as bilevel program:

$$\begin{aligned}
\inf_{\lambda, s_i} \quad & \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i \\
\text{s.t.} \quad & \sup_{\xi \in \mathbb{R}^m} \left(\ell(\xi) - \lambda \|\xi - \widehat{\xi}_i\| - \delta_{\Xi}(\xi) \right) \leq s_i \quad \forall i \leq N \\
& \lambda \geq 0
\end{aligned}$$

Tractability

Take the Fenchel dual of the lower-level problem:[7)]

$$\inf_{\lambda, s_i, v_i, z_i} \quad \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i$$

$$\text{s.t.} \quad [-\ell]^*(z_i - v_i) + \sigma_\Xi(v_i) - z_i^\top \hat{\xi}_i \leq s_i \quad \forall i \leq N$$

$$\|z_i\|_* \leq \lambda$$

dual norm of z_i

convex conjugate of $-\ell$

support function of Ξ

---

7) Ben-Tal, den Hertog & Vial, *Math. Program.*, 2015.

Tractability

The worst-case expectation equals:

$$\inf_{\lambda, s_i, v_i, z_i} \quad \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i$$

$$\text{s.t.} \quad [-\ell]^*(z_i - v_i) + \sigma_{\Xi}(v_i) - z_i^{\top}\widehat{\xi}_i \leq s_i \quad \forall i \leq N$$

$$\|z_i\|_* \leq \lambda$$

- **Finite convex program**
- Problem size grows **polynomially** in input data
- Can be combined with minimization over $x \in \mathcal{X}$: the resulting problem is in the **same complexity class as SAA**