Erick's notes

I wish to warmly thank Daniel Kuhn for sharing this keynote presentation

Data-Driven Distributionally Robust Optimization using the Wasserstein Metric

Daniel Kuhn

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

SWISS NATIONAL SCIENCE FOUNDATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Stochastic Programming

$$\text{SP}: \quad \min_{x \in \mathcal{X}} \mathbb{E}^{\mathbb{P}}\left[\ell(x, \xi)\right]$$

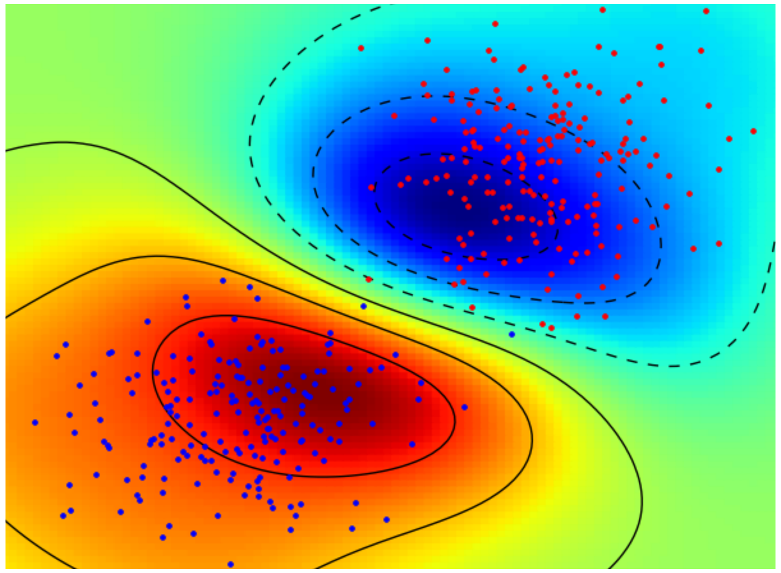
$$J^* = \min \text{SP}$$

$$x^* = \operatorname{argmin} \text{SP}$$

Applications:

Supply Chain Mgmt.

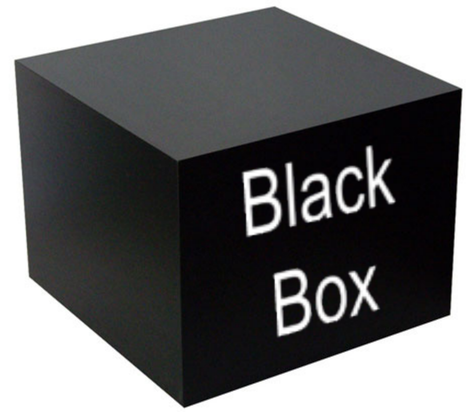
Portfolio Mgmt.



Machine Learning

Data-Driven Stochastic Programming

$$\underbrace{\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N}_{\text{training data} \sim \mathbb{P}^N} \longrightarrow$$



$$\longrightarrow \begin{cases} \widehat{x} & \text{decision} \\ \widehat{J} & \text{certificate} \end{cases}$$

$$\underbrace{\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_M}_{\text{test data} \sim \mathbb{P}^M} \longrightarrow \frac{1}{M}\sum_{i=1}^{M} \ell(\widehat{x}, \widehat{\xi}_i) \to \underbrace{\mathbb{E}^{\mathbb{P}}\left[\ell(\widehat{x}, \xi)\right]}_{\text{out-of-sample performance}} \geq J^*$$

Aim: establish a finite sample guarantee

$$\mathbb{P}^N \left\{ \mathbb{E}^{\mathbb{P}}\left[\ell(\widehat{x}, \xi)\right] \leq \widehat{J} \right\} \geq 1 - \beta$$

Sample Average Approximation (SAA)

$$\text{SAA}: \quad \min_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^{N} \ell(x, \widehat{\xi}_i)$$

$$\underbrace{\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N}_{\text{training data} \sim \mathbb{P}^N} \longrightarrow$$

$$\longrightarrow \begin{cases} \widehat{x} = \text{argmin}\,(\text{SAA}) \\ \widehat{J} = J^* + \varepsilon \end{cases}$$

Finite sample guarantee:[2)]

$$\mathbb{P}^N \left\{ \mathbb{E}^{\mathbb{P}} \left[\ell(\widehat{x}, \xi) \right] \leq \widehat{J} \right\} \geq 1 - \beta \quad \text{if}$$

$$N \geq C \left(\frac{\text{diam}(\mathcal{X})\, \text{lip}(\ell)}{\varepsilon} \right)^2 \left[\text{dim}(x) \log \left(\frac{\text{diam}(\mathcal{X})\, \text{lip}(\ell)}{\varepsilon} \right) + \log \left(\frac{c}{\beta} \right) \right]$$

---

2) Shapiro & Nemirovski, *Springer*, 2005.

SAA with Scarce Data

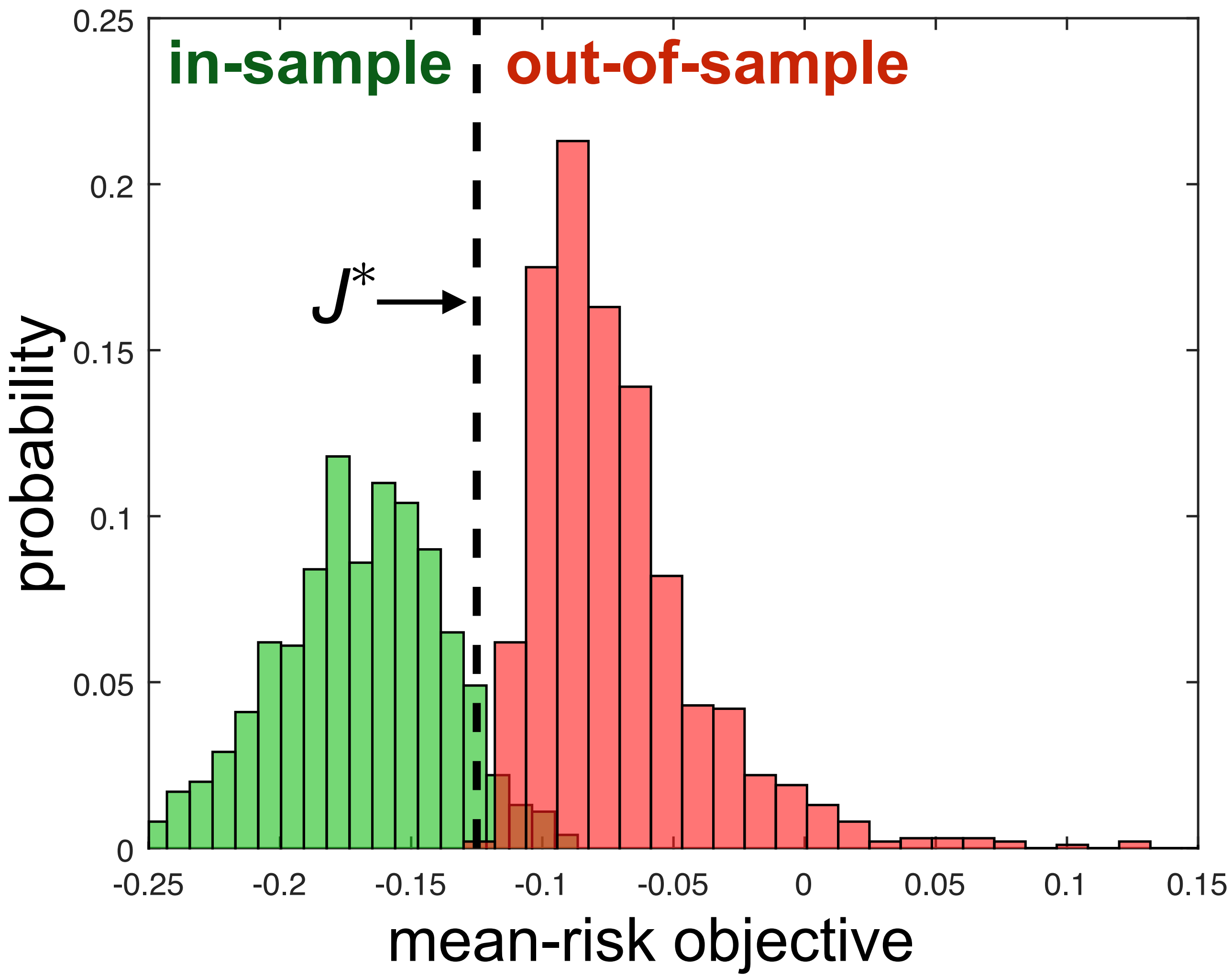
Mean-risk portfolio problem

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

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

Erick's notes
In our notation
1-epsilon=80%

Performance of SAA solution



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

Distributionally Robust Optimization (DRO)

$$\text{DRO}: \quad \min_{x \in \mathcal{X}} \ \sup_{\mathbb{Q} \in \widehat{\mathcal{P}}_N} \ \mathbb{E}^{\mathbb{Q}}\left[\ell(x, \xi)\right]$$

$$\underbrace{\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N}_{\text{training data} \sim \mathbb{P}^N} \longrightarrow$$

$$\longrightarrow \begin{cases} \widehat{x} = \text{argmin (DRO)} \\ \widehat{J} = \min \text{(DRO)} \end{cases}$$

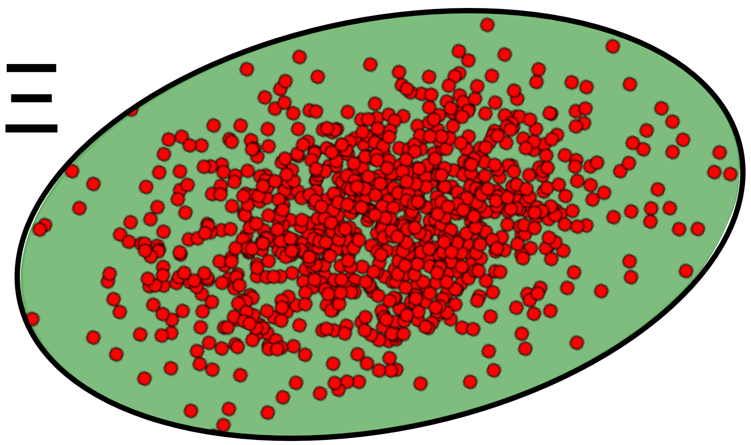
Desired properties:[3]

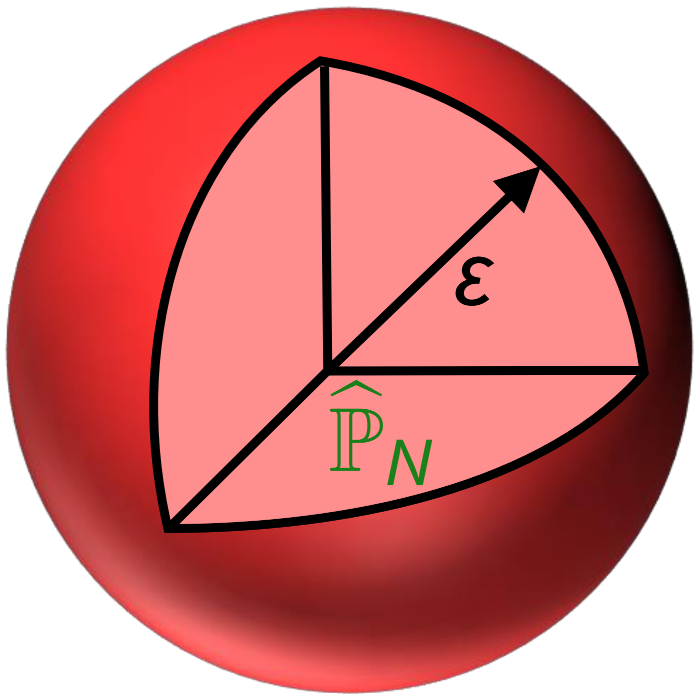
- Finite sample guarantee: $\mathbb{P}^N \left\{ \mathbb{E}^{\mathbb{P}} \left[\ell(\widehat{x}, \xi)\right] \leq \widehat{J} \right\} \geq 1 - \beta$
- Asymptotic guarantee: $\mathbb{P}^\infty \left\{ \lim_{N \to \infty} \widehat{x} = x^* \right\} = 1$
- Tractability: DRO is in the same complexity class as SAA

3) Bertsimas, Gupta & Kallus, *http://arxiv.org*, 2014.

Wasserstein Ambiguity Set[4)]

$$\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N) = \left\{\mathbb{Q} : \mathbb{Q}(\xi \in \Xi) = 1\right\} \cap \left\{\mathbb{Q} : d_{\mathrm{W}}(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq \varepsilon\right\}$$





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

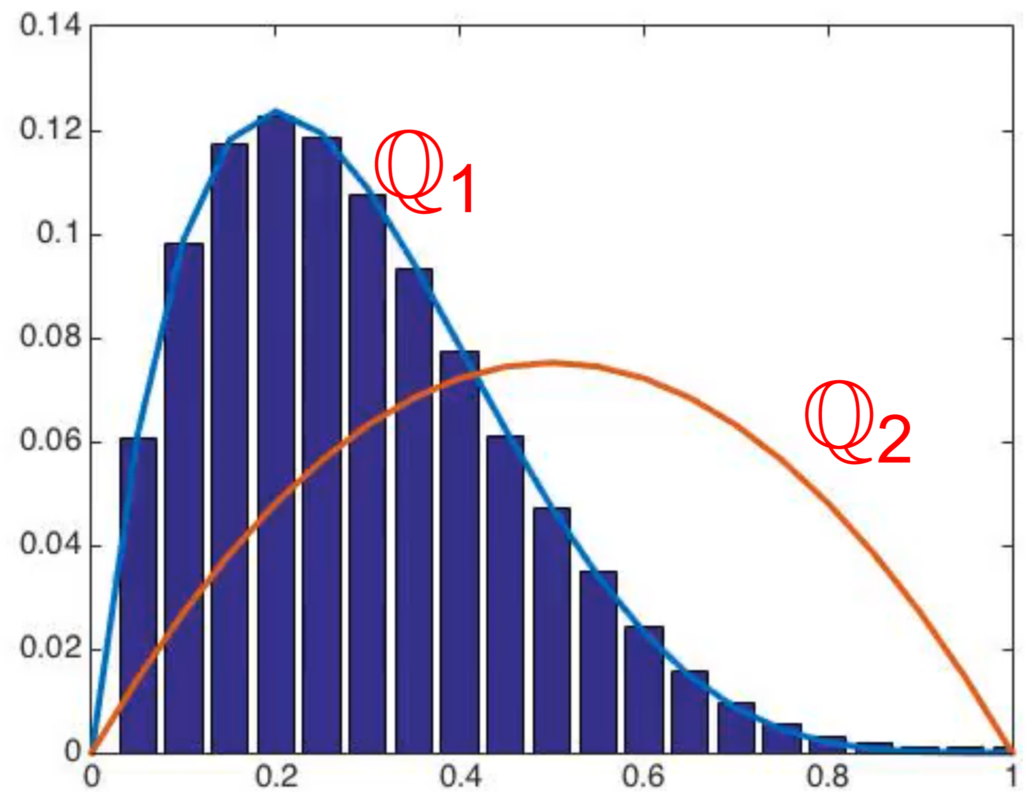
Wasserstein metric: $d_{\mathrm{W}}(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{\mathrm{lip}(f)\leq 1} \mathbb{E}^{\mathbb{Q}_1}[f(\xi)] - \mathbb{E}^{\mathbb{Q}_2}[f(\xi)]$

4) Pflug & Wozabal, *Quant. Finance*, 2007.

Kantorovich-Rubinstein Theorem[6)]

LP duality implies:

$$d_{\mathrm{W}}(\mathbb{Q}_1, \mathbb{Q}_2) = \text{inf} \left\{ \mathbb{E}^{\mathbb{Q}} \left[\|\xi_1 - \xi_2\| \right] : \begin{array}{l} \mathbb{Q} \text{ distribution of } \xi_1 \text{ and } \xi_2 \\ \text{with marginals } \mathbb{Q}_1 \text{ and } \mathbb{Q}_2 \end{array} \right\}$$



$d_{\mathrm{W}}(\mathbb{Q}_1, \mathbb{Q}_2) =$ minimum cost of moving $\mathbb{Q}_1$ to $\mathbb{Q}_2$

6) Kantorovich & Rubinstein, *Vestn. Lenin. U.*, 1958.

Finite-Sample Guarantee

Measure concentration theorem:[5)] If the tail of $\mathbb{P}$ decays exponentially at rate a, then

$$\mathbb{P}^N\left\{d_{\mathrm{W}}(\mathbb{P},\widehat{\mathbb{P}}_N) > \varepsilon\right\} \leq \begin{cases} Ce^{-cN\varepsilon^{\dim(\xi)}} & \text{if } \varepsilon \leq 1, \\ Ce^{-cN\varepsilon^{a}} & \text{if } \varepsilon > 1. \end{cases}$$

$$\varepsilon_N(\beta) = \begin{cases} (\log(C\beta^{-1})/(c\,N))^{\frac{1}{a}} & \text{if } N < \log(C\beta^{-1})/c \\ (\log(C\beta^{-1})/(c\,N))^{\frac{1}{\dim(\xi)}} & \text{if } N \geq \log(C\beta^{-1})/c \end{cases}$$

$$\Longrightarrow \quad \mathbb{P}^N\left\{\mathbb{P} \in \mathbb{B}_{\varepsilon_N(\beta)}(\widehat{\mathbb{P}}_N)\right\} \geq 1-\beta$$

$$\Longrightarrow \quad \mathbb{P}^N\left\{\mathbb{E}^{\mathbb{P}}\left[\ell(\widehat{x},\xi)\right] \leq \widehat{J}\right\} \geq 1-\beta$$

[5)] Fournier & Guillin, *Probab. Theory Rel.*, 2014.

Asymptotic Guarantee

Convergence theorem: If the tail of $\mathbb{P}$ decays exponentially and the Wasserstein radius is $\varepsilon_N(\beta_N)$ with $\beta_N \propto \exp(-N^\delta)$ for some $\delta > 0$, then we have:

- Convergence of optimal values:

$$\mathbb{P}^\infty \left\{ \lim_{N\to\infty} \widehat{J}_N = J^* \right\} = 1$$

- Convergence of optimal solutions:

$$\mathbb{P}^\infty \left\{ \limsup_{N\to\infty} \{\widehat{x}_N\} \subseteq \arg\min(\mathrm{SP}) \right\} = 1$$

Tractability

Convexity assumption:

- $-\ell$ is proper, convex and lsc;
- Ξ is convex and closed.

Tractability

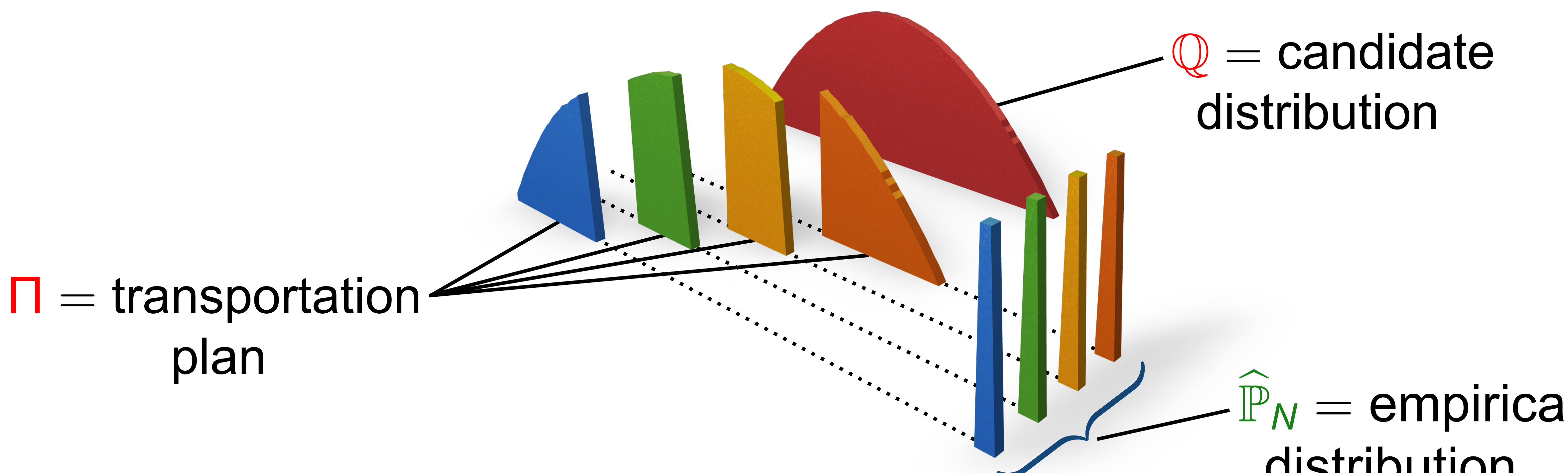
Worst-case expectation problem:

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

Tractability

Use the Kantorovich-Rubinstein theorem:

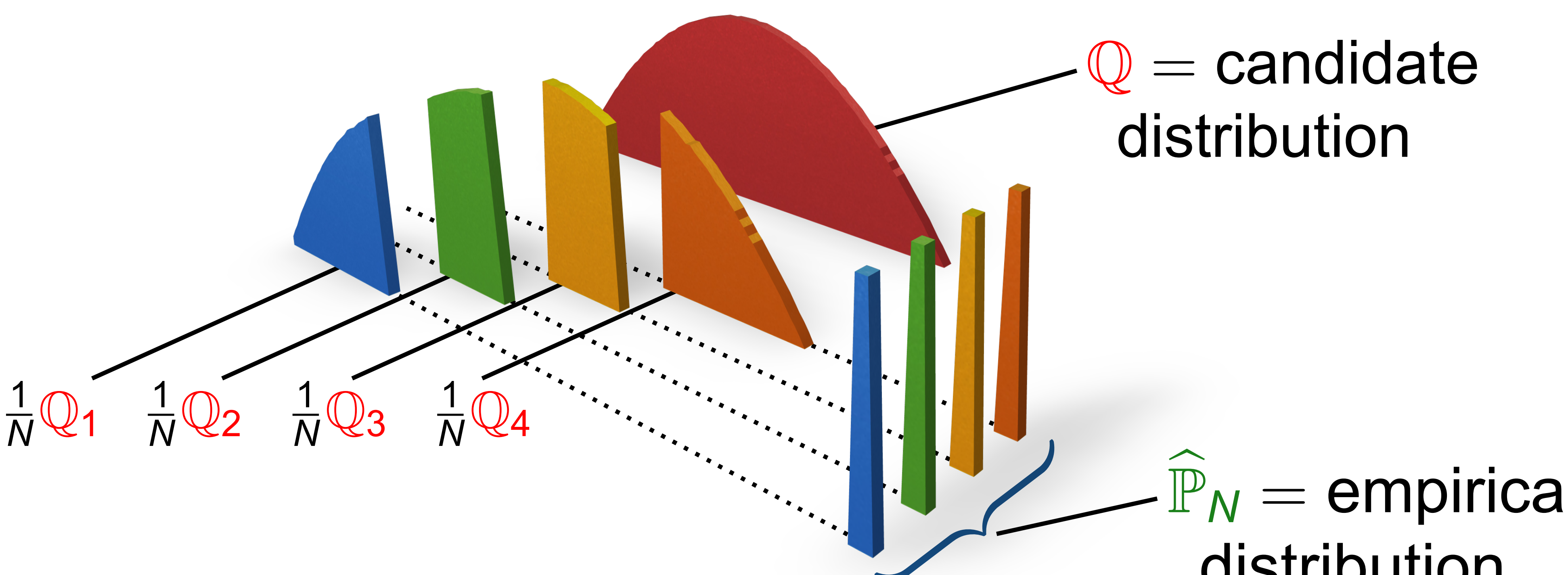
$$\begin{aligned} \sup_{\Pi, \mathbb{Q}} \quad & \int_{\Xi} \ell(\xi)\, \mathbb{Q}(\mathrm{d}\xi) \\ \text{s.t.} \quad & \int_{\Xi^2} \|\xi - \xi'\|\, \Pi(\mathrm{d}\xi, \mathrm{d}\xi') \le \varepsilon \\ & \left\{ \begin{array}{l} \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\ \text{with marginals } \mathbb{Q} \text{ and } \widehat{\mathbb{P}}_N \text{, respectively} \end{array} \right. \end{aligned}$$



Tractability

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

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



Dual of the moment problem is a robust program:

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

Introduce the indicator function of Ξ:

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

Reformulate robust program as bilevel program:

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

Tractability

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

$$
\begin{aligned}
\inf_{\lambda, s_i, v_i, z_i} \quad & \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} s_i \\
\text{s.t.} \quad & \boxed{[-\ell]^*}(z_i - v_i) + \boxed{\sigma_\Xi}(v_i) - z_i{}^\top \widehat{\xi}_i \le s_i \quad \forall i \le N \\
& \boxed{\|z_i\|_*} \le \lambda
\end{aligned}
$$

dual norm of z_i

convex conjugate of $-\ell$

support function of Ξ

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

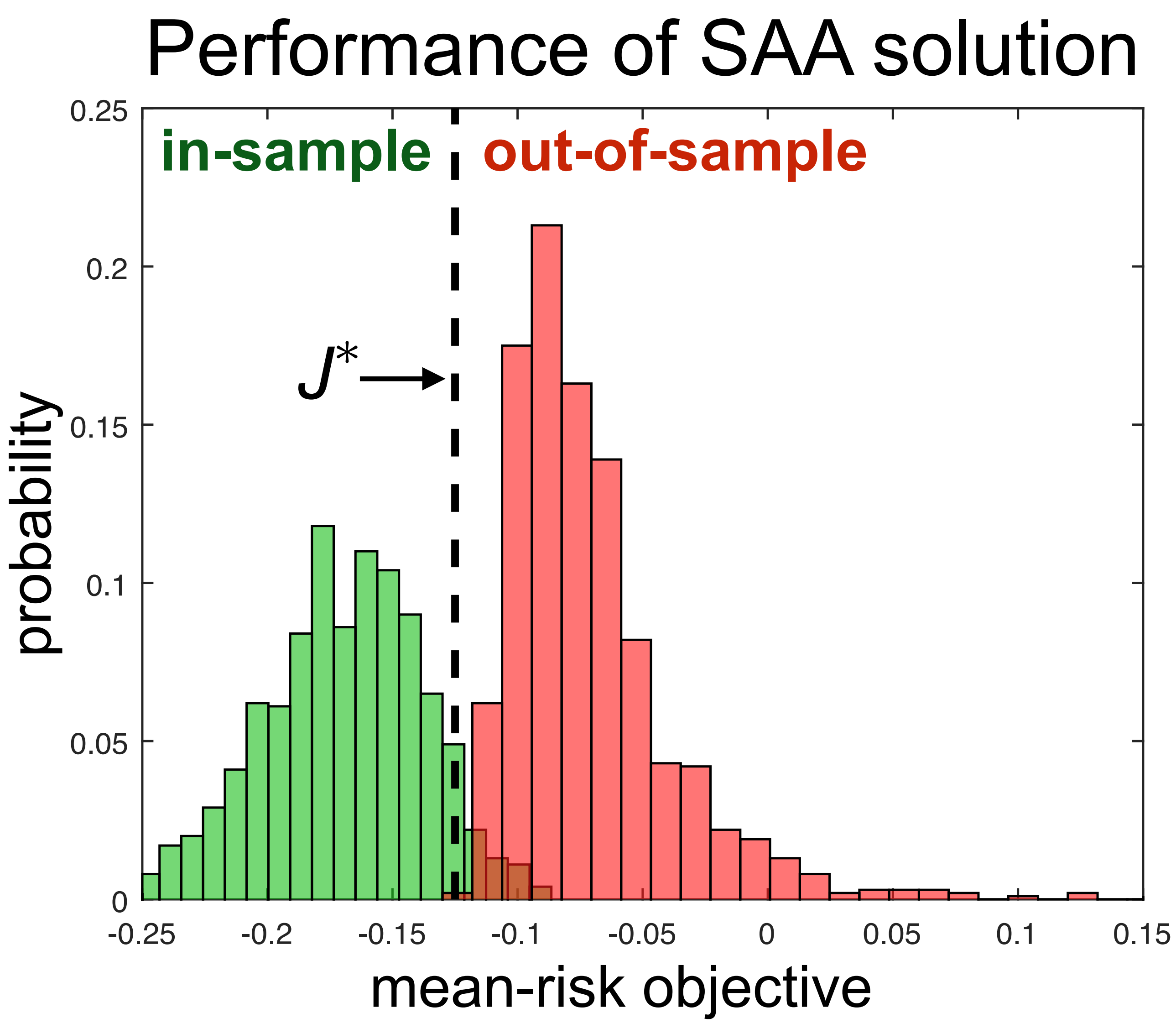
Tractability

The worst-case expectation equals:

$$\begin{aligned} \inf_{\lambda, s_i, v_i, z_i} \quad & \lambda \varepsilon + \frac{1}{N} \sum_{i=1}^{N} s_i \\ \text{s.t.} \quad & [-\ell]^*(z_i - v_i) + \sigma_{\Xi}(v_i) - z_i{}^\top \widehat{\xi}_i \leq s_i \quad \forall i \leq N \\ & \|z_i\|_* \leq \lambda \end{aligned}$$

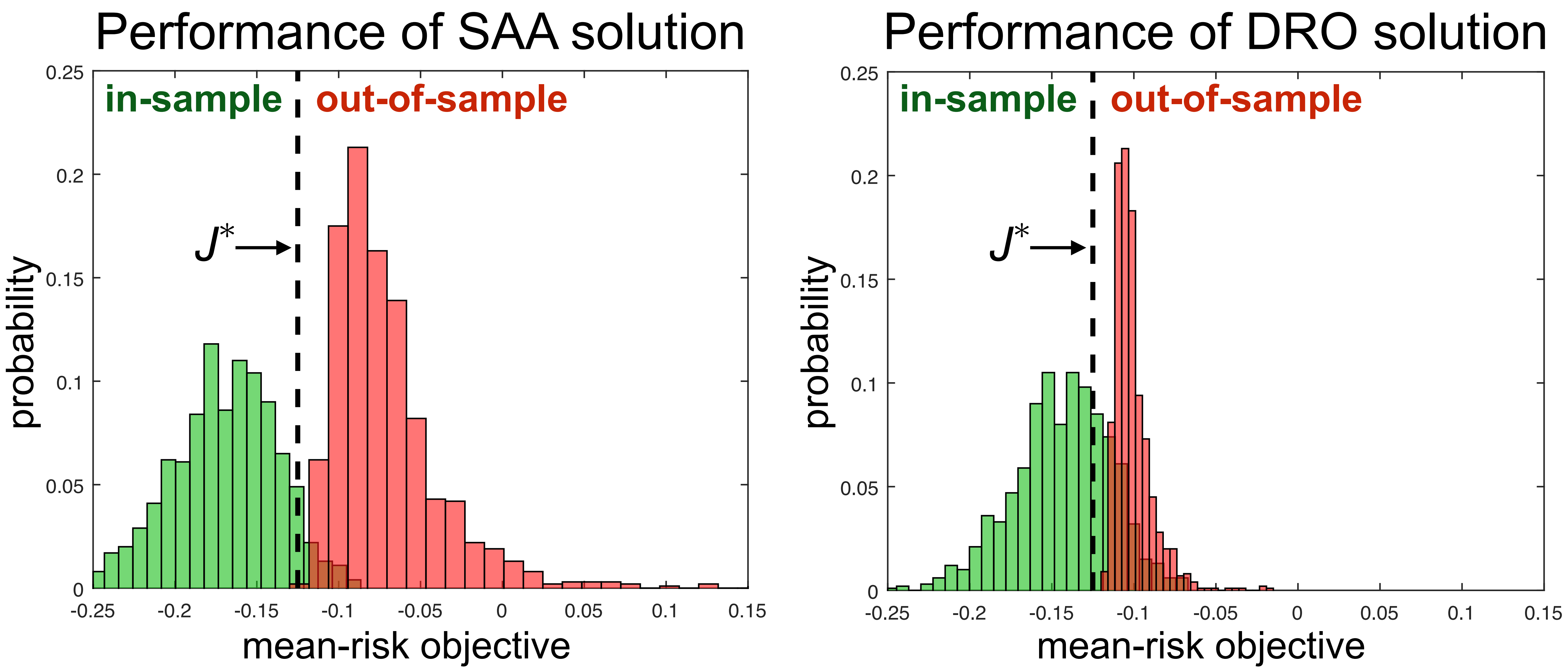
- **Finite convex program**
- Problem size grows **polynomially** in input data
- Can be combined with minimization over $x \in \mathcal{X}$: the resulting problem is in the **same complexity class as SAA**

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**

Worst-Case Distributions

Dualize finite convex program:

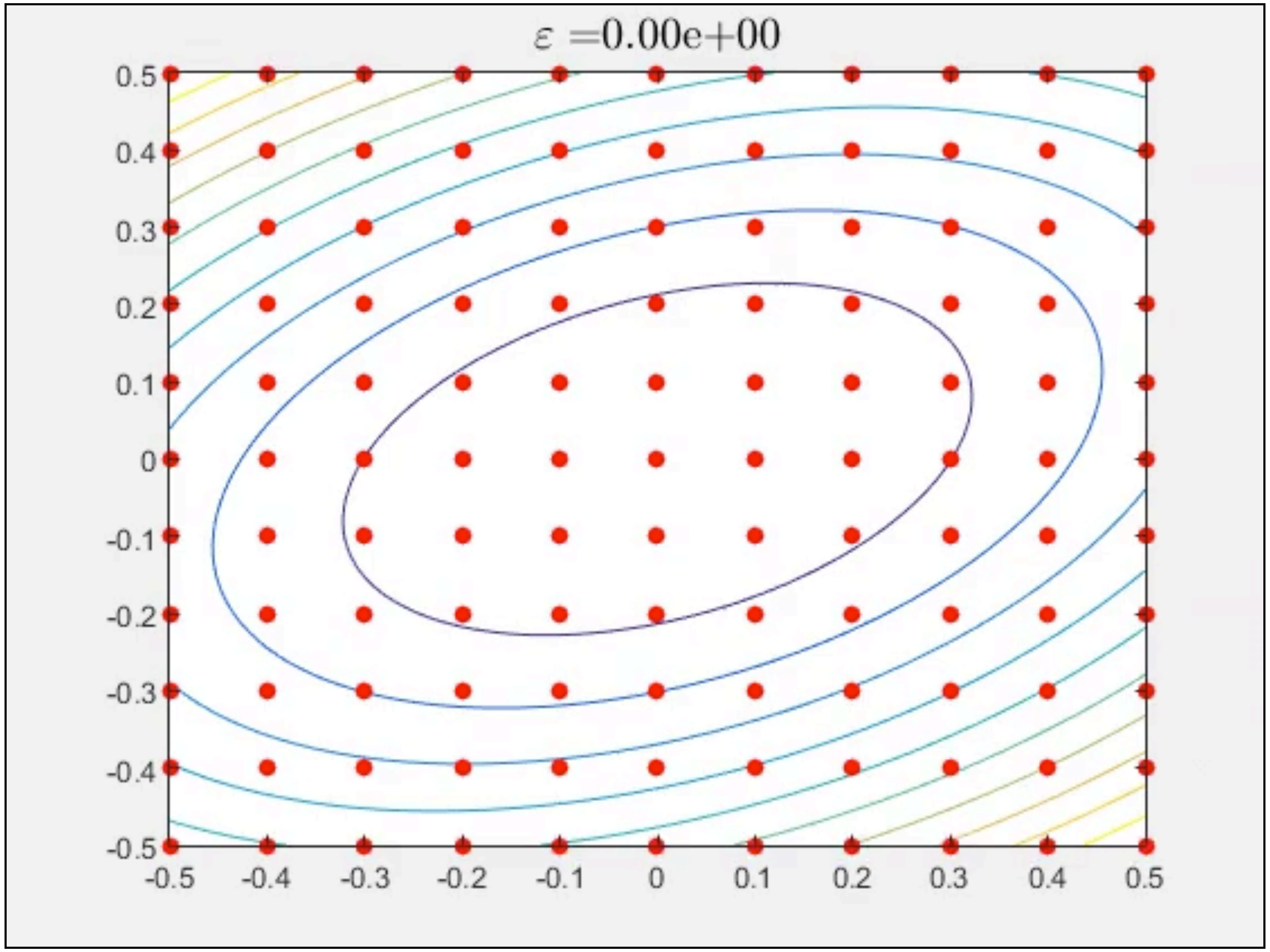
$$\sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}\left[\ell(\xi)\right] \quad = \quad \begin{array}{cl} \sup\limits_{q_i} & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \ell(\widehat{\xi}_i - q_i) \\ \text{s.t.} & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \|q_i\| \leq \varepsilon \\ & \widehat{\xi}_i - q_i \in \Xi \quad \forall i \leq N \end{array}$$

If $q_1^\star, \ldots, q_N^\star$ solves the dual problem, then

$$\mathbb{Q}^\star = \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i - q_i^\star}$$

is a worst-case distribution.

Concave Quadratic Loss Function



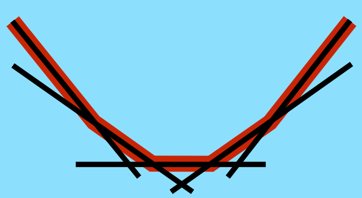
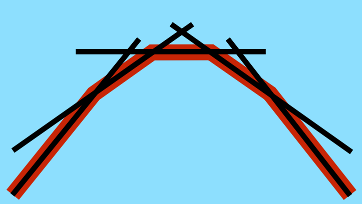
Generalized Loss Functions

Generalized convexity assumption:

- $\ell = \max\{\ell_1, \ldots, \ell_J\}$ where each $-\ell_j$ is proper, convex and lsc.

Maxima & Minima of Affine Functions

Set $\Xi = \{\xi : C\xi \leq d\}$ and $a_j(\xi) = a_j^\top \xi - b_j$ for all $j \leq J$.

$\ell(\xi)$	$\sup_{\mathbb{Q} \in \widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}[\ell(\xi)]$
$\max_{j \leq J} a_j(\xi)$ 	$\begin{array}{lll} \inf_{\lambda, s_i, \gamma_{ij}} & \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i & \\ \text{s.t.} & b_j + a_j^\top \widehat{\xi}_i + \gamma_{ij}^\top (d - C\widehat{\xi}_i) \leq s_i & \forall i \leq N,\ \forall j \leq J \\ & \lVert C^\top \gamma_{ij} - a_j \rVert_* \leq \lambda & \forall i \leq N,\ \forall j \leq J \\ & \gamma_{ij} \geq 0 & \forall i \leq N,\ \forall j \leq J \end{array}$
$\min_{j \leq J} a_j(\xi)$ 	$\begin{array}{lll} \inf_{\lambda, s_i, \gamma_i, \theta_i} & \lambda\varepsilon + \frac{1}{N}\sum_{i=1}^{N} s_i & \\ \text{s.t.} & \theta_i^\top (b - A\widehat{\xi}_i) + \gamma_i^\top (d - C\widehat{\xi}_i) \leq s_i & \forall i \leq N \\ & \lVert C^\top \gamma_i - A^\top \theta_i \rVert_* \leq \lambda & \forall i \leq N \\ & \mathrm{e}^\top \theta_i = 1, \quad \gamma_i \geq 0, \quad \theta_i \geq 0 & \forall i \leq N \end{array}$

Two-Stage Stochastic Programming

Objective uncertainty:

$$\ell(\xi) = \underbrace{\inf_{y} \left\{ y^\top Q \xi \ : \ W y \geq h \right\}}_{\text{concave piecewise affine}}$$

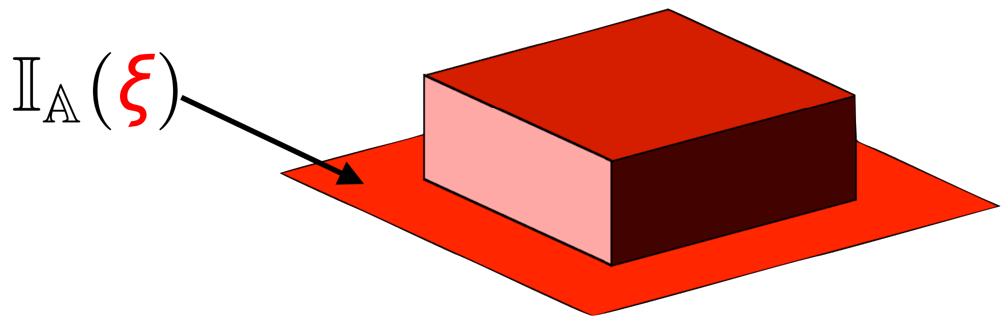
Constraint uncertainty:

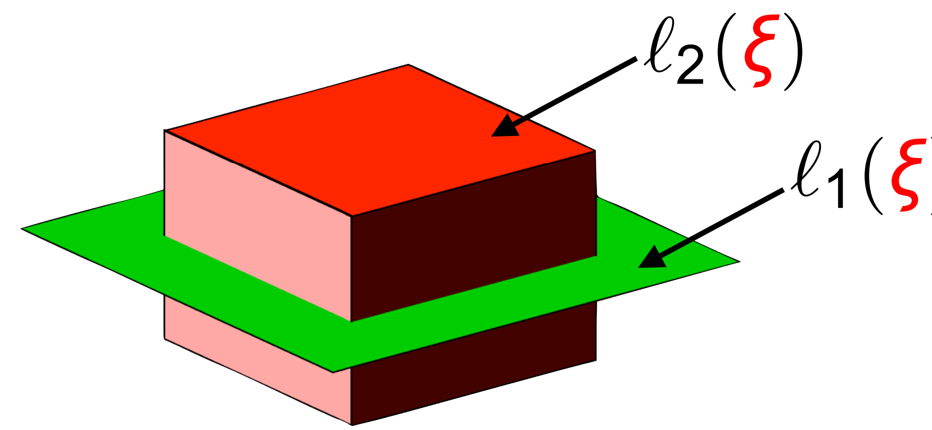
$$\ell(\xi) = \underbrace{\inf_{y} \left\{ q^\top y \ : \ W y \geq H \xi + h \right\}}_{\text{convex piecewise affine}}$$

Uncertainty Quantification

Probability of being inside a polytope:

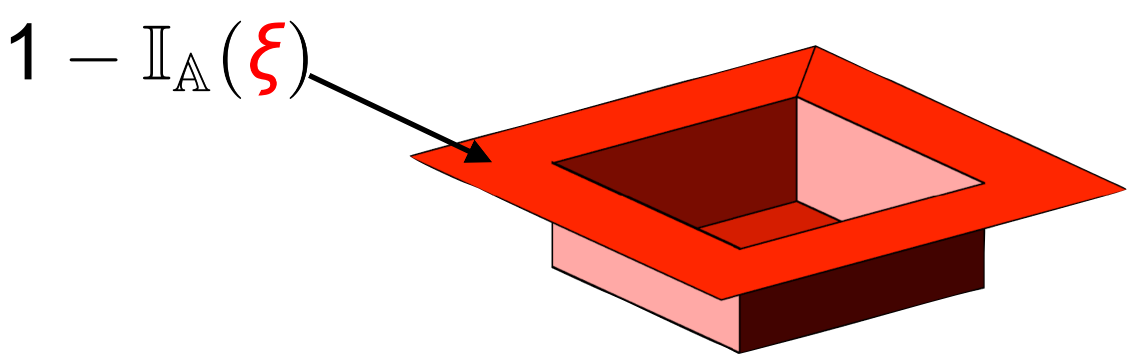
$$\sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{Q}\left[\xi \in \mathbb{A}\right] = \sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}\left[\mathbb{I}_{\mathbb{A}}(\xi)\right] = \sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}\left[\max\{\ell_1(\xi), \ell_2(\xi)\}\right]$$

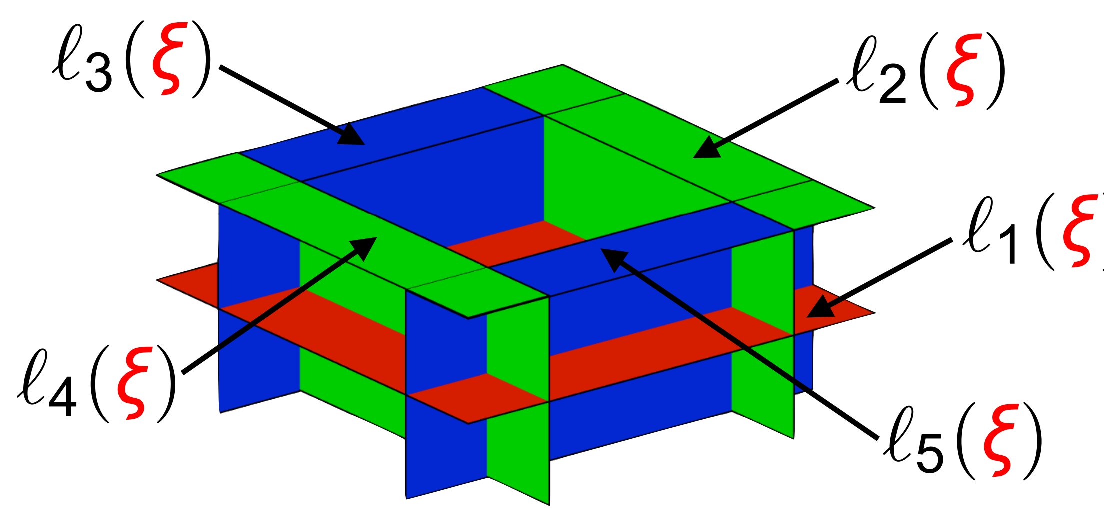




Probability of being outside of a polytope:

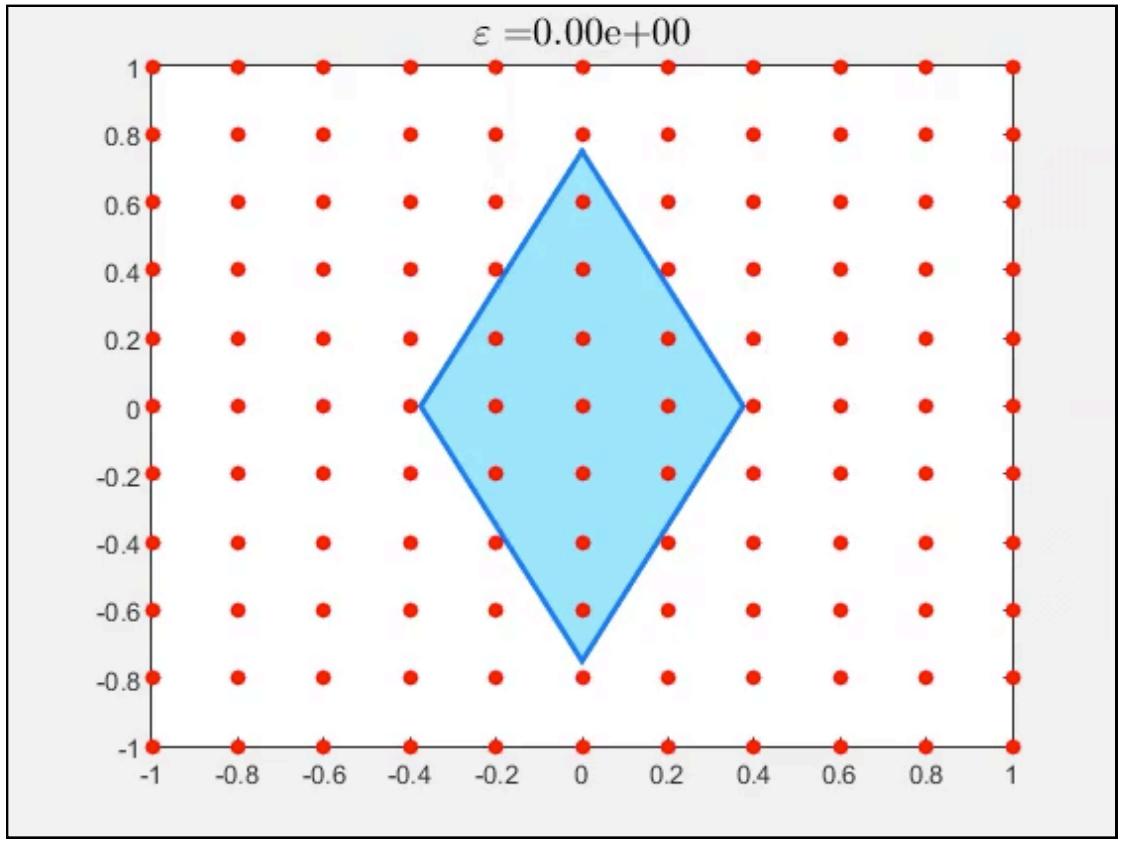
$$\sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{Q}\left[\xi \notin \mathbb{A}\right] = \sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}\left[1 - \mathbb{I}_{\mathbb{A}}(\xi)\right] = \sup_{\mathbb{Q}\in\widehat{\mathcal{P}}_N} \mathbb{E}^{\mathbb{Q}}\left[\max_{j\leq J}\{\ell_j(\xi)\}\right]$$



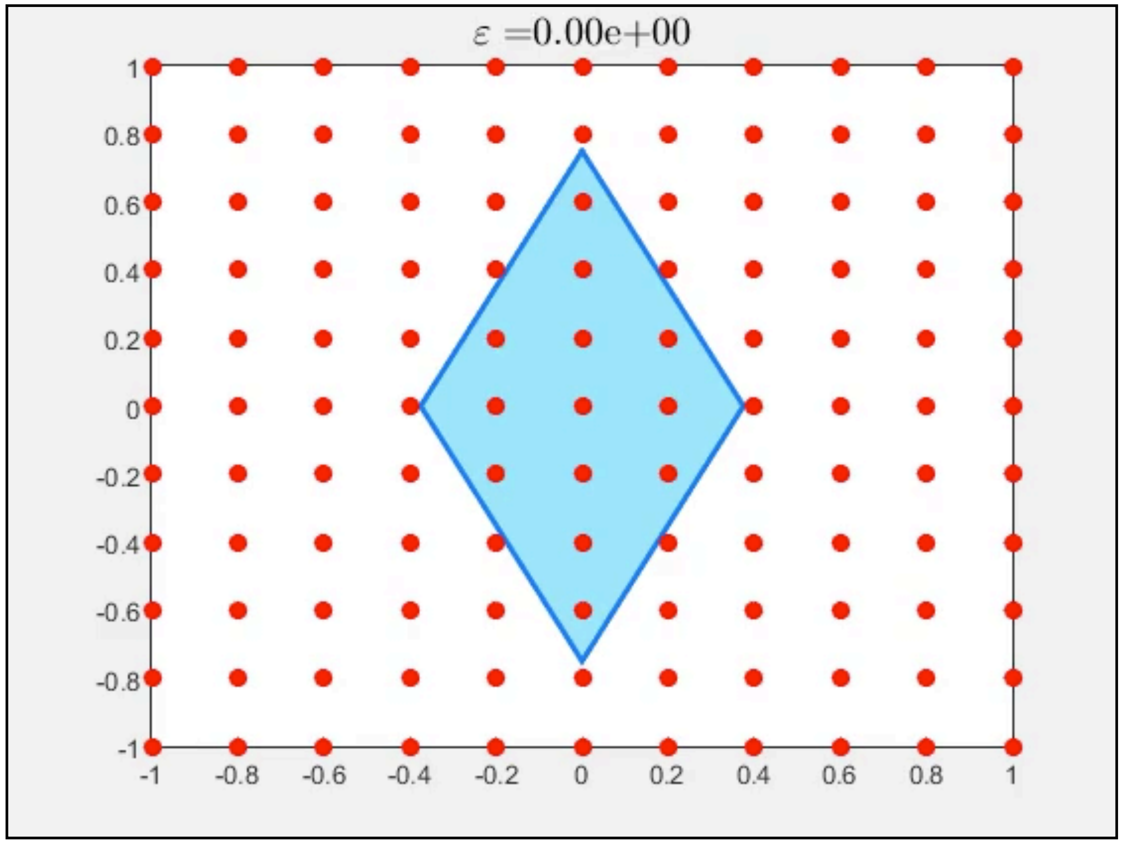


Uncertainty Quantification

Probability of being inside a polytope:



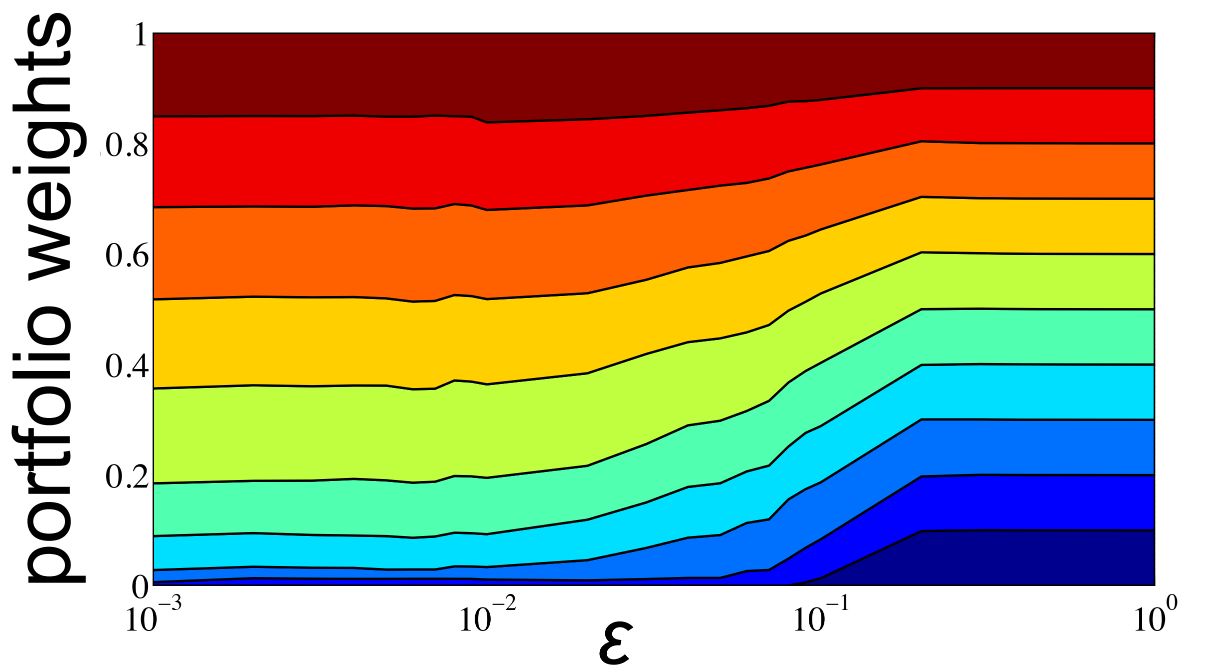
Probability of being outside of a polytope:



Application 1: Portfolio Selection

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

- 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.[8)]

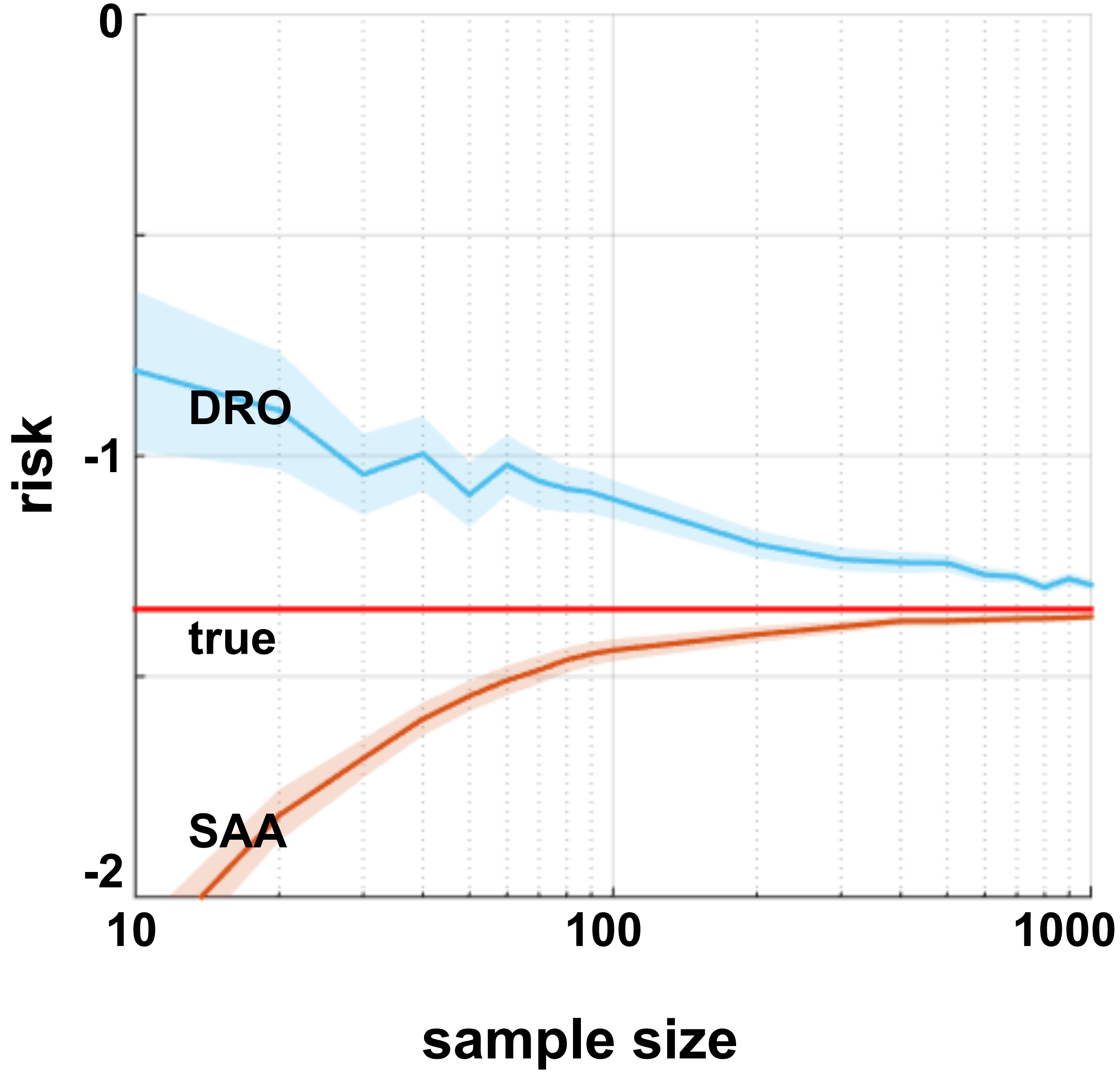
Possible Explanation: It is optimal for $\varepsilon \to \infty$.[9)]

---

8) DeMiguel, Garlappi & Uppal, *Rev. Financ. Stud.*, 2009;
9) Pflug, Pichler & Wozabal, *J. Bank. Financ.*, 2012.

Learning Curves

what we **think** to get …



Learning Curves