Learning for Decision-Making under Uncertainty

Joint work with Irina Wang, Cole Becker, and Bart Van Parys

PRINCETON UNIVERSITY ORFE

Bartolomeo Stellato — IPAM Workshop on Artificial Intelligence and Discrete Optimization, March 1 2023

It is hard to make decisions under uncertainty

Transportation

Finance

Energy

But we have data!

Problem setup with uncertain constraints

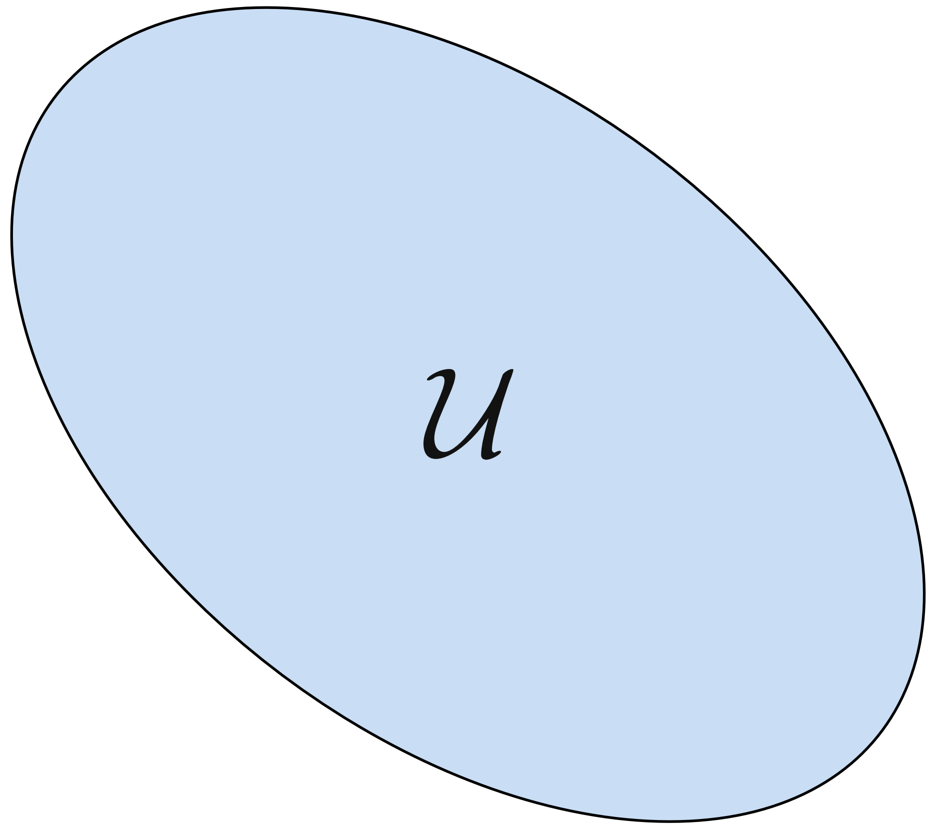
optimization variable

minimize	$f(x)$
subject to	$g(u, x) \leq 0$

uncertain parameter

We want to guarantee constraint satisfaction

Robust optimization



Recipe

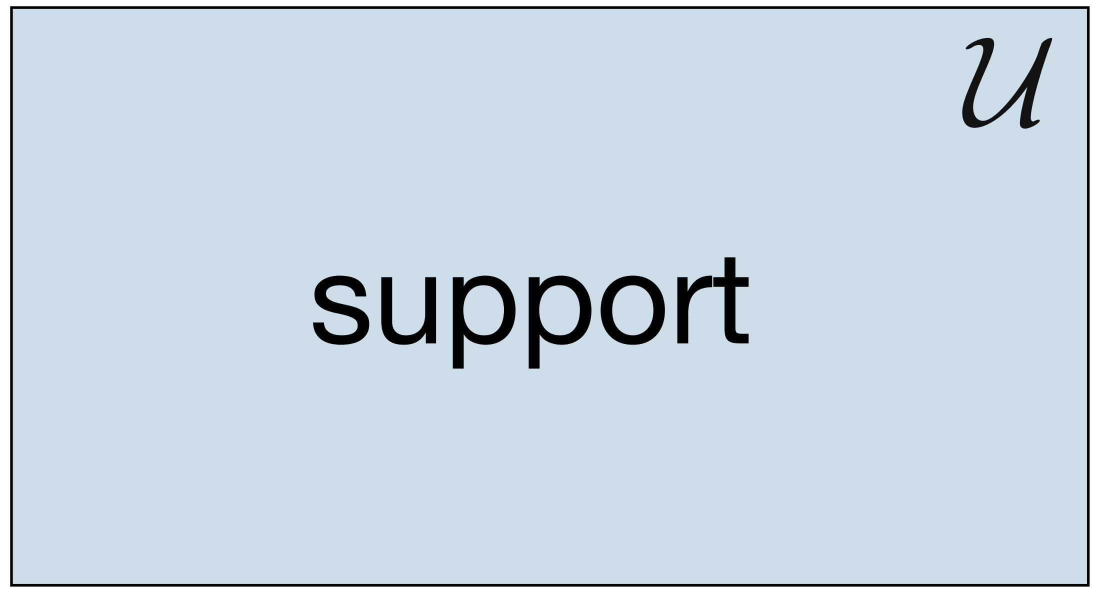
1. Pick uncertainty set $\mathcal{U}$
2. Ensure constraint satisfaction $\forall u \in \mathcal{U}$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(u, x) \leq 0, \quad \forall u \in \mathcal{U} \end{array}$$

How do we pick the uncertainty set?

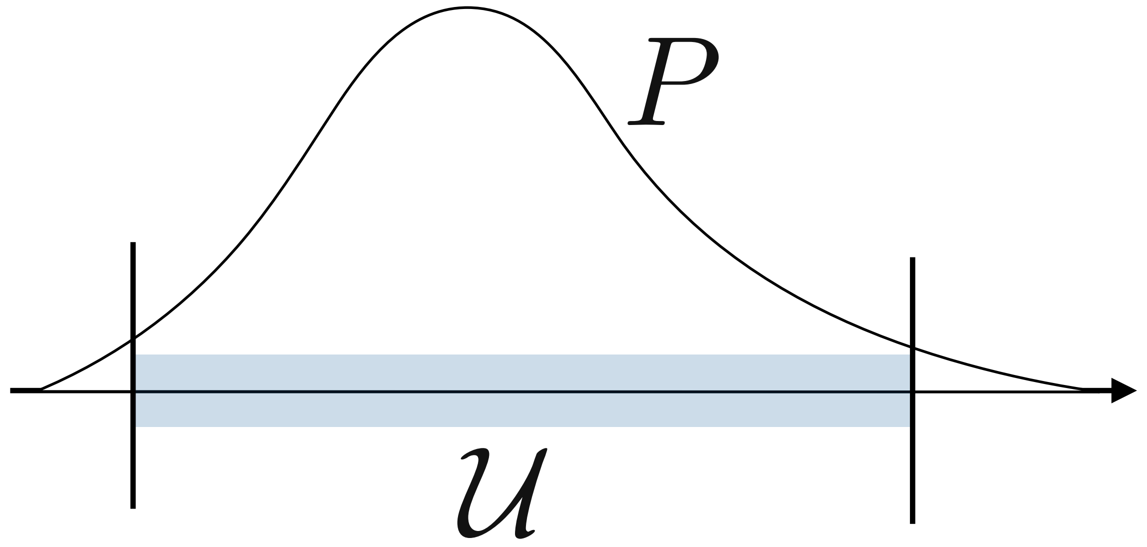
Picking the uncertainty set

Worst-case approach



❌ Very conservative

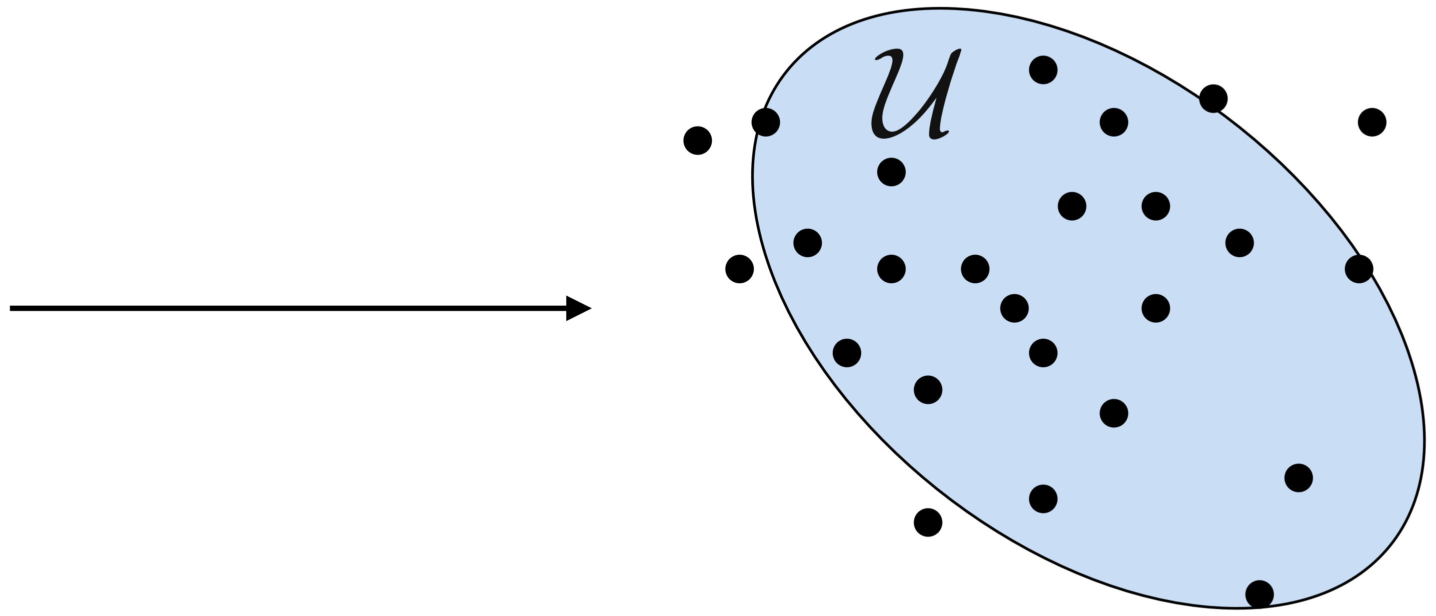
Probabilistic approach



❌ nobody knows P

Data-driven approach

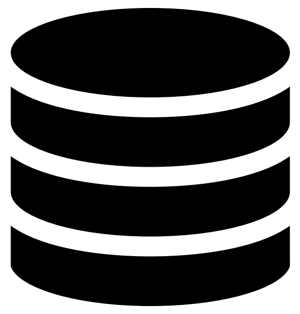
$\mathcal{D} = \{d_i\}_{i=1}^N$



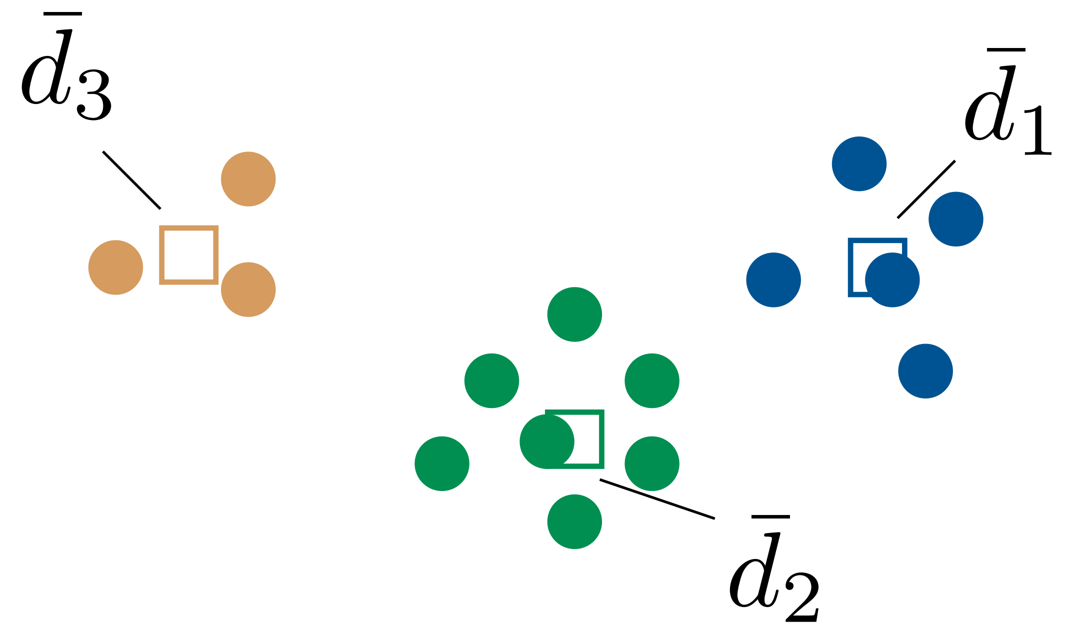
“Probabilistic Guarantees in Robust Optimization”, D. Bertsimas, D. den Hertog, and J. Pauphilet (2019)
“Data-driven robust optimization”, D. Bertsimas, V. Gupta, and N. Kallus (2014)
“Learning-Based Robust Optimization: Procedures and Statistical Guarantees”, L. Jeff Hong, Z. Huang, and H. Lam (2021)

Machine learning to make better decisions under uncertainty

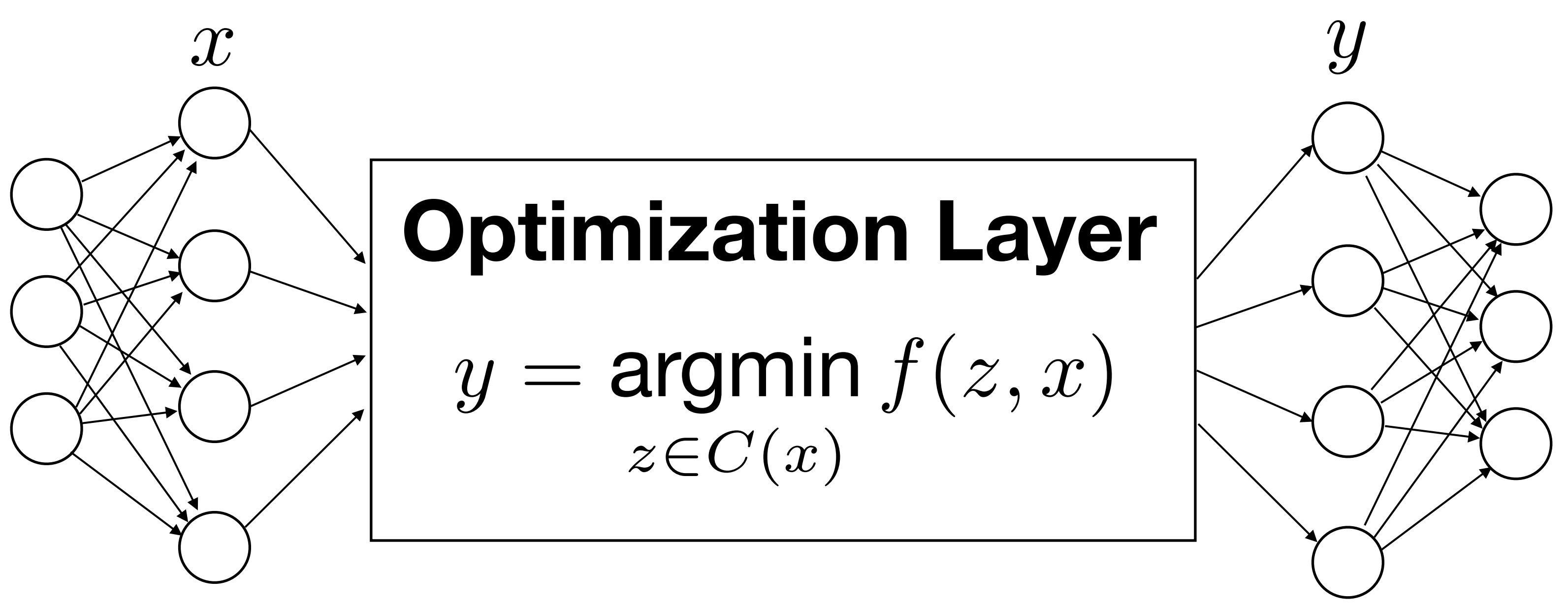
How can we use data to build
tractable and **high-performance**
uncertainty sets?



Clustering



Differentiable Optimization

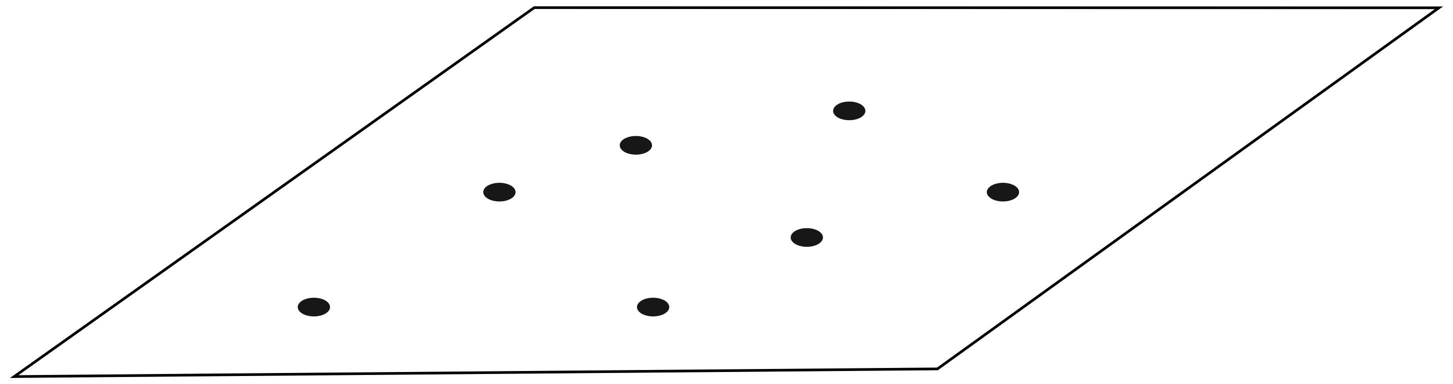


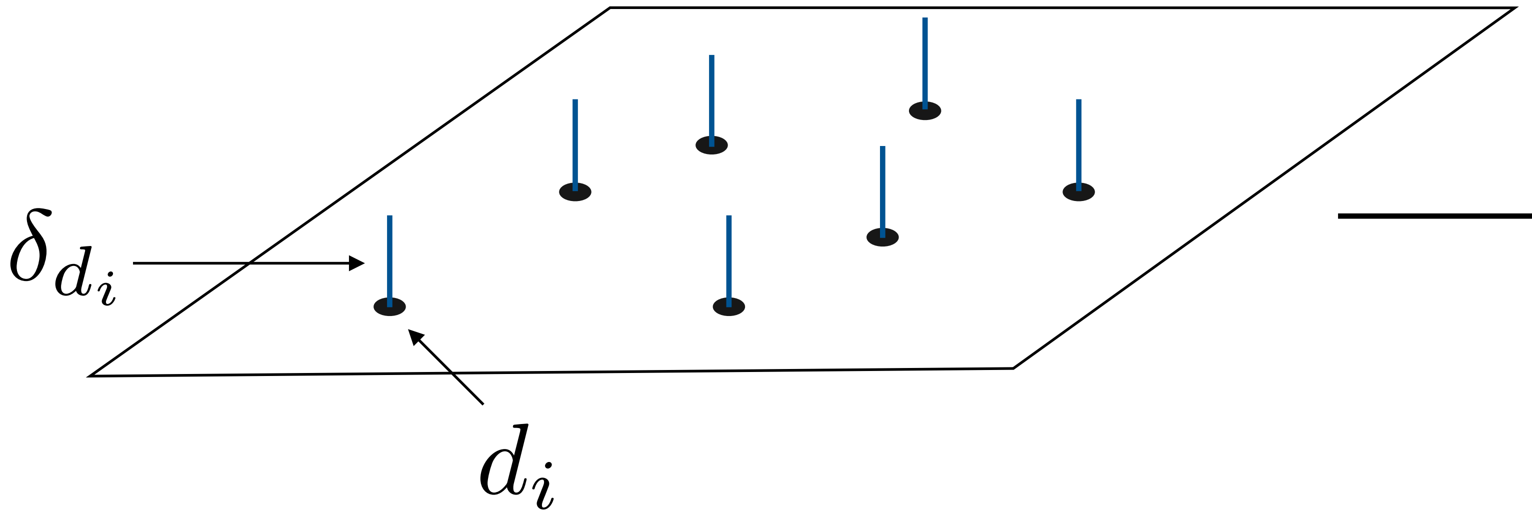
Estimating true distribution from data

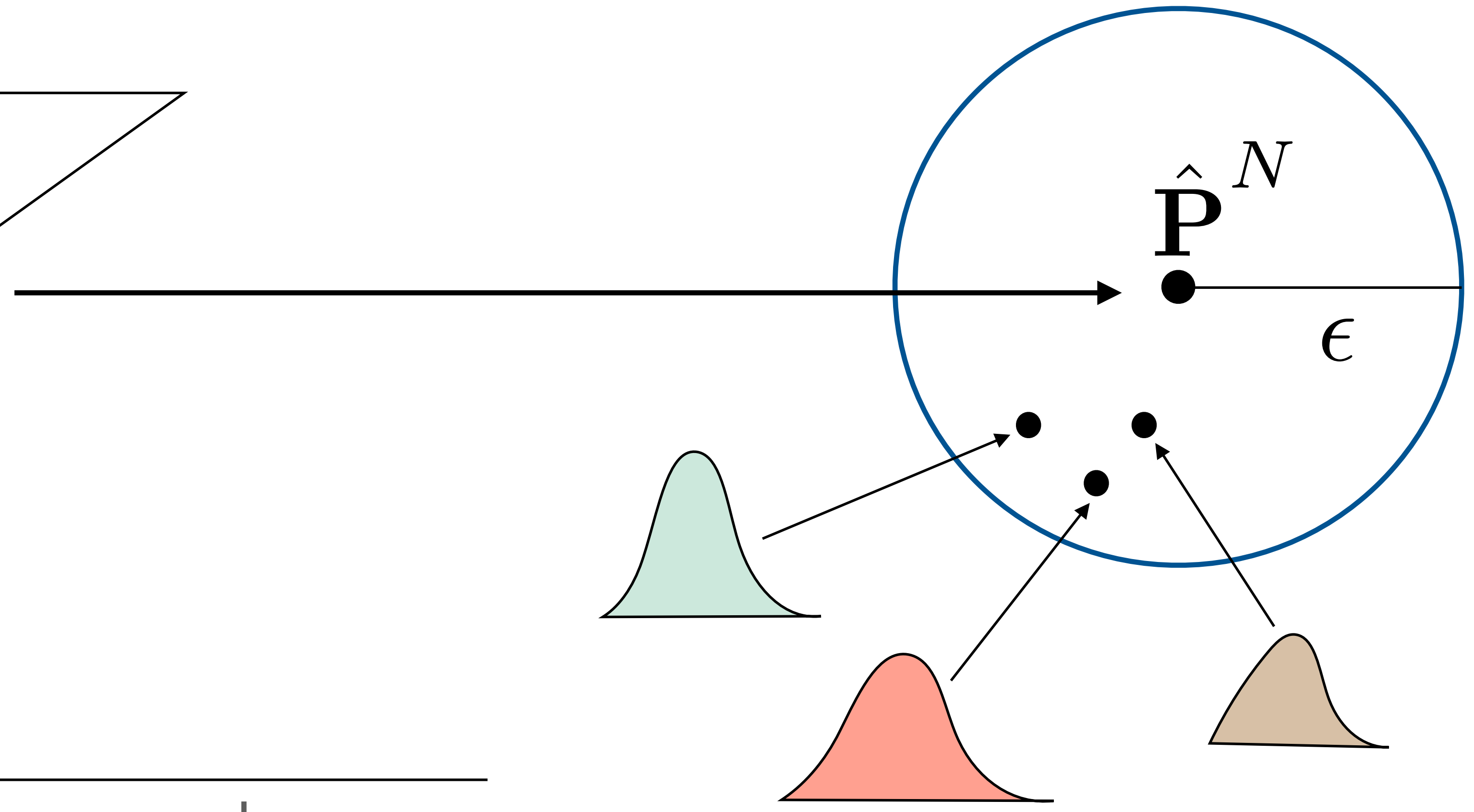
Data-Driven Distributionally Robust Optimization

Data		Empirical Distribution		Ambiguity Set
$\mathcal{D} = \{d_i\}_{i=1}^N$	$\longrightarrow$	$\hat{\mathbf{P}}^N = \frac{1}{N}\sum_{i=1}^N \delta_{d_i}$	$\longrightarrow$	$\mathcal{P}_N = \left\{ W_p(\hat{\mathbf{P}}^N, \mathbf{P}) \leq \epsilon \right\}$

Wasserstein distance

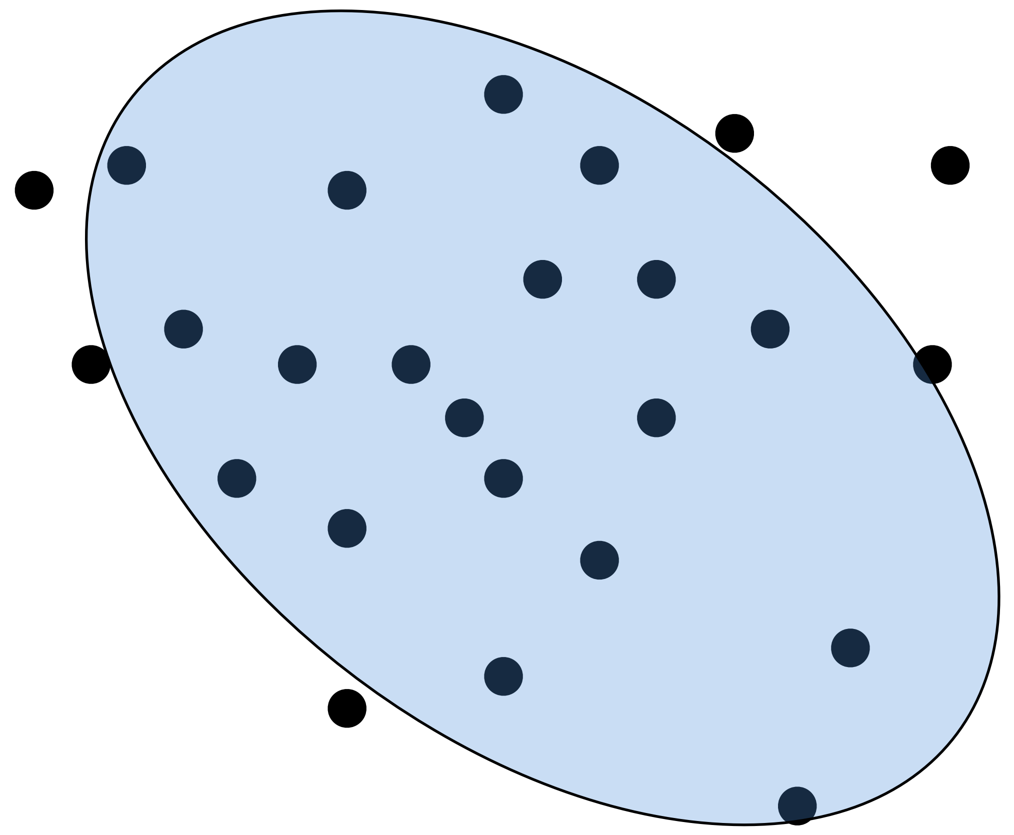






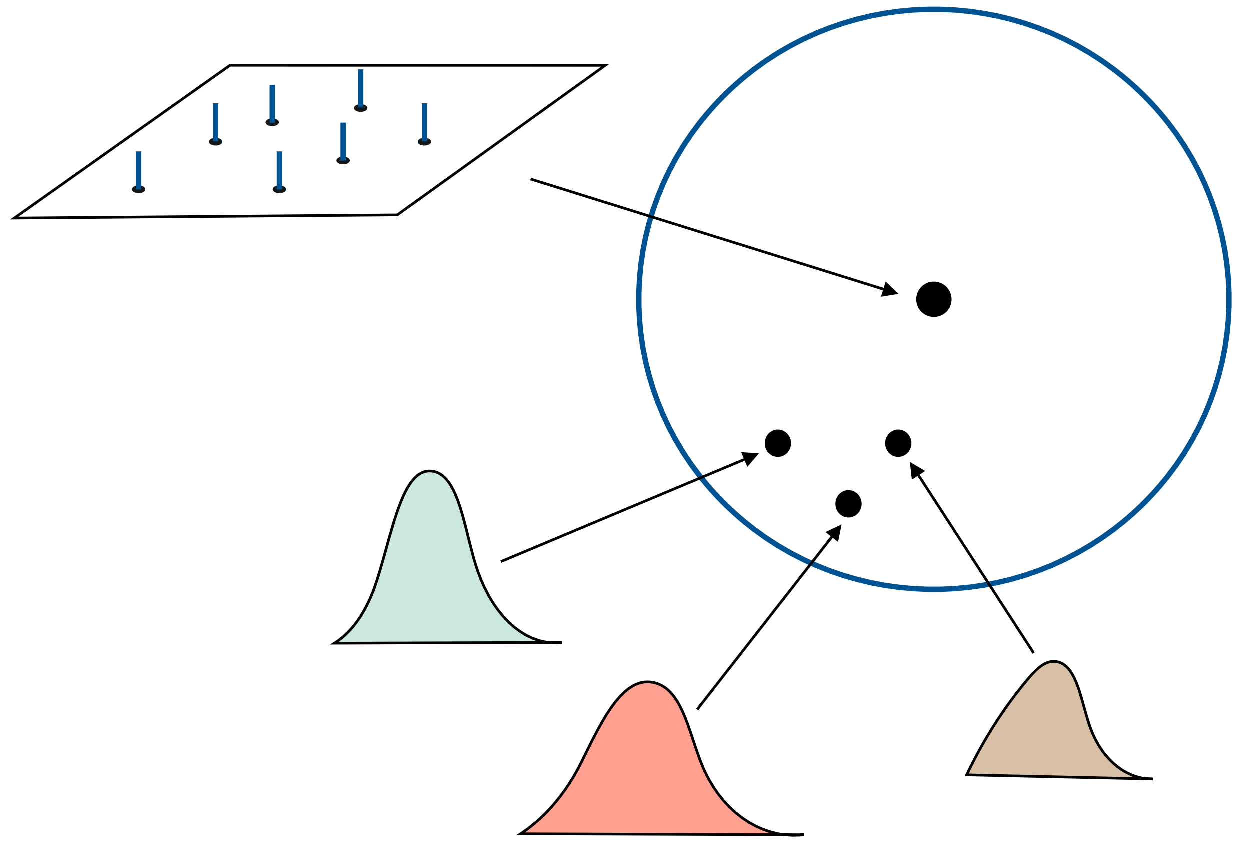
Robust vs Distributionally Robust Optimization

Robust optimization
(RO)



✓ Tractable, expressive
❌ Can be conservative

Distributionally Robust Optimization
(DRO)



✓ Can be less conservative
❌ Computationally expensive

Can we get the best of both worlds?

Probabilistic guarantees

$$\mathbf{E}(g(u, x)) \leq 0$$

concave function

$u \sim P$

(but we never know P!)

Data

$\mathcal{D} = \{d_i\}_{i=1}^N$

Data-driven probabilistic guarantees

Product Distribution

$$\mathbf{P}^N \left(\mathbf{E}(g(u, \hat{x}_N)) \leq 0\right) \geq 1 - \beta$$

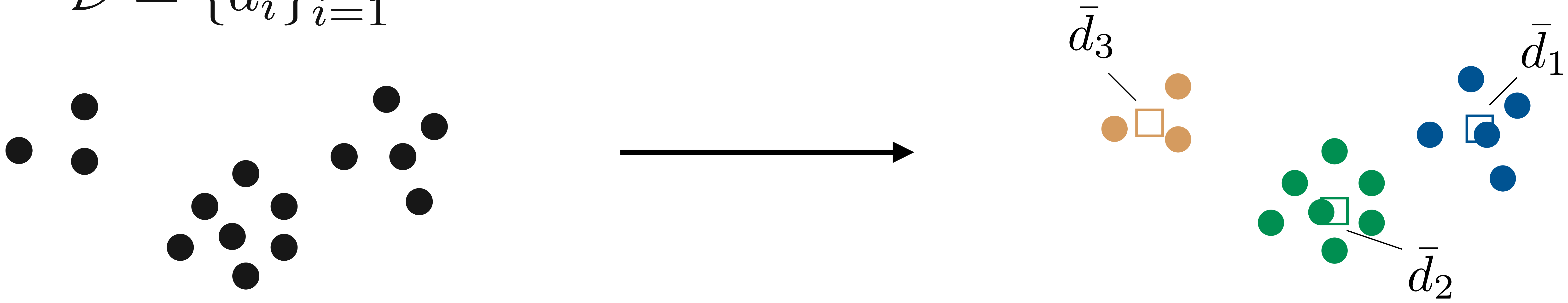
probability of constraint satisfaction

data-driven solution

Machine Learning Clustering

Data

$$\mathcal{D} = \{d_i\}_{i=1}^{N}$$



Goal

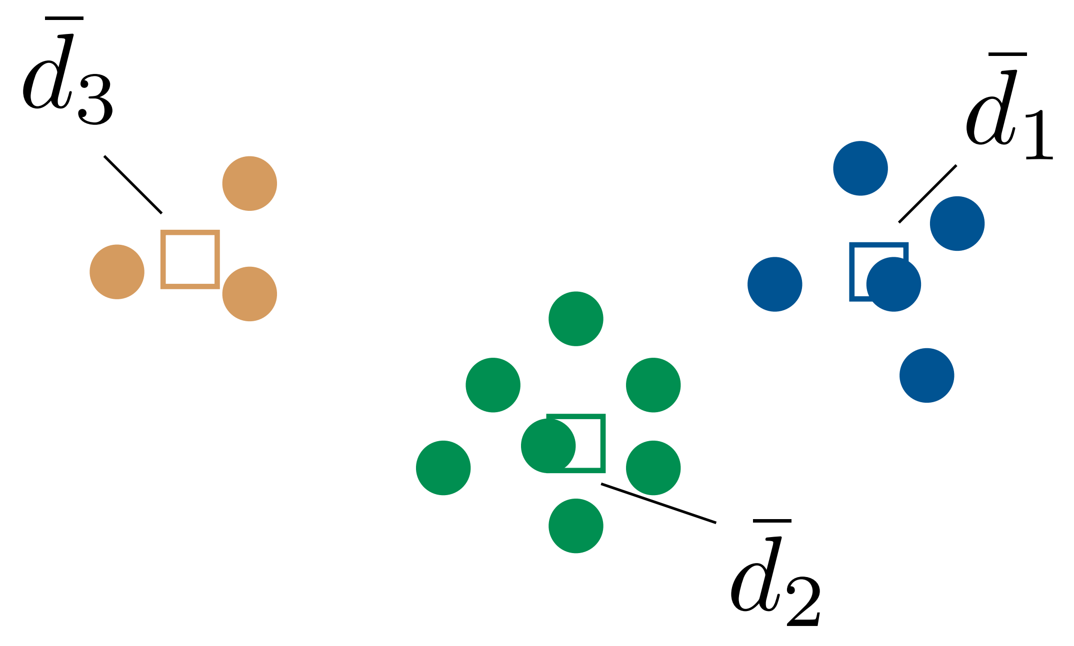
$$\text{minimize} \quad \sum_{k=1}^{K} \sum_{i \in C_k} \|d_i - \bar{d}_k\|^2$$

cluster centers

Mean Robust Optimization (MRO)

Data

$\mathcal{D}_N$ → **Machine Learning Clustering** → Uncertainty Set $\mathcal{U}(K, \epsilon)$ → **Mean Robust Optimization Problem** → Data-driven solution $\hat{x}_N$

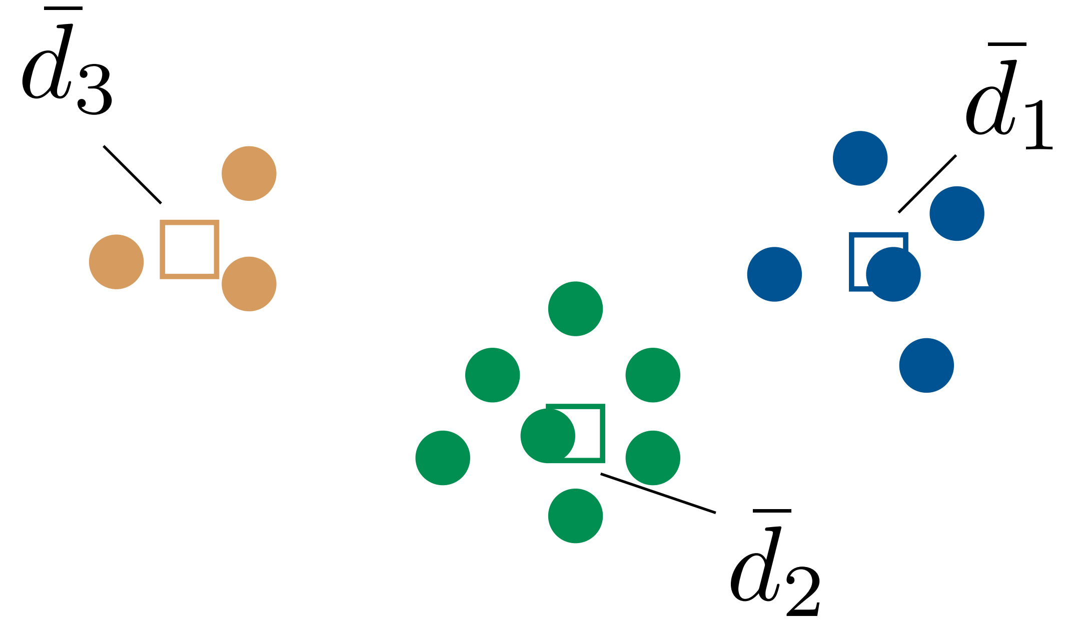


$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \bar{g}(u, x) \leq 0 \quad \forall u \in \mathcal{U}(K, \epsilon) \end{array}$$

constraint function

uncertainty set

Uncertainty set



cluster weights

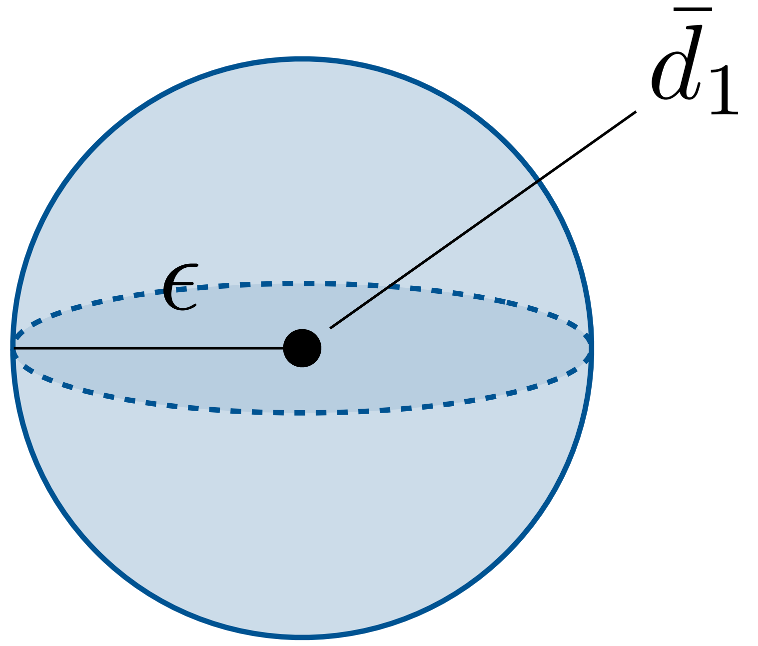
order

$$\mathcal{U}(K, \epsilon) = \left\{ u = (v_1, \dots, v_K) \;\middle|\; \sum_{k=1}^{K} w_k \|v_k - \bar{d}_k\|^p \leq \epsilon^p \right\}$$

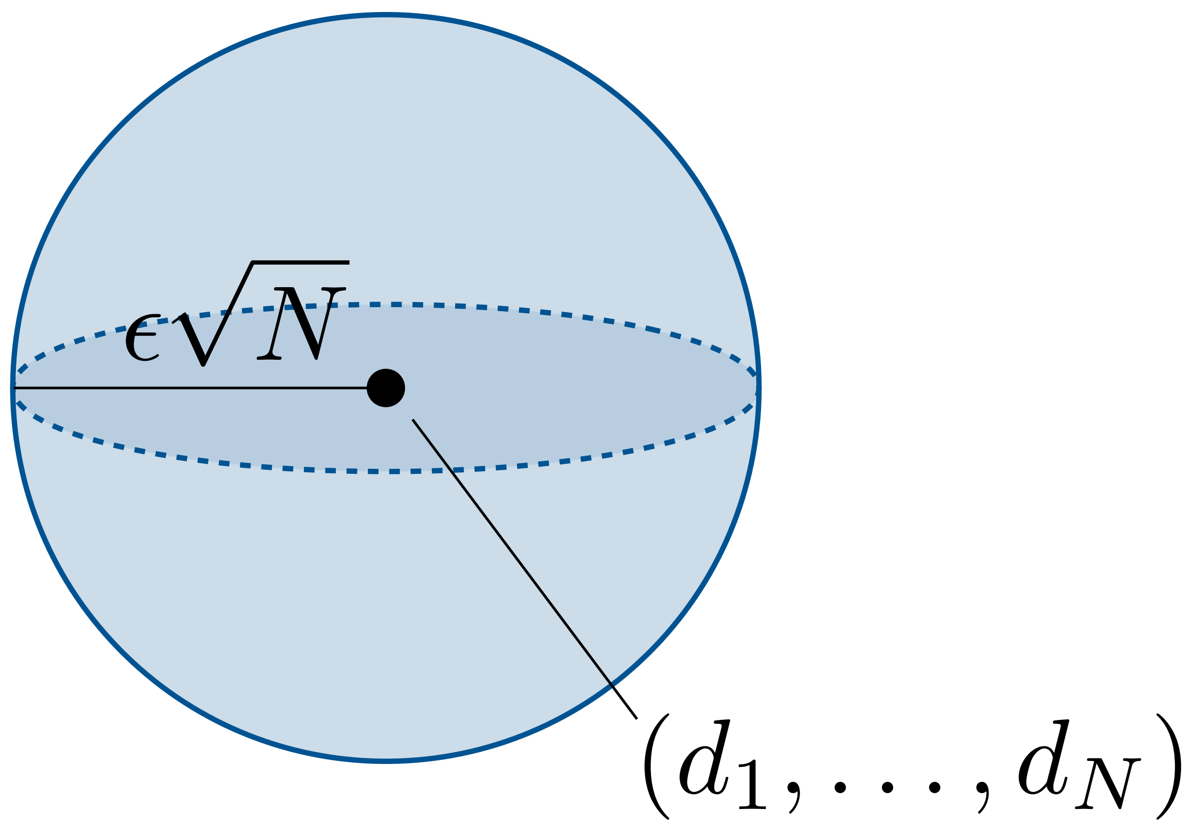
cluster centers

Examples

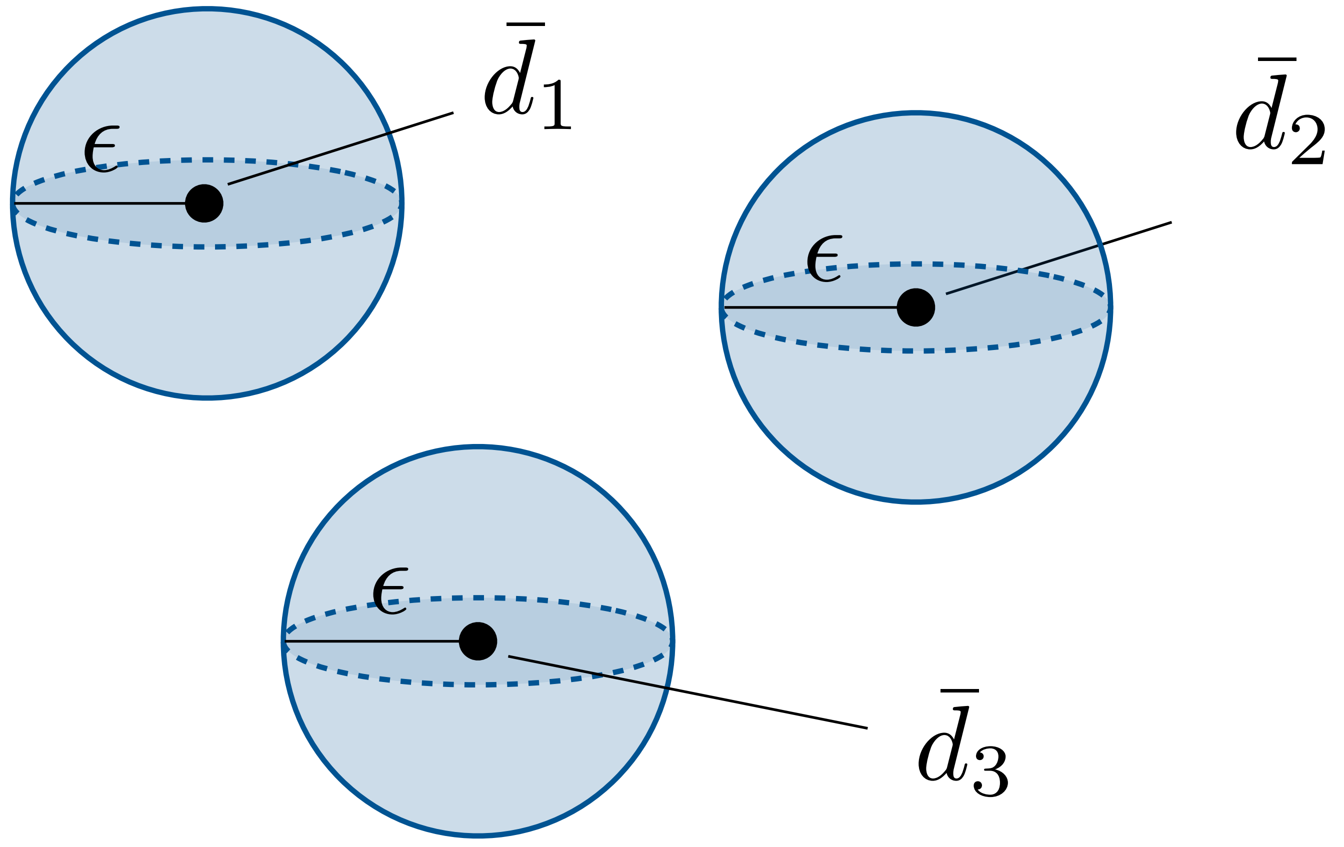
$K = 1$



$K = N, p = 2$



$K = 3, p = \infty$



Mean Robust Optimization Problem

uncertainty set

$$\left\{ \sum_{k=1}^{K} w_k \|v_k - \bar{d}_k\|^p \leq \epsilon^p \right\}$$

Uncertain variable lifting

$u = (v_1, \ldots, v_K)$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & \bar{g}(u, x) \leq 0 \quad \forall u \in \mathcal{U}(K, \epsilon) \end{array}$$

constraint function

$$\sum_{k=1}^{K} w_k g(v_k, x)$$

Solving the MRO problem

Dualize constraint $\bar{g}(u,x) \leq 0, \; \forall u \in \mathcal{U}(K,\epsilon)$

$$
\begin{array}{ll}
\text{minimize} & f(x) \\
\text{subject to} & \sum_{k=1}^K w_k s_k \leq 0 \\
& [-g]^*(z_k, x) - z_k^T \bar{d}_k + \phi(p)\lambda\|z_k/\lambda\|_*^{p/(p-1)} + \lambda\epsilon^p \leq s_k, \qquad k = 1,\dots,K \\
& \lambda \geq 0
\end{array}
$$

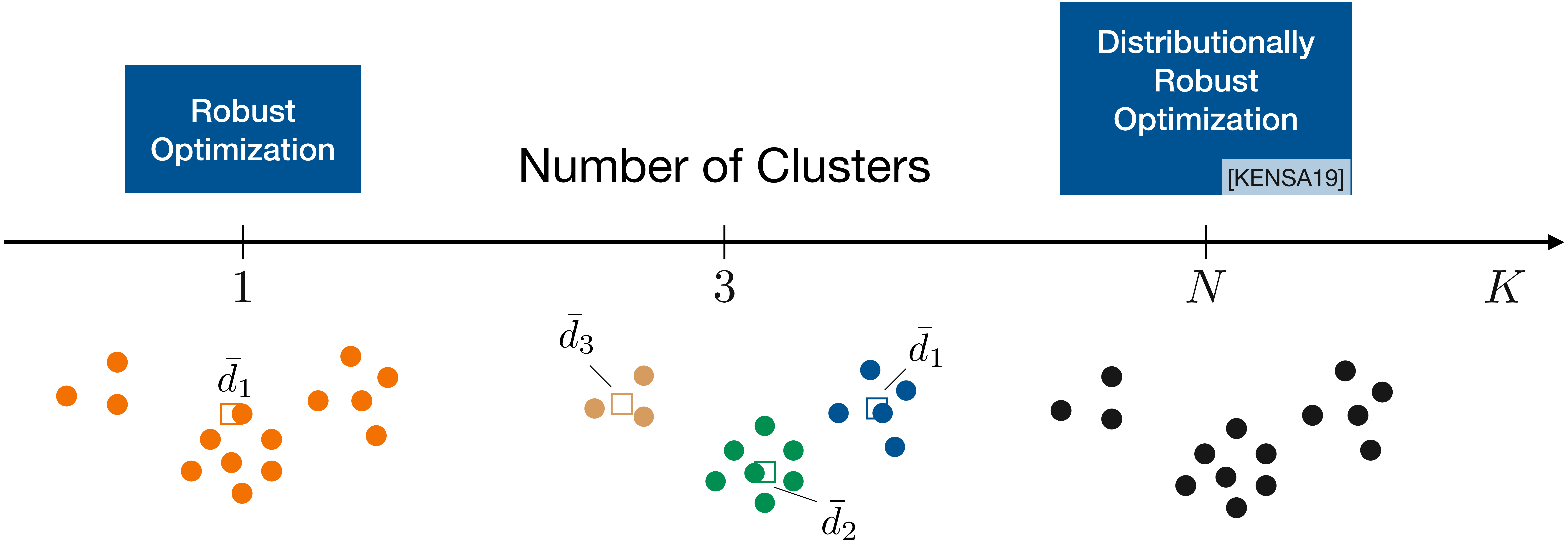
conjugate function

cluster centers

function of $p \geq 1$
$\phi(p) \to 1$ as $p \to \infty$
$\phi(1) = 0$

It can be very expensive when K is large (e.g., $K = N$)

MRO bridges between RO and DRO



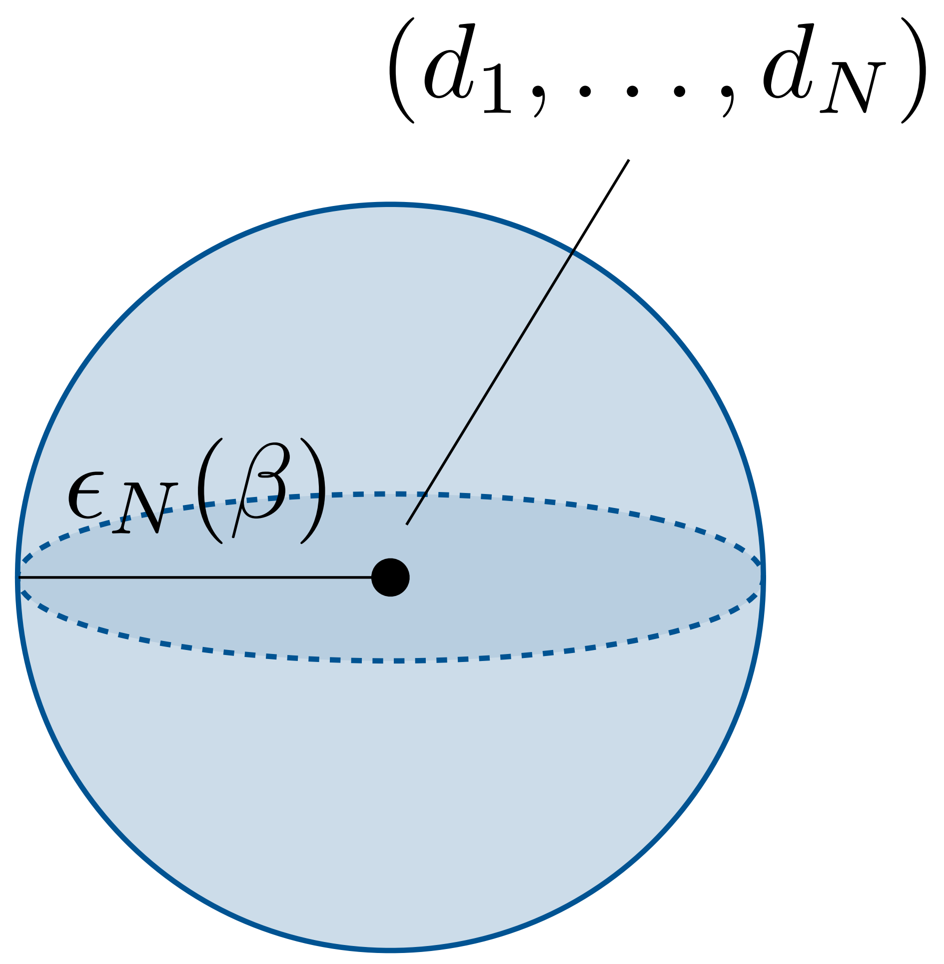
[KENSA19] D Kuhn, P M Esfahani, V A Nguyen, and S Shafieezadeh-Abadeh, "Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning

Satisfying the probabilistic guarantees

probability of constraint satisfaction

$$\mathbf{P}^N\left(\mathbf{E}(g(u,\hat{x}_N)) \leq 0\right) \geq 1-\beta \xrightarrow{\text{light-tailed}} \mathcal{U}(N, \epsilon_N(\beta))$$

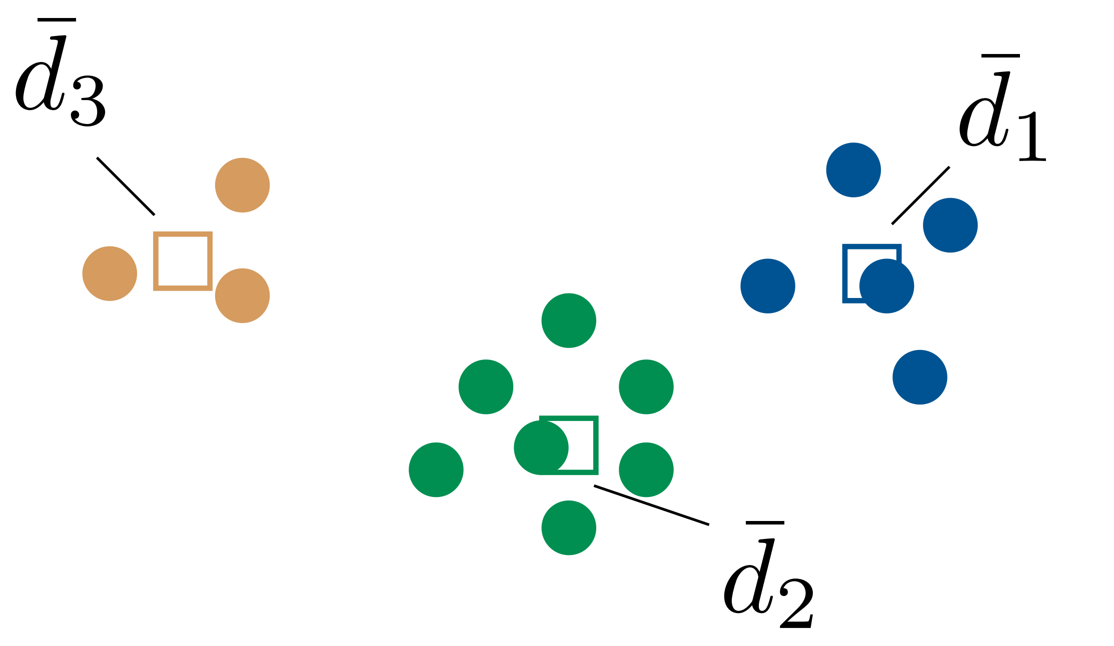
uncertainty set radius



MRO clustering

$$\mathcal{U}(K, \epsilon_N(\beta) + \eta_N(K))$$

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{d_i \in C_k} \|d_i - \bar{d}_k\|^p$$



Quite conservative bounds… can we do better?

Bounding the conservatism

Worst-case values

$$\bar{g}^{N}(x) = \underset{u \in \mathcal{U}(N,\epsilon)}{\text{maximize}} \quad \bar{g}(u,x)$$

$$\bar{g}^{K}(x) = \underset{u \in \mathcal{U}(K,\epsilon)}{\text{maximize}} \quad \bar{g}(u,x)$$

MRO constraint

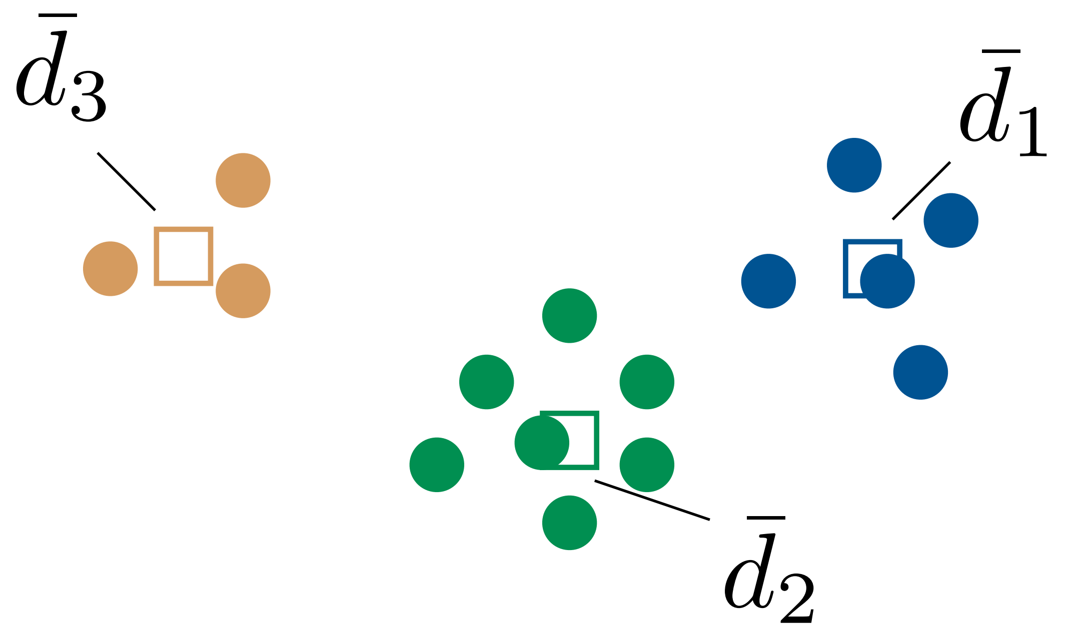
$$\bar{g}(u,x) \le 0 \quad \forall u \in \mathcal{U}(K,\epsilon)$$

Theorem

If $-g$ is L-smooth in u, we have

$$\bar{g}^{N}(x) \le \bar{g}^{K}(x) \le \bar{g}^{N}(x) + \frac{L}{2} D(K) \longleftarrow \min \; \frac{1}{N} \sum_{k=1}^{K} \sum_{d_i \in C_k} \|d_i - \bar{d}_k\|^2$$

clustering objective



When g is affine in u ($L = 0$), clustering makes no difference to the optimal value or optimal solution

Example MRO with linear constraints

$p = \infty$

$$(a + Pu)^T x \leq b \longrightarrow g(u, x) = (a + Pu)^T x - b$$

$$[-g]^*(z_k, x) = \sup_u z_k^T u - (a + Pu)^T x + b = \begin{cases} a^T x - b & \text{if } \boxed{z_k + P^T x = 0} \\ \infty & \text{otherwise} \end{cases}$$

$$z_1 = z_2 = \cdots = z_k = -P^T x$$

Convex reformulation

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & a^T x - b + (P^T x)^T \boxed{\sum_{k=1}^K w_k \bar{d}_k} + \epsilon \|P^T x\|_* \leq 0 \end{array}$$

(clustering makes no difference)

average $\bar{d}$

Classical Robust Optimization reformulation

Capital budgeting example

Problem

total
net present value (NPV)

$$\begin{array}{ll} \text{maximize} & \eta(u)^T x \\ \text{subject to} & a^T x \leq b \\ & x \in \{0,1\}^n \end{array}$$

budget
constraint

$$g(u, x, \tau) = -\eta(u)^T x - \tau$$

MRO formulation

$$\begin{array}{ll} \text{minimize} & \tau \\ \text{subject to} & \bar{g}(u, x, \tau) \leq 0, \quad \forall u \in \mathcal{U}(K, \epsilon) \\ & a^T x \leq b \\ & x \in \{0,1\}^n \end{array}$$

NPV of project j

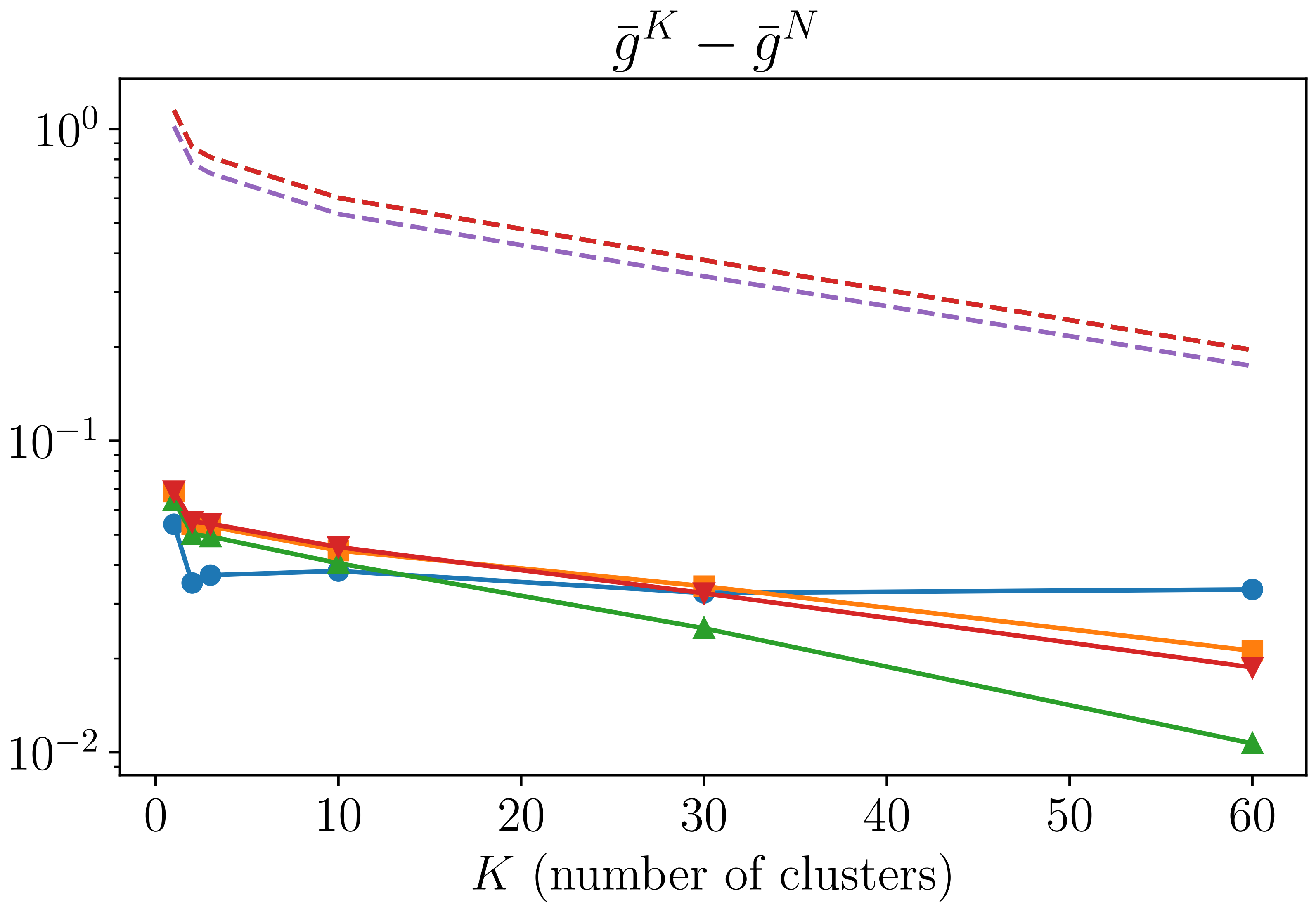
$$\eta_j(u) = \sum_{t=1}^{T} \frac{F_{jt}}{(1 + u_j)^t}$$

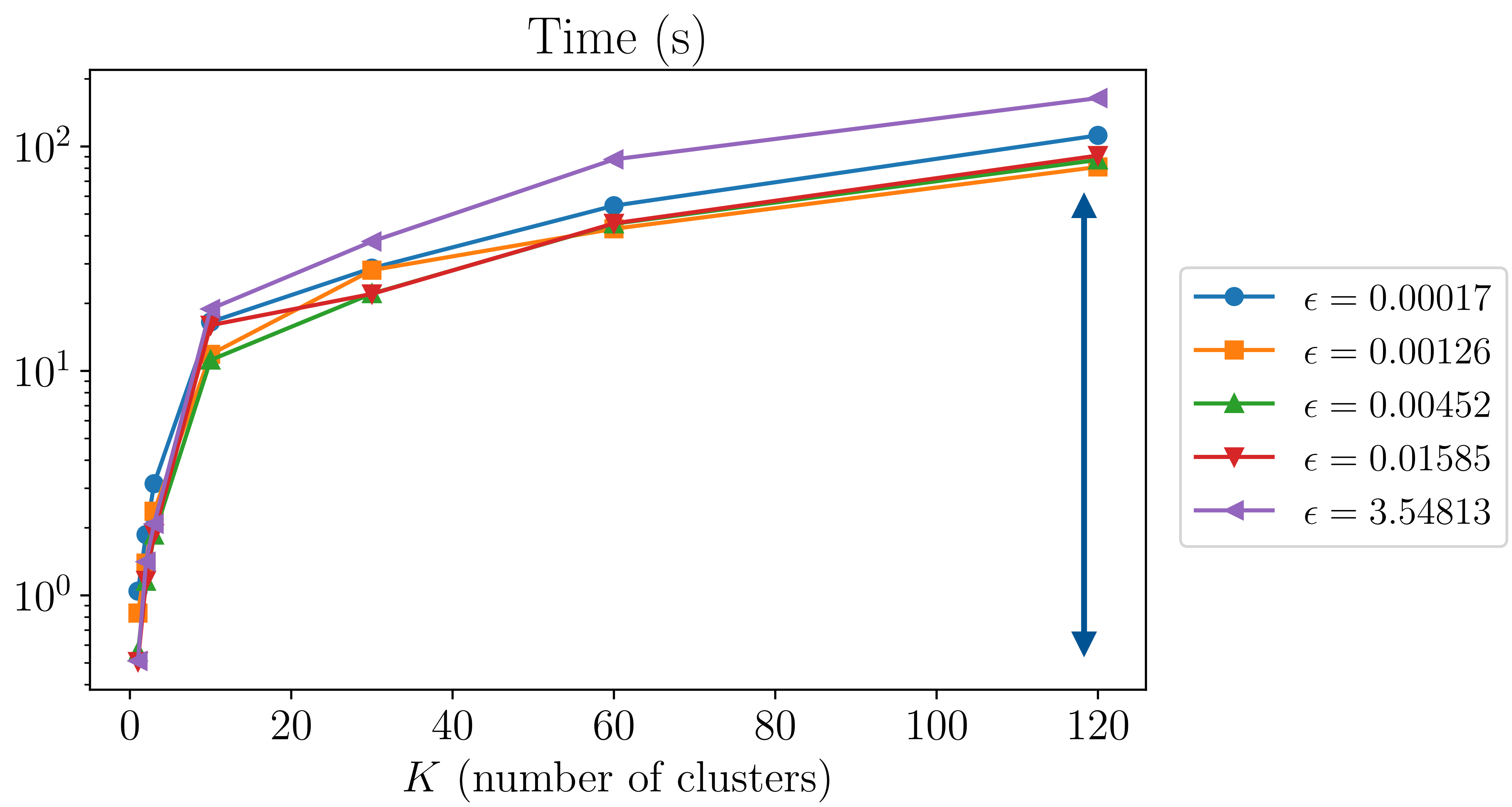
cash
flow

discount
rate

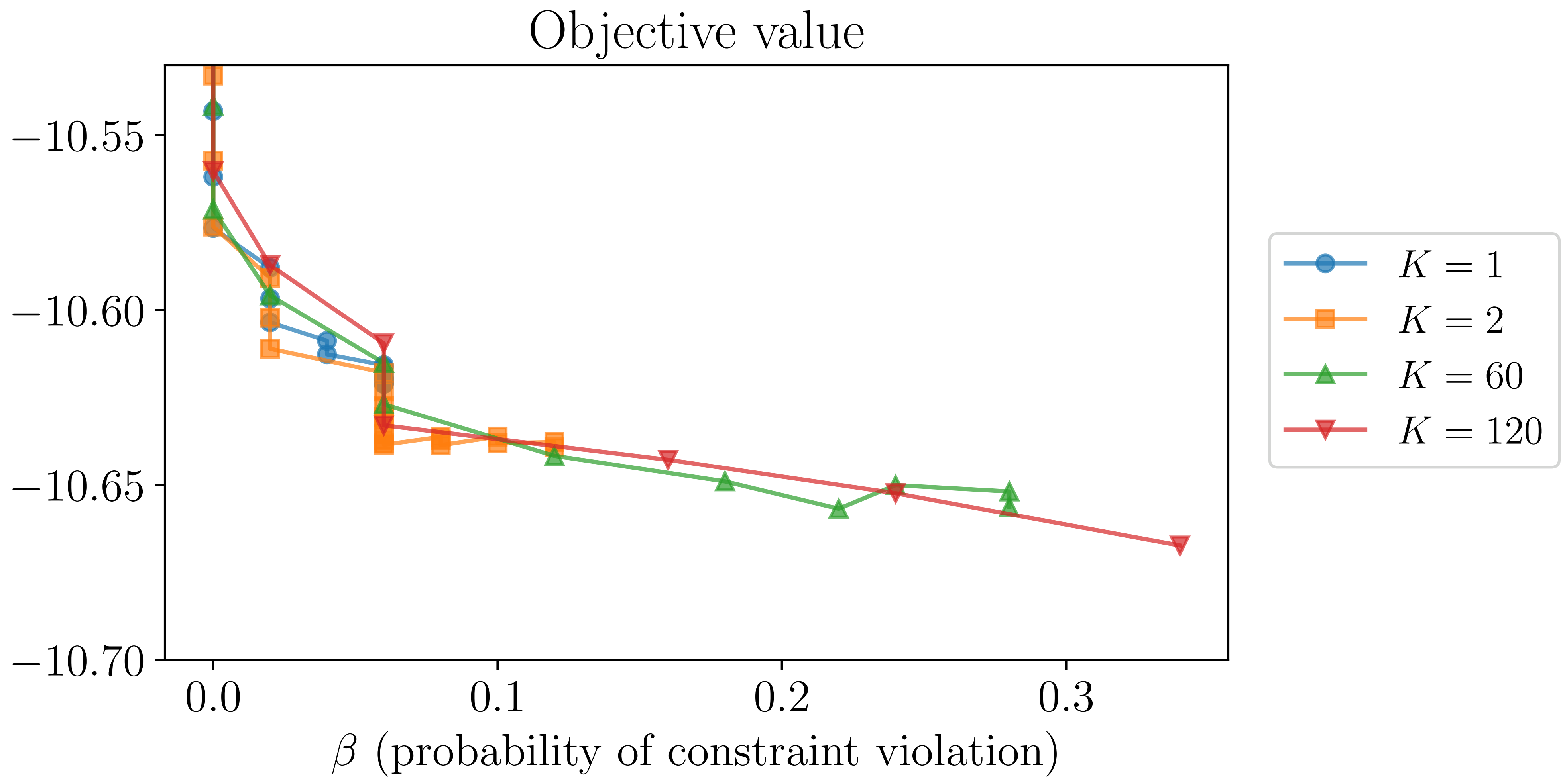
Capital budgeting results

$n = 20, N = 120, T = 5$





100x speedups!



2 clusters give near-optimal performance

Mean Robust Optimization

- Bridge between RO and DRO

- Clustering effect
  - g affine in u → zero clustering effect!
  - g concave in u → performance bound

- Multiple orders of magnitude speedups

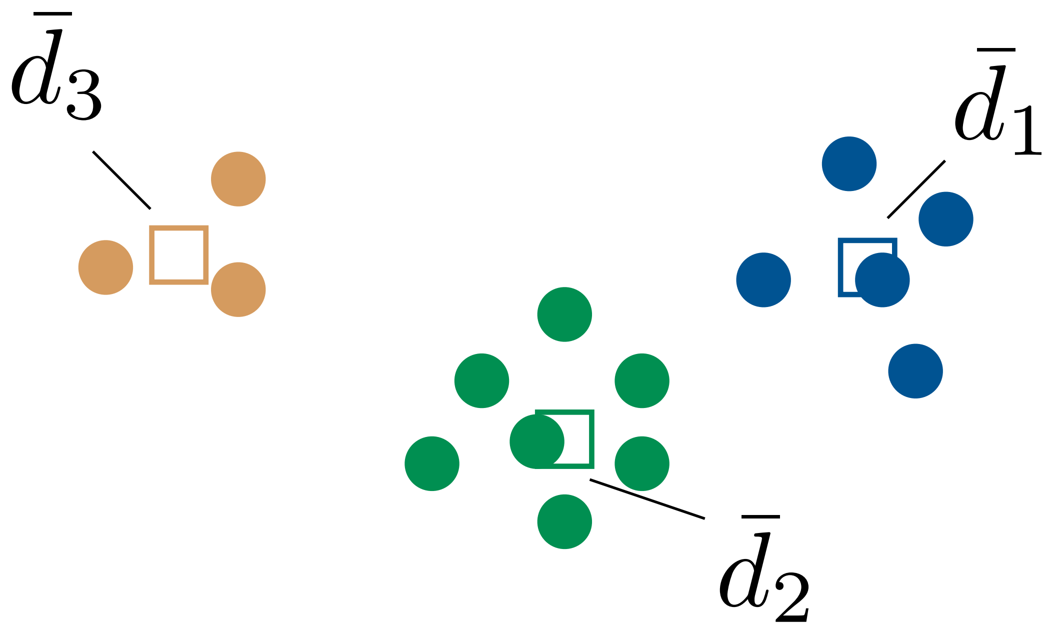
https://github.com/stellatogrp/mro_experiments

https://arxiv.org/abs/2207.10820 ← INFORMS Computing Society Student Paper Award Winner

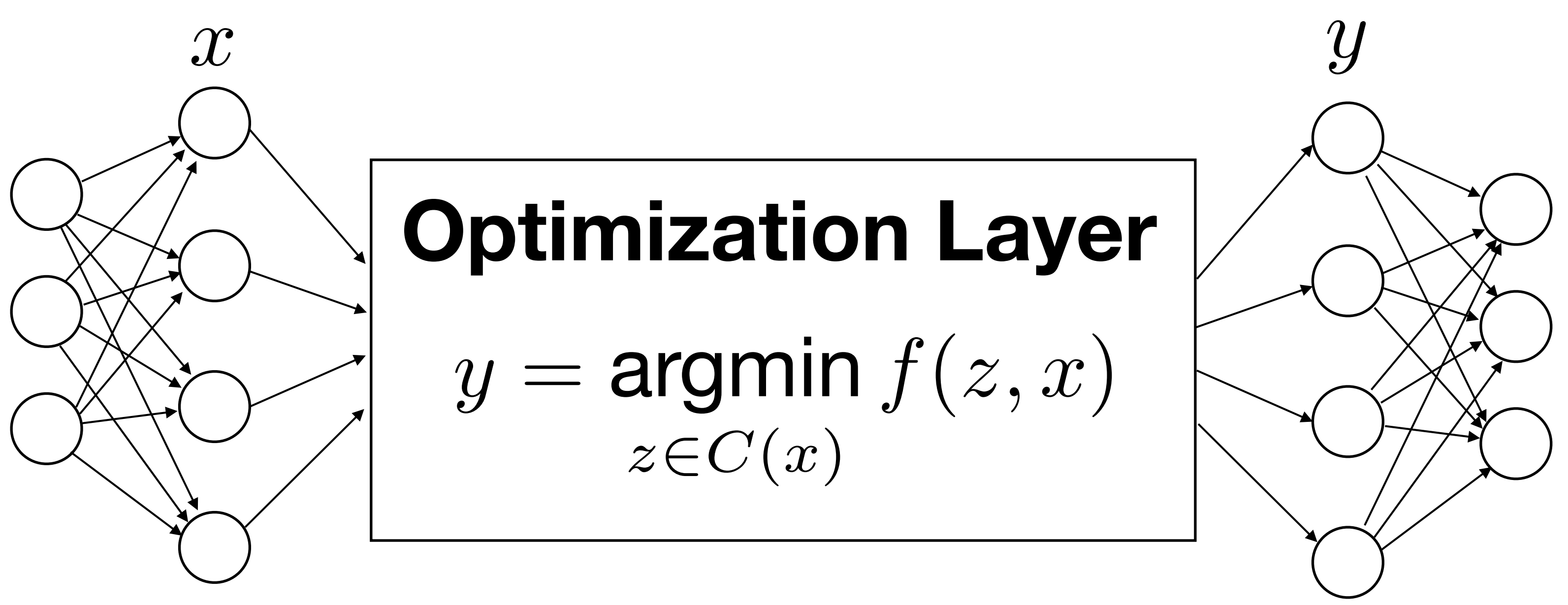
Machine learning to make better decisions under uncertainty

How can we use data to build
tractable and **high-performance**
uncertainty sets?

Clustering



Differentiable Optimization



Simple newsvendor problem

“A newsboy selling math papers in UCLA,” Dall-E

Objective

order price

sell price

$$k^T y - p^T \min(y, u)$$

uncertain demand

$\log u \sim \mathcal{N}(\mu, \Sigma)$

orders

Robust Reformulation

$$\begin{array}{ll} \text{minimize} & \tau \\ \text{subject to} & k^T y + \max\{-p^T y, -p^T u\} \leq \tau, \quad \forall u \in \mathcal{U}(\epsilon) \\ & y \geq 0 \end{array}$$

constraint

$g(x, u) \leq 0$

with $x = (y, \tau)$

how do we pick the uncertainty set?

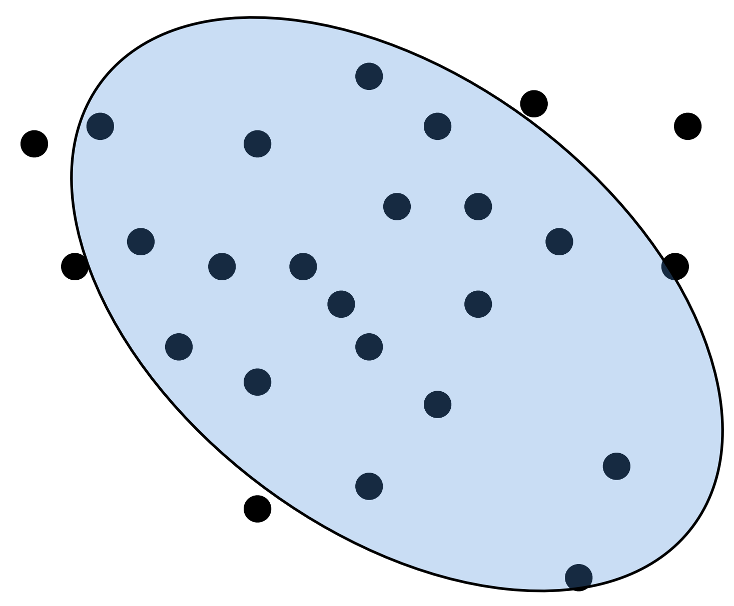
Uncertainty sets for newsvendor

Standard uncertainty set

$$\mathcal{U}^{\mathrm{st}}(\epsilon) = \{\hat{\mu} + \hat{\Sigma}^{1/2} u \mid \|u\|_2 \leq \epsilon\} = \{\|A^{\mathrm{st}} u + b^{\mathrm{st}}\|_2 \leq \epsilon\}$$

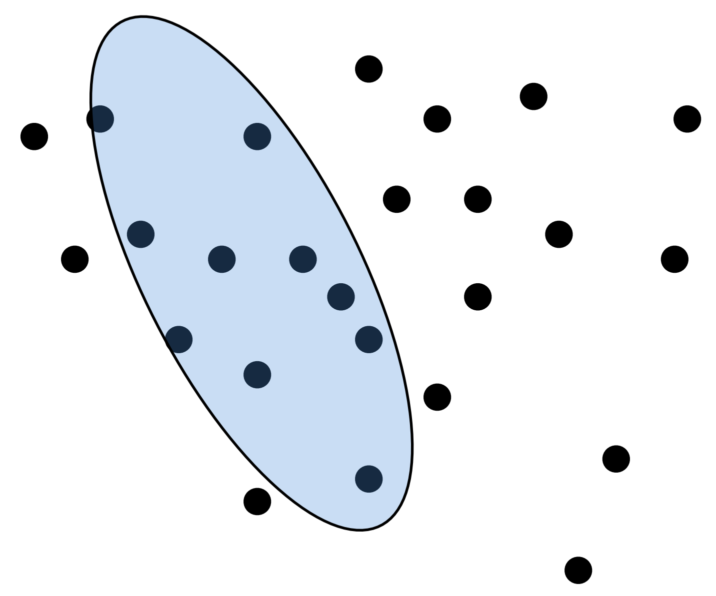
empirical
mean and covariance

$\hat{\Sigma}^{-1/2}$ $\quad$ $\hat{\Sigma}^{-1/2}\hat{\mu}$



Reshaped uncertainty set

$$\mathcal{U}^{\mathrm{re}}(\epsilon) = \{\|A^{\mathrm{re}} u + b^{\mathrm{re}}\|_2 \leq \epsilon\}$$



How do we pick the shape and size?

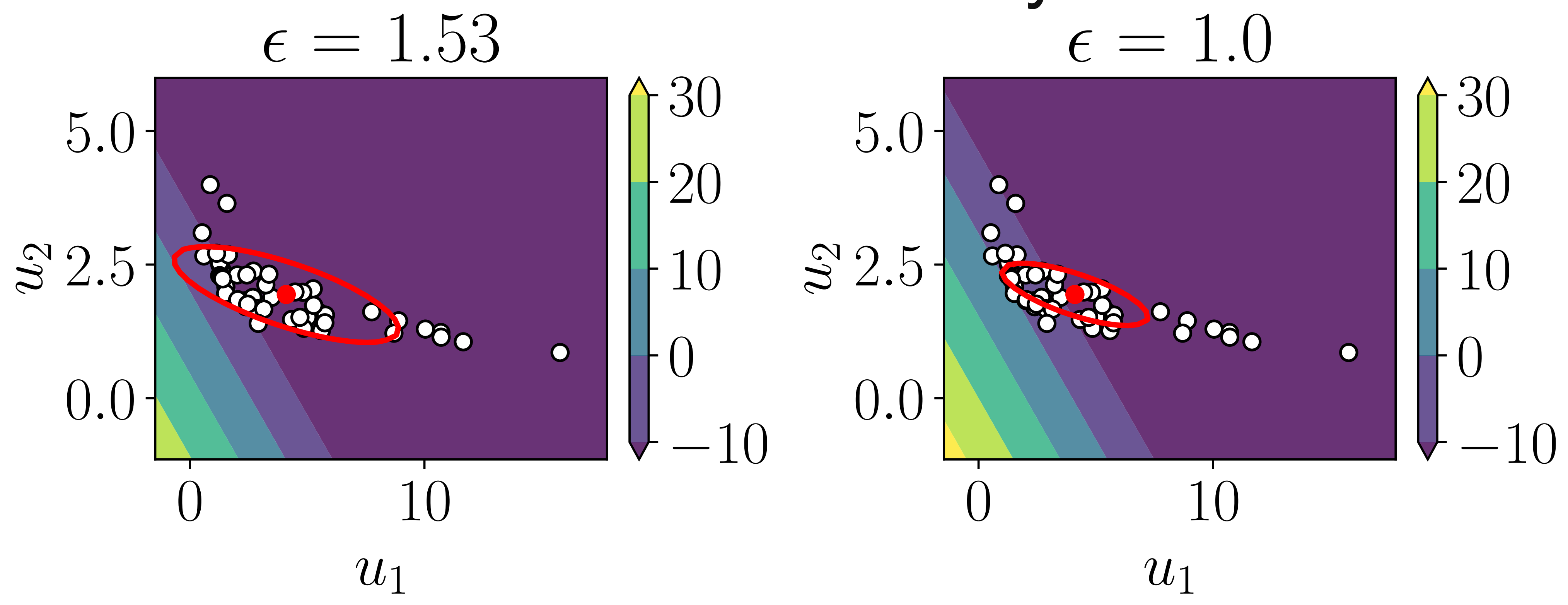
Can the reshaped set be better?

Classic newsvendor

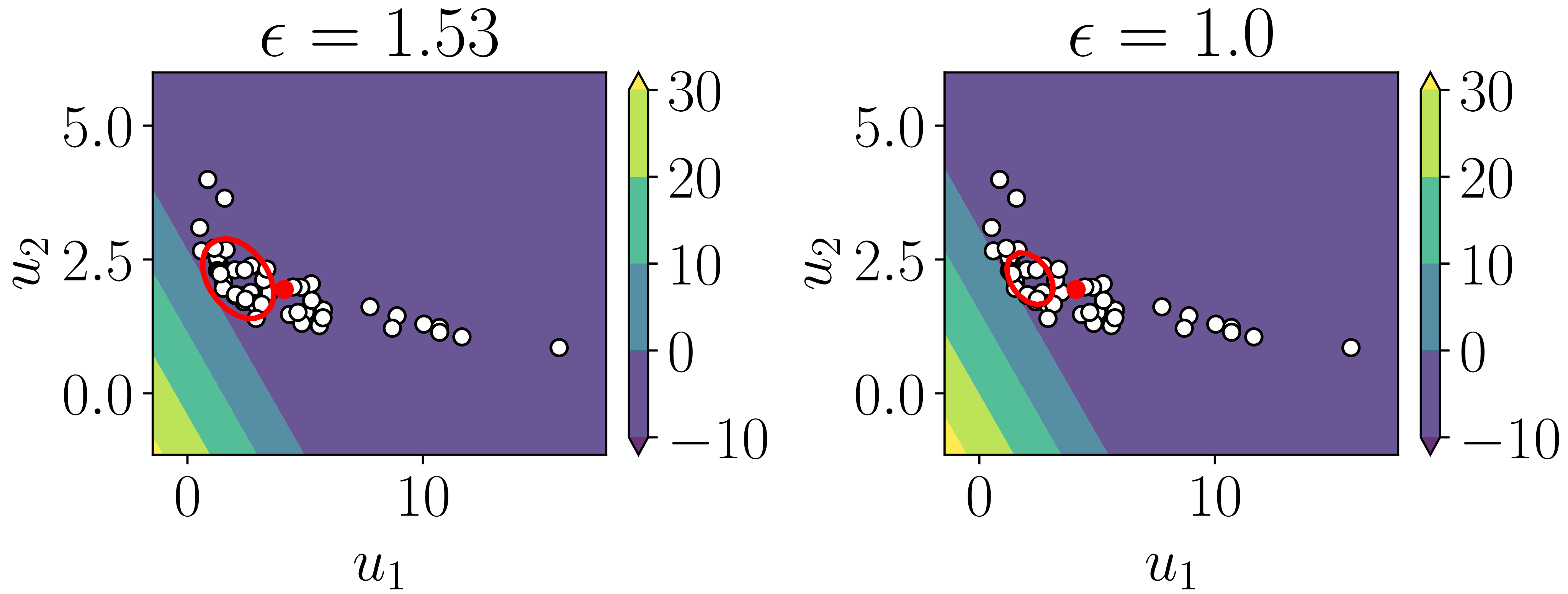
$$\epsilon \Rightarrow \mathcal{U}(\epsilon) \Rightarrow g(u, x^\star(\epsilon))$$

$\downarrow$ Level sets

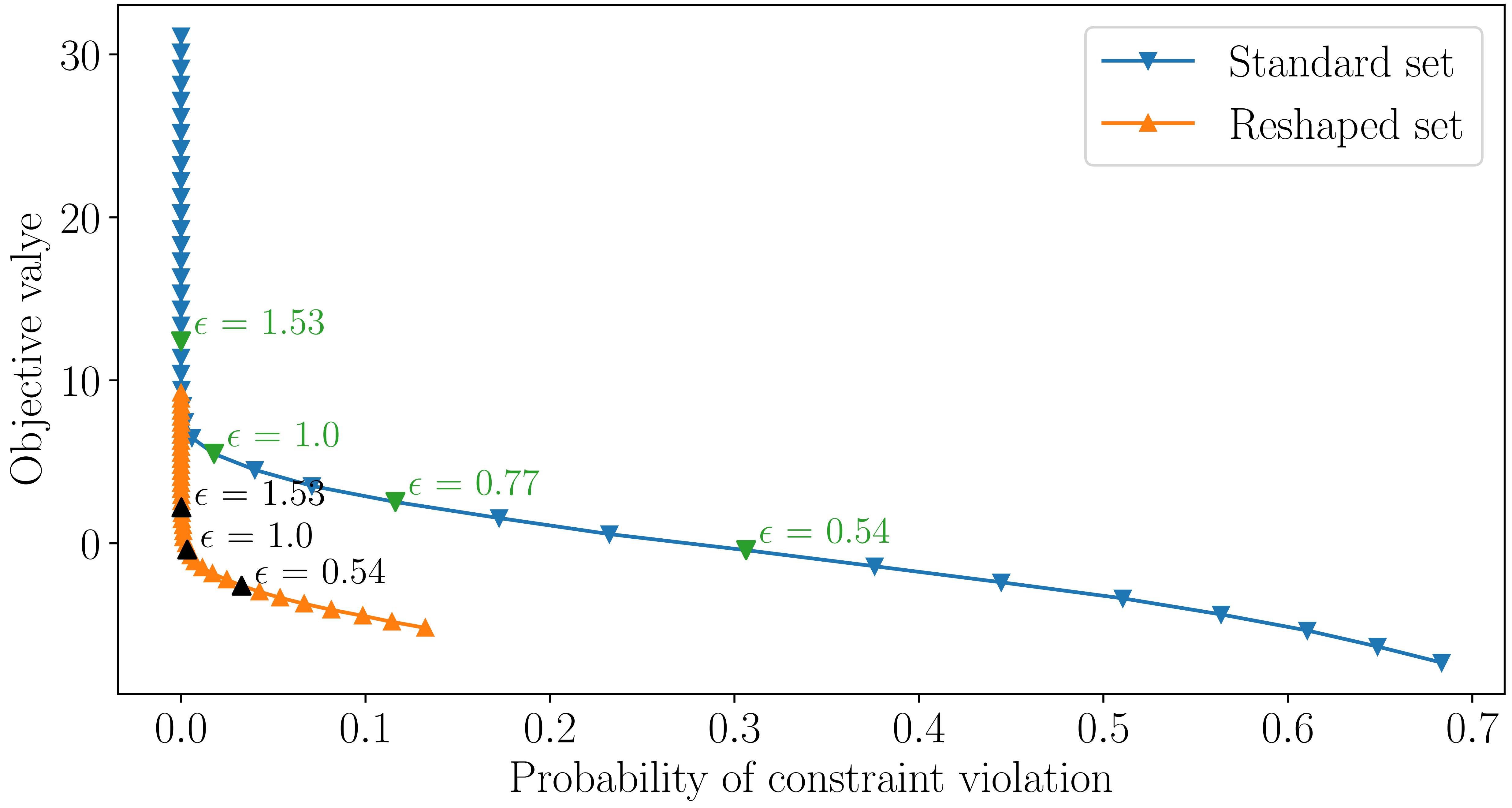
Standard uncertainty set $\mathcal{U}^{\mathrm{st}}$



Reshaped uncertainty set $\mathcal{U}^{\mathrm{re}}$



Grid search over size



Reshaped set does not capture empirical distribution…

…but it performs much better!

Parametric uncertainty sets

$$
\begin{array}{ll}
\text{minimize} & f(x) \\
\text{subject to} & g(u,x) \leq 0, \quad \forall u \in \mathcal{U}(\theta)
\end{array}
$$

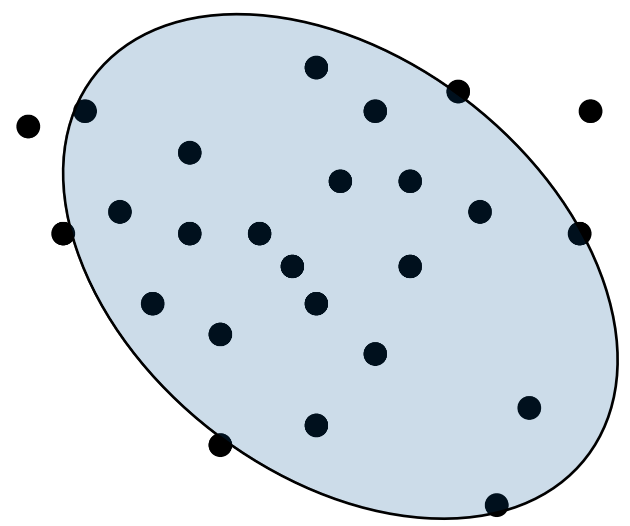
key parameters

max of concave

Example
Ellipsoidal set

$$\mathcal{U}_{\text{ellip}}(\theta) = \{\|Au + b\|_2 \leq 1\}$$

$$\theta = (A, b)$$



Linear constraint

$$g(u,x) = u^T x \leq 0, \quad \forall u \in \mathcal{U}_{\text{ellip}}(\theta) \longrightarrow$$

Robust counterpart

$$
\begin{aligned}
-b^T y + \|y\|_2 &\leq 0 \\
A^T y &= x
\end{aligned}
$$

Probabilistic guarantees

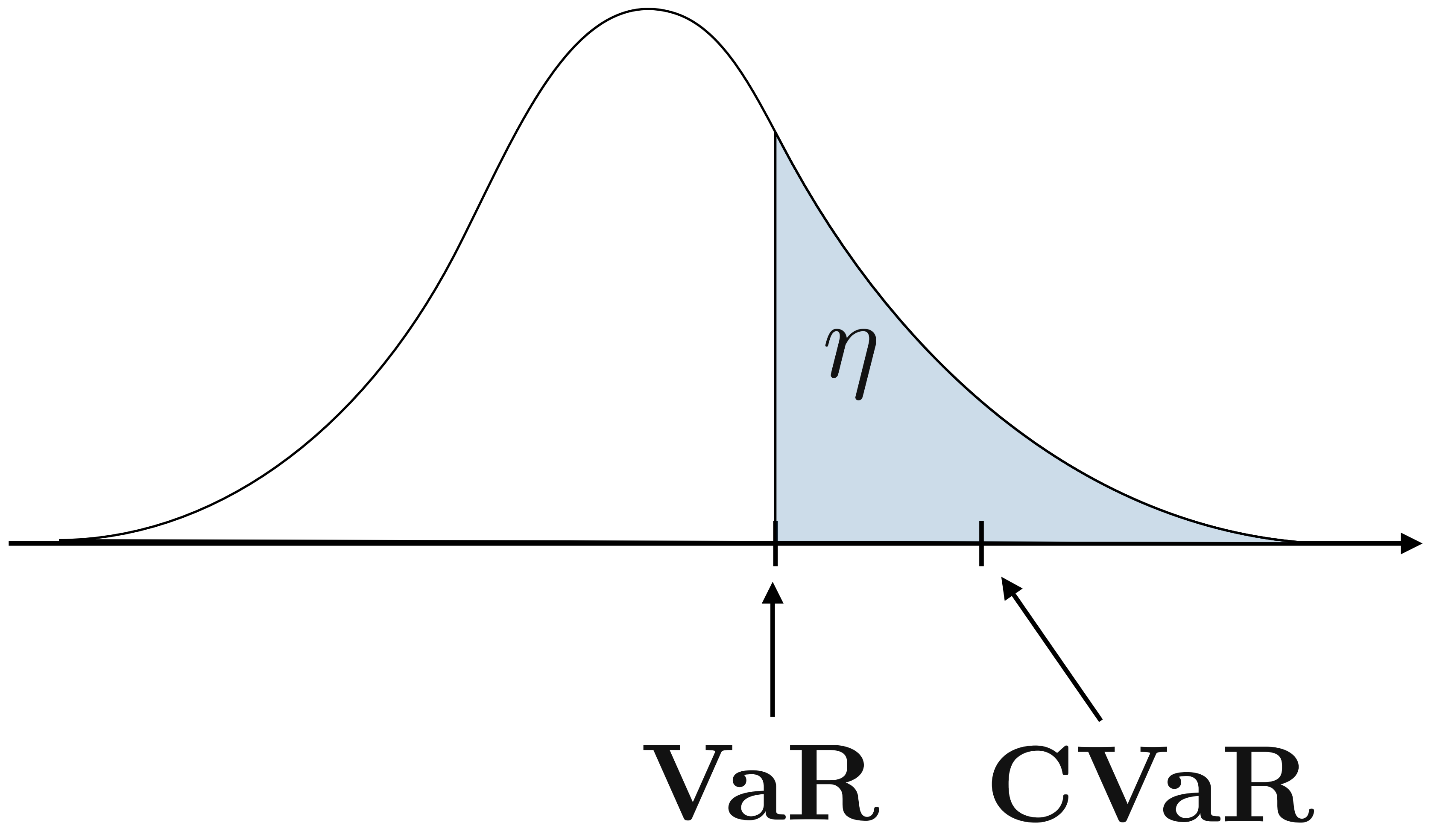
$$\mathbf{P}(g(u,\hat{x}(\theta)) \leq 0) \geq 1-\eta$$

Value-at-Risk

$$\mathbf{VaR}(g(u,\hat{x}(\theta)),\eta) = \inf\{\gamma \mid \mathbf{P}(g(u,\hat{x}(\theta)) \leq \gamma) \geq 1-\eta\}$$

Conditional Value-at-Risk

$$\mathbf{CVaR}(g(u,\hat{x}(\theta)),\eta) = \inf_{\alpha}\{\mathbf{E}((1/\eta)(g(u,\hat{x}(\theta))-\alpha)_{+})+\alpha\}$$



Work with CVaR

$$\boxed{\mathbf{CVaR}(g(u,x),\eta) \leq 0} \quad \Longrightarrow \quad \mathbf{VaR}(g(u,x),\eta) \leq 0 \quad \Longrightarrow \quad \mathbf{P}(g(u,x) \leq 0) \geq 1-\eta$$

From CVaR to data (heuristic)

$$\mathbf{CVaR}(g(u,x),\eta) \leq 0$$

Turn into constraint

$$\mathbf{E}\left(\frac{g(u,\hat{x}(\theta)) - \alpha)_+}{\eta} + \alpha\right) \leq 0$$

Data-driven approximation

constraint satisfaction condition

$$h(\hat{x}(\theta),\alpha,\eta) \longrightarrow \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{(g(d_i,\hat{x}(\theta)) - \alpha)_+}{\eta} + \alpha\right\} \leq 0$$

training data

Our approach to learn the uncertainty sets

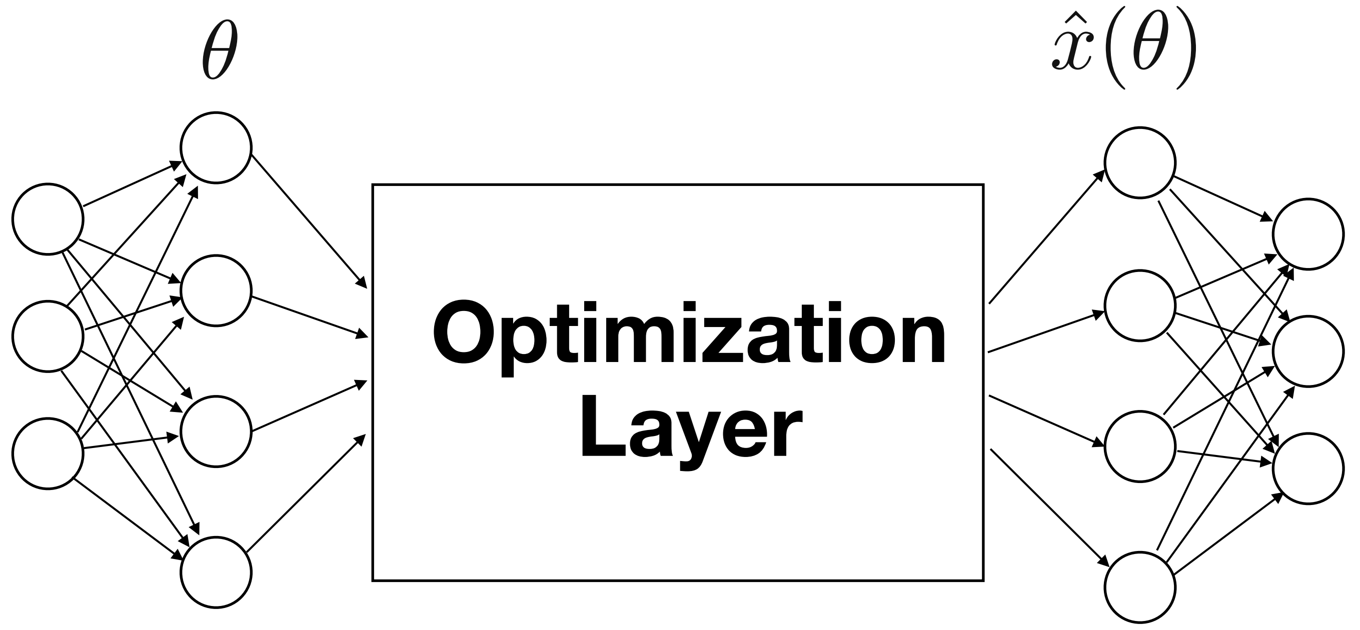
Training problem

$$
\begin{array}{ll}
\text{minimize} & f(\hat{x}(\theta)) \\
\text{subject to} & h(\hat{x}(\theta), \eta, \alpha) = 0 \\
 & \hat{x}(\theta) \in \Phi(\theta)
\end{array}
$$

tight at optimality

Robust problem

$$
\Phi(\theta) = \begin{array}{ll}
\text{argmin} & f(x) \\
\text{subject to} & g(u, x) \leq 0 \quad \forall u \in \mathcal{U}(\theta)
\end{array}
$$



Stochastic Augmented Lagrangian Method

Augmented Lagrangian

$$L(\hat{x}(\theta), \alpha, \lambda, \mu) = f(\hat{x}(\theta)) + \lambda h(\hat{x}(\theta), \eta, \alpha) + (\mu/2)\boxed{h(\hat{x}(\theta), \eta, \alpha)^2}$$

quadratic penalty

for $k = 1, \ldots, k_{\max}$ **do**

$\nabla_\theta L, \nabla_\alpha L \leftarrow$ **implicit differentiation** $\xleftarrow{D\hat{x}(\theta)}$ Differentiate inner problem KKT optimality conditions

$\theta^{k+1} \leftarrow \theta^k - r\nabla_\theta L(x^k, \alpha^k, \lambda^k, \mu^k)$

$x^{k+1} \leftarrow$ solve **inner robust problem** given θ^{k+1} $\leftarrow$ primal variables update

$\alpha^{k+1} \leftarrow \alpha^k - r\nabla_\alpha L(x^k, \alpha^k, \lambda^k, \mu^k)$

$\lambda^{k+1} \leftarrow \lambda^k + \mu^k h(x^{k+1}, \mu^k, \alpha^{k+1})$

Choose $\mu^{k+1} \geq \mu^k$

"Differentiable Convex Optimization Layers", A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and J. Zico Kolter (2019)
"Differentiating Through a Conic Program", A. Agrawal, S. Barratt, S. Boyd, E. Busseti, W. M. Moursi (2019)

Back to the newsvendor example

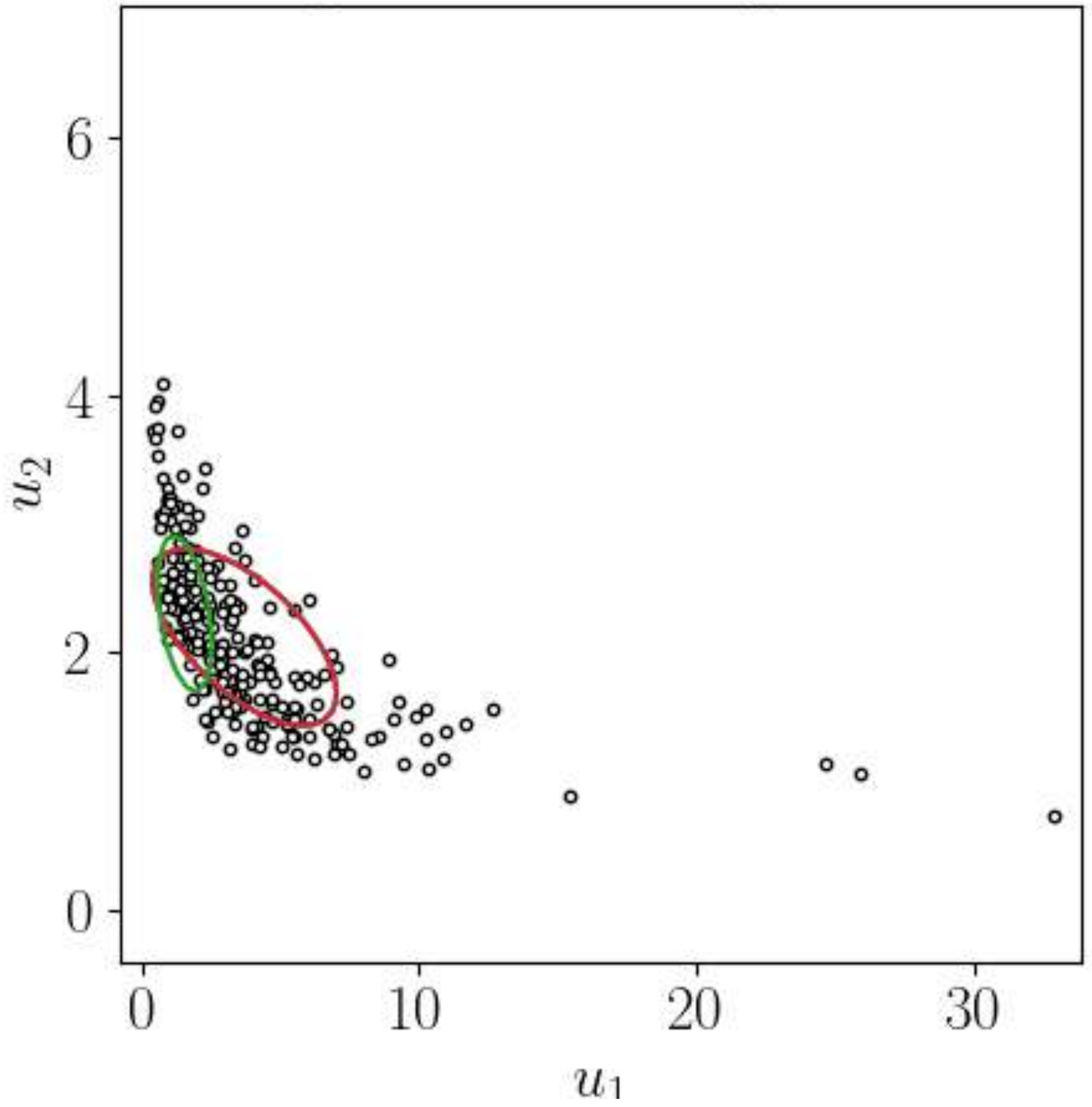
"A newsboy selling math papers in UCLA," Dall-E

Robust Reformulation

minimize τ

subject to $k^T y + \max\{-p^T y, -p^T u\} \leq \tau, \quad \forall u \in \mathcal{U}(\theta)$

$y \geq 0$



Inventory Management

transportation and holding costs

stocking decisions

sales prices

sales decisions

$$
\begin{array}{ll}
\text{minimize} & \tau \\
\text{subject to} & (t+h)^T s - r^T y(u) \le \tau, \quad \forall u \in \mathcal{U}(\theta) \\
& y(u) \le s, \quad \forall u \in \mathcal{U}(\theta) \\
& y(u) \le \bar{d} + Qu, \quad \forall u \in \mathcal{U}(\theta) \\
& \mathbf{1}^T s = C \\
& 0 \le s \le c
\end{array}
$$

demand

Two-stage adjustable optimization

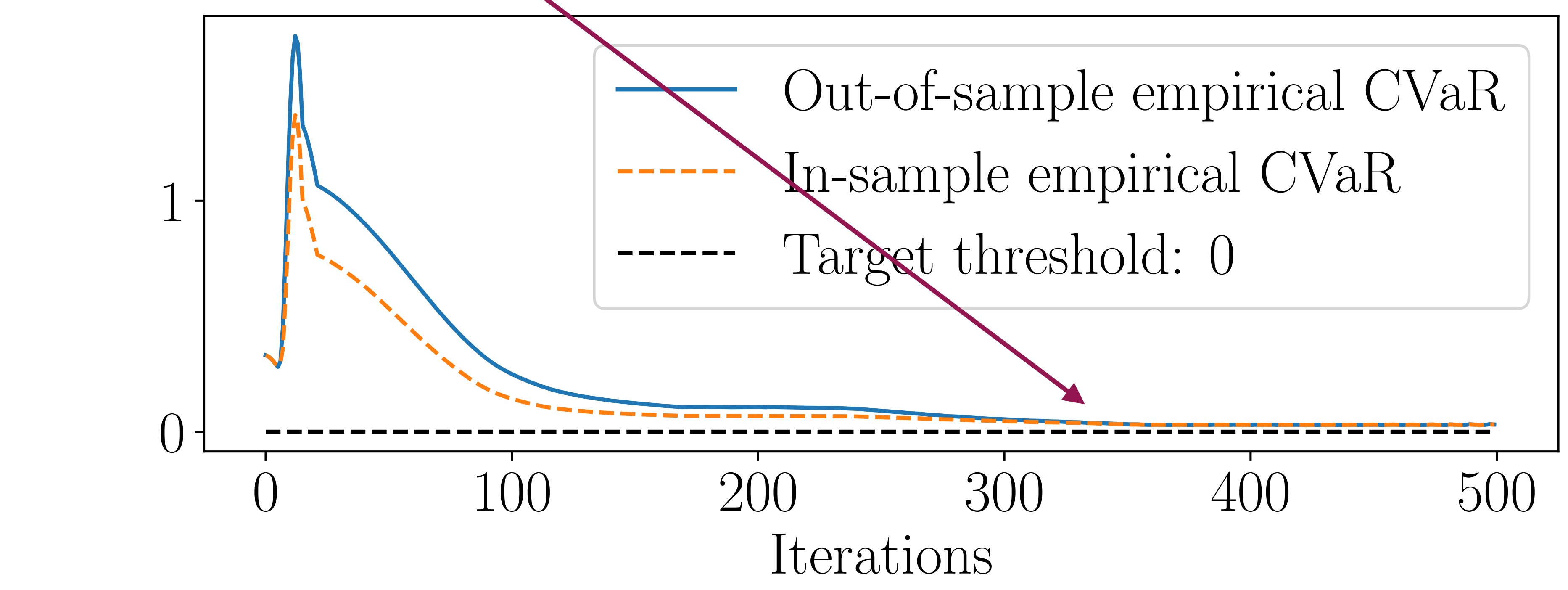
(linear decision rules for $y(u)$)

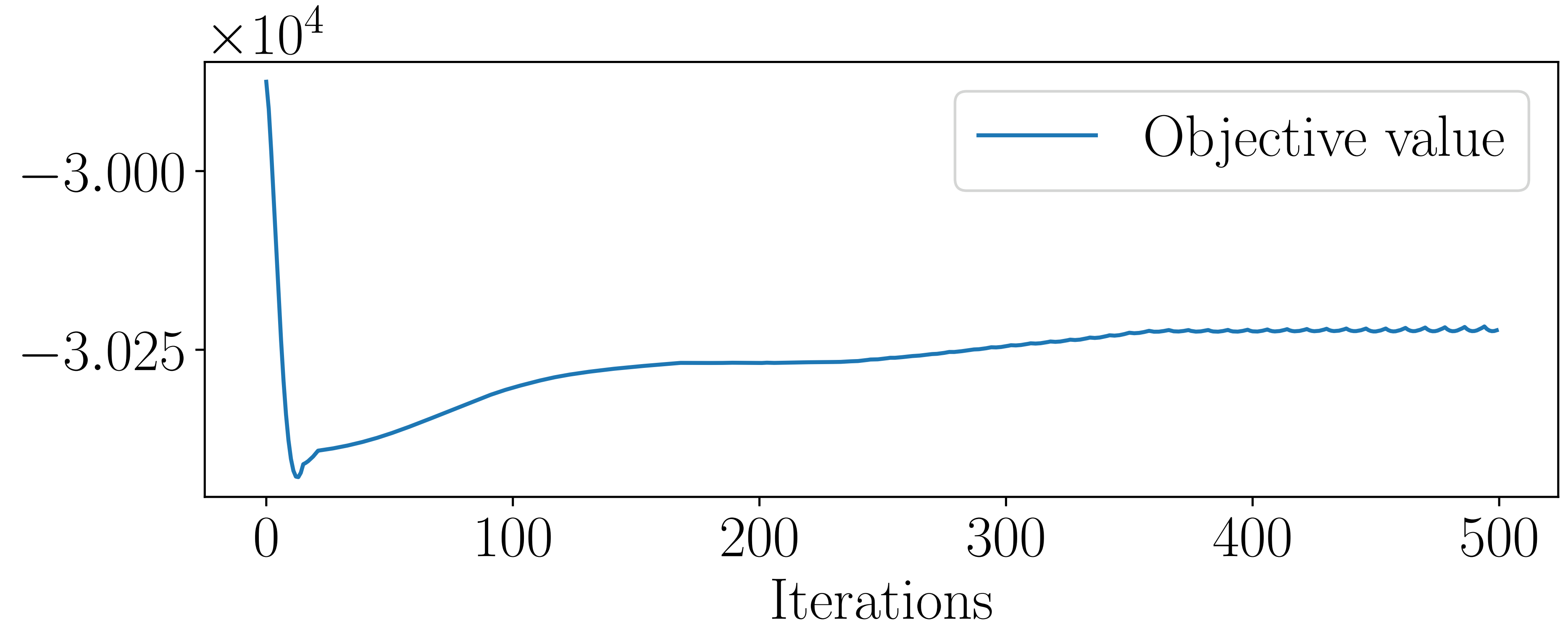
"Pareto Efficiency in Robust Optimization" D. A. Iancu, N. Trichakis (2014)

Inventory management results

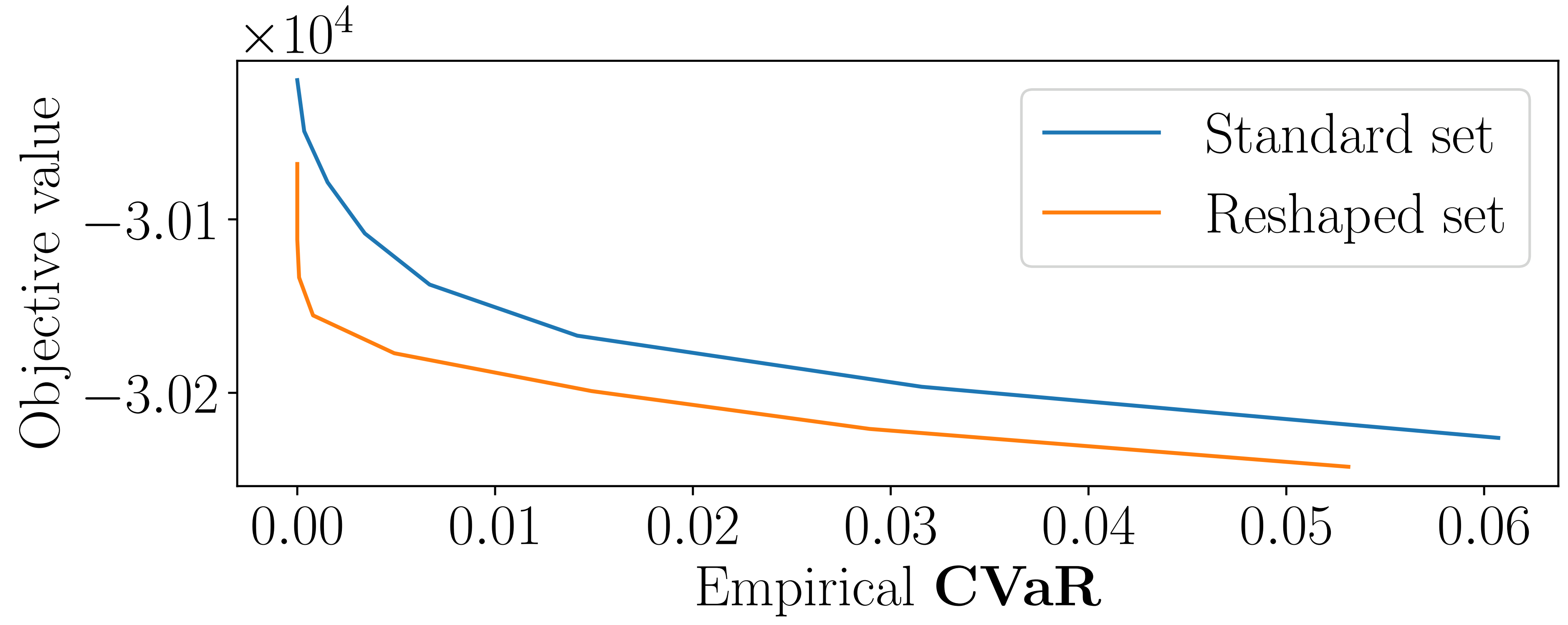
CVaR reaches target level

Training

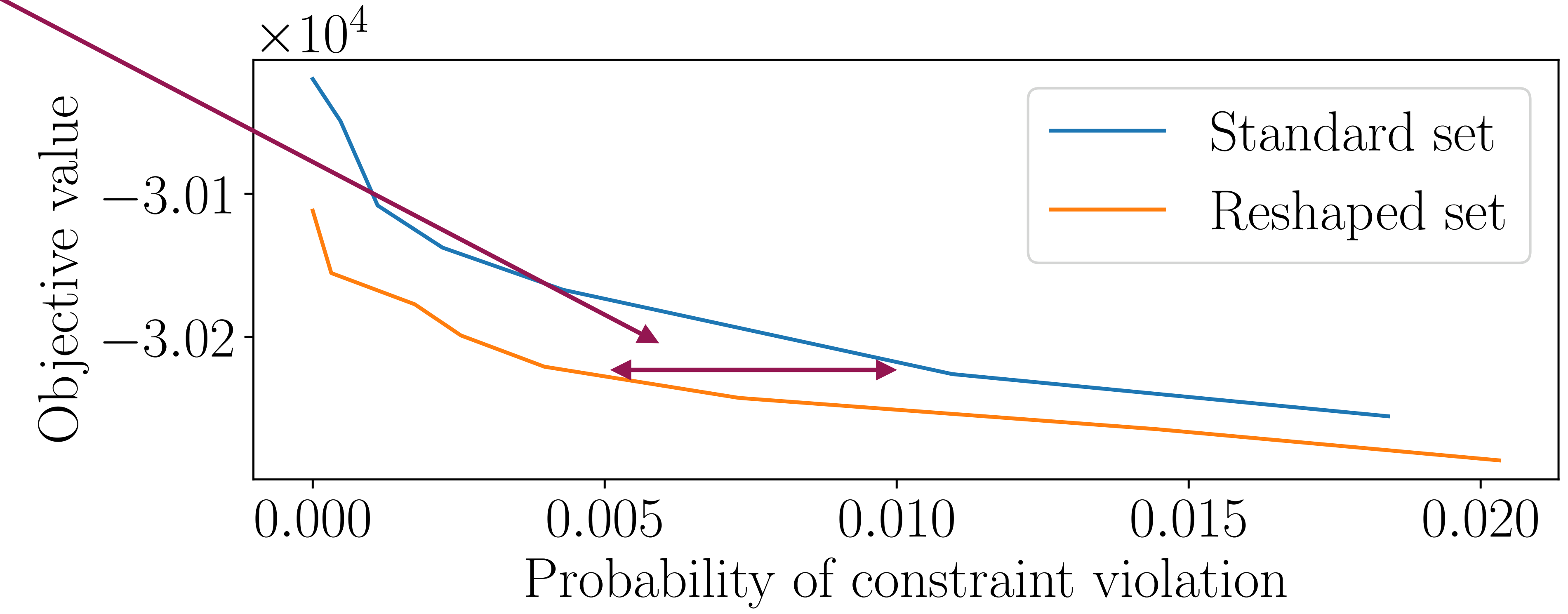




Resizing comparison (out-of-sample)

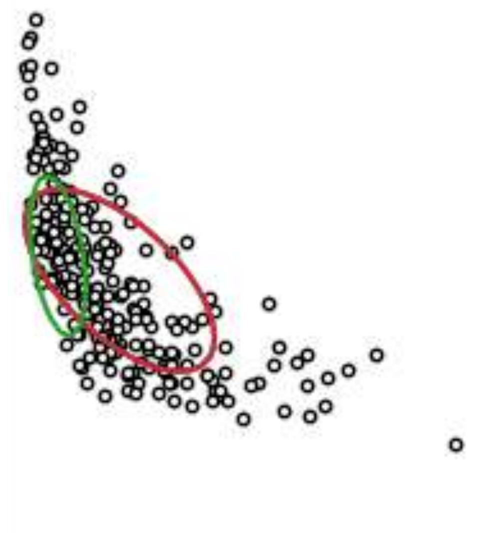


Reduction in constraint violation

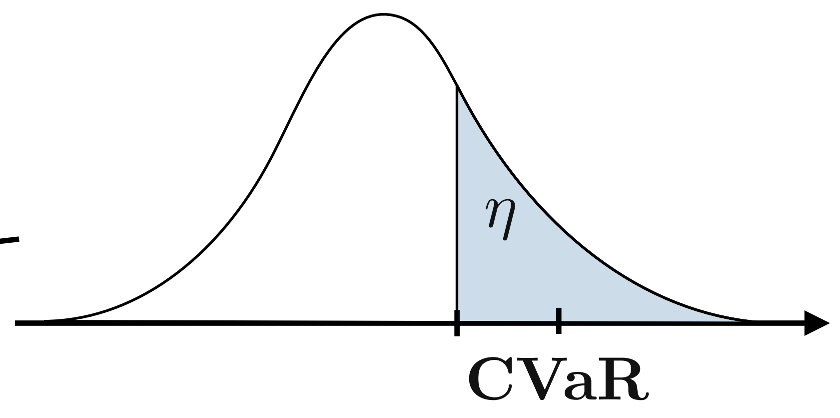


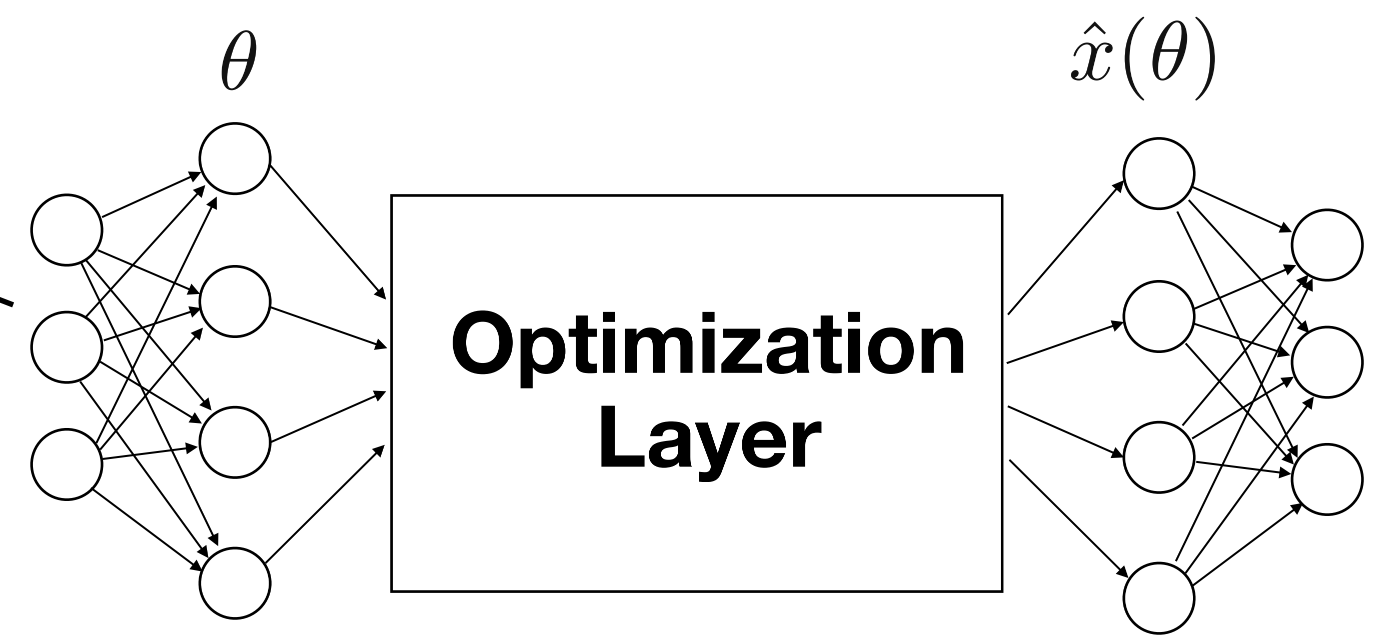
Learning Uncertainty sets in Robust Optimization

- Optimize **shape and size** at the same time



- **Bi-level optimization** formulation
  - Empirical CVaR as constraints
  - Differentiable optimization to compute derivatives





- **Improvements over standard uncertainty sets**

arXiv Soon! Check out https://stellato.io

Conclusions

Acknowledgements

Irina Wang

Princeton

Cole Becker

Princeton

Bart Van Parys

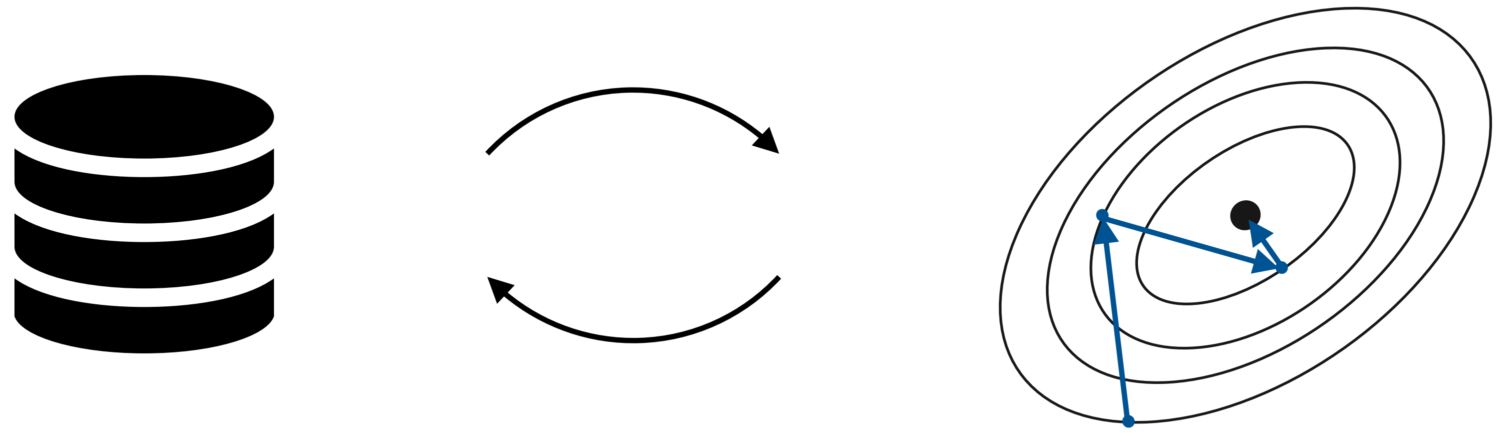
MIT

Conclusion

Machine Learning can help us
solving decision-making problems under uncertainty

We should rethink
building robust optimization models
as a **training/validation procedure**

Data Optimization



stellato.io bstellato@princeton.edu @b_stellato