CWI

Centrum Wiskunde & Informatica

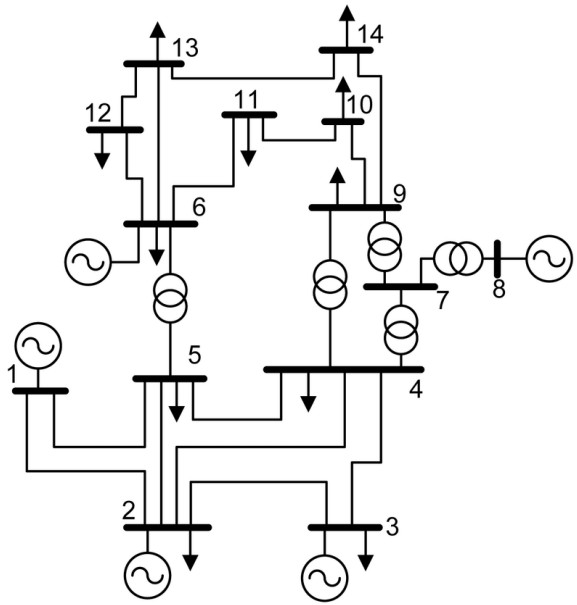
Optimal Decisions with Messy Data and Uncertainty

Bart P.G. Van Parys

CWI Amsterdam

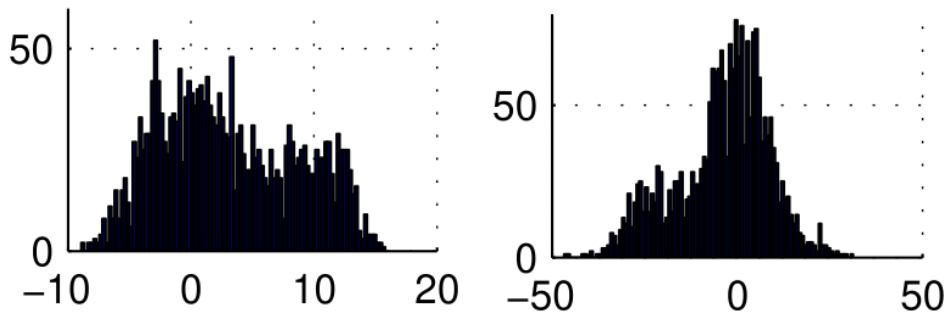
27 October 2025

Practical Decisions require Dealing with Uncertainty and Data



Optimal Power Flow (Roald et al., 2015)

- **minimize:** operation cost
- **decide:** generation commitment
- **uncertain:** forecast error



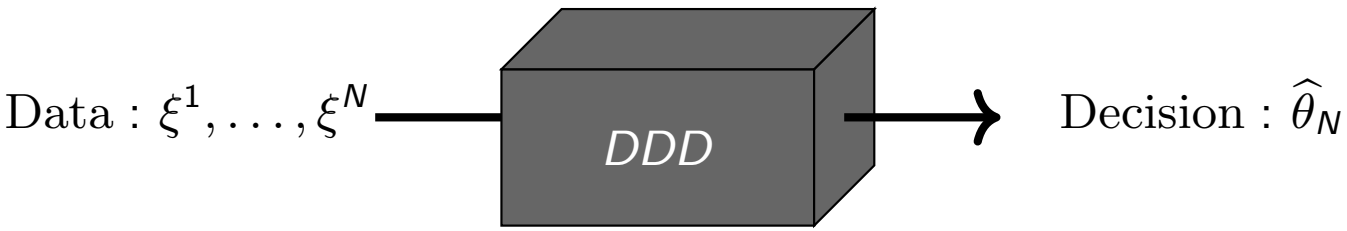
Data-Driven Decisions

Classical approaches for decisions in the face of uncertainty

$$\min_{\theta \in \Theta} \mathbf{E}_{\mathbb{P}} \left[\ell(\theta, \xi) \right]$$

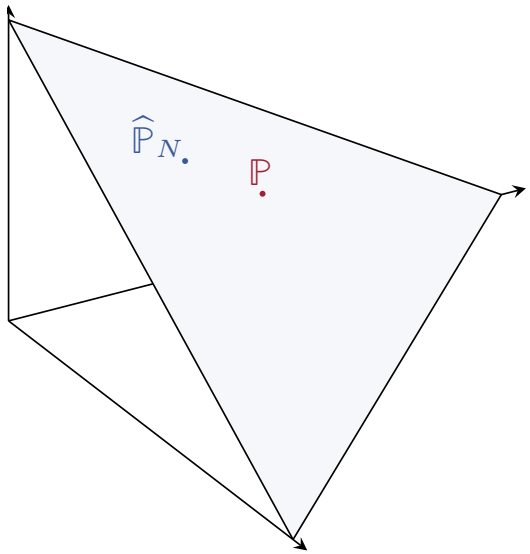
use the **distribution** $\mathbb{P}$ of $\xi \in \Xi$ as a primitive.

Modern approaches try to learn good decisions with **data**



Empirical Risk Minimization (ERM/SAA)

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

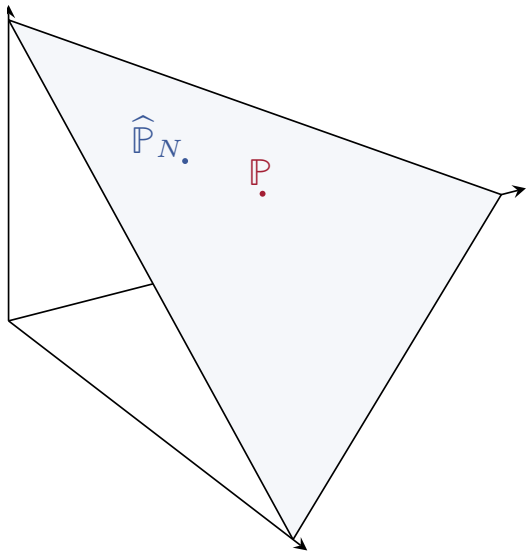


- Empirical risk minimization (Birge and Louveaux, 2011)

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \mathbf{E}_{\widehat{\mathbb{P}}_N}\left[\ell(\theta,\xi)\right]$$
$$\in \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \ell(\theta,\xi^i).$$

Empirical Risk Minimization (ERM/SAA)

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

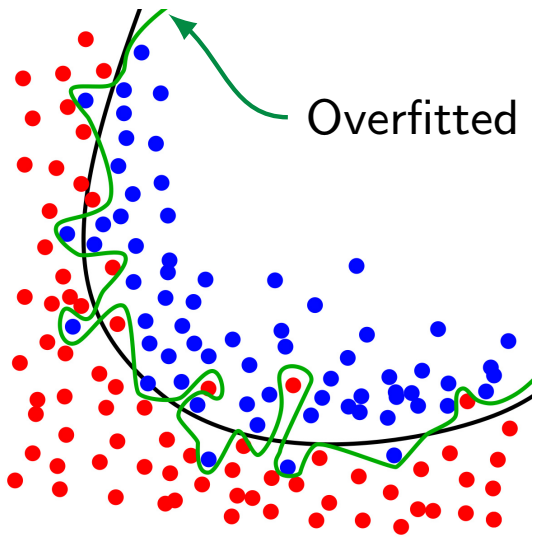


- Empirical risk minimization (Birge and Louveaux, 2011)

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \mathbf{E}_{\widehat{\mathbb{P}}_N}\left[\ell(\theta,\xi)\right]$$
$$\in \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \ell(\theta,\xi^i).$$

- "Overfitting", "Error Maximization Effect", "Optimizer's Curse" (Michaud, 1989)

$$\underbrace{\mathbf{E}_{\mathbb{P}}\left[\ell(\widehat{\theta}_N,\xi)\right]}_{\textit{actual cost}} > \underbrace{\mathbf{E}_{\widehat{\mathbb{P}}_N}\left[\ell(\widehat{\theta}_N,\xi)\right]}_{\textit{predicted cost}}$$



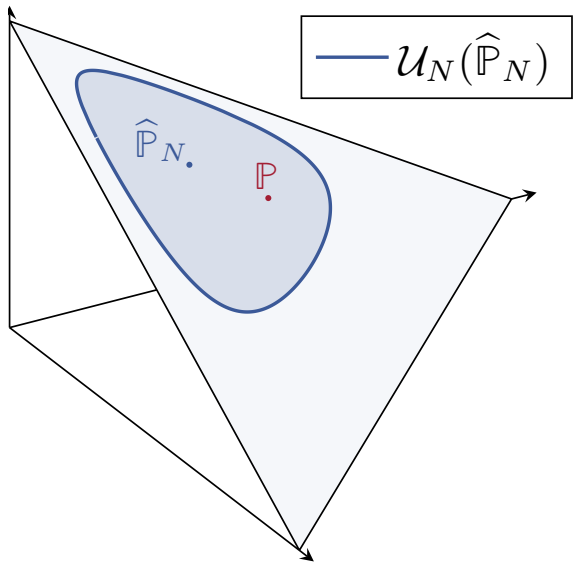
Worst-Case Formulations

Worst-case formulations

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \max_{\mathbb{Q}\in\mathcal{U}_N(\widehat{\mathbb{P}}_N)} \mathbf{E}_{\mathbb{Q}}\left[\ell(\theta,\xi)\right]$$

and guard against overfitting, i.e.,

$$\mathbb{P}\in\mathcal{U}_N(\widehat{\mathbb{P}}_N) \implies \underbrace{\max_{\mathbb{Q}\in\mathcal{U}_N(\widehat{\mathbb{P}}_N)} \mathbf{E}_{\mathbb{Q}}\left[\ell(\widehat{\theta}_N,\xi)\right]}_{\textit{predicted cost}} \geq \underbrace{\mathbf{E}_{\mathbb{P}}\left[\ell(\widehat{\theta}_N,\xi)\right]}_{\textit{actual cost}}.$$



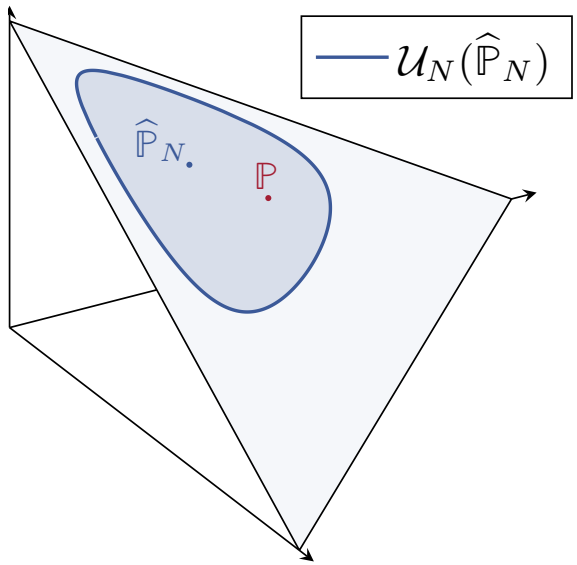
Worst-Case Formulations

Worst-case formulations

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \max_{\mathbb{Q}\in\mathcal{U}_N(\widehat{\mathbb{P}}_N)} \mathbf{E}_{\mathbb{Q}}\left[\ell(\theta,\xi)\right]$$

and guard against overfitting, i.e.,

$$\mathbb{P}\in\mathcal{U}_N(\widehat{\mathbb{P}}_N) \implies \underbrace{\max_{\mathbb{Q}\in\mathcal{U}_N(\widehat{\mathbb{P}}_N)} \mathbf{E}_{\mathbb{Q}}\left[\ell(\widehat{\theta}_N,\xi)\right]}_{\textit{predicted cost}} \geq \underbrace{\mathbf{E}_{\mathbb{P}}\left[\ell(\widehat{\theta}_N,\xi)\right]}_{\textit{actual cost}}.$$



Tractability depends on whether the **distributional optimization** problem

$$\max_{\mathbb{Q}\in\mathcal{U}_N(\widehat{\mathbb{P}}_N)} \mathbf{E}_{\mathbb{Q}}\left[\ell(\theta,\xi)\right]$$

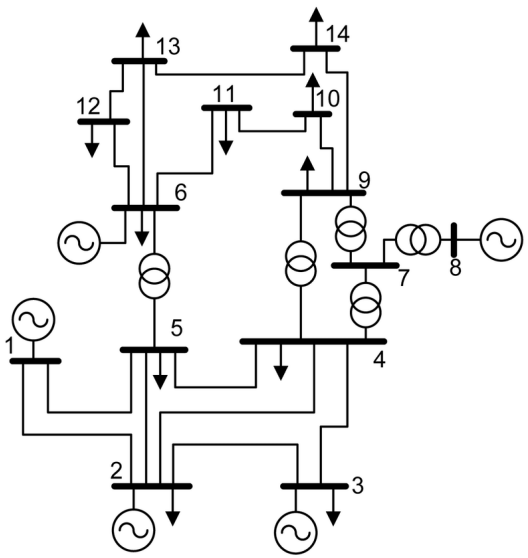
can be solved efficiently.

Table of Contents

Example: Power Grid Safety (Roald et al., 2015)

Power Flow Problem:

- Generators (⏦) generate power to satisfy demand (⏚)



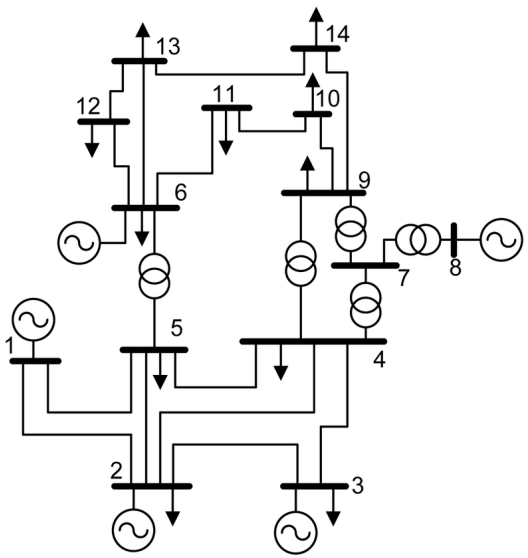
Example: Power Grid Safety (Roald et al., 2015)

Power Flow Problem:

- Generators (⏦) generate power to satisfy demand (⬇)
- Power through every line is limited

$$p_k = \underbrace{p_{nom,k}}_{nominal} < \underbrace{p_{\max,k}}_{trip\ \ limit} \quad \forall k$$

if generators honor their power commitment



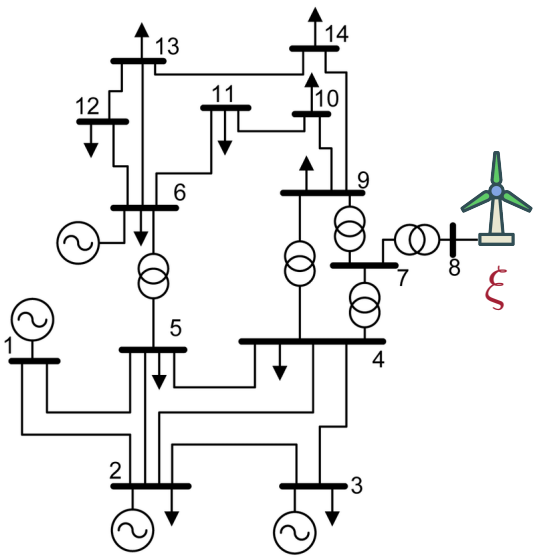
Example: Power Grid Safety (Roald et al., 2015)

Power Flow Problem:

- Generators (⏦) generate power to satisfy demand (⬇)
- Power through every line is limited

$$p_k = \underbrace{p_{nom,k}}_{nominal} + \underbrace{a_k^\top \xi}_{correction} < \underbrace{p_{\max,k}}_{trip\ \ limit} \quad \forall k$$

- Renewable generators suffer **forecast errors** ξ



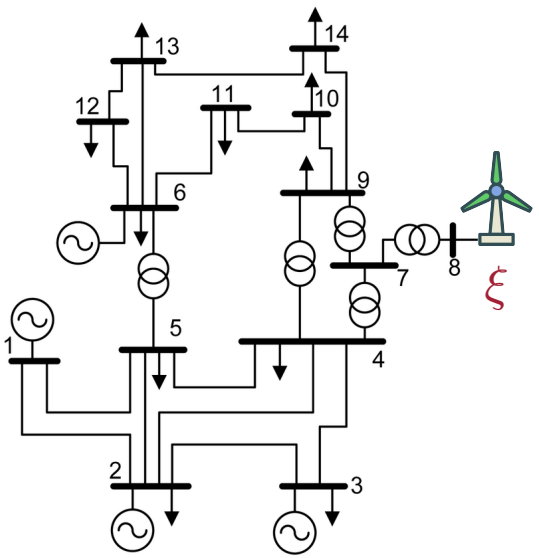
Example: Power Grid Safety (Roald et al., 2015)

Power Flow Problem:

- Generators (⏦) generate power to satisfy demand (⬇)
- Power through every line is limited

$$P = \{\xi : \underbrace{p_{nom,k}}_{\text{nominal}} + \underbrace{a_k^\top \xi}_{\text{correction}} < \underbrace{p_{\max,k}}_{\text{trip limit}} \quad \forall k\}$$

- Renewable generators suffer **forecast errors** ξ



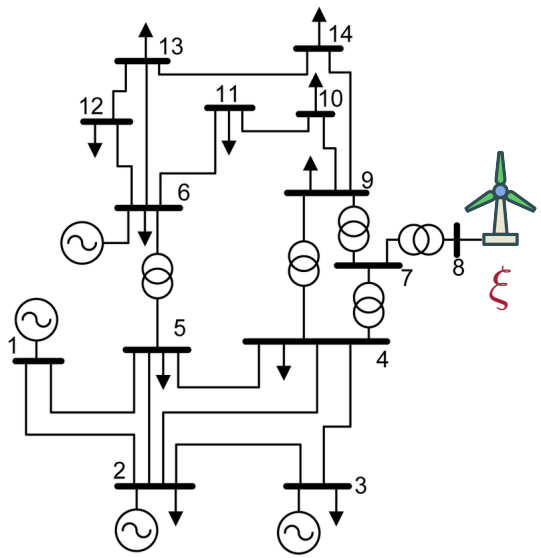
Probability of a line tripping is $\mathbb{P}(\xi \notin P)$:

Example: Power Grid Safety (Roald et al., 2015)

Power Flow Problem:

- Generators (⏦) generate power to satisfy demand (🡻)
- Power through every line is limited

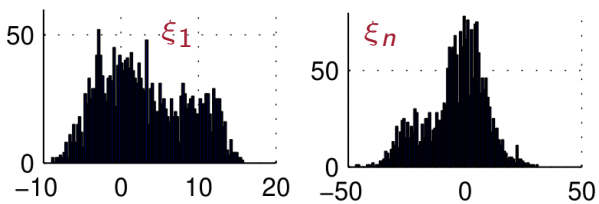
$$P = \{\xi : \underbrace{p_{nom,k}}_{\text{nominal}} + \underbrace{a_k^\top \xi}_{\text{correction}} < \underbrace{p_{\max,k}}_{\text{trip limit}} \quad \forall k\}$$



- Renewable generators suffer **forecast errors** ξ

Probability of a line tripping is $\mathbb{P}(\xi \notin P)$:

- Distribution $\xi \sim \mathbb{P}$ is not known

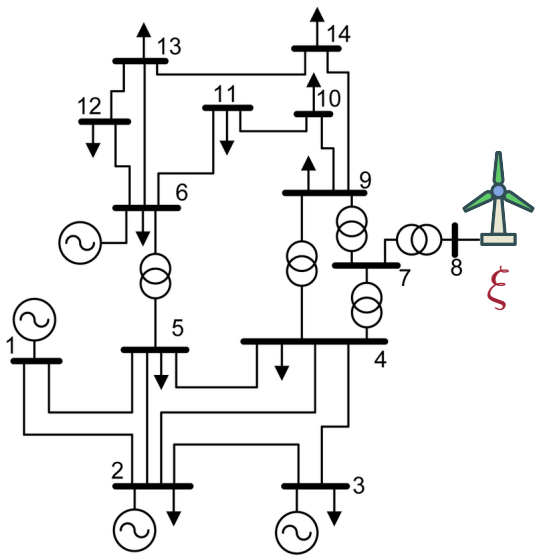


Example: Power Grid Safety (Roald et al., 2015)

Power Flow Problem:

- Generators (⏦) generate power to satisfy demand (🡻)
- Power through every line is limited

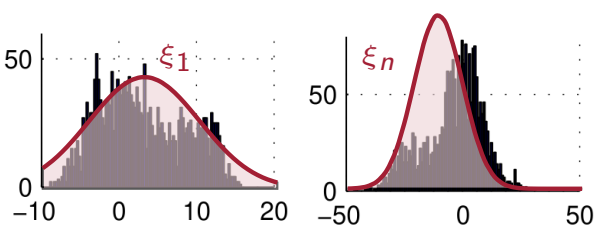
$$P = \{\xi : \underbrace{p_{nom,k}}_{\text{nominal}} + \underbrace{a_k^\top \xi}_{\text{correction}} < \underbrace{p_{\max,k}}_{\text{trip limit}} \quad \forall k\}$$



- Renewable generators suffer **forecast errors** ξ

Probability of a line tripping is $\mathbb{P}(\xi \notin P)$:

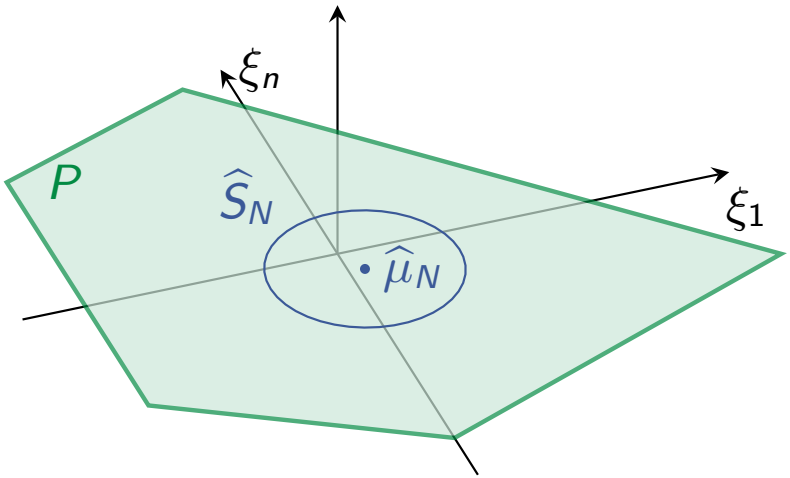
- Distribution $\xi \sim \mathbb{P}$ is not known
- $\mathbb{P}$ is normal may **underestimate** trip probability



Uncertainty Quantification Problems with Moments

Worst-Case Probability:

$$\begin{aligned} \max \quad & \mathbb{Q}(\xi \notin P) \\ \text{s.t.} \quad & \mathbb{Q} \in \mathcal{P}, \\ & \int \xi \mathrm{d}\mathbb{Q} = \widehat{\mu}_N, \\ & \int \xi\xi^\top \mathrm{d}\mathbb{Q} = \widehat{S}_N \end{aligned}$$



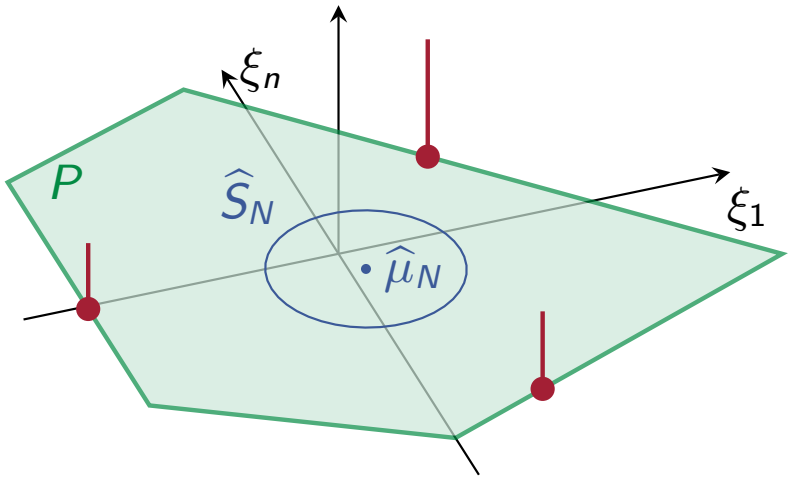
Properties:

- Reduces to a tractable convex optimization problem (Bertsimas and Popescu, 2005)

Uncertainty Quantification Problems with Moments

Worst-Case Probability:

$$\begin{aligned} \max \quad & \mathbb{Q}(\xi \notin P) \\ \text{s.t.} \quad & \mathbb{Q} \in \mathcal{P}, \\ & \int \xi \mathrm{d}\mathbb{Q} = \widehat{\mu}_N, \\ & \int \xi\xi^\top \mathrm{d}\mathbb{Q} = \widehat{S}_N \end{aligned}$$



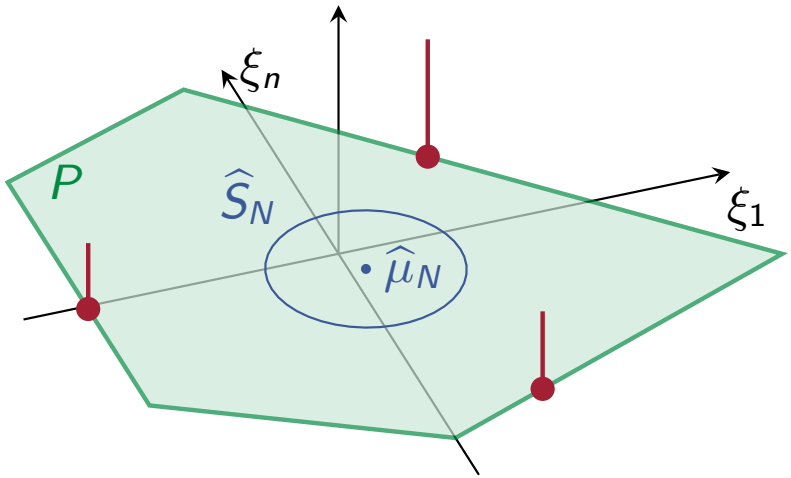
Properties:

- Reduces to a tractable convex optimization problem (Bertsimas and Popescu, 2005)
- **Pathological** worst-case distribution $\mathbb{Q}^\star$

Uncertainty Quantification Problems with Moments

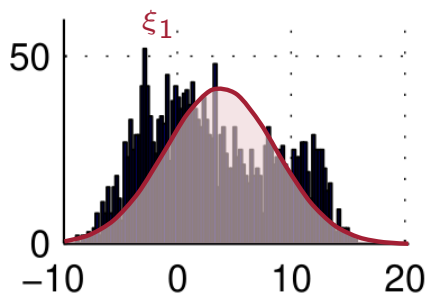
Worst-Case Probability:

$$\begin{aligned} \max \quad & \mathbb{Q}(\xi \notin P) \\ \text{s.t.} \quad & \mathbb{Q} \in \mathcal{P}, \\ & \int \xi \mathrm{d}\mathbb{Q} = \widehat{\mu}_N, \\ & \int \xi\xi^\top \mathrm{d}\mathbb{Q} = \widehat{S}_N \end{aligned}$$

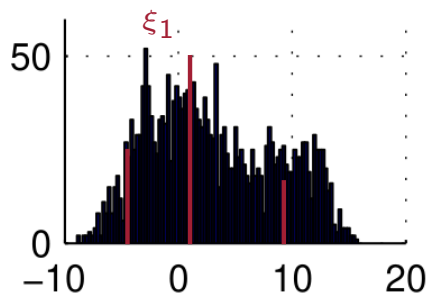


Properties:

- Reduces to a tractable convex optimization problem (Bertsimas and Popescu, 2005)
- **Pathological** worst-case distribution $\mathbb{Q}^\star$



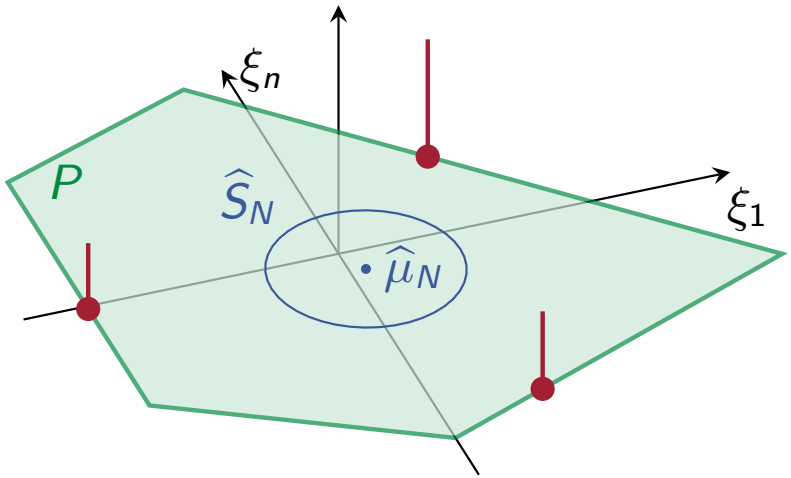
?



Uncertainty Quantification Problems with Moments

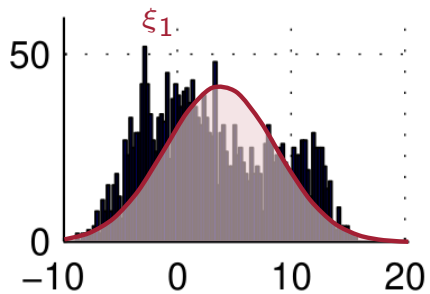
Worst-Case Probability:

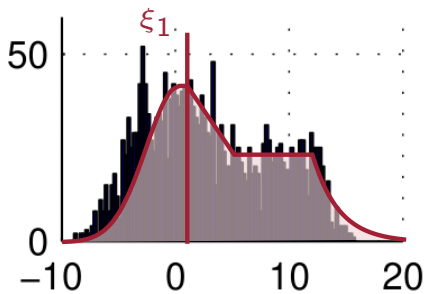
$$\begin{aligned} \max \quad & \mathbb{Q}(\xi \notin P) \\ \text{s.t.} \quad & \mathbb{Q} \in \mathcal{P}, \\ & \int \xi \mathrm{d}\mathbb{Q} = \widehat{\mu}_N, \\ & \int \xi\xi^\top \mathrm{d}\mathbb{Q} = \widehat{S}_N \end{aligned}$$

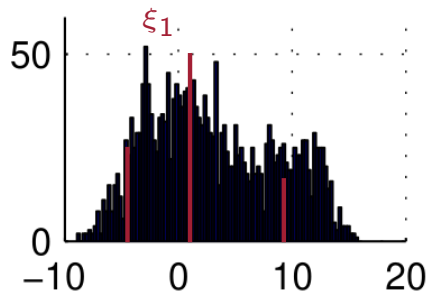


Properties:

- Reduces to a tractable convex optimization problem (Bertsimas and Popescu, 2005)
- **Pathological** worst-case distribution $\mathbb{Q}^\star$
- Unimodal distribution $\mathbb{Q}^\star$ (Van Parys, P. J. Goulart, et al., 2016)







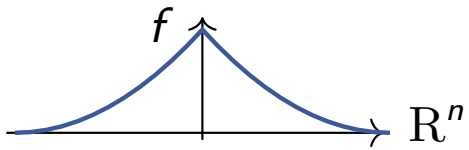
Generalizations

$$\max \mathbb{Q}(\xi \notin P)$$
$$\text{s.t. } \mathbb{Q} \in \mathcal{S},\ \int \xi \mathrm{d}\mathbb{Q} = \widehat{\mu}_N,\ \int \xi\xi^\top \mathrm{d}\mathbb{Q} = \widehat{S}_N$$

Generalizations

$$\begin{aligned} &\max \ \mathbf{E}_{\mathbb{Q}}\left[\ell(\theta, \xi)\right] \\ &\text{s.t. } \mathbb{Q} \in \mathcal{S},\ \int \xi \mathrm{d}\mathbb{Q} = \widehat{\mu}_N, \int \xi\xi^{\top} \mathrm{d}\mathbb{Q} = \widehat{S}_N \end{aligned}$$

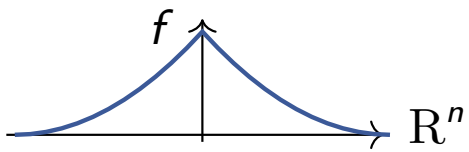
- conditional value-at-risk quantification in financial risk management
- distributions with common tail properties (Van Parys, P. Goulart, and Morari, 2019)



Generalizations

$$\begin{aligned} &\max \mathbf{E}_{\mathbb{Q}}\left[\ell(\theta, \xi)\right] \\ &\text{s.t. } \mathbb{Q} \in \mathcal{S},\ \int \xi \mathrm{d}\mathbb{Q} = \widehat{\mu}_N, \int \xi\xi^{\top} \mathrm{d}\mathbb{Q} = \widehat{S}_N \end{aligned}$$

- conditional value-at-risk quantification in financial risk management
- distributions with common tail properties (Van Parys, P. Goulart, and Morari, 2019)

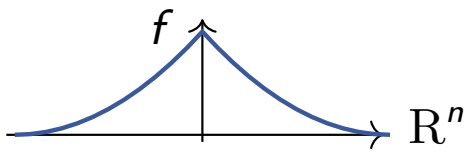


\+ Transfer information to tail regions where little data is observed

Generalizations

$$\max \mathbf{E}_{\mathbb{Q}}\left[\ell(\theta, \xi)\right]$$
$$\text{s.t. } \mathbb{Q} \in \mathcal{S},\ \int \xi \mathrm{d}\mathbb{Q} = \widehat{\mu}_N, \int \xi\xi^{\top} \mathrm{d}\mathbb{Q} = \widehat{S}_N$$

- conditional value-at-risk quantification in financial risk management
- distributions with common tail properties (Van Parys, P. Goulart, and Morari, 2019)



\+ Transfer information to tail regions where little data is observed

\- Learns from data **indirectly** through its moments

Table of Contents

Divergence Balls $\mathcal{U}_N(\widehat{\mathbb{P}}_N) = \{\mathbb{Q} \ : \ D(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r\}$

Formulations

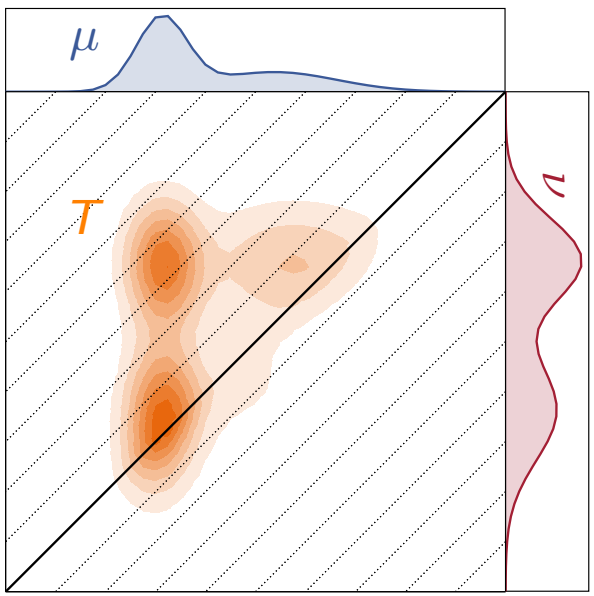
$$\widehat{\theta}_N \in \arg\min_{\theta \in \Theta} \max_{D(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r} \mathbf{E}_{\mathbb{Q}} \left[\ell(\theta, \xi)\right]$$

based on tractable statistical distances can learn **directly** from data.

Examples:

- **Wasserstein distance** is defined as

$$W(\mu, \nu) := \inf_{T \in \mathcal{T}(\mu, \nu)} \int \left\| \xi - \xi' \right\| \mathrm{d}T$$



Divergence Balls $\mathcal{U}_N(\widehat{\mathbb{P}}_N) = \{\mathbb{Q} \ : \ D(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r\}$

Formulations

$$\widehat{\theta}_N \in \arg\min_{\theta \in \Theta} \max_{D(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r} \mathbf{E}_{\mathbb{Q}} \left[\ell(\theta, \xi)\right]$$

based on tractable statistical distances can learn **directly** from data.

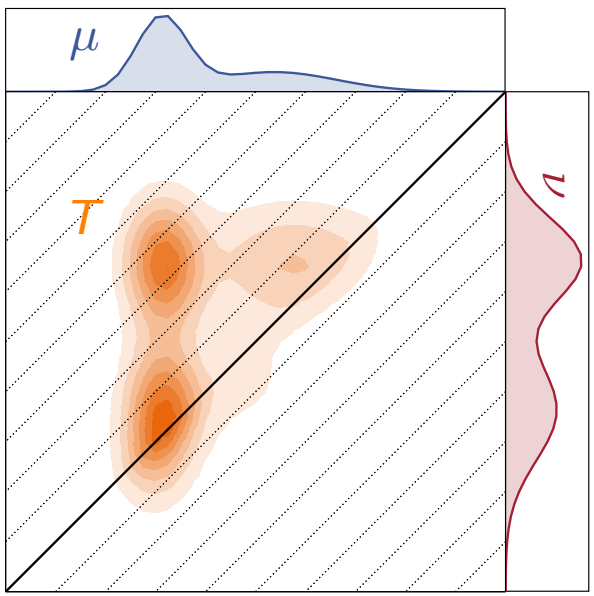
Examples:

- **Wasserstein distance** is defined as

$$W(\mu, \nu) := \inf_{T \in \mathcal{T}(\mu, \nu)} \int \left\| \xi - \xi' \right\| \mathrm{d}T$$

- **KL-Divergence** is defined as

$$\mathrm{KL}(\mu, \nu) = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu$$



Sizing Ambiguity Sets

Example: Wasserstein DRO

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \max_{W(\widehat{\mathbb{P}}_N,\mathbb{Q})\leq r} \mathbf{E}_{\mathbb{Q}}[\ell(\theta,\xi)]\,.$$

Sizing Ambiguity Sets

Example: Wasserstein DRO

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \max_{W(\widehat{\mathbb{P}}_N,\mathbb{Q})\leq r} \mathbf{E}_{\mathbb{Q}}[\ell(\theta,\xi)]\,.$$

- Assume data $\{\xi^i \in \mathrm{R}^n\}_{i=1}^N$ is drawn independently from $\mathbb{P}$.

Sizing Ambiguity Sets

Example: Wasserstein DRO

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \max_{W(\widehat{\mathbb{P}}_N,\mathbb{Q})\leq r} \mathbf{E}_{\mathbb{Q}}[\ell(\theta,\xi)]\,.$$

- Assume data $\{\xi^i \in \mathrm{R}^n\}_{i=1}^N$ is drawn independently from $\mathbb{P}$.
- For Wasserstein balls we can show the **concentration inequality** (Fournier and Guillin, 2015)

$$\mathrm{Prob}_{\mathbb{P}}\left[W(\widehat{\mathbb{P}}_N,\mathbb{P}) \leq r\right] \geq 1 - c_1 \exp(-c_2 N r^{\max(n,2)})$$

with $c_1 > 0$ and $c_2 > 0$ positive constants.

Sizing Ambiguity Sets

Example: Wasserstein DRO

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \max_{W(\widehat{\mathbb{P}}_N,\mathbb{Q})\le r} \mathbf{E}_{\mathbb{Q}}[\ell(\theta,\xi)]\,.$$

- Assume data $\{\xi^i \in \mathrm{R}^n\}_{i=1}^N$ is drawn independently from $\mathbb{P}$.
- For Wasserstein balls we can show the **out-of-sample** (OOS) guarantee

$$\mathrm{Prob}_{\mathbb{P}}\Big[\underbrace{\max_{W(\widehat{\mathbb{P}}_N,\mathbb{Q})\le r} \mathbf{E}_{\mathbb{Q}}[\ell(\theta,\xi)]}_{\textit{predicted cost}} \ge \underbrace{\mathbf{E}_{\mathbb{P}}[\ell(\theta,\xi)]}_{\textit{actual cost}}\ \forall\theta\in\Theta\Big] \ge 1 - c_1\exp(-c_2 N r^{\max(n,2)})$$

with $c_1 > 0$ and $c_2 > 0$ positive constants.

Why Robust Optimization?

There are many ways to protect decisions against overfitting:

- OR: Robustification,
- Statistics : Regularization,
- Deep Learning: Early stopping, drop out, ...

Why Robust Optimization?

There are many ways to protect decisions against overfitting:

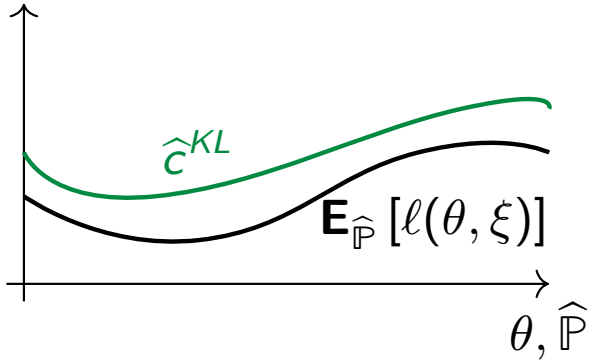
- OR: Robustification,
- Statistics : Regularization,
- Deep Learning: Early stopping, drop out, ...

Why should we use robust optimization?

Kullback-Leibler Robust Formulation

Kullback-Leibler (KL) formulation:

$$\begin{aligned}\widehat{c}^{KL}(\theta, \widehat{\mathbb{P}}) = \max\ & \mathbf{E}_{\mathbb{Q}}\left[\ell(\theta, \xi)\right] \\ \text{s.t. } & \mathbb{Q} \in \mathcal{P}, \\ & \mathrm{KL}(\widehat{\mathbb{P}}, \mathbb{Q}) \leq r.\end{aligned}$$



Thm (Van Parys, Esfahani, et al., 2020): $\Xi \subset \mathrm{R}^n$ is compact, $\{\xi \mapsto \ell(\theta, \xi)\}_{\theta \in \Theta}$ is equicontinuous.

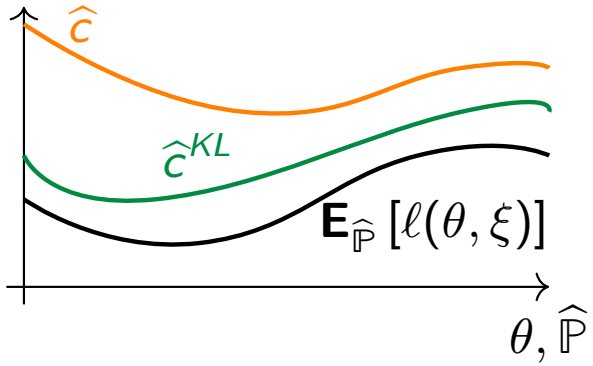
- (OOS Guarantee) we have

$$\mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}^{KL}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right]\ \ \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N))\ \ \forall \mathbb{P}.$$

Kullback-Leibler Robust Formulation

Kullback-Leibler (KL) formulation:

$$\widehat{c}^{KL}(\theta, \widehat{\mathbb{P}}) = \max \mathbf{E}_{\mathbb{Q}}\left[\ell(\theta, \xi)\right]$$
$$\text{s.t. } \mathbb{Q} \in \mathcal{P},$$
$$\mathrm{KL}(\widehat{\mathbb{P}}, \mathbb{Q}) \leq r.$$



Thm (Van Parys, Esfahani, et al., 2020): $\Xi \subset \mathrm{R}^n$ is compact, $\{\xi \mapsto \ell(\theta, \xi)\}_{\theta \in \Theta}$ is equicontinuous.

- (OOS Guarantee) we have

$$\mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}^{KL}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right] \ \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall \mathbb{P}.$$

- (Efficiency) Furthermore, let $\widehat{c}$ satisfy

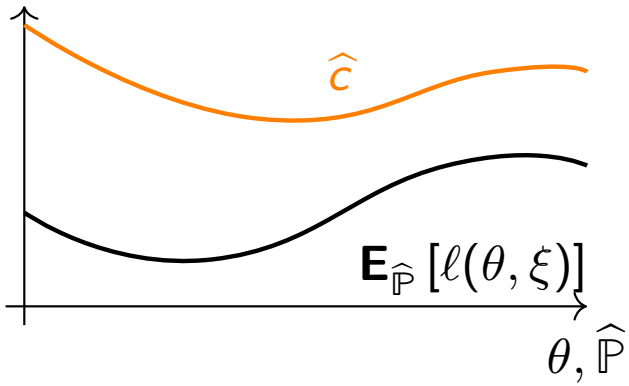
$$\mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right] \ \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall \mathbb{P}.$$

Then, $\widehat{c}(\theta, \widehat{\mathbb{P}}) \geq \widehat{c}^{KL}(\theta, \widehat{\mathbb{P}})$ for all $\theta \in \Theta$ and $\widehat{\mathbb{P}} \in \mathcal{P}$.

Meta Optimization Problem

Data driven formulations

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \widehat{c}(\theta, \widehat{\mathbb{P}}_N)$$



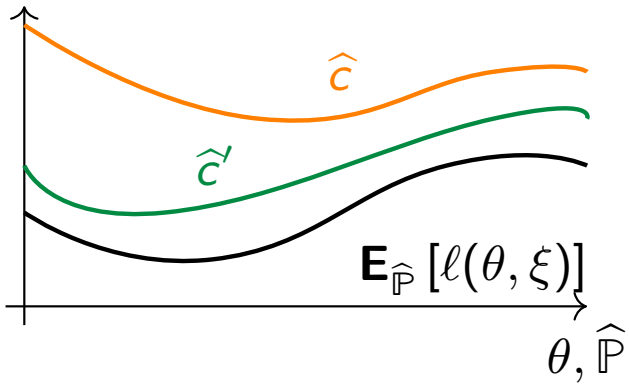
Feasibility: Consider formulations who enjoy **OOS guarantee**

$$\text{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}[\ell(\theta,\xi)] \ \ \forall\theta\in\Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall\mathbb{P}.$$

Meta Optimization Problem

Data driven formulations

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \widehat{c}(\theta, \widehat{\mathbb{P}}_N)$$



Feasibility: Consider formulations who enjoy **OOS guarantee**

$$\text{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}[\ell(\theta,\xi)] \ \ \forall\theta\in\Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall\mathbb{P}.$$

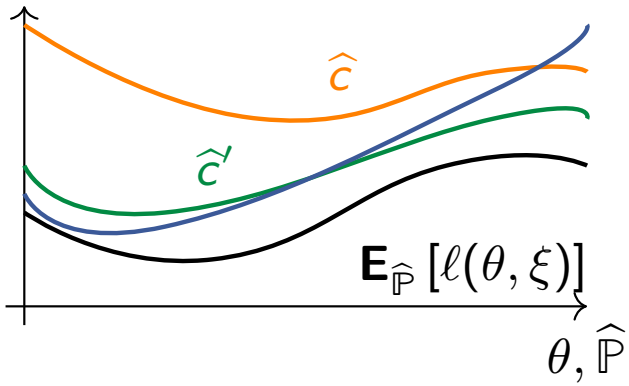
Conservatism: Formulation $\widehat{c}'$ is **less conservative** than $\widehat{c}$ if and only if

$$\widehat{c}' \preceq \widehat{c} \iff \widehat{c}'(\theta, \widehat{\mathbb{P}}) \leq \widehat{c}(\theta, \widehat{\mathbb{P}}) \ \ \forall\theta, \widehat{\mathbb{P}}.$$

Meta Optimization Problem

Data driven formulations

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \widehat{c}(\theta, \widehat{\mathbb{P}}_N)$$



Feasibility: Consider formulations who enjoy **OOS guarantee**

$$\text{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}[\ell(\theta,\xi)] \ \ \forall\theta\in\Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall\mathbb{P}.$$

Conservatism: Formulation $\widehat{c}'$ is **less conservative** than $\widehat{c}$ if and only if

$$\widehat{c}' \preceq \widehat{c} \iff \widehat{c}'(\theta, \widehat{\mathbb{P}}) \leq \widehat{c}(\theta, \widehat{\mathbb{P}}) \ \ \forall\theta, \widehat{\mathbb{P}}.$$

Meta Optimization Problem

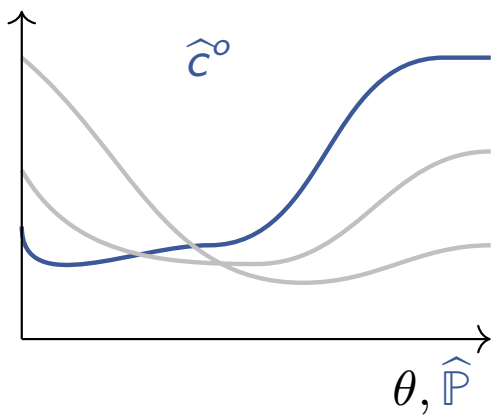
Among all predictors with **out-of-sample guarantee** find the **least conservative** one

$$\begin{cases} \min\limits_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right]\ \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N))\ \ \forall \mathbb{P} \end{cases}$$

Meta Optimization Problem

Among all predictors with **out-of-sample guarantee** find the **least conservative** one

$$\begin{cases} \min\limits_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right]\ \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N))\ \ \forall \mathbb{P} \end{cases}$$

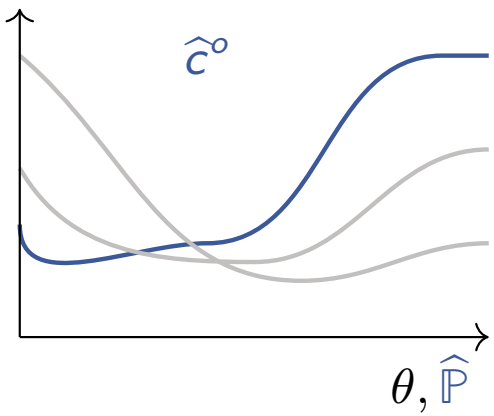


Pareto optimal solution

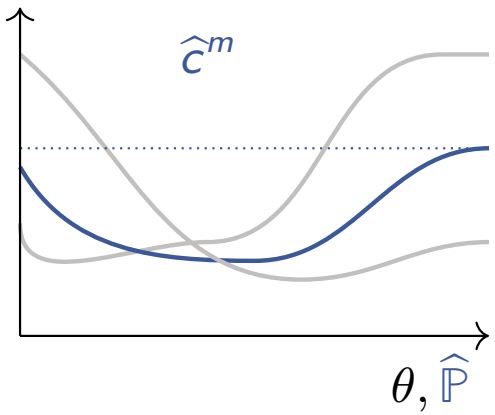
Meta Optimization Problem

Among all predictors with **out-of-sample guarantee** find the **least conservative** one

$$\begin{cases} \min_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right]\ \ \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N))\ \ \forall \mathbb{P} \end{cases}$$



Pareto optimal solution

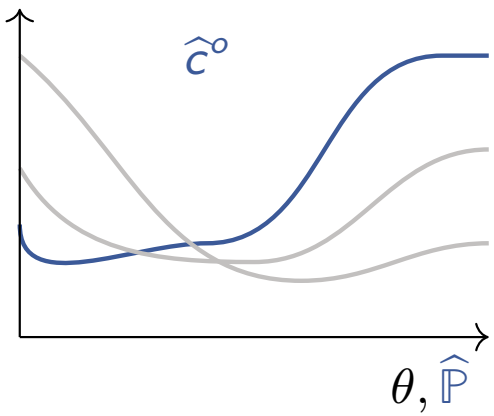


Minimax solution

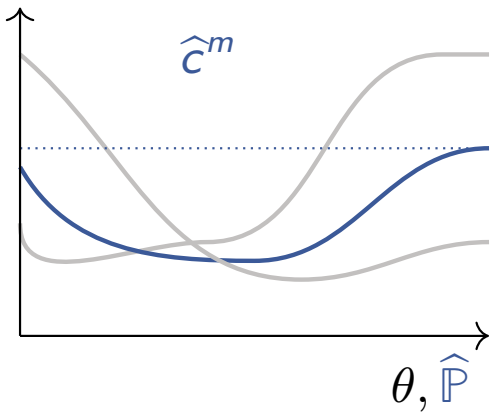
Meta Optimization Problem

Among all predictors with **out-of-sample guarantee** find the **least conservative** one

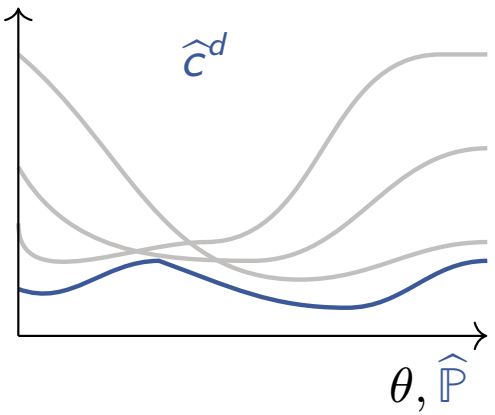
$$\begin{cases} \min_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right]\ \ \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N))\ \ \forall \mathbb{P} \end{cases}$$



Pareto optimal solution



Minimax solution



Pareto dominant solution

Pareto Dominant Solution

Thm (Van Parys, Esfahani, et al., 2020): The meta optimization problem

$$\begin{cases} \min\limits_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta,\xi)\right]\ \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall \mathbb{P} \end{cases}$$

admits the **Pareto dominant** solution $\widehat{c}^{KL}$.

Proof Crux: Indirect consequence of Sanov's theorem.

Pareto Dominant Solution

Thm (Van Parys, Esfahani, et al., 2020): The meta optimization problem

$$\begin{cases} \min\limits_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta,\xi)\right]\ \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall \mathbb{P} \end{cases}$$

admits the **Pareto dominant** solution $\widehat{c}^{KL}$.

Proof Crux: Indirect consequence of Sanov's theorem.

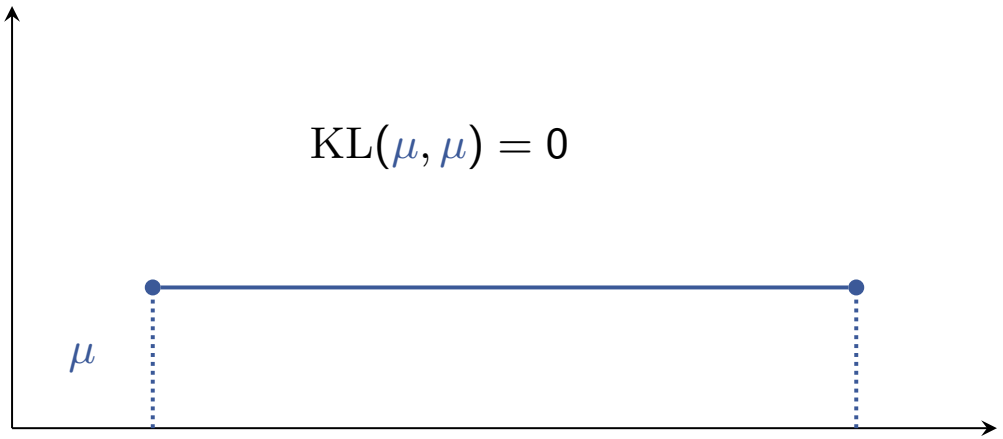
Generalizations:

- IID data assumption can be relaxed to account for dependence (Sutter et al., 2020)
- Out-of-sample guarantee can be relaxed (Bennouna and Van Parys, 2021)

Table of Contents

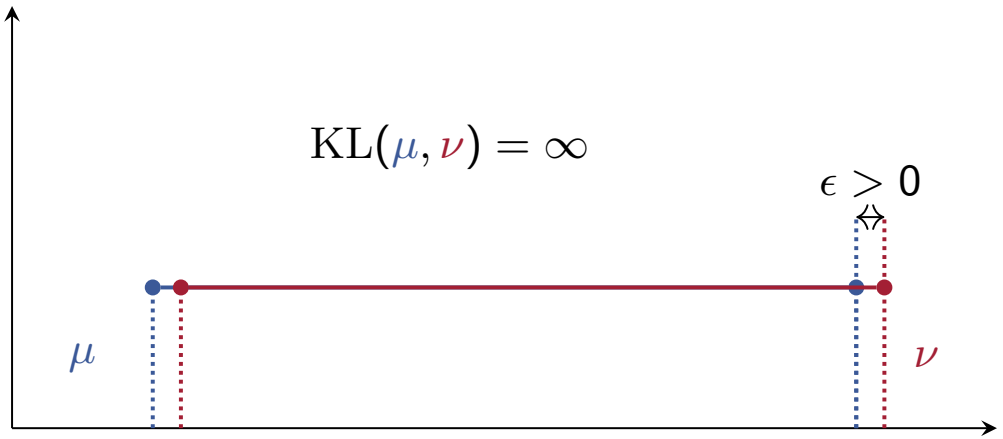
A Paradoxical Observation

Observation: KL-divergence lacks continuity



A Paradoxical Observation

Observation: KL-divergence lacks continuity

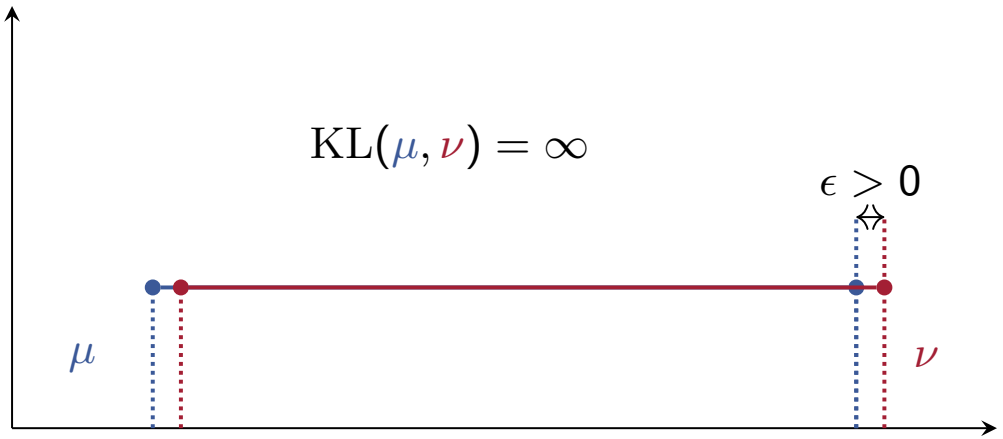


Remarks:

- KL robustness offers no protection against data corruption

A Paradoxical Observation

Observation: KL-divergence lacks continuity



Remarks:

- KL robustness offers no protection against data corruption
- Wasserstein distance does not suffer this shortcoming!

Practical Overfitting

Statistical Error

Randomness of the sampled data

$$\xi \sim \mathbb{P}$$

$$\Downarrow$$

$$\{\xi^1, \ldots, \xi^N\}$$

Practical Overfitting

Statistical Error	**Noise**	**Contamination** (Huber, 1964)
Randomness of the sampled data	Small ϵ-perturbation on each sample	Small fraction α of all samples is wholly corrupted
$\xi \sim \mathbb{P}$	$\{\xi^1, \ldots, \xi^N\}$	$\{\cancel{\xi^1}, \ldots, \xi^N\}$
$\Downarrow$	$\Downarrow$	$\Downarrow$
$\{\xi^1, \ldots, \xi^N\}$	$\{\xi^1 + \delta^1, \ldots, \xi^N + \delta^N\}$ $\|\delta^1\| \leq \epsilon, \ldots, \|\delta^N\| \leq \epsilon$	$\{\xi'^1, \ldots, \xi^N\}$

Existing Formulations Protect Against a Specific Overfitting Source

KL DRO $\max \left\{ \mathbf{E}_{\mathbb{Q}} [\ell(\theta, \xi)] : \mathrm{KL}(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r \right\}$ → **Statistical Error**

Noise

Contamination

Existing Formulations Protect Against a Specific Overfitting Source

KL DRO $\quad \max \left\{ \mathbf{E}_{\mathbb{Q}} [\ell(\theta, \xi)] : \mathrm{KL}(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r \right\}$ → **Statistical Error**

W DRO $\quad \max \left\{ \mathbf{E}_{\mathbb{Q}} [\ell(\theta, \xi)] : W(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq \epsilon \right\}$ → **Noise**

$\overset{\star}{=} \max \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\theta, \xi^i + \delta^i) : \frac{1}{N} \sum_{i=1}^{N} \left\| \delta^i \right\| \leq \epsilon \right\}$

Contamination

[$\star$] Assuming $\xi \mapsto \ell(\theta, \xi)$ is concave for all θ.

Existing Formulations Protect Against a Specific Overfitting Source

KL DRO $\quad \max \left\{ \mathbf{E}_{\mathbb{Q}} [\ell(\theta, \xi)] : \mathrm{KL}(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r \right\}$

W DRO $\quad \max \left\{ \mathbf{E}_{\mathbb{Q}} [\ell(\theta, \xi)] : W(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq \epsilon \right\}$

$\quad \overset{\star}{=} \max \left\{ \frac{1}{N} \sum_{i=1}^{N} \ell(\theta, \xi^i + \delta^i) : \frac{1}{N} \sum_{i=1}^{N} \left\| \delta^i \right\| \leq \epsilon \right\}$

Statistical Error

Noise

Contamination

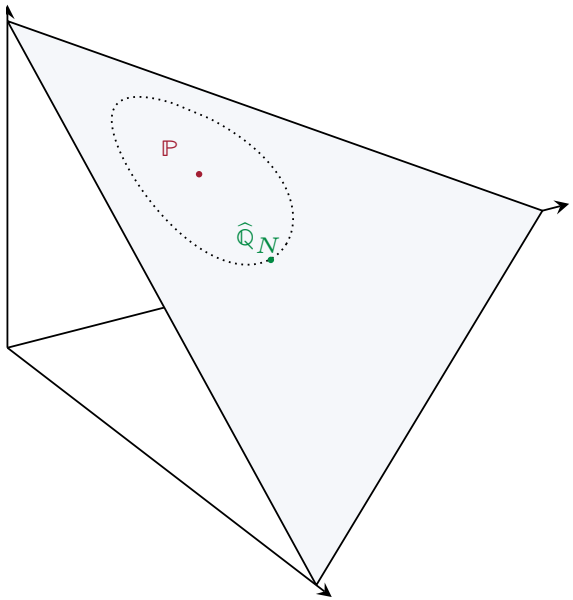
[$\star$] Assuming $\xi \mapsto \ell(\theta, \xi)$ is concave for all θ.

Kullback-Leibler Distance

$$\xi \sim \mathbb{P}$$

$$\Downarrow \quad \text{Statistical Error}$$

$$\{\xi^1, \xi^2 \ldots, \xi^N\} \qquad \widehat{\mathbb{Q}}_N$$



Thm: (Sanov, 1958) If $\widehat{\mathbb{Q}}_N$ is an empirical distribution of independent samples from $\mathbb{P}$ then

$$\mathrm{Prob}_{\mathbb{P}}\left[\mathrm{KL}(\widehat{\mathbb{Q}}_N, \mathbb{P}) \leq r\right] = 1 - \exp(-rN + o(N)).$$

Lévy-Prokhorov Distance

$$\xi \sim \mathbb{P}$$

$$\Downarrow \quad \text{Statistical Error}$$

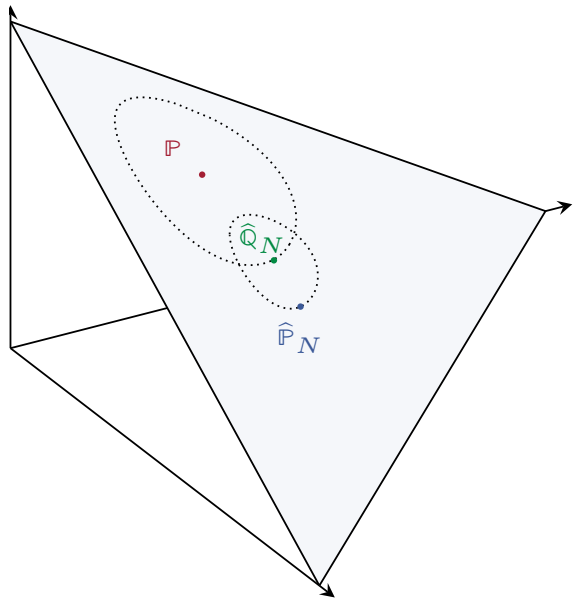
$$\{\xi^1, \xi^2 \ldots, \xi^N\} \qquad \widehat{\mathbb{Q}}_N$$

$$\Downarrow \quad \epsilon\text{-Noise}$$

$$\{\xi^1 + \delta^1, \xi^2 + \delta^2, \ldots, \xi^N + \delta^N\}$$

$$\Downarrow \quad \alpha\text{-Contamination}$$

$$\{\xi^1 + \delta^1, \xi'^2, \ldots, \xi^N + \delta^N\} \qquad \widehat{\mathbb{P}}_N$$



Prop: If $\widehat{\mathbb{P}}_N$ and $\widehat{\mathbb{Q}}_N$ differ by ϵ-norm bounded noise and α-contamination. Then

$$\mathrm{LP}_\epsilon(\widehat{\mathbb{P}}_N, \widehat{\mathbb{Q}}_N) := \inf_{T \in \mathcal{T}(\widehat{\mathbb{P}}_N, \widehat{\mathbb{Q}}_N)} \int \mathbb{1}\{\|\xi - \xi'\| \geq \epsilon\} \, \mathrm{d}T \leq \alpha.$$

Lévy-Prokhorov Distance

$$\xi \sim \mathbb{P}$$

$$\Downarrow \quad \text{Statistical Error}$$

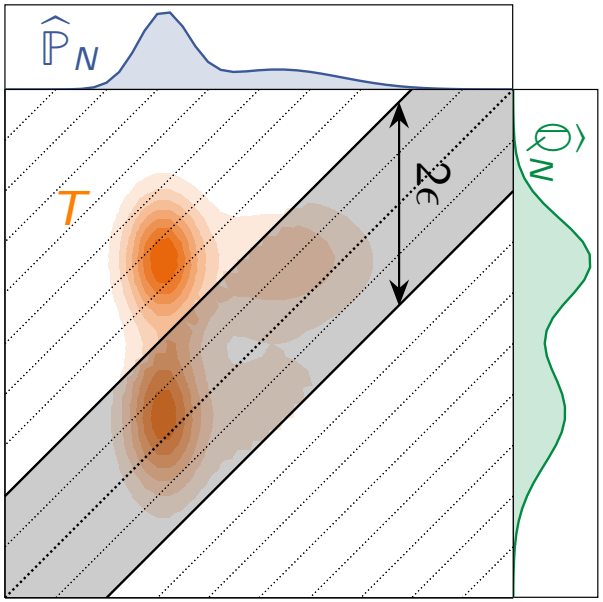
$$\{\xi^1, \xi^2 \ldots, \xi^N\} \qquad \widehat{\mathbb{Q}}_N$$

$$\Downarrow \quad \epsilon\text{-Noise}$$

$$\{\xi^1 + \delta^1, \xi^2 + \delta^2, \ldots, \xi^N + \delta^N\}$$

$$\Downarrow \quad \alpha\text{-Contamination}$$

$$\{\xi^1 + \delta^1, \xi'^2, \ldots, \xi^N + \delta^N\} \qquad \widehat{\mathbb{P}}_N$$



Prop: If $\widehat{\mathbb{P}}_N$ and $\widehat{\mathbb{Q}}_N$ differ by ϵ-norm bounded noise and α-contamination. Then

$$\mathrm{LP}_\epsilon(\widehat{\mathbb{P}}_N, \widehat{\mathbb{Q}}_N) := \inf_{T \in \mathcal{T}(\widehat{\mathbb{P}}_N, \widehat{\mathbb{Q}}_N)} \int \mathbb{1}\{\|\xi - \xi'\| \geq \epsilon\} \, \mathrm{d}T \leq \alpha.$$

Distributions with marginals $\widehat{\mathbb{P}}_N$ and $\widehat{\mathbb{Q}}_N$

Lévy-Prokhorov Distance

$$\xi \sim \mathbb{P}$$

$$\Downarrow \quad \text{Statistical Error}$$

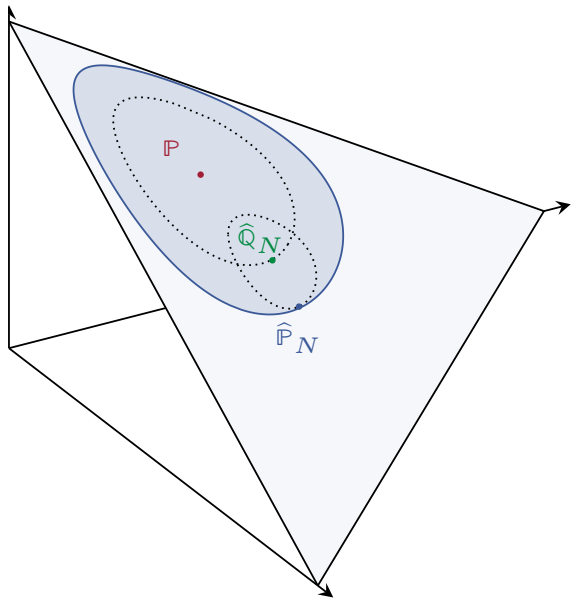
$$\{\xi^1, \xi^2 \ldots, \xi^N\} \qquad \widehat{\mathbb{Q}}_N$$

$$\Downarrow \quad \epsilon\text{-Noise}$$

$$\{\xi^1 + \delta^1, \xi^2 + \delta^2, \ldots, \xi^N + \delta^N\}$$

$$\Downarrow \quad \alpha\text{-Contamination}$$

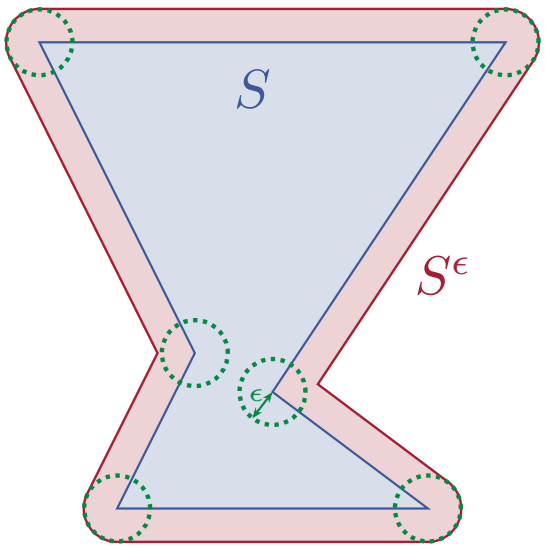
$$\{\xi^1 + \delta^1, \xi'^2, \ldots, \xi^N + \delta^N\} \qquad \widehat{\mathbb{P}}_N$$



Prop: If $\widehat{\mathbb{P}}_N$ and $\widehat{\mathbb{Q}}_N$ differ by ϵ-norm bounded noise and α-contamination. Then

$$\mathrm{LP}_\epsilon(\widehat{\mathbb{P}}_N, \widehat{\mathbb{Q}}_N) := \inf_{T \in \mathcal{T}(\widehat{\mathbb{P}}_N, \widehat{\mathbb{Q}}_N)} \int \mathbb{1}\{\|\xi - \xi'\| \geq \epsilon\} \, \mathrm{d}T \leq \alpha.$$

Intermezzo : Lévy-Prokhorov Distance

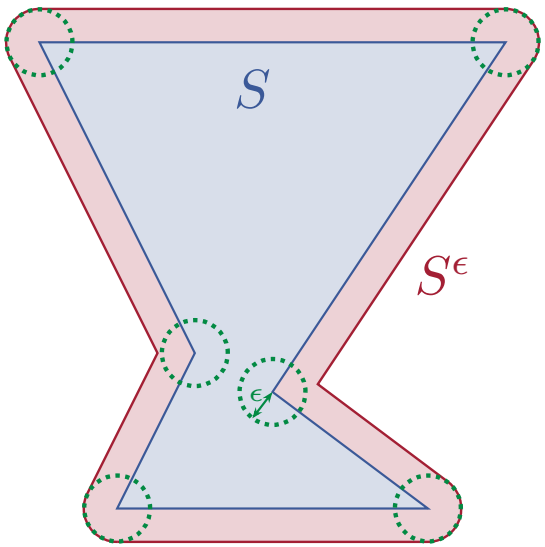


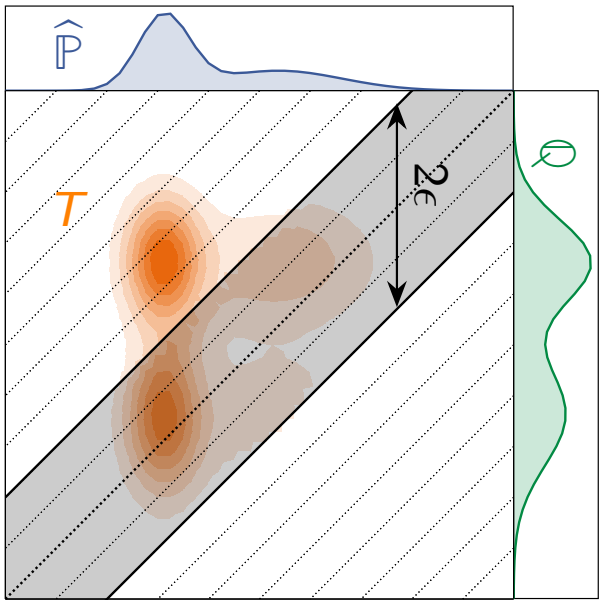
- **Prokhorov** in 1956: Metrize the topology of weak convergence on metric space using

$$\left\{\mathbb{Q} \ : \ \mathrm{LP}_\epsilon(\widehat{\mathbb{P}}, \mathbb{Q}) < \alpha\right\} = \left\{\mathbb{Q} \ : \ \mathbb{Q}(S) \leq \widehat{\mathbb{P}}(S^\epsilon) + \alpha \ \ \forall S\right\}$$

where $S^\epsilon := \{\xi' \in \Xi \ : \ \exists \xi \in S, \ \|\xi' - \xi\| \leq \epsilon\}$ and $\alpha = \epsilon$.

Intermezzo : Lévy-Prokhorov Distance





- **Prokhorov** in 1956: Metrize the topology of weak convergence on metric space using

$$\left\{\mathbb{Q} \;:\; \mathrm{LP}_\epsilon(\widehat{\mathbb{P}},\mathbb{Q}) < \alpha\right\} = \left\{\mathbb{Q} \;:\; \mathbb{Q}(S) \le \widehat{\mathbb{P}}(S^\epsilon) + \alpha \;\; \forall S\right\}$$
$$= \left\{\mathbb{Q} \;:\; \exists T \in \mathcal{T}(\widehat{\mathbb{P}},\mathbb{Q}),\; \int \mathbb{1}\{\|\xi - \xi'\| > \epsilon\}\, \mathrm{d}\, T(\xi,\xi') < \alpha\right\}.$$

  where $S^\epsilon := \{\xi' \in \Xi \;:\; \exists \xi \in S,\; \|\xi' - \xi\| \le \epsilon\}$ and $\alpha = \epsilon$.

- **Strassen** in 1965: Established an optimal transport characterization

Holistic Robust Formulation

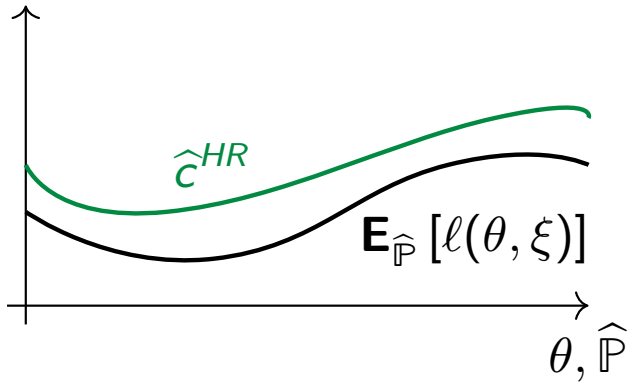
Holistic Robust (HR) Formulation:

$$\widehat{c}^{HR}(\theta, \widehat{\mathbb{P}}) = \max \mathbf{E}_{\mathbb{P}'}\left[\ell(\theta, \xi)\right]$$
$$\text{s.t. } \mathbb{Q}, \mathbb{P}' \in \mathcal{P},$$
$$\mathrm{LP}_{\epsilon}(\widehat{\mathbb{P}}, \mathbb{Q}) \leq \alpha, \quad \text{Noise and contamination}$$
$$\mathrm{KL}(\mathbb{Q}, \mathbb{P}') \leq r. \quad \text{Statistical error}$$

Holistic Robust Formulation

Holistic Robust (HR) Formulation:

$$\begin{aligned}\widehat{c}^{HR}(\theta, \widehat{\mathbb{P}}) = \max\ & \mathbf{E}_{\mathbb{P}'}\left[\ell(\theta, \xi)\right] \\ \text{s.t. } & \mathbb{Q}, \mathbb{P}' \in \mathcal{P}, \\ & \mathrm{LP}_{\epsilon}(\widehat{\mathbb{P}}, \mathbb{Q}) \leq \alpha, \\ & \mathrm{KL}(\mathbb{Q}, \mathbb{P}') \leq r.\end{aligned}$$



Thm (Bennouna and Van Parys, 2022): Consider ϵ-norm bounded noise and α-contamination. Then,

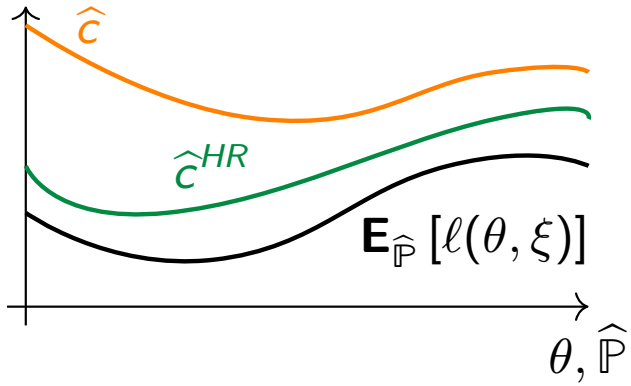
- (OOS Guarantee) we have

$$\mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}^{HR}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right] \quad \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall \mathbb{P}.$$

Holistic Robust Formulation

Holistic Robust (HR) Formulation:

$$\begin{aligned}\widehat{c}^{HR}(\theta, \widehat{\mathbb{P}}) = \max\ & \mathbf{E}_{\mathbb{P}'}\left[\ell(\theta, \xi)\right] \\ \text{s.t. } & \mathbb{Q}, \mathbb{P}' \in \mathcal{P}, \\ & \mathrm{LP}_{\epsilon}(\widehat{\mathbb{P}}, \mathbb{Q}) \leq \alpha, \\ & \mathrm{KL}(\mathbb{Q}, \mathbb{P}') \leq r.\end{aligned}$$



Thm (Bennouna and Van Parys, 2022): Consider ϵ-norm bounded noise and α-contamination. Then,

- (OOS Guarantee) we have

$$\mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}^{HR}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right] \quad \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall \mathbb{P}.$$

- (Efficiency) Furthermore, let $\widehat{c}$ satisfy

$$\mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right] \quad \forall \theta \in \Theta\right] \geq 1 - \exp(-rN + o(N)) \quad \forall \mathbb{P}.$$

Then, $\widehat{c}(\theta, \widehat{\mathbb{P}}) \geq \widehat{c}^{HR}(\theta, \widehat{\mathbb{P}})$ for all θ and $\widehat{\mathbb{P}}$.

Tractability

Even for $r = 0$, $\alpha = \epsilon > 0$ no exact algorithm was known (Erdoğan and Iyengar, 2006).

Tractability

Even for $r = 0$, $\alpha = \epsilon > 0$ no exact algorithm was known (Erdoğan and Iyengar, 2006).

HR Robust formulation is as tractable as Wasserstein robust formulation.

Tractability

Even for $r = 0$, $\alpha = \epsilon > 0$ no exact algorithm was known (Erdoğan and Iyengar, 2006).

HR Robust formulation is as tractable as Wasserstein robust formulation.

Thm: We have

$$\begin{aligned}\widehat{c}^{HR}(\theta, \widehat{\mathbb{P}}) = \max\ & \textstyle\sum_{i=1}^{N} p_i' \ell^{\epsilon}(\theta, \xi^i) + p_0' \ell^{\infty}(\theta) \\ \text{s.t. } & p' \in \mathrm{R}_+^{N+1},\ q \in \mathrm{R}_+^{N+1},\ s \in \mathrm{R}_+^{N}, \\ & \textstyle\sum_{i=0}^{N} p_i' = \sum_{i=0}^{N} q_i = 1,\ q_i + s_i = \widehat{\mathbb{P}}(\xi^i) \quad \forall i = 1, \ldots, N, \\ & \textstyle\sum_{i=0}^{N} q_i \log(\frac{q_i}{p_i'}) \leq r,\ \sum_{i=1}^{N} s_i \leq \alpha.\end{aligned}$$

with

- Inflated loss $\ell^{\epsilon}(\theta, \xi) = \max_{\|\xi' - \xi\| \leq \epsilon,\, \xi' \in \Xi} \ell(\theta, \xi')$.
- Worst-case loss $\ell^{\infty}(\theta) = \max_{\xi' \in \Xi} \ell(\theta, \xi')$

Table of Contents

Huber's TV Contamination

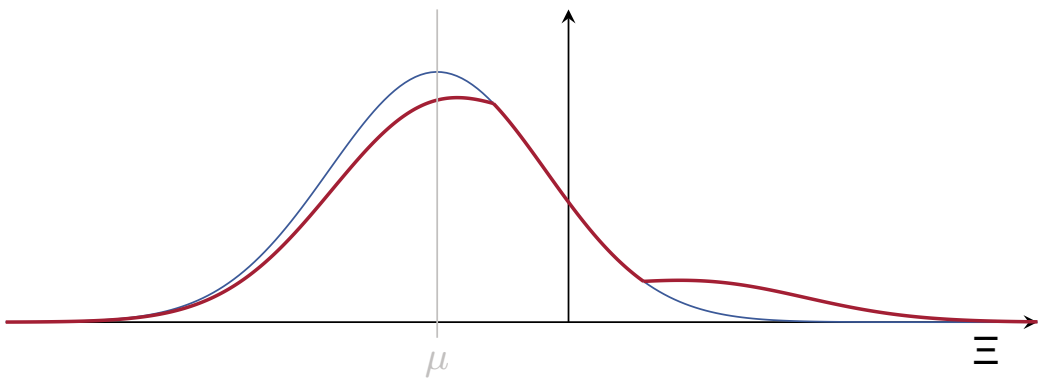
Contamination

Small fraction $\alpha \in [0, \frac{1}{2})$ of probability mass can be shifted

$$\xi \sim \mathbb{G}_\mu$$

$$\Downarrow$$

$$\xi \sim \mathbb{P}^c$$



$$\mathrm{TV}(\mathbb{G}_\mu, \mathbb{P}^c) = \tfrac{1}{2} \int |g_\mu(\xi) - p^c(\xi)| \, \mathrm{d}\xi \leq \alpha$$

Statistical Error

Randomness of the sampled data

$$\xi \sim \mathbb{P}^c$$

$$\Downarrow$$

$$\{\xi^1, \ldots, \xi^N\}$$

Location Estimation (Huber, 1964)

Let $\{\xi^i \sim \mathbb{P}^c\}_{i=1}^N$.

- Classical location estimate is

$$\widehat{\mu}^{\text{avg}}(\widehat{\mathbb{P}}_N) = \sum_{i=1}^N \frac{1}{n} \cdot \xi^i$$

- Very sensitive to corruption, i.e., mapping $\widehat{\mathbb{P}} \to \mu^{\text{avg}}(\widehat{\mathbb{P}})$ is discontinuous.

Location Estimation (Huber, 1964)

Let $\{\xi^i \sim \mathbb{P}^c\}_{i=1}^N$.

- Classical location estimate is

$$\widehat{\mu}^{\text{avg}}(\widehat{\mathbb{P}}_N) = \sum_{i=1}^N \tfrac{1}{n} \cdot \xi^i \iff \sum_{i=1}^N \tfrac{1}{n} \cdot (\xi^i - \widehat{\mu}^{\text{avg}}(\widehat{\mathbb{P}}_N)) = 0.$$

- Very sensitive to corruption, i.e., mapping $\widehat{\mathbb{P}} \to \mu^{\text{avg}}(\widehat{\mathbb{P}})$ is discontinuous.

Location Estimation (Huber, 1964)

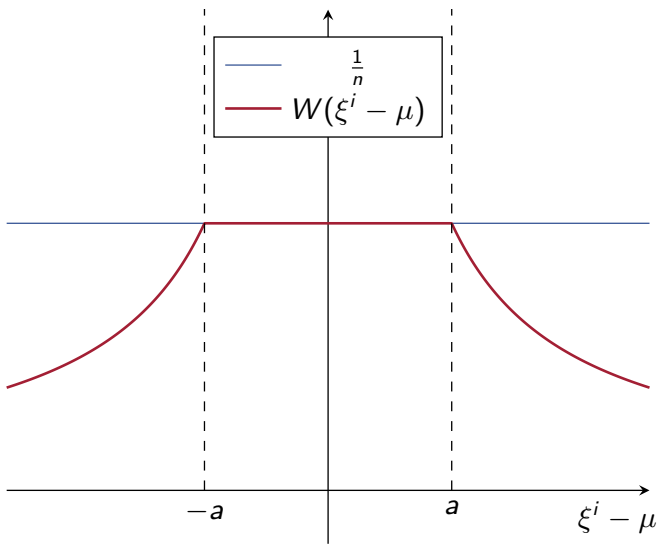
Let $\{\xi^i \sim \mathbb{P}^c\}_{i=1}^N$.

- Classical location estimate is

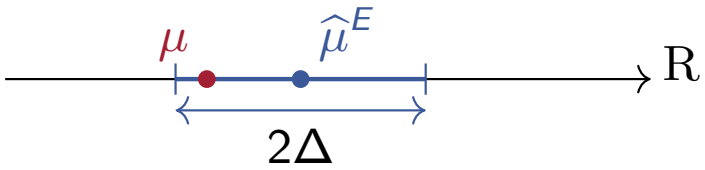
$$\widehat{\mu}^{\text{avg}}(\widehat{\mathbb{P}}_N) = \sum_{i=1}^N \tfrac{1}{n} \cdot \xi^i \iff \sum_{i=1}^N \tfrac{1}{n} \cdot (\xi^i - \widehat{\mu}^{\text{avg}}(\widehat{\mathbb{P}}_N)) = 0.$$

- Very sensitive to corruption, i.e., mapping $\widehat{\mathbb{P}} \to \mu^{\text{avg}}(\widehat{\mathbb{P}})$ is discontinuous.
- Huber's proposes instead

$$\sum_{i=1}^N W(\xi^i - \widehat{\mu}^{\text{hub}}(\widehat{\mathbb{P}}_N)) \cdot (\xi^i - \widehat{\mu}^{\text{hub}}(\widehat{\mathbb{P}}_N)) = 0.$$



Error Margin

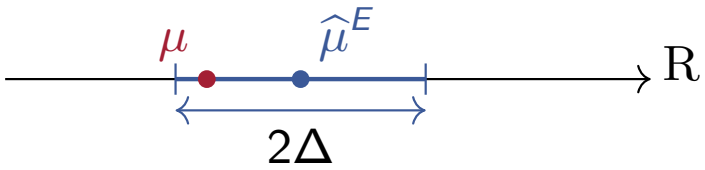


Take the **error margin** $\Delta(\widehat{\mu}^E)$ of an estimator $\widehat{\mu}^E$ to be the smallest $\Delta \geq 0$ for which

$$\text{Prob}_{\mathbb{P}^c}[|\widehat{\mu}^E(\widehat{\mathbb{P}}_N) - \mu| \leq \Delta] \geq 1 - \exp(-rN + o(N))$$

for all $\mu \in \mathrm{R}$, $\mathbb{P}^c \in \mathcal{P}$ such that $\mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha$.

Error Margin



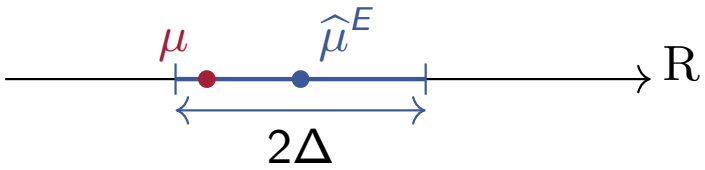
Take the **error margin** $\Delta(\widehat{\mu}^E)$ of an estimator $\widehat{\mu}^E$ to be the smallest $\Delta \geq 0$ for which

$$\mathrm{Prob}_{\mathbb{P}^c}[|\widehat{\mu}^E(\widehat{\mathbb{P}}_N) - \mu| \leq \Delta] \geq 1 - \exp(-rN + o(N))$$

for all $\mu \in \mathrm{R}$, $\mathbb{P}^c \in \mathcal{P}$ such that $\mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha$.

Remark: For $\alpha > 0$ we get $+\infty = \Delta(\widehat{\mu}^{avg}) > \Delta(\widehat{\mu}^{median})$.

Error Margin



Take the **error margin** $\Delta(\widehat{\mu}^E)$ of an estimator $\widehat{\mu}^E$ to be the smallest $\Delta \geq 0$ for which

$$\mathrm{Prob}_{\mathbb{P}^c}[|\widehat{\mu}^E(\widehat{\mathbb{P}}_N) - \mu| \leq \Delta] \geq 1 - \exp(-rN + o(N))$$

for all $\mu \in \mathrm{R}$, $\mathbb{P}^c \in \mathcal{P}$ such that $\mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha$.

Remark: For $\alpha > 0$ we get $+\infty = \Delta(\widehat{\mu}^{avg}) > \Delta(\widehat{\mu}^{median})$.

Thm: (Huber, 1964) We have that $\Delta(\widehat{\mu}^E) \geq \Delta(\widehat{\mu}^{\mathrm{hub}})$ for all estimators $\widehat{\mu}^E$.

- Huber's estimator yields "minimal" confidence **intervals**.

Confidence Set Estimators

Our holistic robust perspective suggests the confidence **set**

$$\widehat{S}^{hr}(\widehat{\mathbb{P}}_N) = \left\{ \mu \in \mathrm{R} \ : \begin{array}{l} \exists \mathbb{P}^c \in \mathcal{P}, \\ \mathrm{KL}(\widehat{\mathbb{P}}_N, \mathbb{P}^c) \leq r, \\ \mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha \end{array} \right\}.$$

Confidence Set Estimators

Our holistic robust perspective suggests the confidence **set**

$$\widehat{S}^{hr}(\widehat{\mathbb{P}}_N) = \left\{ \mu \in \mathrm{R} \ : \begin{array}{l} \exists \mathbb{P}^c \in \mathcal{P}, \\ \mathrm{KL}(\widehat{\mathbb{P}}_N, \mathbb{P}^c) \leq r, \\ \mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha \end{array} \right\}.$$

Confidence Set Estimators

Our holistic robust perspective suggests the confidence **set**

$$\widehat{S}^{hr}(\widehat{\mathbb{P}}_N) = \left\{ \mu \in \mathrm{R} \ : \begin{array}{l} \exists \mathbb{P}^c \in \mathcal{P}, \\ \mathrm{KL}(\widehat{\mathbb{P}}_N, \mathbb{P}^c) \leq r, \\ \mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha \end{array} \right\}.$$

Thm (Informal):

- (OOS Guarantee) We have

$$\mathrm{Prob}_{\mathbb{P}^c}[\mu \in \widehat{S}^{hr}(\widehat{\mathbb{P}}_N)] \geq 1 - \exp(-rN + o(N))$$

for all $\mu \in \mathrm{R}$, $\mathbb{P}^c \in \mathcal{P}$ so that $\mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha$.

Confidence Set Estimators

Our holistic robust perspective suggests the confidence **set**

$$\widehat{S}^{hr}(\widehat{\mathbb{P}}_N) = \left\{ \mu \in \mathrm{R} \ : \begin{array}{l} \exists \mathbb{P}^c \in \mathcal{P}, \\ \mathrm{KL}(\widehat{\mathbb{P}}_N, \mathbb{P}^c) \leq r, \\ \mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha \end{array} \right\}.$$

Thm (Informal):

- (OOS Guarantee) We have

$$\mathrm{Prob}_{\mathbb{P}^c}[\mu \in \widehat{S}^{hr}(\widehat{\mathbb{P}}_N)] \geq 1 - \exp(-rN + o(N))$$

  for all $\mu \in \mathrm{R}$, $\mathbb{P}^c \in \mathcal{P}$ so that $\mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha$.

- (Efficiency) Furthermore, let $\widehat{S}$ satisfy

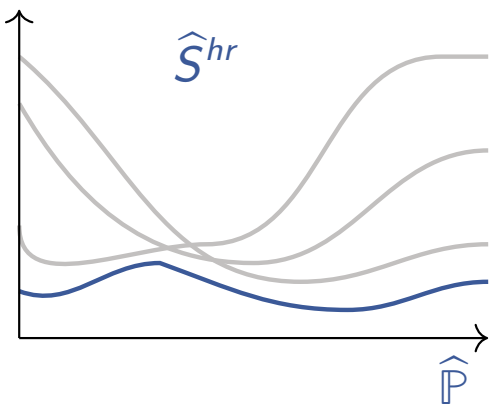
$$\mathrm{Prob}_{\mathbb{P}^c}[\mu \in \widehat{S}(\widehat{\mathbb{P}}_N)] \geq 1 - \exp(-rN + o(N))$$

  for all $\mu \in \mathrm{R}$, $\mathbb{P}^c \in \mathcal{P}$ so that $\mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha$. Then, $\widehat{S}^{hr}(\widehat{\mathbb{P}}) \subseteq \widehat{S}(\widehat{\mathbb{P}})$ for all $\widehat{\mathbb{P}}$.

Meta Optimization Problem

Among all confidence sets with **out-of-sample guarantee** find the **smallest** one

$$\begin{cases} \min_{\widehat{S}}[\subseteq]\ \widehat{S} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}^c}[\mu \in \widehat{S}(\widehat{\mathbb{P}}_n)] \geq 1 - \exp(-rN + o(N)) \\ \qquad\qquad \forall \mu \in \mathrm{R},\ \mathbb{P}^c \in \mathcal{P} \text{ st } \mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha. \end{cases}$$

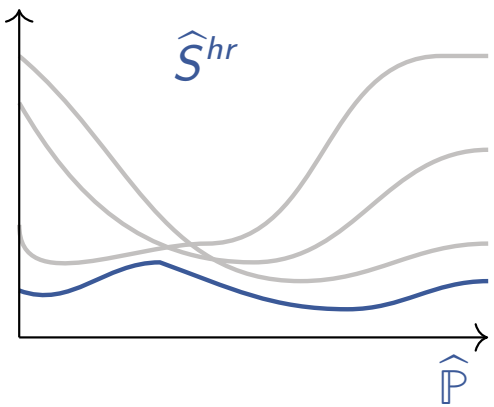


Pareto dominant solution

Meta Optimization Problem

Among all confidence sets with **out-of-sample guarantee** find the **smallest** one

$$\begin{cases} \min_{\widehat{S}}[\subseteq]\ \widehat{S} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}^c}[\mu \in \widehat{S}(\widehat{\mathbb{P}}_n)] \geq 1 - \exp(-rN + o(N)) \\ \qquad\qquad \forall \mu \in \mathrm{R},\ \mathbb{P}^c \in \mathcal{P} \text{ st } \mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha. \end{cases}$$

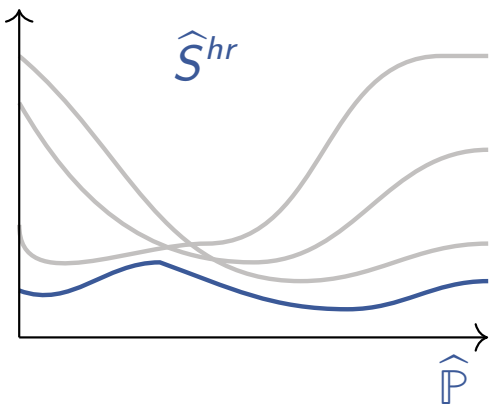


Pareto dominant solution

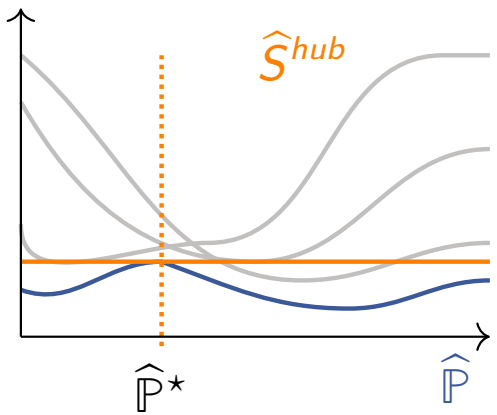
Meta Optimization Problem

Among all confidence sets with **out-of-sample guarantee** find the **smallest** one

$$\begin{cases} \min_{\widehat{S}}[\subseteq]\ \widehat{S} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}^c}[\mu \in \widehat{S}(\widehat{\mathbb{P}}_n)] \geq 1 - \exp(-rN + o(N)) \\ \qquad\qquad \forall \mu \in \mathrm{R},\ \mathbb{P}^c \in \mathcal{P} \text{ st } \mathrm{TV}(\mathbb{P}^c, \mathbb{G}_\mu) \leq \alpha. \end{cases}$$



Pareto dominant solution

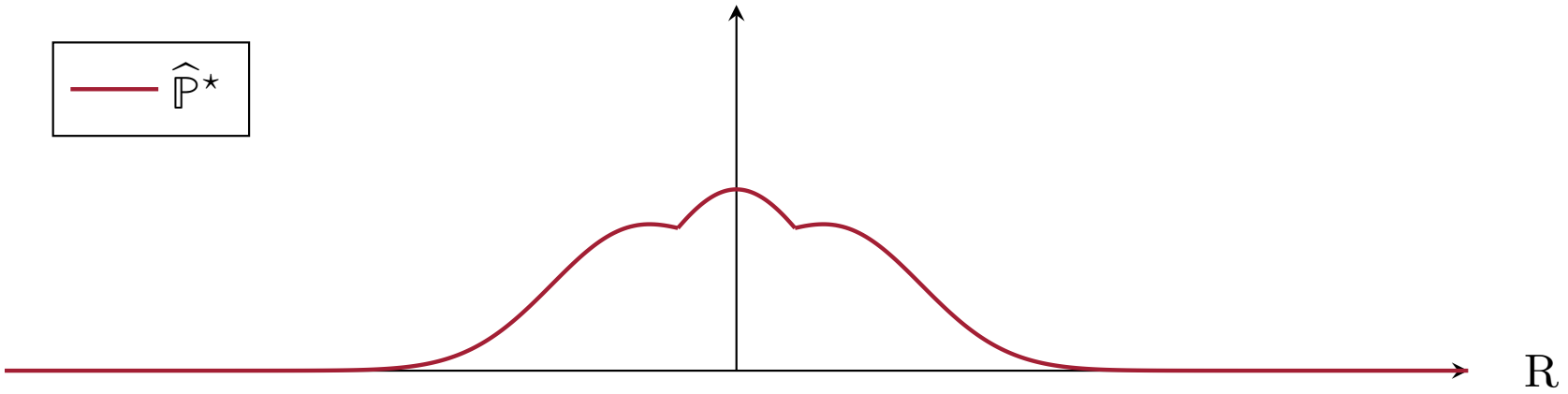


Minimax solution

Minimax Solution

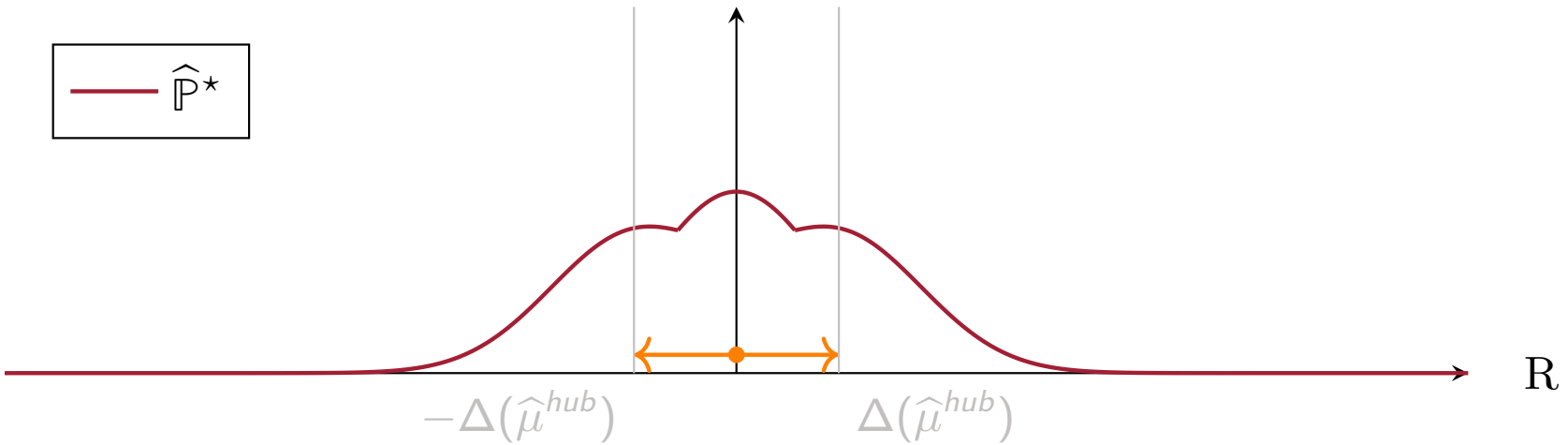
Thm: There exists $\widehat{\mathbb{P}}^\star$ so that we have $\mathrm{diam}(\widehat{S}^{hr}(\widehat{\mathbb{P}}^\star)) = 2\Delta(\widehat{\mu}^{hub})$.

Minimax Solution



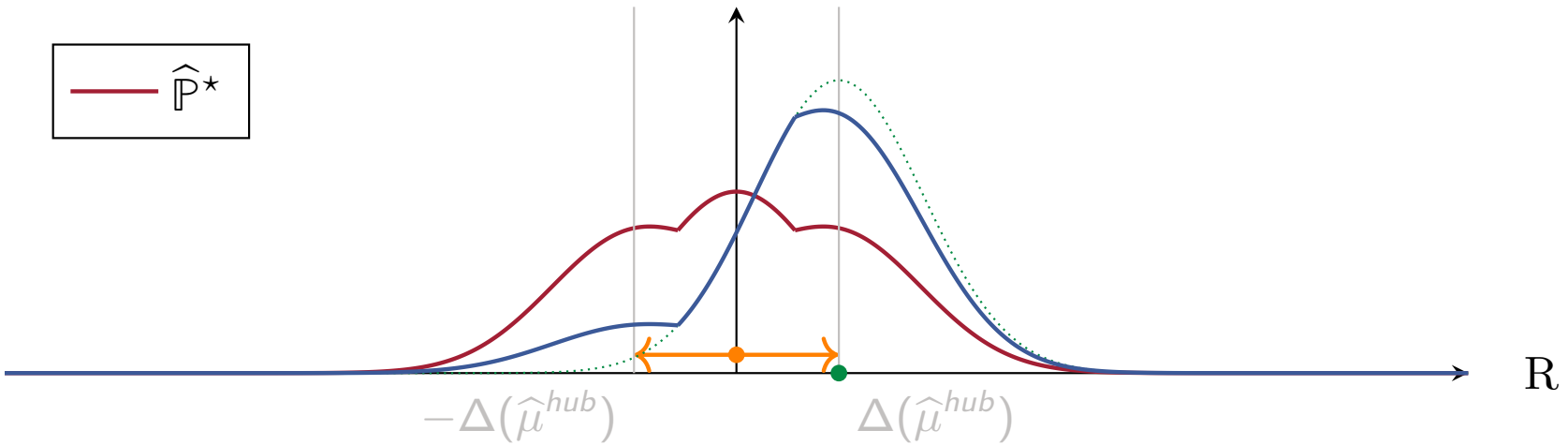
Thm: There exists $\widehat{\mathbb{P}}^\star$ so that we have $\mathrm{diam}(\widehat{S}^{hr}(\widehat{\mathbb{P}}^\star)) = 2\Delta(\widehat{\mu}^{hub})$.

Minimax Solution



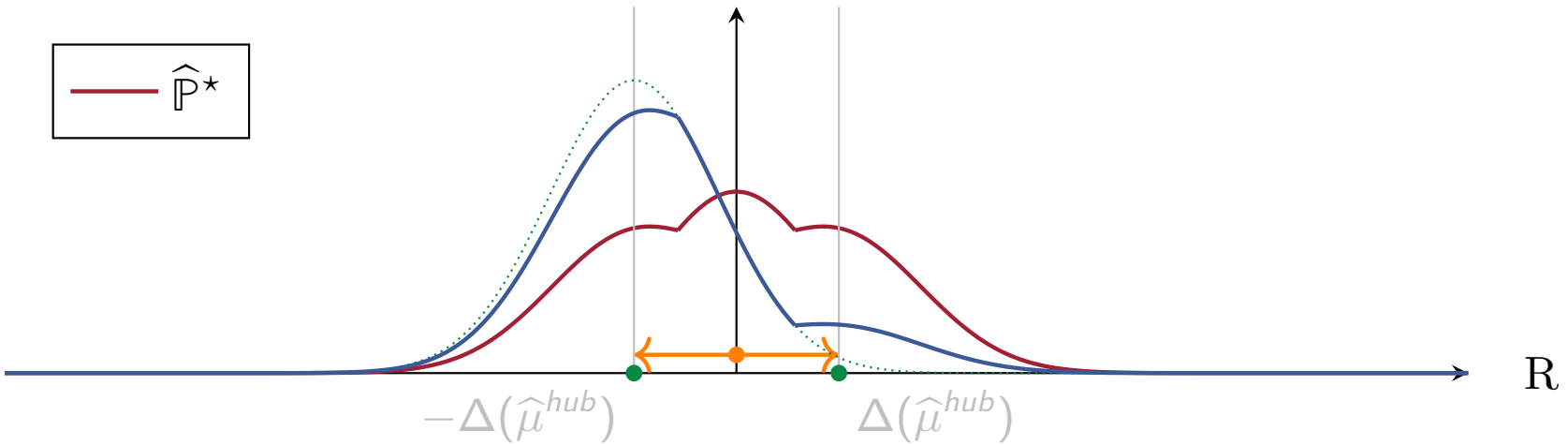
Thm: There exists $\widehat{\mathbb{P}}^\star$ so that we have $\mathrm{diam}(\widehat{S}^{hr}(\widehat{\mathbb{P}}^\star)) = 2\Delta(\widehat{\mu}^{hub})$.

Minimax Solution



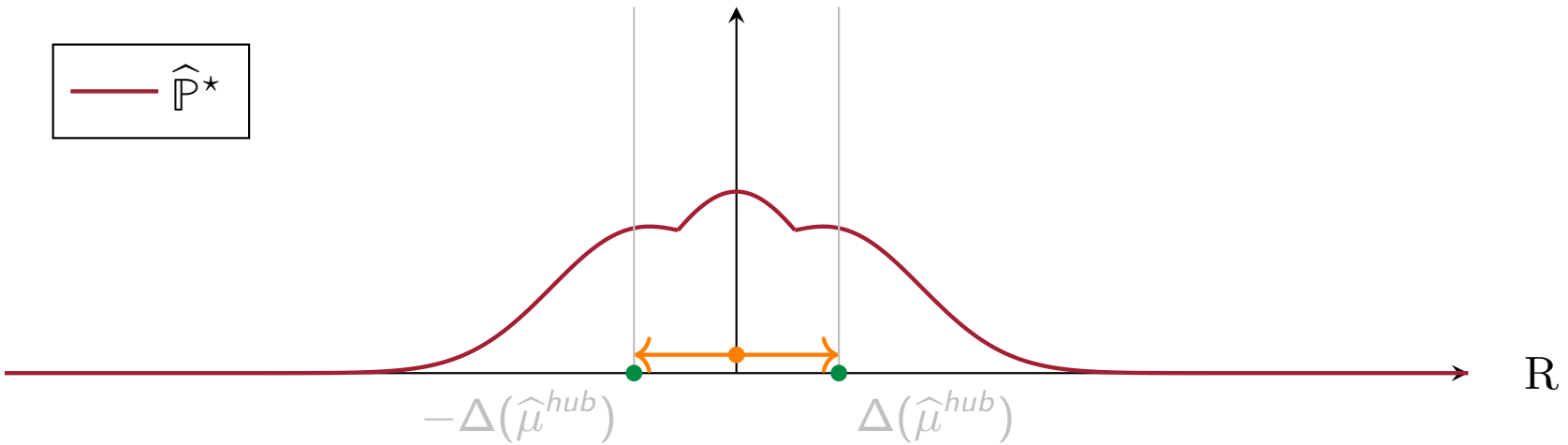
Thm: There exists $\widehat{\mathbb{P}}^\star$ so that we have $\mathrm{diam}(\widehat{S}^{hr}(\widehat{\mathbb{P}}^\star)) = 2\Delta(\widehat{\mu}^{hub})$.

Minimax Solution



Thm: There exists $\widehat{\mathbb{P}}^\star$ so that we have $\mathrm{diam}(\widehat{S}^{hr}(\widehat{\mathbb{P}}^\star)) = 2\Delta(\widehat{\mu}^{hub})$.

Minimax Solution



Thm: There exists $\widehat{\mathbb{P}}^\star$ so that we have $\mathrm{diam}(\widehat{S}^{hr}(\widehat{\mathbb{P}}^\star)) = 2\Delta(\widehat{\mu}^{hub})$.

Remarks

- We may have $\widehat{S}^{hr}(\widehat{\mathbb{P}}) \subset \mu^{hub}(\widehat{\mathbb{P}}) + [-\Delta(\widehat{\mu}^{hub}), \Delta(\widehat{\mu}^{hub})]$ for $\widehat{\mathbb{P}} \neq \widehat{\mathbb{P}}^\star$.
- However, the estimator $\widehat{S}^{hr}$ may be hard to evaluate.

Table of Contents

Summary

1. Robust optimization can make data-driven decisions efficiently

2. Holistic robust optimization additionally protects decisions against corruption

3. Parametric DRO is intimately related to classical robust statistics

References I

Bennouna, Amine and Bart Van Parys (2021). "Learning and decision-making with data: Optimal formulations and phase transitions". In: **arXiv preprint arXiv:2109.06911**.

Bennouna, Amine and Bart Van Parys (2022). "Holistic Robust Data-Driven Decisions". In: **arXiv preprint arXiv:2207.09560**.

Bertsimas, Dimitris and Ioana Popescu (2005). "Optimal inequalities in probability theory: A convex optimization approach". In: **SIAM Journal on Optimization** 15.3, pp. 780–804.

Birge, John R and Francois Louveaux (2011). **Introduction to stochastic programming**. Springer Science & Business Media.

Dharmadhikari, Sudhakar and Kumar Joag-Dev (1988). **Unimodality, convexity, and applications**. Elsevier.

Erdoğan, Emre and Garud Iyengar (2006). "Ambiguous chance constrained problems and robust optimization". In: **Mathematical Programming** 107, pp. 37–61.

Fournier, Nicolas and Arnaud Guillin (2015). "On the rate of convergence in Wasserstein distance of the empirical measure". In: **Probability theory and related fields** 162.3-4, pp. 707–738.

References II

- Huber, Peter J. (1964). "Robust estimation of a location parameter". In: **Ann. Math. Statist.** 35, pp. 73–101.
- Michaud, Richard O. (1989). "The Markowitz optimization enigma: Is "optimized" optimal?" In: **Financial analysts journal** 45.1, pp. 31–42.
- Roald, Line, Frauke Oldewurtel, Bart Van Parys, and Göran Andersson (2015). "Security constrained optimal power flow with distributionally robust chance constraints". In: **arXiv preprint arXiv:1508.06061**.
- Sanov, Ivan N (1958). **On the probability of large deviations of random variables**. United States Air Force, Office of Scientific Research.
- Strassen, Volker (1965). "The existence of probability measures with given marginals". In: **The Annals of Mathematical Statistics** 36.2, pp. 423–439.
- Sutter, Tobias, Bart Van Parys, and Daniel Kuhn (2020). "A general framework for optimal data-driven optimization". In: **arXiv preprint arXiv:2010.06606**.
- Van Parys, Bart, P. Esfahani Mohajerin Esfahani, and D. Kuhn (2020). "From data to decisions: Distributionally robust optimization is optimal". In: **Management Science**.
- Van Parys, Bart, Paul Goulart, and Paul Embrechts (2016). "Fréchet inequalities via convex optimization". In: **Available on Optimization Online**.

References III


Van Parys, Bart, Paul Goulart, and Manfred Morari (2019). "Distributionally robust expectation inequalities for structured distributions". In: **Mathematical Programming** 173, pp. 251–280.

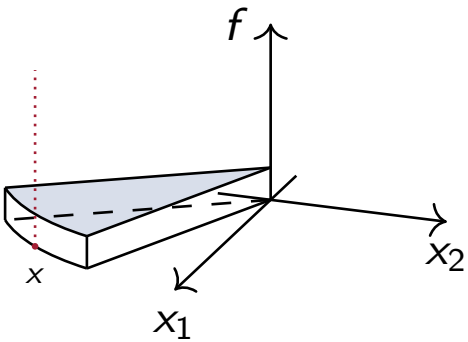

Van Parys, Bart, Paul J. Goulart, and Daniel Kuhn (2016). "Generalized Gauss inequalities via semidefinite programming". In: **Mathematical Programming** 156.1-2, pp. 271–302.

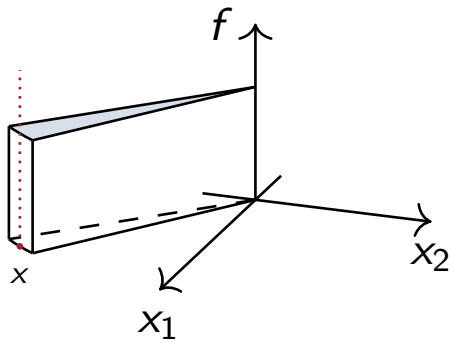

Van Parys, Bart, Daniel Kuhn, Paul J Goulart, and Manfred Morari (2015). "Distributionally robust control of constrained stochastic systems". In: **IEEE Transactions on Automatic Control** 61.2, pp. 430–442.


Van Parys, Bart, Bing Feng Ng, Paul Goulart, and Rafael Palacios (2014). "Optimal control for load alleviation in wind turbines". In: **32nd ASME Wind Energy Symposium**, p. 1222.

Table of Contents

Extreme Unimodal Distributions





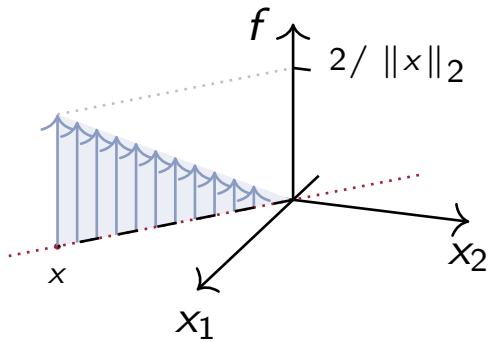
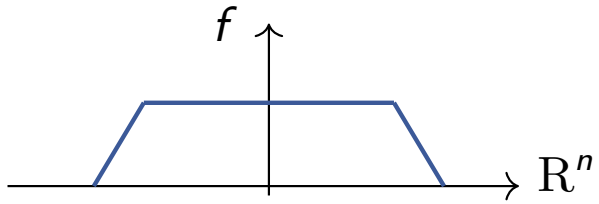


Table of Contents

Structured Moment Sets $\mathcal{U}_\alpha(\mu, S) := \{\mathbb{Q} \in \mathcal{S}_\alpha : \int \xi \mathrm{d}\mathbb{Q} = \mu,\ \int \xi\xi^\top \mathrm{d}\mathbb{Q} = S\}$

- Reduce pessimism by imposing additional structure

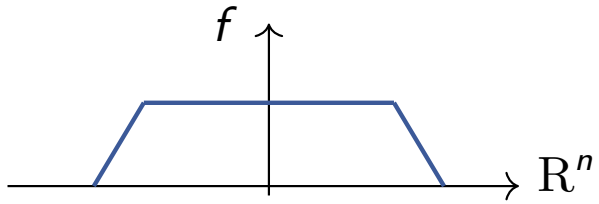


BvP, Goulart, Kuhn (2016). "Generalized Gauss inequalities via semidefinite programming". **Mathematical Programming**

BvP, Goulart, Morari (2019). "Distributionally robust expectation inequalities for structured distributions". **Mathematical Programming**

Structured Moment Sets $\mathcal{U}_\alpha(\mu, S) := \{\mathbb{Q} \in \mathcal{S}_\alpha : \int \xi \mathrm{d}\mathbb{Q} = \mu,\ \int \xi\xi^\top \mathrm{d}\mathbb{Q} = S\}$

- Reduce pessimism by imposing additional structure



- **Definition:** If $\mathbb{P}$ has a continuous density function f, then $\mathbb{P}$ is **unimodal** if

$$t \mapsto f(tx)$$

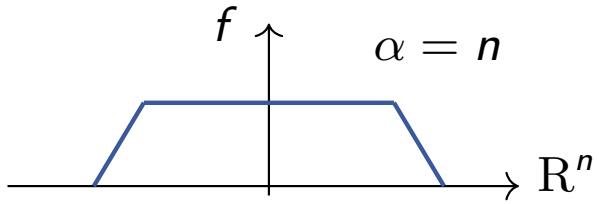
is non-increasing for any $x \in \mathrm{R}^n$ (Dharmadhikari and Joag-Dev, 1988).

BvP, Goulart, Kuhn (2016). "Generalized Gauss inequalities via semidefinite programming". **Mathematical Programming**

BvP, Goulart, Morari (2019). "Distributionally robust expectation inequalities for structured distributions". **Mathematical Programming**

Structured Moment Sets $\mathcal{U}_\alpha(\mu, S) := \{\mathbb{Q} \in \mathcal{S}_\alpha : \int \xi \mathrm{d}\mathbb{Q} = \mu,\ \int \xi\xi^\top \mathrm{d}\mathbb{Q} = S\}$

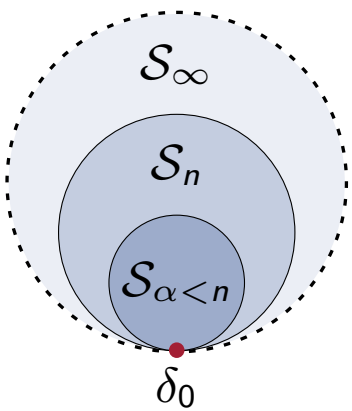
- Reduce pessimism by imposing additional structure



- **Definition:** If $\mathbb{P}$ has a continuous density function f, then $\mathbb{P}$ is α-**unimodal** if

$$t \mapsto f(tx)/t^{\alpha-n}$$

is non-increasing for any $x \in \mathrm{R}^n$ (Dharmadhikari and Joag-Dev, 1988).

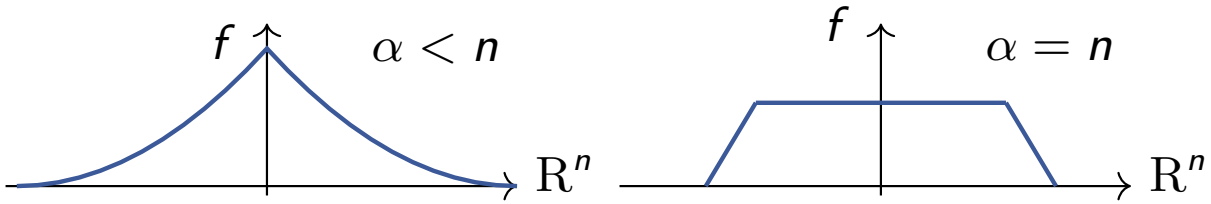


BvP, Goulart, Kuhn (2016). "Generalized Gauss inequalities via semidefinite programming". **Mathematical Programming**

BvP, Goulart, Morari (2019). "Distributionally robust expectation inequalities for structured distributions". **Mathematical Programming**

Structured Moment Sets $\mathcal{U}_\alpha(\mu, S) := \{\mathbb{Q} \in \mathcal{S}_\alpha : \int \xi \mathrm{d}\mathbb{Q} = \mu,\ \int \xi\xi^\top \mathrm{d}\mathbb{Q} = S\}$

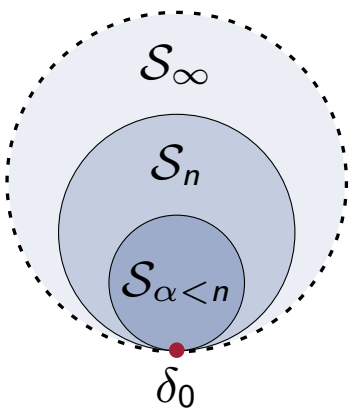
- Reduce pessimism by imposing additional structure



- **Definition:** If $\mathbb{P}$ has a continuous density function f, then $\mathbb{P}$ is α-**unimodal** if

$$t \mapsto f(tx)/t^{\alpha-n}$$

is non-increasing for any $x \in \mathrm{R}^n$ (Dharmadhikari and Joag-Dev, 1988).

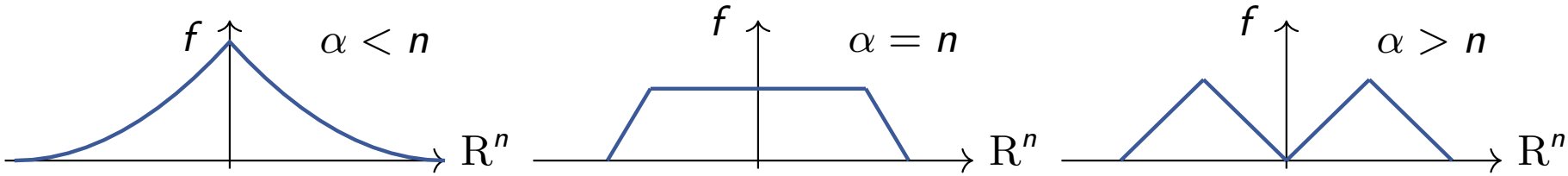


BvP, Goulart, Kuhn (2016). "Generalized Gauss inequalities via semidefinite programming". **Mathematical Programming**

BvP, Goulart, Morari (2019). "Distributionally robust expectation inequalities for structured distributions". **Mathematical Programming**

Structured Moment Sets $\mathcal{U}_\alpha(\mu, S) := \{\mathbb{Q} \in \mathcal{S}_\alpha : \int \xi \mathrm{d}\mathbb{Q} = \mu,\ \int \xi\xi^\top \mathrm{d}\mathbb{Q} = S\}$

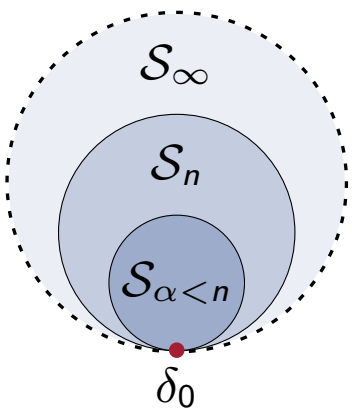
- Reduce pessimism by imposing additional structure



- **Definition:** If $\mathbb{P}$ has a continuous density function f, then $\mathbb{P}$ is α-**unimodal** if

$$t \mapsto f(tx)/t^{\alpha-n}$$

is non-increasing for any $x \in \mathrm{R}^n$ (Dharmadhikari and Joag-Dev, 1988).



BvP, Goulart, Kuhn (2016). "Generalized Gauss inequalities via semidefinite programming". **Mathematical Programming**

BvP, Goulart, Morari (2019). "Distributionally robust expectation inequalities for structured distributions". **Mathematical Programming**

Structured Moment Problems

Thm: Let $P := \left\{\xi \in \mathrm{R}^n \ : \ a_k^\top \xi - b_k < 0 \ \ \forall k = 1, \ldots, K\right\}$ be a polytope. Then,

$$\begin{aligned} &\max \ \mathbb{Q}(\xi \notin P) \\ &\text{s.t.} \ \mathbb{Q} \in \mathcal{U}_\alpha(\widehat{\mu}_N, \widehat{S}_N) \end{aligned}$$

is a convex **tractable optimization** problem.

Structured Moment Problems

Thm: Let $P := \left\{\xi \in \mathrm{R}^n \;:\; a_k^\top \xi - b_k < 0 \;\; \forall k = 1, \ldots, K\right\}$ be a polytope. Then,

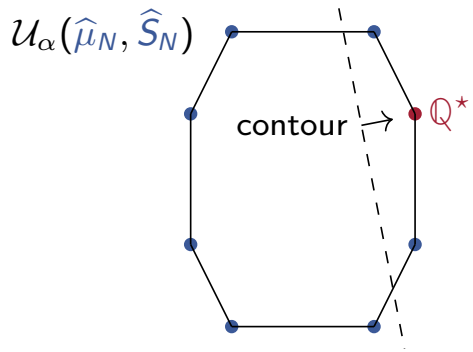
$$\begin{array}{lcll} \max \mathbb{Q}(\xi \notin P) & = & \max & \sum_{k=1}^K \lambda_k \left(1 - \left(b_k \lambda_k / a_k^\top z_k\right)^\alpha\right) \\ \text{s.t. } \mathbb{Q} \in \mathcal{U}_\alpha(\widehat{\mu}_N, \widehat{S}_N) & & \text{s.t.} & \sum_{k=1}^K \begin{pmatrix} Z_k & z_k \\ z_k^\top & \lambda_k \end{pmatrix} \preceq \begin{pmatrix} (\alpha+2)\widehat{S}_n/\alpha & (\alpha+1)\widehat{\mu}_n/\alpha \\ (\alpha+1)\widehat{\mu}_n^\top/\alpha & 1 \end{pmatrix} \\ & & & \begin{pmatrix} Z_k & z_k \\ z_k^\top & \lambda_k \end{pmatrix} \succeq 0 \quad \forall k = 1, \ldots, K \end{array}$$

is a convex **tractable optimization** problem.

Proof Sketch

- We only need to optimize over **extreme distributions**

$$\max \mathbb{Q}(\xi \in P) \qquad = \max \mathbb{Q}(\xi \in P)$$
$$\text{s.t. } \mathbb{Q} \in \mathcal{U}_\alpha(\widehat{\mu}_N, \widehat{S}_N) \qquad \text{s.t. } \mathbb{Q} \in \text{ex}\mathcal{U}_\alpha(\widehat{\mu}_N, \widehat{S}_N)$$



- The set $\mathcal{S}_\alpha$ is a **Choquet Simplex** extreme points $\{\mathbb{e}_x \ : \ x \in \mathrm{R}^n\}$.
- Via Caratheodory's theorem we can establish that in fact

$$\mathbb{Q}^\star = \sum_{k=0}^{K} p_k \mathbb{e}_{x_k} \text{ with } a_k^\top x_k \geq b_k \ \forall k = 1, \ldots, K.$$

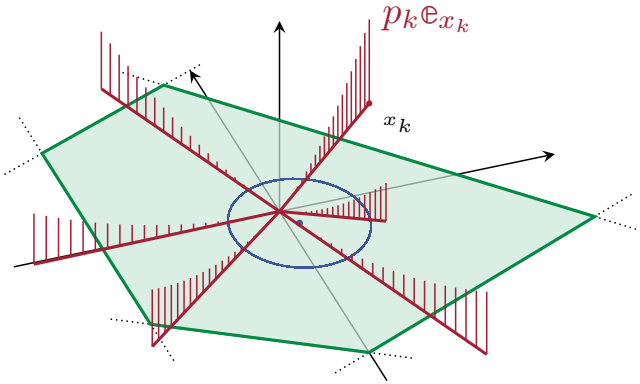
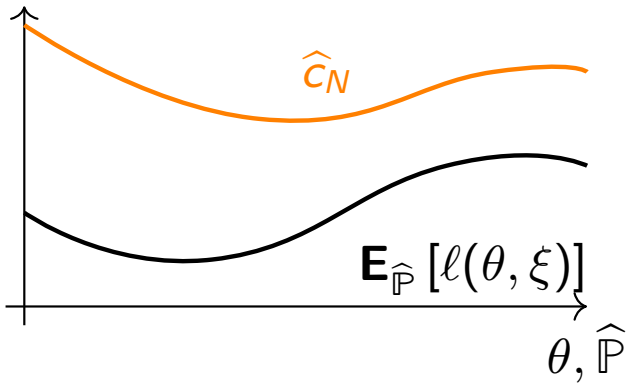


Table of Contents

Meta Optimization Problem

Data driven formulations

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \widehat{c}_N(\theta, \widehat{\mathbb{P}}_N)$$



Feasibility: Consider formulations so that

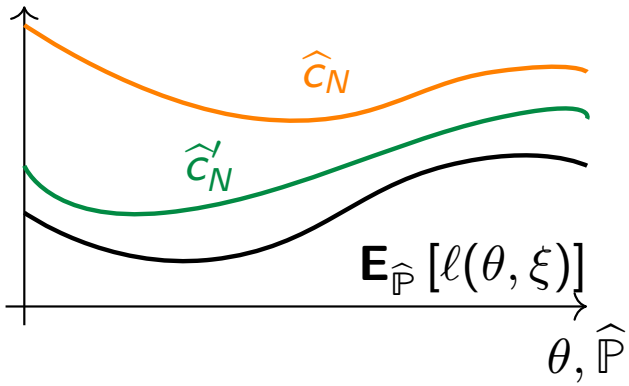
$$\text{Prob}_{\mathbb{P}}\left[\widehat{c}_N(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}[\ell(\theta,\xi)] \ \ \forall\theta\in\Theta\right] \geq 1 - \exp(-a_N + o(a_N)) \quad \forall\mathbb{P}.$$

Bennouna, **BvP** (2023). "Learning and decision-making with data: Optimal formulations and phase transitions". **Mathematical Programming** (Major Revision)

Meta Optimization Problem

Data driven formulations

$$\widehat{\theta}_N \in \arg\min_{\theta\in\Theta} \widehat{c}_N(\theta, \widehat{\mathbb{P}}_N)$$



Feasibility: Consider formulations so that

$$\text{Prob}_{\mathbb{P}}\left[\widehat{c}_N(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}[\ell(\theta,\xi)] \ \ \forall\theta\in\Theta\right] \geq 1 - \exp(-a_N + o(a_N)) \quad \forall\mathbb{P}.$$

Conservatism: Formulation $\widehat{c}'$ is less conservative then $\widehat{c}$ if and only if

$$\widehat{c}' \preceq \widehat{c} \iff \lim_{N\to\infty} \frac{|\widehat{c}'_N(\theta,\widehat{\mathbb{P}}) - \mathbf{E}_{\widehat{\mathbb{P}}}[\ell(\theta,\xi)]|}{|\widehat{c}_N(\theta,\widehat{\mathbb{P}}) - \mathbf{E}_{\widehat{\mathbb{P}}}[\ell(\theta,\xi)]|} \leq 1 \ \ \forall\theta,\widehat{\mathbb{P}}.$$

Bennouna, **BvP** (2023). "Learning and decision-making with data: Optimal formulations and phase transitions". **Mathematical Programming** (Major Revision)

Meta Optimization Problem

Among all predictors with **out-of-sample guarantee** find the **least conservative** one

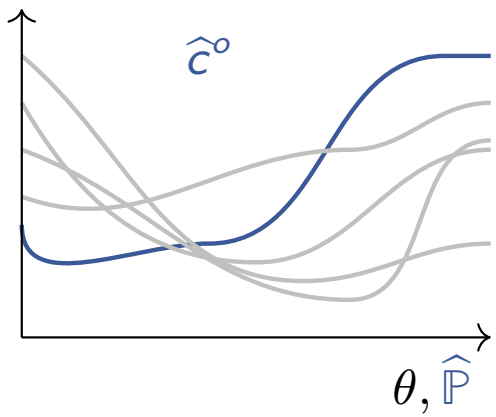
$$\begin{cases} \min\limits_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}_N(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right]\ \forall \theta \in \Theta\right] \geq 1 - \exp(-a_N + o(a_N))\ \ \forall \mathbb{P} \end{cases}$$

Bennouna, **BvP** (2023). "Learning and decision-making with data: Optimal formulations and phase transitions". **Mathematical Programming** (Major Revision)

Meta Optimization Problem

Among all predictors with **out-of-sample guarantee** find the **least conservative** one

$$\begin{cases} \min\limits_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}_N(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right]\ \forall \theta \in \Theta\right] \geq 1 - \exp(-a_N + o(a_N))\ \ \forall \mathbb{P} \end{cases}$$



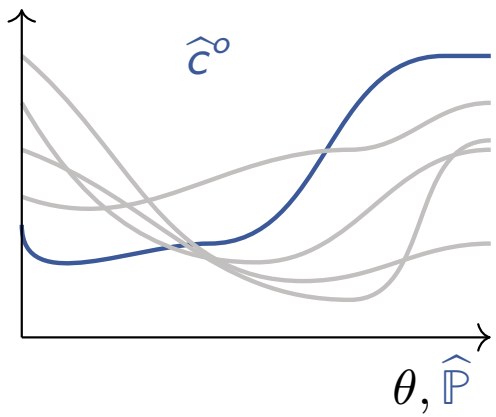
Pareto optimal solution

Bennouna, **BvP** (2023). "Learning and decision-making with data: Optimal formulations and phase transitions". **Mathematical Programming** (Major Revision)

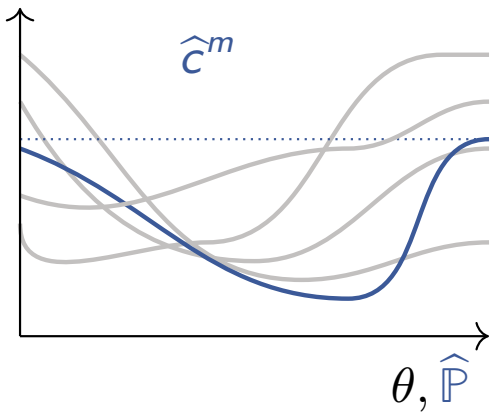
Meta Optimization Problem

Among all predictors with **out-of-sample guarantee** find the **least conservative** one

$$\begin{cases} \min\limits_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}_N(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta,\xi)\right]\ \forall \theta \in \Theta\right] \geq 1 - \exp(-a_N + o(a_N))\ \ \forall \mathbb{P} \end{cases}$$



Pareto optimal solution



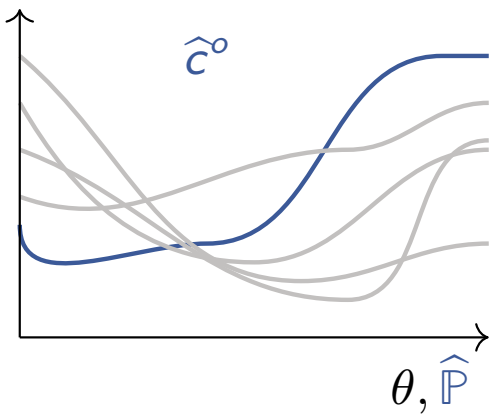
Minimax solution

Bennouna, **BvP** (2023). "Learning and decision-making with data: Optimal formulations and phase transitions". **Mathematical Programming** (Major Revision)

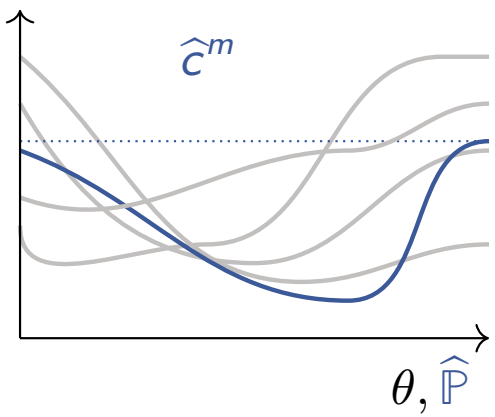
Meta Optimization Problem

Among all predictors with **out-of-sample guarantee** find the **least conservative** one

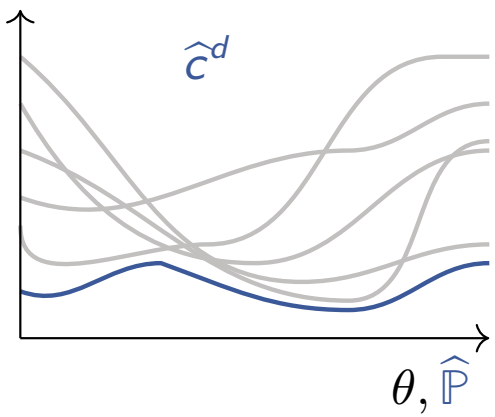
$$\begin{cases} \min\limits_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \text{Prob}_{\mathbb{P}}\left[\widehat{c}_N(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta,\xi)\right]\ \forall \theta \in \Theta\right] \geq 1 - \exp(-a_N + o(a_N))\ \ \forall \mathbb{P} \end{cases}$$



Pareto optimal solution



Minimax solution



Pareto dominant solution

Bennouna, **BvP** (2023). "Learning and decision-making with data: Optimal formulations and phase transitions". **Mathematical Programming** (Major Revision)

Pareto Dominant Solution

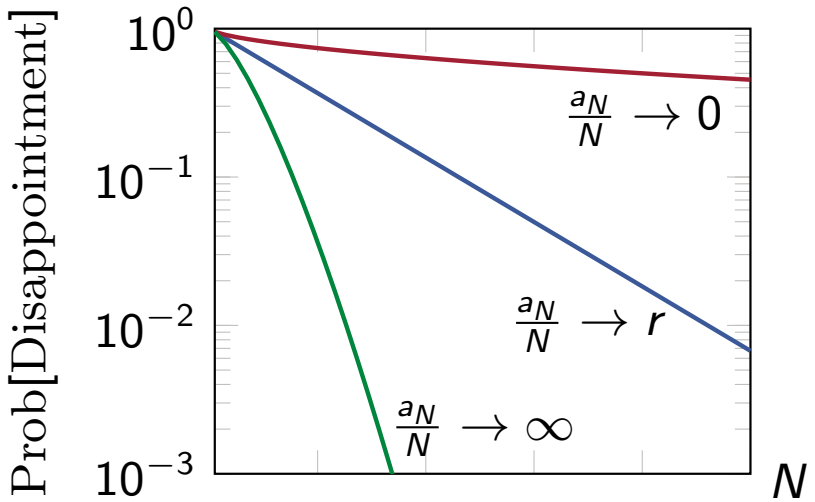
Thm: The meta optimization problem

$$\begin{cases} \min\limits_{\widehat{c}}[\preceq]\ \widehat{c} \\ \quad \text{s.t. } \mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}_N(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta,\xi)\right]\ \forall \theta \in \Theta\right] \geq 1 - \exp(-a_N + o(a_N)) \quad \forall \mathbb{P} \end{cases}$$

admits a **Pareto dominant** solution. We observe three regimes depending on rate a_N.

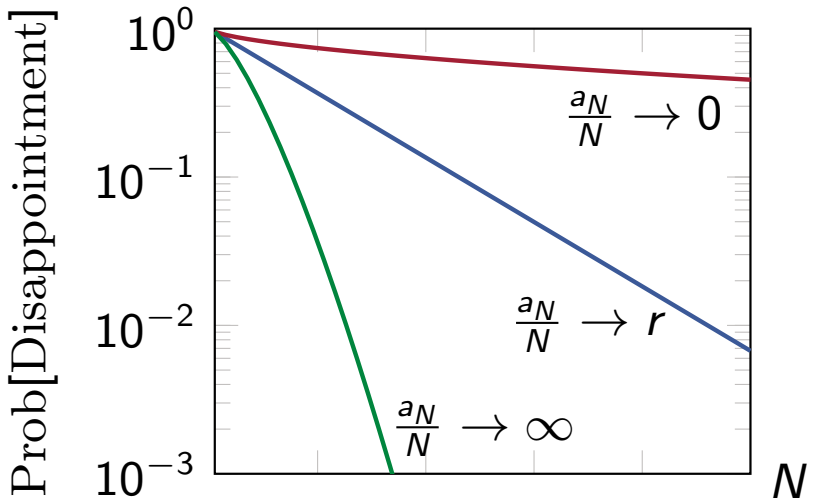
Bennouna, **BvP** (2023). "Learning and decision-making with data: Optimal formulations and phase transitions". **Mathematical Programming** (Major Revision)

Phase Transitions



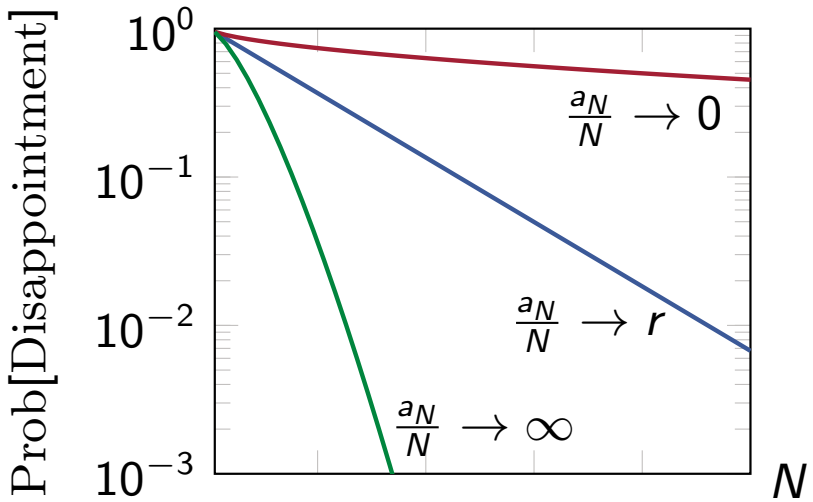
$a_N \gg N$	$a_N \sim rN$	$a_N \ll N$

Phase Transitions



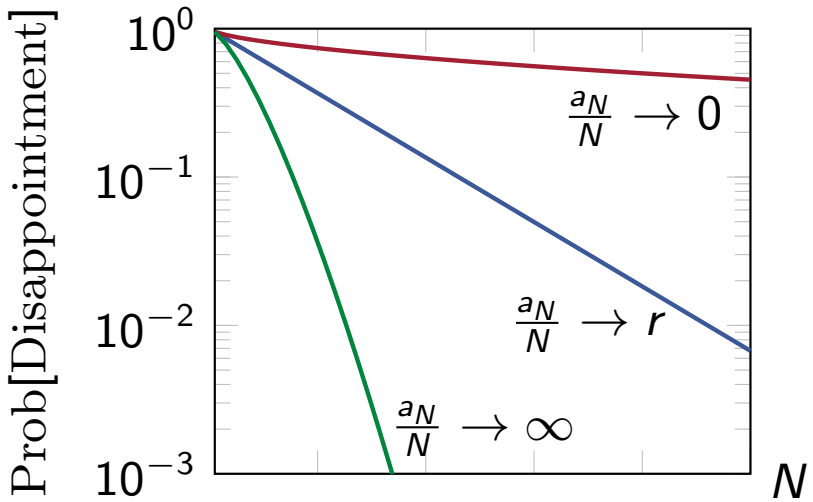
$a_N \gg N$	$a_N \sim rN$	$a_N \ll N$
	DRO: $\widehat{c}_N^d(\theta, \widehat{\mathbb{P}}_N)$ $= \max_{KL(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r} \mathbf{E}_{\mathbb{Q}}[\ell(\theta, \xi)]$	
	Inconsistent	

Phase Transitions



$a_N \gg N$	$a_N \sim rN$	$a_N \ll N$
RO: $\widehat{c}_N^d(\theta, \widehat{\mathbb{P}}_N)$ $= \max_{KL(\widehat{\mathbb{P}}_n, \mathbb{Q}) \leq \infty} \mathbf{E}_{\mathbb{Q}}[\ell(\theta, \xi)]$ $= \max_{\xi \in \Xi} \ell(\theta, \xi)$	DRO: $\widehat{c}_N^d(\theta, \widehat{\mathbb{P}}_N)$ $= \max_{KL(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r} \mathbf{E}_{\mathbb{Q}}[\ell(\theta, \xi)]$	
Inconsistent	Inconsistent	

Phase Transitions



$a_N \gg N$	$a_N \sim rN$	$a_N \ll N$
RO: $\widehat{c}_N^d(\theta, \widehat{\mathbb{P}}_N)$ $= \max_{KL(\widehat{\mathbb{P}}_n, \mathbb{Q}) \leq \infty} \mathbf{E}_{\mathbb{Q}}[\ell(\theta, \xi)]$ $= \max_{\xi \in \Xi} \ell(\theta, \xi)$	DRO: $\widehat{c}_N^d(\theta, \widehat{\mathbb{P}}_N)$ $= \max_{KL(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq r} \mathbf{E}_{\mathbb{Q}}[\ell(\theta, \xi)]$	VR: $\widehat{c}_N^d(\theta, \widehat{\mathbb{P}}_N)$ $= \max_{KL(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leq \frac{a_N}{N}} \mathbf{E}_{\mathbb{Q}}[\ell(\theta, \xi)]$ $\equiv \mathbf{E}_{\widehat{\mathbb{P}}_N}[\ell(\theta, \xi)] + \sqrt{\frac{2a_N}{N}} \mathbf{V}_{\widehat{\mathbb{P}}_N}[\ell(x, \xi)]$
Inconsistent	Inconsistent	Consistent

Table of Contents

Prescriptors without Predictors

Not all data driven **prescriptors** are of the form

$$\widehat{\theta}_N(\widehat{\mathbb{P}}_N) \in \arg\min_{\theta\in\Theta} \widehat{c}_N(\theta, \widehat{\mathbb{P}}_N).$$

Prescriptors without Predictors

Not all data driven **prescriptors** are of the form

$$\widehat{\theta}_N(\widehat{\mathbb{P}}_N) \in \arg\min_{\theta\in\Theta} \widehat{c}_N(\theta, \widehat{\mathbb{P}}_N).$$

Assumption: $\theta \mapsto \ell(\theta, \xi)$ is L-Lipschitz. Consider a **prescriptor** $\widehat{\theta}_N : \mathcal{P}(\Xi) \to \Theta$ which enjoys some **out-of-sample guarantee**

$$\mathrm{Prob}_{\mathbb{P}}\left[\widehat{g}_N(\widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\widehat{\theta}_N(\widehat{\mathbb{P}}_N), \xi)\right]\right] \geq 1 - \exp(-a_N + o(a_N)) \quad \forall \mathbb{P}.$$

for some function $\widehat{g}_N$.

Reduction

Any prescriptor $\widehat{\theta}_N(\widehat{\mathbb{P}}_N)$ can be **represented** as minimizing the proxy cost function

$$\widehat{c}_N(\theta, \widehat{\mathbb{P}}_N) := \widehat{g}_N(\widehat{\mathbb{P}}_N) + L\|\theta - \theta_N(\widehat{\mathbb{P}}_N)\|$$

Furthermore, observe that

$$\widehat{g}_N(\widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\widehat{\theta}_N(\widehat{\mathbb{P}}_N), \xi)\right] \implies g_N(\widehat{\mathbb{P}}_N) + L\|\theta - \widehat{\theta}_N(\widehat{\mathbb{P}}_N)\| \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right] \quad \forall \theta \in \Theta$$

and consequently

$$\mathrm{Prob}_{\mathbb{P}}\left[\widehat{c}_N(\theta, \widehat{\mathbb{P}}_N) \geq \mathbf{E}_{\mathbb{P}}\left[\ell(\theta, \xi)\right] \ \forall \theta \in \Theta\right] \geq 1 - \exp(-a_N + o(a_N)) \quad \forall \mathbb{P}.$$

Dominance

Recall the **KL prescriptor**

$$\widehat{\theta}_N^{\mathrm{KL}} \in \arg\min_{\theta\in\Theta} \{c_N^{\mathrm{KL}}(\theta, \widehat{\mathbb{P}}_N) := \max_{D_{\mathrm{KL}}(\widehat{\mathbb{P}}_N, \mathbb{Q}) \le a_N/N} \mathbf{E}_{\mathbb{Q}}[\ell(\theta, \xi)]\}.$$

From the **Pareto dominance** of $\widehat{c}_N^d$ we have $\forall\theta \in \Theta$ and $\forall\widehat{\mathbb{P}}$ that $\widehat{c}_N(\theta, \widehat{\mathbb{P}}) \ge \widehat{c}_N^{\mathrm{KL}}(\theta, \widehat{\mathbb{P}})$ and in particular

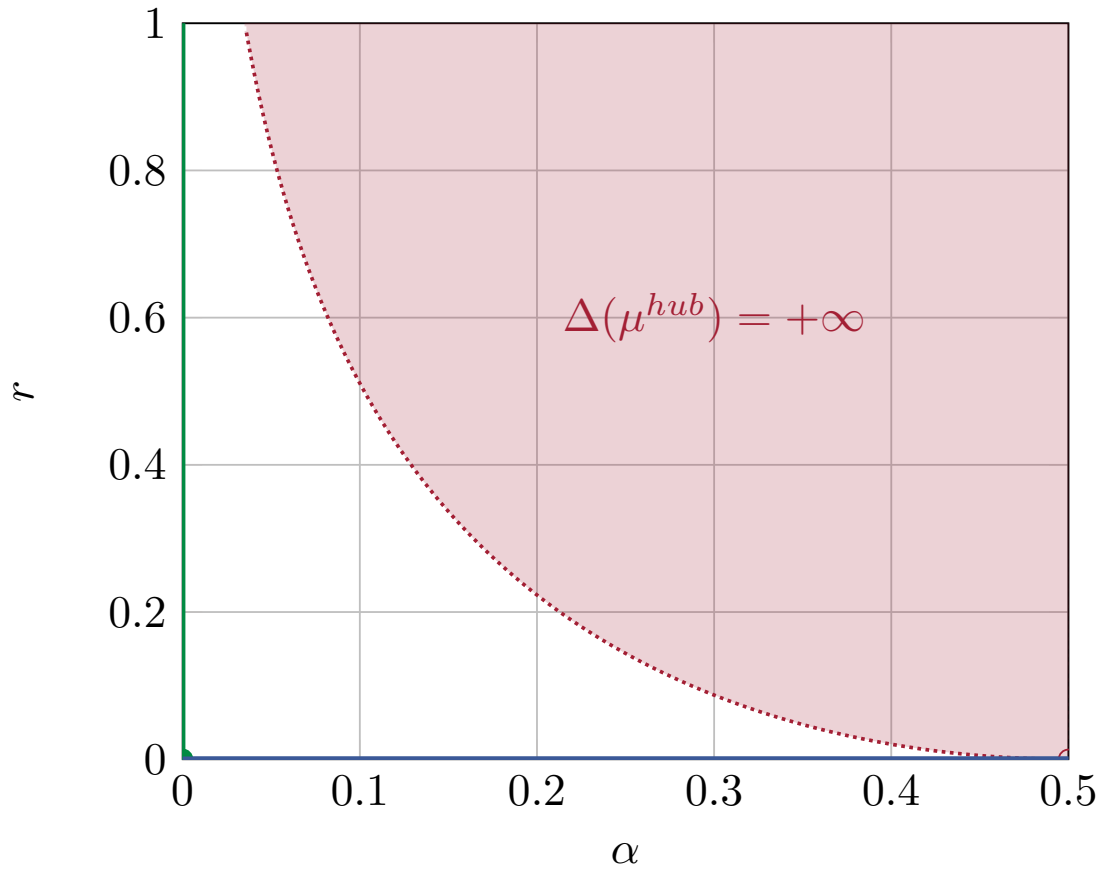
$$\widehat{g}_N(\widehat{\mathbb{P}}_N) = \widehat{c}_N(\widehat{\theta}_N(\widehat{\mathbb{P}}_N), \widehat{\mathbb{P}}_N) \ge \widehat{c}_N^{\mathrm{KL}}(\widehat{\theta}_N(\widehat{\mathbb{P}}_N), \widehat{\mathbb{P}}_N) \ge \widehat{c}_N^{\mathrm{KL}}(\widehat{\theta}_{KL}(\widehat{\mathbb{P}}_N), \widehat{\mathbb{P}}_N)$$

The KL prescriptor $\widehat{\theta}_N^{KL}$ also dominates $\widehat{\theta}_N$. That is,

$$\mathrm{Prob}_{\mathbb{P}}\left[\widehat{g}_N(\widehat{\mathbb{P}}_N) \ge \widehat{c}_N^{\mathrm{KL}}(\widehat{\theta}_{KL}(\widehat{\mathbb{P}}_N), \widehat{\mathbb{P}}_N) \ge \mathbf{E}_{\mathbb{P}}\left[\ell(\widehat{\theta}_{KL}(\widehat{\mathbb{P}}_N), \xi)\right]\right] \ge 1 - \exp(-a_N + o(a_N)) \quad \forall\mathbb{P}.$$

Table of Contents

Phase Diagram



g

- Green: Mean is optimal.
- Blue: Median is optimal.
- Red: Error margin is unbounded.