Data-driven distributionally robust optimization: Classification of ambiguity sets

Tobias Sutter[1], Bart van Parys[2] & Daniel Kuhn[3]

Data Driven Combinatorial Optimization,
Ecole des Ponts, Paris, October 4, 2023

[1]University of Konstanz
[2]MIT Sloan School of Management
[3]Risk Analytics and Optimization Chair, EPFL

Data-driven decision making

Stochastic optimization	Family of prob. measures	Data-gen. process
$\min_x c(x, \theta)$	$\{\mathbb{P}_\theta : \theta \in \Theta\}$	$\{\xi_t\}_{t \in \mathbb{N}}$

Data-driven decision making

Stochastic optimization	Family of prob. measures	Data-gen. process
$\min_x c(x, \theta)$	$\{\mathbb{P}_\theta : \theta \in \Theta\}$	$\{\xi_t\}_{t\in\mathbb{N}}$

Examples:

- Expected loss $c(x, \theta) = \mathbb{E}_\theta[\ell(x, \xi)]$
- Risk of loss $c(x, \theta) = \rho_\theta[\ell(x, \xi)]$
- Covariate information $c(x, \theta) = \mathbb{E}_\theta[\ell(x, \xi) | C\xi \in B]$
- Long-run average loss $c(x, \theta) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_\theta[\ell(\pi_x(s_t), s_t)]$

Data-driven decision making

Stochastic optimization	Family of prob. measures	Data-gen. process
$\min_x c(x, \theta)$	$\{\mathbb{P}_\theta : \theta \in \Theta\}$	$\{\xi_t\}_{t \in \mathbb{N}}$

Assumptions:

- All measures defined on $(\Omega, \mathcal{F})$
- $\Theta \subseteq \mathbb{R}^d$ open

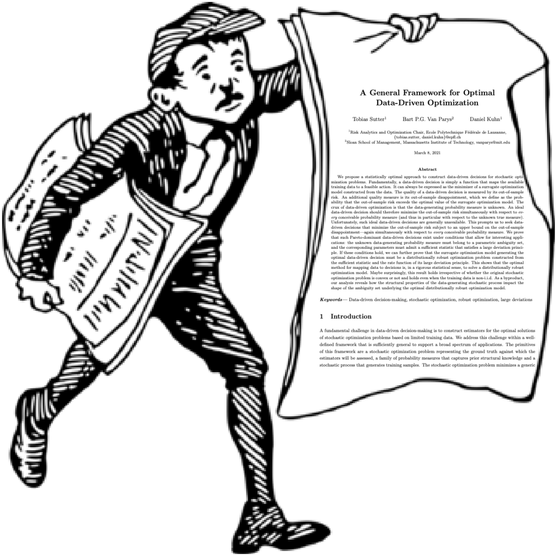
Data-driven decision making

Stochastic optimization	Family of prob. measures	Data-gen. process
$\min_x c(x, \theta)$	$\{\mathbb{P}_\theta : \theta \in \Theta\}$	$\{\xi_t\}_{t\in\mathbb{N}}$

Examples:

- Finite-state i.i.d. processes
- Finite-state Markov chains
- Vector-autoregressive processes
- I.i.d. processes with parametric distribution function

Motivating example — newsvendor problem



A General Framework for Optimal Data-Driven Optimization

Tobias Sutter[1] Bart P.G. Van Parys[2] Daniel Kuhn[1]

[1]Risk Analytics and Optimization Chair, Ecole Polytechnique Fédérale de Lausanne, {tobias.sutter, daniel.kuhn}@epfl.ch
[2]Sloan School of Management, Massachusetts Institute of Technology, vanparys@mit.edu

March 8, 2021

Abstract

We propose a statistically optimal approach to construct data-driven decisions for stochastic optimization problems. Fundamentally, a data-driven decision is simply a function that maps the available training data to a feasible action. It can always be expressed as the minimizer of a surrogate optimization model constructed from the data. The quality of a data-driven decision is measured by its out-of-sample risk. An additional quality measure is its out-of-sample disappointment, which we define as the probability that the out-of-sample risk exceeds the optimal value of the surrogate optimization model. The crux of data-driven optimization is that the data-generating probability measure is unknown. An ideal data-driven decision should therefore minimize the out-of-sample risk simultaneously with respect to *every* conceivable probability measure (and thus in particular with respect to the unknown true measure). Unfortunately, such ideal data-driven decisions are generally unavailable. This prompts us to seek data-driven decisions that minimize the out-of-sample risk subject to an upper bound on the out-of-sample disappointment—again simultaneously with respect to *every* conceivable probability measure. We prove that such Pareto-dominant data-driven decisions exist under conditions that allow for interesting applications: the unknown data-generating probability measure must belong to a parametric ambiguity set, and the corresponding parameters must admit a sufficient statistic that satisfies a large deviation principle. If these conditions hold, we can further prove that the surrogate optimization model generating the optimal data-driven decision must be a distributionally robust optimization problem constructed from the sufficient statistic and the rate function of its large deviation principle. This shows that the optimal method for mapping data to decisions is, in a rigorous statistical sense, to solve a distributionally robust optimization model. Maybe surprisingly, this result holds irrespective of whether the original stochastic optimization problem is convex or not and holds even when the training data is non-i.i.d. As a byproduct, our analysis reveals how the structural properties of the data-generating stochastic process impact the shape of the ambiguity set underlying the optimal distributionally robust optimization model.

Keywords— Data-driven decision-making, stochastic optimization, robust optimization, large deviations

1 Introduction

A fundamental challenge in data-driven decision-making is to construct estimators for the optimal solutions of stochastic optimization problems based on limited training data. We address this challenge within a well-defined framework that is sufficiently general to support a broad spectrum of applications. The primitives of this framework are a stochastic optimization problem representing the ground truth against which the estimators will be assessed, a family of probability measures that captures prior structural knowledge and a stochastic process that generates training samples. The stochastic optimization problem minimizes a generic

Newsvendor problem

Stochastic optimization	
$\min_{x \in X} c(x, \theta)$	

Stochastic optimization

$$\min_{x \in X} c(x, \theta)$$

- Order quantities $x \in X = \{1, 2, \cdots, d\}$
- Demand $\xi \in \Xi = \{1, 2, \cdots, d\}$
- Objective $c(x, \theta) = \mathbb{E}_\theta[kx - p \min\{x, \xi\}]$

Data-gen. process

$$\{\xi_t\}_{t \in \mathbb{N}}$$

- Historical demand $\xi_t \in \Xi$

Family of prob. measures

$$\{\mathbb{P}_\theta : \theta \in \Theta\}$$

- $\{\xi_t\}_{t \in \mathbb{N}}$ i.i.d. process under $\mathbb{P}_\theta$
- $\mathbb{P}_\theta(\xi_t = i) = \theta_i$ for $i \in \Xi$

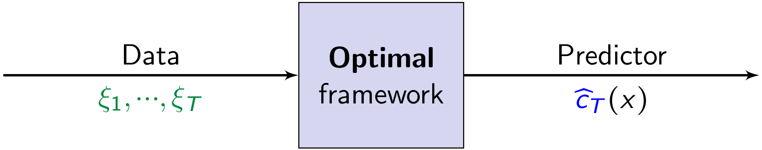
Surrogate optimization models

Original optimization problem

$$\min_{x \in X} c(x, \theta)$$

Surrogate optimization problem

$$\min_{x \in X} \widehat{c}_T(x)$$



Surrogate optimization models

Original optimization problem

$$\min_{x \in X} c(x, \theta)$$

Surrogate optimization problem

$$\min_{x \in X} \widehat{c}_T(x)$$

Construction of $\widehat{c}_T$

- Sample average approximation[1]
- Predict-then-optimize approach[2]
- Neural network model[3]
- Distributionally robust optimization model[4]
- etc.

---

[1] Shapiro, Annals of Statistics, 1989; [2] Elmachtoub & Grigas, Management Science, 2021; [3] Donti et al., NeurIPS, 2017; [4] Delage & Ye, Operations Research, 2010

Terminology

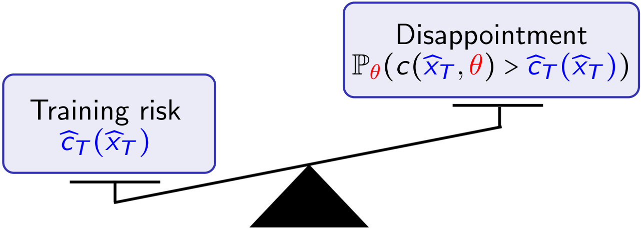
As a function of the training data $\xi_1, \cdots \xi_T$ we denote

- Data-driven predictor $\widehat{c}_T$
- Data-driven prescriptor $\widehat{x}_T = \arg\min_{x \in X} \widehat{c}_T(x)$

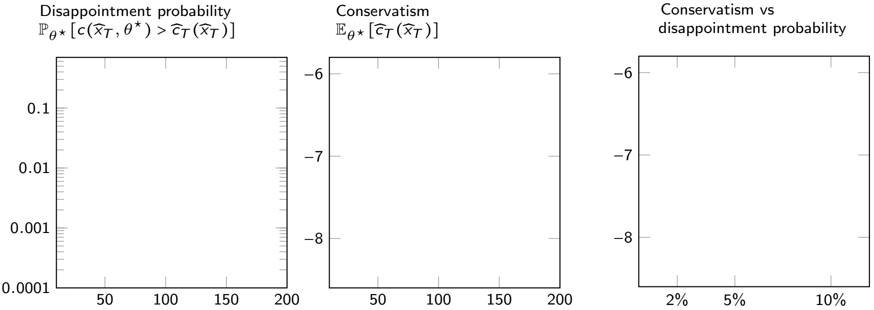
Performance measures

1. In-sample (training) risk $\widehat{c}_T(\widehat{x}_T)$
2. Out-of-sample (generalization) risk $c(\widehat{x}_T, \theta)$
3. Out-of-sample disappointment $\mathbb{P}_\theta(c(\widehat{x}_T, \theta) > \widehat{c}_T(\widehat{x}_T))$

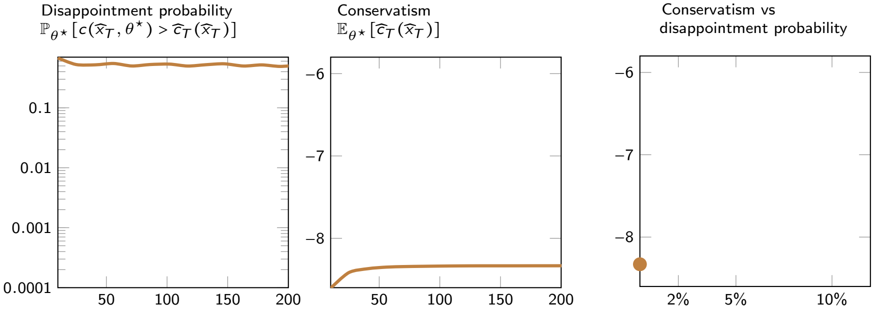
A basic tradeoff



Data-driven newsvendor problem



Data-driven newsvendor problem

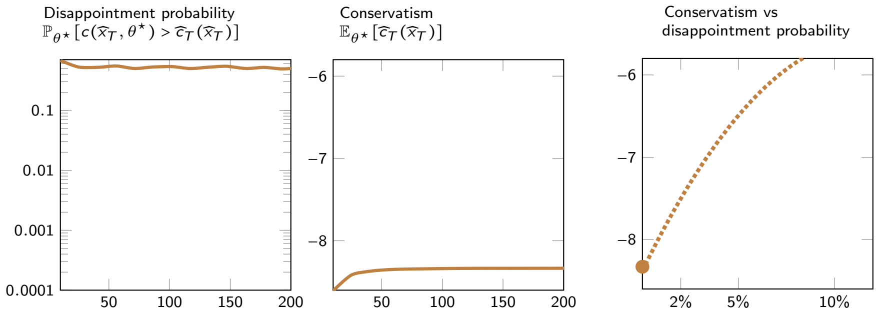


Model 1: SAA model[2]

$$\widehat{c}_T(x) = c(x, \widehat{\theta}_T)$$

[2]Shapiro, Annals of Statistics, 1989

Data-driven newsvendor problem

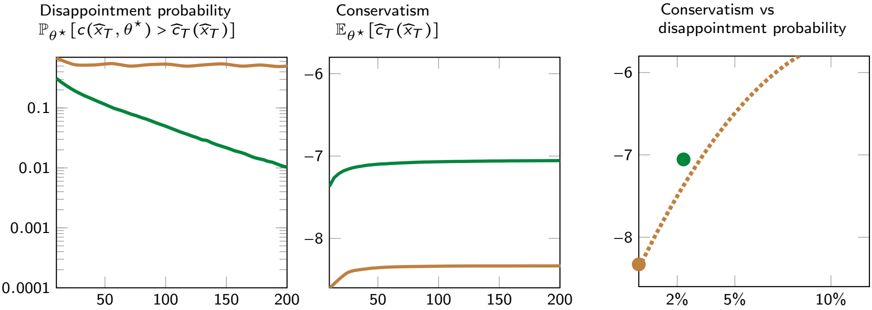


Model 1: SAA model[2]

$$\widehat{c}_T(x) = c(x, \widehat{\theta}_T) + r$$

[2]Shapiro, Annals of Statistics, 1989

Data-driven newsvendor problem

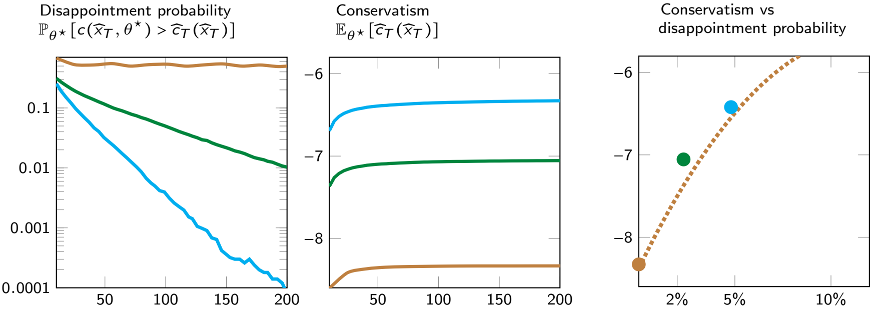


Model 2: DRO model with moment ambiguity set[2]

$$\widehat{c}_T(x) = \sup_{\theta \in \Theta} \left\{ c(x, \theta) \ : \ |\mathbb{E}_{\widehat{\theta}_T}[\xi^j] - \mathbb{E}_\theta[\xi^j]| \le r \quad \forall j = 1, \ldots, 4 \right\}$$

[2]Delage & Ye, Operations Research, 2010

Data-driven newsvendor problem

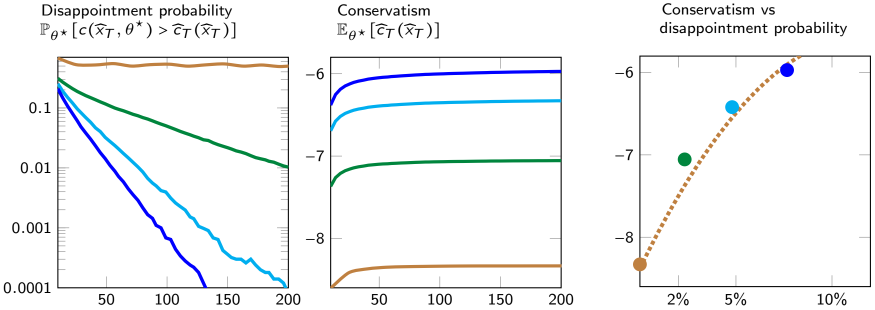


Model 2: DRO model with moment ambiguity set[2]

$$\widehat{c}_T(x) = \sup_{\theta \in \Theta} \left\{ c(x, \theta) \ : \ |\mathbb{E}_{\widehat{\theta}_T}[\xi^j] - \mathbb{E}_{\theta}[\xi^j]| \leq r \quad \forall j = 1, \ldots, 4 \right\}$$

[2] Delage & Ye, Operations Research, 2010

Data-driven newsvendor problem

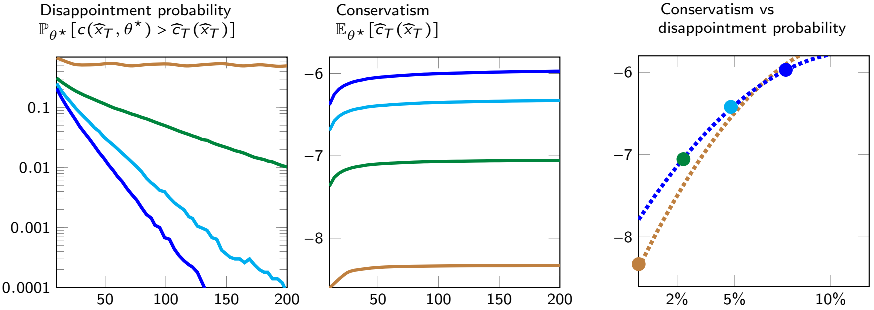


Model 2: DRO model with moment ambiguity set[2]

$$\widehat{c}_T(x) = \sup_{\theta \in \Theta} \left\{ c(x, \theta) \ : \ |\mathbb{E}_{\widehat{\theta}_T}[\xi^j] - \mathbb{E}_{\theta}[\xi^j]| \le r \quad \forall j = 1, \ldots, 4 \right\}$$

[2]Delage & Ye, Operations Research, 2010

Data-driven newsvendor problem

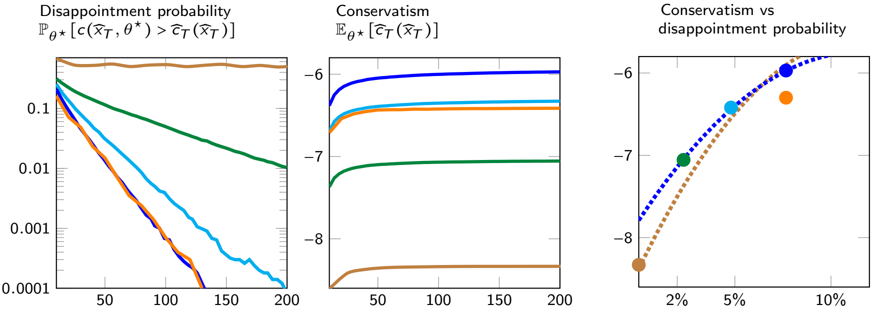


Model 2: DRO model with moment ambiguity set[2]

$$\widehat{c}_T(x) = \sup_{\theta \in \Theta} \left\{ c(x, \theta) \ : \ |\mathbb{E}_{\widehat{\theta}_T}[\xi^j] - \mathbb{E}_{\theta}[\xi^j]| \leq r \quad \forall j = 1, \ldots, 4 \right\}$$

[2]Delage & Ye, Operations Research, 2010

Data-driven newsvendor problem

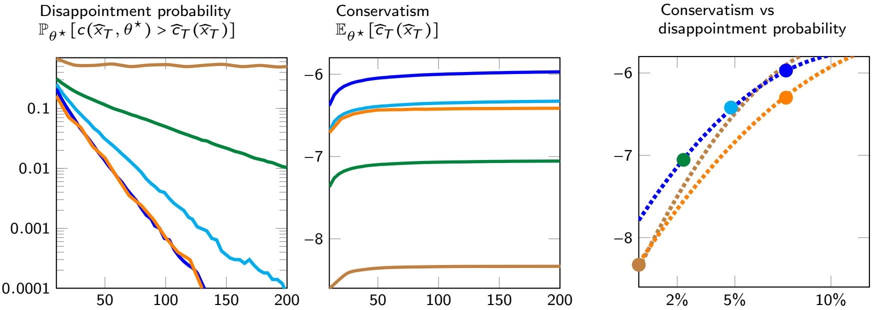


Model 3: DRO model with Wasserstein ambiguity set[2]

$$\widehat{c}_T(x) = \sup_{\theta \in \Theta} \left\{ c(x, \theta) \ : \mathrm{d_W}(\widehat{\theta}_T, \theta) \le r \right\}$$

[2] Mohajerin Esfahani & Kuhn, Mathematical Programming, 2018

Data-driven newsvendor problem

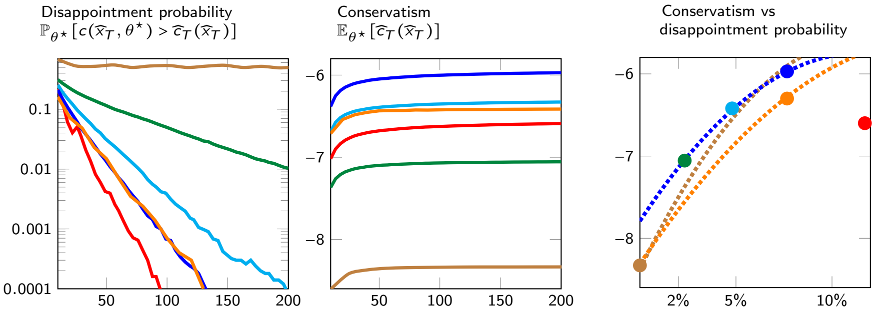


Model 3: DRO model with Wasserstein ambiguity set[2]

$$\widehat{c}_T(x) = \sup_{\theta \in \Theta} \left\{ c(x, \theta) \ : \mathrm{d_W}(\widehat{\theta}_T, \theta) \leq r \right\}$$

[2]Mohajerin Esfahani & Kuhn, Mathematical Programming, 2018

Data-driven newsvendor problem

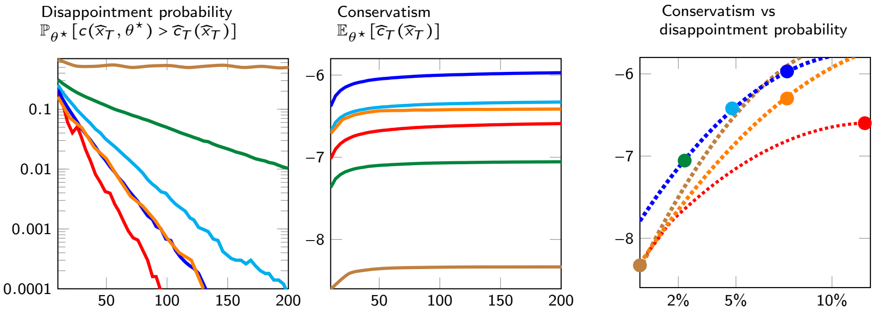


Model 4: DRO model with KL-ambiguity set[2]

$$\widehat{c}_T(x) = \sup_{\theta \in \Theta} \left\{ c(x, \theta) \ : \mathrm{D}\left(\widehat{\theta}_T \,\middle\|\, \theta \right) \le r \right\}$$

[2]Ben-Tal et al., Management Science, 2013

Data-driven newsvendor problem

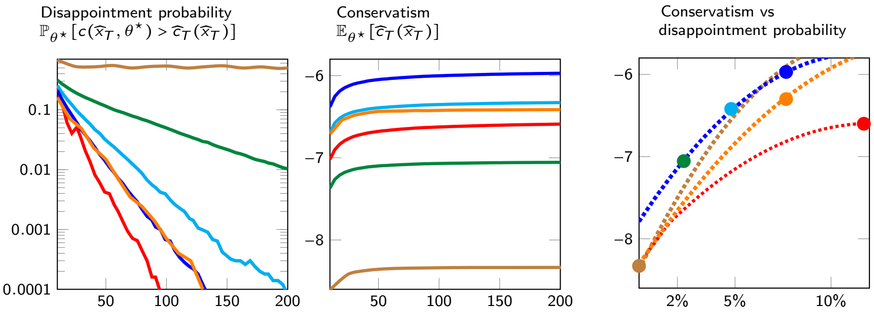


Model 4: DRO model with KL-ambiguity set[2]

$$\widehat{c}_T(x) = \sup_{\theta \in \Theta} \left\{ c(x, \theta) \ : \mathrm{D}\left(\widehat{\theta}_T \middle\| \theta\right) \leq r \right\}$$

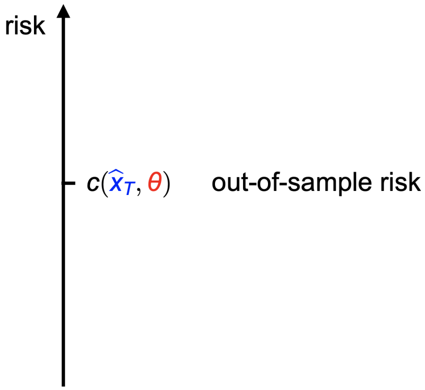
[2]Ben-Tal et al., Management Science, 2013

Data-driven newsvendor problem

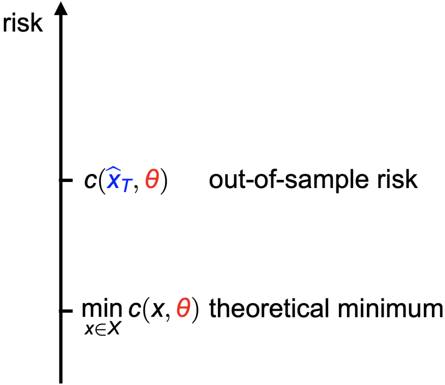


Which method is optimal?

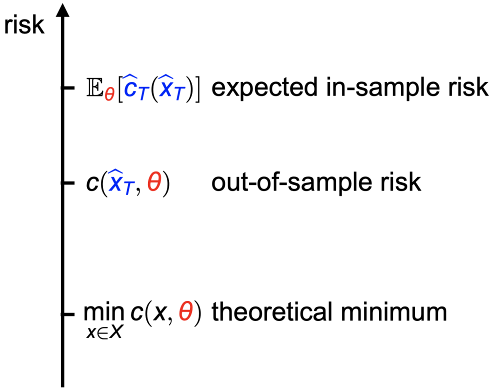
Optimal data-driven decision making



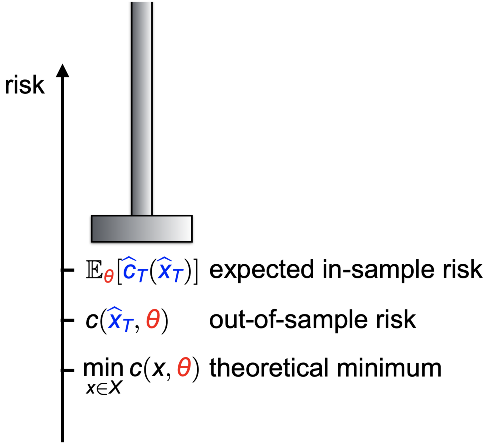
Optimal data-driven decision making



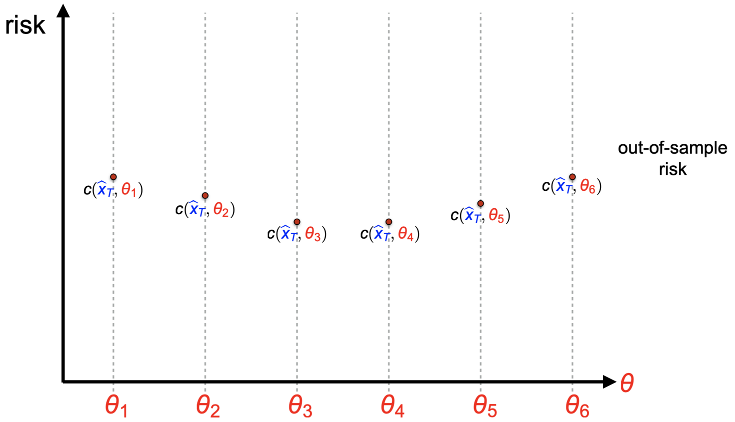
Optimal data-driven decision making



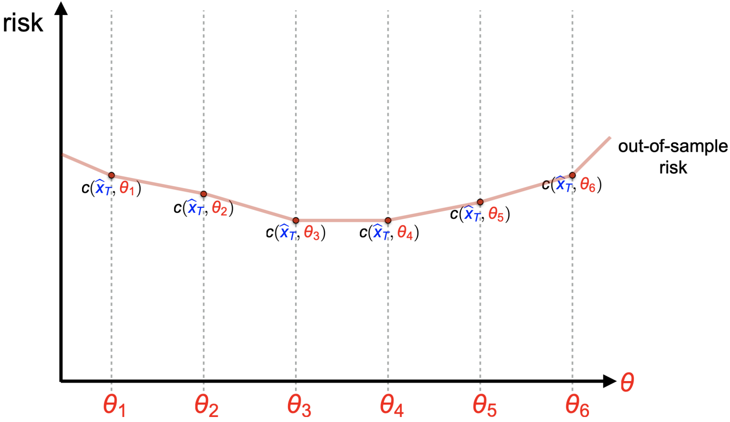
Optimal data-driven decision making



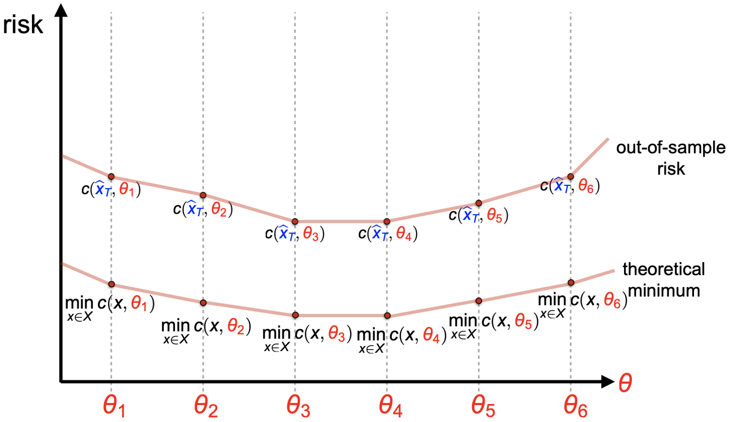
Optimal data-driven decision making



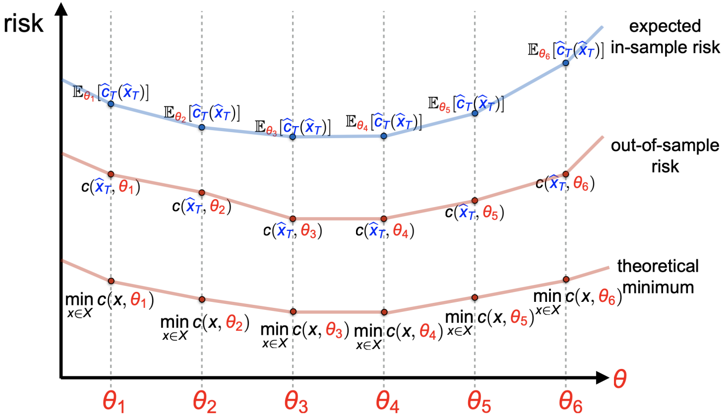
Optimal data-driven decision making



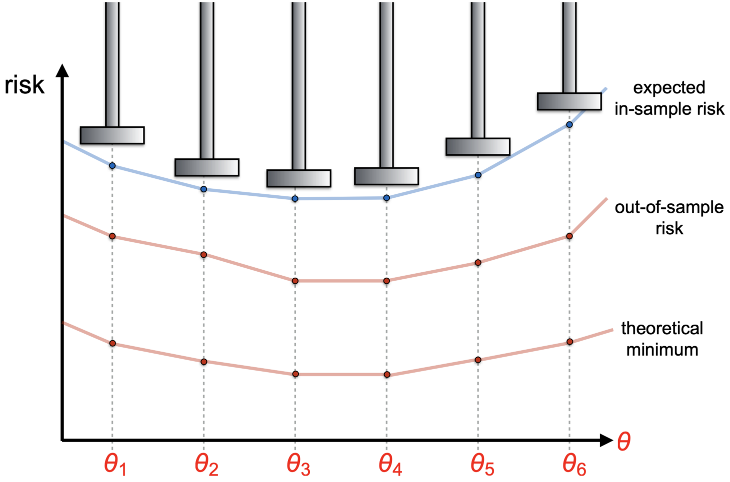
Optimal data-driven decision making



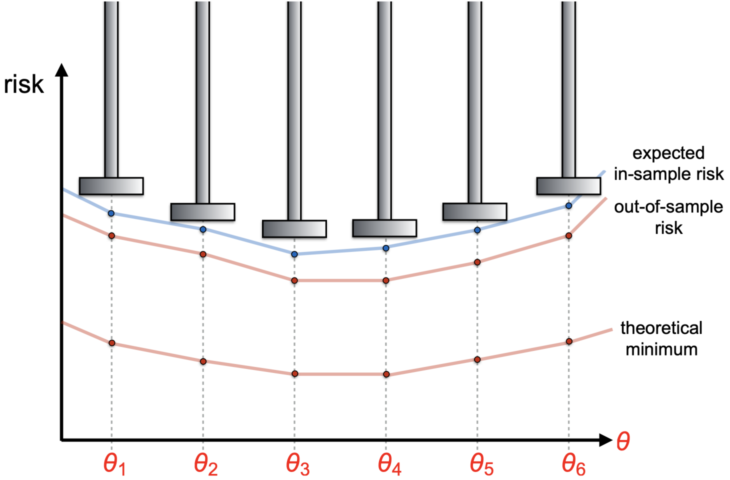
Optimal data-driven decision making



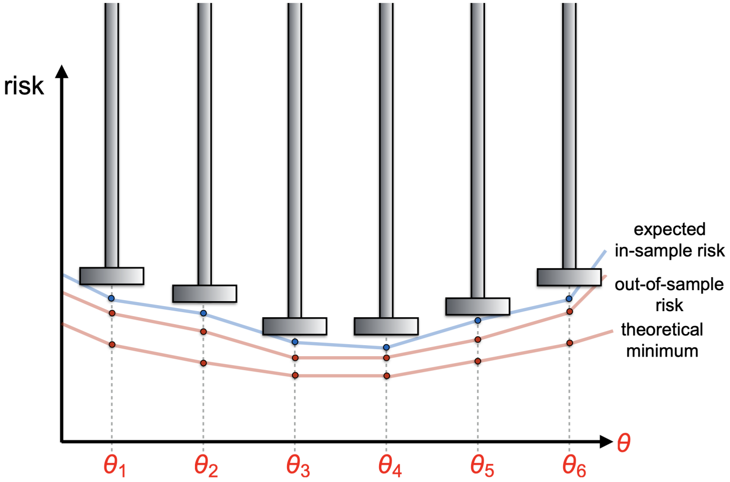
Optimal data-driven decision making



Optimal data-driven decision making



Optimal data-driven decision making



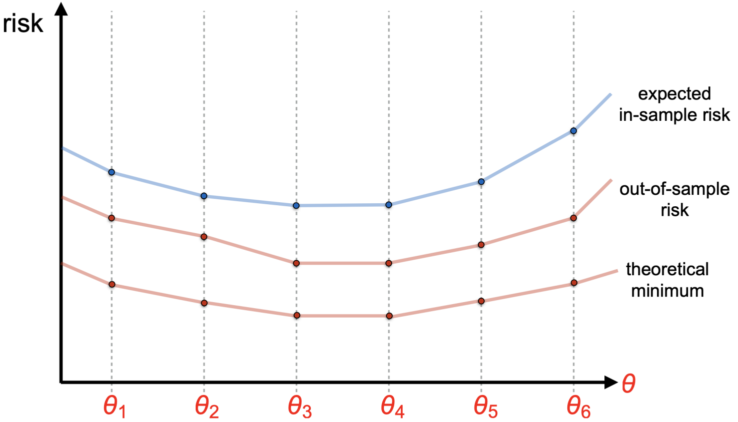
Optimal data-driven decision making (cont'd)

$$(\bigstar)\begin{cases} \min\limits_{\widehat{c},\,\widehat{x}} & \left\{\lim\limits_{T\to\infty} \mathbb{E}_{\theta}[\widehat{c}_T(\widehat{x}_T)]\right\}_{\theta\in\Theta} \\ \text{s. t.} & \lim\limits_{T\to\infty} \frac{1}{T}\log \mathbb{P}_{\theta}\left(c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)\right) \leq -r \quad \forall \theta\in\Theta \end{cases}$$

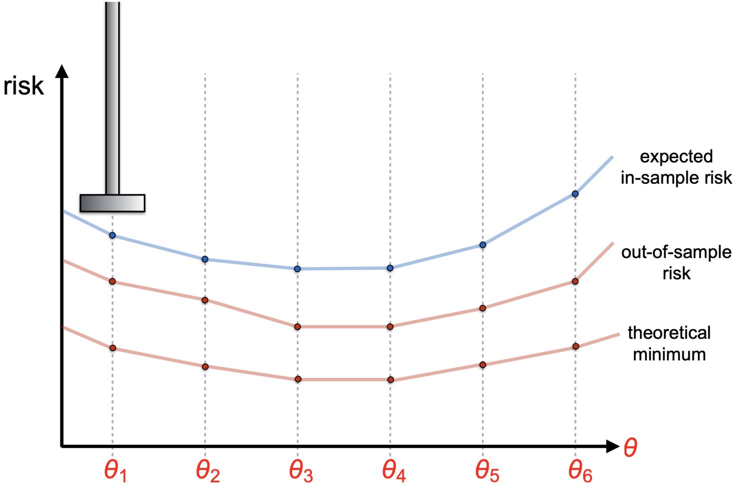
Interpretation: Among all predictors and prescriptors with **"small" disappointment** find the **least conservative** one

$$\mathbb{P}_{\theta}\left(c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)\right) \leq e^{-rT+o(T)}$$

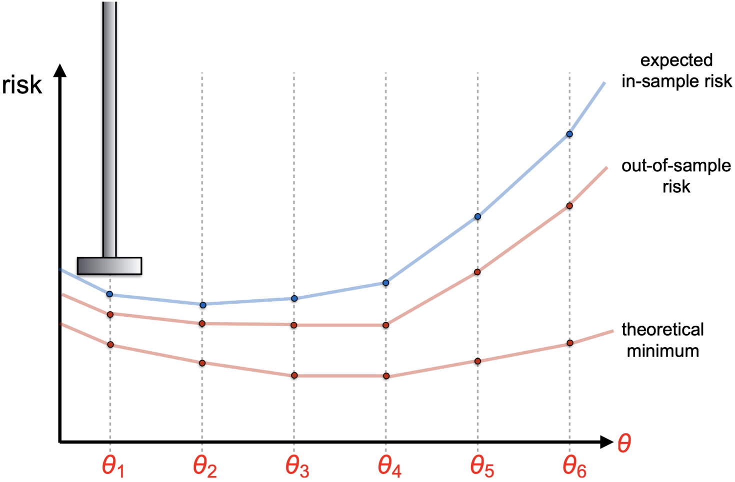
Optimal data-driven decision making (cont'd)



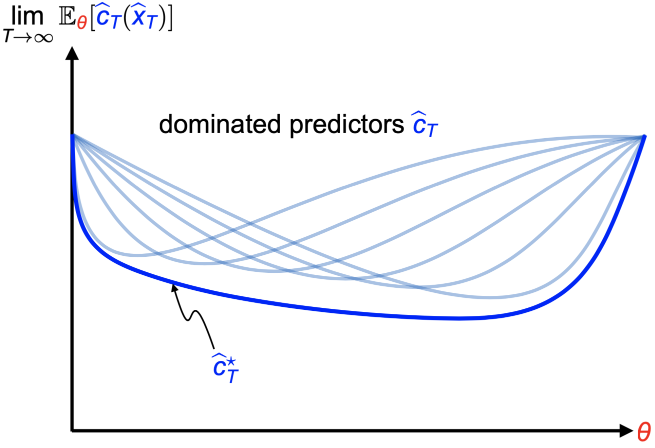
Optimal data-driven decision making (cont'd)



Optimal data-driven decision making (cont'd)



Pareto dominant solutions



$\widehat{c}^\star_T$ minimizes the in-sample risk simultaneously for *every* θ

Optimizing over ALL surrogate models

$$(\bigstar)\begin{cases} \min\limits_{\widehat{c},\,\widehat{x}} & \left\{\lim\limits_{T\to\infty} \mathbb{E}_{\theta}[\widehat{c}_T(\widehat{x}_T)]\right\}_{\theta\in\Theta} \\ \text{s.t.} & \lim\limits_{T\to\infty} \frac{1}{T}\log \mathbb{P}_{\theta}\left(c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)\right) \le -r \quad \forall \theta \in \Theta \end{cases}$$

Optimizing over ALL surrogate models

$$(\bigstar)\begin{cases} \min\limits_{\widehat{c},\,\widehat{x}} & \left\{\lim\limits_{T\to\infty} \mathbb{E}_{\theta}[\widehat{c}_T(\widehat{x}_T)]\right\}_{\theta\in\Theta} \\ \text{s. t.} & \lim\limits_{T\to\infty} \frac{1}{T}\log \mathbb{P}_{\theta}\left(c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)\right) \le -r \quad \forall\theta\in\Theta \end{cases}$$

Strenghts

- proxy for optimizing the out-of-sample risk
- admits a Pareto dominant solution in closed form
- errs on the side of caution
- facilitates separation of estimation and optimization

Optimizing over ALL surrogate models

$$(\bigstar)\begin{cases} \min\limits_{\widehat{c},\,\widehat{x}} & \left\{\lim\limits_{T\to\infty} \mathbb{E}_{\theta}[\widehat{c}_T(\widehat{x}_T)]\right\}_{\theta\in\Theta} \\ \text{s. t.} & \lim\limits_{T\to\infty} \frac{1}{T}\log \mathbb{P}_{\theta}\left(c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)\right) \le -r \quad \forall\theta\in\Theta \end{cases}$$

Weaknesses

- performance criteria are asymptotic
- choice of r is subjective
- feasible/optimal models are biased

Optimizing over ALL surrogate models

$$(\bigstar)\begin{cases} \min_{\widehat{c},\,\widehat{x}} & \left\{\lim_{T\to\infty} \mathbb{E}_{\theta}[\widehat{c}_T(\widehat{x}_T)]\right\}_{\theta\in\Theta} \\ \text{s. t.} & \lim_{T\to\infty} \frac{1}{T}\log \mathbb{P}_{\theta}\left(c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)\right) \le -r \quad \forall \theta \in \Theta \end{cases}$$

Space of all possible predictors and prescriptors is large

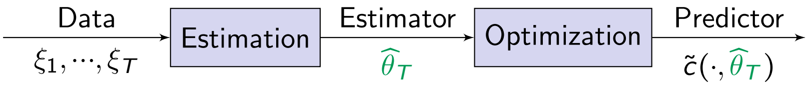
- $\widehat{c}_T$, $\widehat{x}_T$ can be any[2] function depending on the available training data $\xi_1, \cdots \xi_T$
- Can we restrict ourselves to smaller class of functions without loosing optimality?

[2] Some technical details, see arXiv:2010:06606

Optimizing over ALL surrogate models

$$(\bigstar)\begin{cases} \min\limits_{\widehat{c},\,\widehat{x}} & \left\{\lim\limits_{T\to\infty} \mathbb{E}_{\theta}[\widehat{c}_T(\widehat{x}_T)]\right\}_{\theta\in\Theta} \\ \text{s. t.} & \lim\limits_{T\to\infty} \frac{1}{T}\log \mathbb{P}_{\theta}\left(c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)\right) \le -r \quad \forall\theta\in\Theta \end{cases}$$

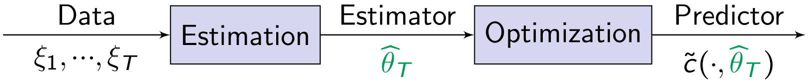
Key idea: Separation of estimation and optimization



1. Which estimator should one pick?
2. Is this separation without loss of optimality? Can we represent the (strong) solution to $(\bigstar)$ as

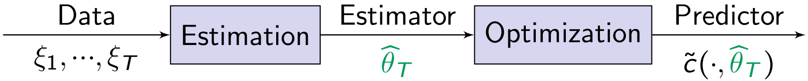
$$\widehat{c}_T(x) = \tilde{c}(x, \widehat{\theta}_T)$$

Separation principle - intuition



When can this separation be without loss of optimality?

Separation principle - intuition

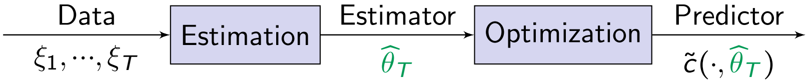


When can this separation be without loss of optimality?

(i) **No statistical information** about θ **is lost** when considering an estimator, i.e.,

$$\theta \longrightarrow \widehat{\theta}_T \longrightarrow \xi_1, \ldots, \xi_T \quad \text{forms a Markov chain}$$

Separation principle - intuition



When can this separation be without loss of optimality?

(i) **No statistical information** about θ **is lost** when considering an estimator, i.e.,

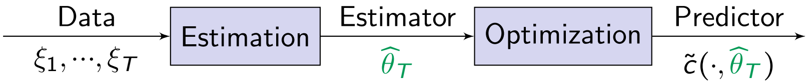
$$\theta \longrightarrow \widehat{\theta}_T \longrightarrow \xi_1, \ldots, \xi_T \quad \text{forms a Markov chain}$$

(ii) **Estimator concentrates fast enough** around true model θ $\Rightarrow \widehat{\theta}_T$ satisfies a large deviation principle

Restricted optimization problem

Original problem (★)

$$\begin{cases} \min\limits_{\widehat{c},\,\widehat{x}} & \left\{ \lim\limits_{T\to\infty} \mathbb{E}_{\theta}\left[\widehat{c}_T(\widehat{x}_T)\right] \right\}_{\theta\in\Theta} \\ \text{s.t.} & \lim\limits_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\theta}\left(c(\widehat{x}_T, \theta) > \widehat{c}_T(\widehat{x}_T) \right) \le -r \;\; \forall \theta \end{cases}$$



Restricted problem (★★)

$$\begin{cases} \min\limits_{\tilde{c},\,\tilde{x}} & \left\{ \tilde{c}(\tilde{x}(\theta), \theta) \right\}_{\theta\in\Theta} \\ \text{s.t.} & \lim\limits_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\theta}\left(c(\tilde{x}(\widehat{\theta}_T), \theta) > \tilde{c}(\tilde{x}(\widehat{\theta}_T), \widehat{\theta}_T) \right) \le -r \;\; \forall \theta \end{cases}$$

Large Deviations Theory

Definition: A sequence $\{\widehat{\theta}_T\}_{T\in\mathbb{N}}$ satisfies a **Large Deviation Principle** (LDP) if there is a "*distance*" function $I(\theta',\theta)$ such that for any Borel set $\mathcal{D}\subset\Theta'$

$$-\underbrace{\inf_{\theta'\in\mathrm{int}\mathcal{D}} I(\theta',\theta)}_{r} \le \liminf_{T\to\infty} \frac{1}{T}\log\mathbb{P}_\theta\left(\widehat{\theta}_T\in\mathcal{D}\right)$$

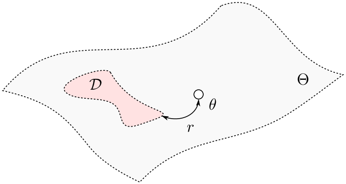
$$\le \limsup_{T\to\infty} \frac{1}{T}\log\mathbb{P}_\theta\left(\widehat{\theta}_T\in\mathcal{D}\right) \le -\underbrace{\inf_{\theta'\in\mathrm{cl}\mathcal{D}} I(\theta',\theta)}_{r}$$

Large Deviations Theory

Definition: A sequence $\{\widehat{\theta}_T\}_{T\in\mathbb{N}}$ satisfies a **Large Deviation Principle** (LDP) if there is a "*distance*" function $I(\theta',\theta)$ such that for any Borel set $\mathcal{D}\subset\Theta'$

$$-\underbrace{\inf_{\theta'\in\mathrm{int}\mathcal{D}} I(\theta',\theta)}_{r} \le \liminf_{T\to\infty} \frac{1}{T}\log\mathbb{P}_\theta\left(\widehat{\theta}_T\in\mathcal{D}\right)$$

$$\le \limsup_{T\to\infty} \frac{1}{T}\log\mathbb{P}_\theta\left(\widehat{\theta}_T\in\mathcal{D}\right) \le -\underbrace{\inf_{\theta'\in\mathrm{cl}\mathcal{D}} I(\theta',\theta)}_{r}$$



$$\mathbb{P}_\theta\left(\widehat{\theta}_T\in\mathcal{D}\right) = e^{-r\cdot T+o(T)}$$

Varadhan: *Rare events do occur every day. Someone always wins a lottery!*

Large Deviations Theory (cont'd)

Definition: $I : \Theta' \times \mathrm{cl}\Theta \to [0, \infty]$ is called a **regular rate function** if it is

(i) **Radially monotonic** in θ, i.e.,

$$\{\theta \in \mathrm{cl}\Theta : I(\theta', \theta) \leq r\} \subseteq \mathrm{cl}\{\theta \in \Theta : I(\theta', \theta) < r\}$$

(ii) **Continuous** on $\Theta' \times \Theta$

(iii) **Level-compact**, i.e.,
$\{(\theta, \theta') \in \mathrm{cl}\Theta \times \mathrm{cl}\Theta' : I(\theta', \theta) \leq r\}$ is compact $\forall r > 0$

Large Deviations Theory (cont'd)

Definition: $I : \Theta' \times \mathrm{cl}\Theta \to [0, \infty]$ is called a **regular rate function** if it is

(i) **Radially monotonic** in θ, i.e.,

$$\{\theta \in \mathrm{cl}\Theta : I(\theta', \theta) \leq r\} \subseteq \mathrm{cl}\{\theta \in \Theta : I(\theta', \theta) < r\}$$

(ii) **Continuous** on $\Theta' \times \Theta$

(iii) **Level-compact**, i.e.,
$\{(\theta, \theta') \in \mathrm{cl}\Theta \times \mathrm{cl}\Theta' : I(\theta', \theta) \leq r\}$ is compact $\forall r > 0$

Examples:

- Relative entropy $\Theta = \Theta' = \Delta_d$, $I(\theta', \theta) = D(\theta' \| \theta)$
- Ellipsoid $\Theta = \Theta = \mathbb{R}$, $I(\theta', \theta) = (\theta - \theta')^\top \Sigma^{-1} (\theta - \theta')$
- many more ...

Pareto-dominant solution to restricted optimization problem

Restricted problem (★★)

$$\begin{cases} \min\limits_{\tilde{c},\, \tilde{x}} & \{\tilde{c}(\tilde{x}(\theta), \theta)\}_{\theta \in \Theta} \\ \text{s.t.} & \lim\limits_{T \to \infty} \frac{1}{T} \log \mathbb{P}_{\theta} \left(c(\tilde{x}(\widehat{\theta}_T), \theta) > \tilde{c}(\tilde{x}(\widehat{\theta}_T), \widehat{\theta}_T) \right) \leq -r \;\; \forall \theta \end{cases}$$

Assumption: $\widehat{\theta}_T$ satisfies an LDP with regular rate function I

Theorem. The Pareto-dominant solution to (★★) is given by the distributionally robust predictor

$$\tilde{c}(x, \widehat{\theta}_T) = \begin{cases} \max\limits_{\theta \in \Theta} & c(x, \theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \leq r \end{cases}$$

DRO is optimal

Assumption: $\widehat{\theta}_T$ satisfies an LDP with regular rate function I

Theorem. The Pareto-dominant solution to (★★) is given by the distributionally robust predictor

$$\tilde{c}(x, \widehat{\theta}_T) = \begin{cases} \max\limits_{\theta \in \Theta} & c(x, \theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \leq r \end{cases}$$

- Shape of the ambiguity set determined by $\widehat{\theta}_T$
- Radius of ambiguity set determines the desired decay rate

Separation principle

Assumptions:

- $\widehat{\theta}_T$ satisfies an LDP with regular rate function I
- $\widehat{\theta}_T$ is a sufficient statistic for θ

Theorem. The Pareto-dominant solution to (★) is given by the distributionally robust predictor

$$\widehat{c}_T(x) = \begin{cases} \max\limits_{\theta \in \Theta} & c(x, \theta) \\ \text{s. t.} & I(\widehat{\theta}_T, \theta) \leq r \end{cases}$$

- Problems (★) and (★★) have the same optimal solution $\Rightarrow$ Separation of estimation and optimization is without loss of optimality
- Sufficiency restricts to exponential fam. of distributions for $\mathbb{P}_\theta$
- Non-convex Rao-Blackwell type result

Conclusions of the Separation Theorem

① **DRO predictors are optimal** in a wide sense

$$\widehat{c}_T(x) = \begin{cases} \max_{\theta \in \Theta} & c(x,\theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \le r \end{cases}$$

Conclusions of the Separation Theorem

① **DRO predictors are optimal** in a wide sense

$$\widehat{c}_T(x) = \begin{cases} \max\limits_{\theta \in \Theta} & c(x,\theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \le r \end{cases}$$

② **Ambiguity set** is **induced** by the choice of **estimator**

- $I(\cdot, \theta)$ is the rate function related to the estimator $\widehat{\theta}_T$
- size of the ambiguity set r quantifies the decay rate of the disappointment probability

Conclusions of the Separation Theorem

① **DRO predictors are optimal** in a wide sense

$$\widehat{c}_T(x) = \begin{cases} \max_{\theta \in \Theta} & c(x,\theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \le r \end{cases}$$

② **Ambiguity set** is **induced** by the choice of **estimator**

- $I(\cdot,\theta)$ is the rate function related to the estimator $\widehat{\theta}_T$
- size of the ambiguity set r quantifies the decay rate of the disappointment probability

③ **Invariance principle**

- $\psi : \Theta' \to \Theta'$ homeomorphism
- $\psi(\widehat{\theta}_T)$ satisfies LDP with rate function $I^{\psi}(\theta',\theta) = I(\psi^{-1}(\theta'),\theta)$
- DRO predictor is invariant

$$\begin{cases} \max_{\theta \in \Theta} & c(x,\theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \le r \end{cases} = \begin{cases} \max_{\theta \in \Theta} & c(x,\theta) \\ \text{s.t.} & I^{\psi}(\psi(\widehat{\theta}_T), \theta) \le r \end{cases}$$

Revisit newsvendor problem

- # copies stocked: $x \in \mathbb{X} = \{0, 1, \cdots, d\}$
- random daily demand: $\xi \in \Xi = \{0, 1, \cdots, d\}$
- model: $\mathbb{P}_{\theta}[\xi \in i] = \theta_i$
- cost: $c(x, \theta) = \sum_{i=0}^{d} \theta_i(-p \min\{x, i\}) + kx$
- estimator: $(\widehat{\theta}_T)_i = \frac{1}{T} \sum_{t=1}^{T} 1_{\xi_t = i}, \quad i = 0, \cdots, d$

Revisit newsvendor problem

- # copies stocked: $x \in \mathbb{X} = \{0, 1, \cdots, d\}$
- random daily demand: $\xi \in \Xi = \{0, 1, \cdots, d\}$
- model: $\mathbb{P}_{\theta}[\xi \in i] = \theta_i$
- cost: $c(x, \theta) = \sum_{i=0}^{d} \theta_i(-p \min\{x, i\}) + kx$
- estimator: $(\widehat{\theta}_T)_i = \frac{1}{T} \sum_{t=1}^{T} 1_{\xi_t = i}, \quad i = 0, \cdots, d$

Sanov's Theorem. The estimator $\widehat{\theta}_T$ satisfies a large deviation principle with regular rate function $I(\widehat{\theta}_T, \theta) = \mathrm{D}(\widehat{\theta}_T \| \theta)$

Revisit newsvendor problem

- # copies stocked: $x \in \mathbb{X} = \{0, 1, \cdots, d\}$
- random daily demand: $\xi \in \Xi = \{0, 1, \cdots, d\}$
- model: $\mathbb{P}_\theta[\xi \in i] = \theta_i$
- cost: $c(x, \theta) = \sum_{i=0}^{d} \theta_i(-p \min\{x, i\}) + kx$
- estimator: $(\widehat{\theta}_T)_i = \frac{1}{T}\sum_{t=1}^{T} 1_{\xi_t = i}, \quad i = 0, \cdots, d$

Sanov's Theorem. The estimator $\widehat{\theta}_T$ satisfies a large deviation principle with regular rate function $I(\widehat{\theta}_T, \theta) = \mathrm{D}(\widehat{\theta}_T \| \theta)$

- $\widehat{\theta}_T$ is a sufficient statistic (Fisher-Neyman)
- DRO predictor with relative entropy ambiguity set is optimal

$$\widehat{c}_T(x) = \tilde{c}(x, \widehat{\theta}_T) = \begin{cases} \max\limits_{\theta \in \Theta} & c(x, \theta) \\ \text{s. t.} & \mathrm{D}(\widehat{\theta}_T \| \theta) \le r \end{cases}$$

Correlated setting: finite state Markov chain

- $\{\xi_t\}_{t \in \mathbb{N}}$ stationary Markov chain, initial state σ

Correlated setting: finite state Markov chain

- $\{\xi_t\}_{t\in\mathbb{N}}$ stationary Markov chain, initial state σ
- pairwise description $\theta_{ij} = \mathbb{P}[\xi_t = i, \xi_{t+1} = j]$

Correlated setting: finite state Markov chain

- $\{\xi_t\}_{t\in\mathbb{N}}$ stationary Markov chain, initial state σ
- pairwise description $\theta_{ij} = \mathbb{P}[\xi_t = i, \xi_{t+1} = j]$
- Models considered (irreducible)

$$\Theta = \left\{ \theta \in \mathbb{R}^{d\times d}_{++} : \sum_{i,j\in\Xi} \theta_{ij} = 1, \ \sum_{j\in\Xi} \theta_{ij} = \sum_{j\in\Xi} \theta_{ji} \ \forall i \in \Xi \right\}$$

Correlated setting: finite state Markov chain

- $\{\xi_t\}_{t\in\mathbb{N}}$ stationary Markov chain, initial state σ
- pairwise description $\theta_{ij} = \mathbb{P}[\xi_t = i, \xi_{t+1} = j]$
- Models considered (irreducible)

$$\Theta = \left\{\theta \in \mathbb{R}_{++}^{d\times d} : \sum_{i,j\in\Xi} \theta_{ij} = 1,\ \sum_{j\in\Xi} \theta_{ij} = \sum_{j\in\Xi} \theta_{ji}\ \ \forall i \in \Xi\right\}$$

- Estimator: $(\widehat{\theta}_T)_{ij} = \frac{1}{T}(1_{\sigma=i}1_{\xi_1=j} + \sum_{t=1}^{T-1} 1_{\xi_t=i}1_{\xi_{t+1}=j})$
  sufficient statistic for θ (Fisher-Neyman)

Correlated setting: finite state Markov chain

- $\{\xi_t\}_{t\in\mathbb{N}}$ stationary Markov chain, initial state σ
- pairwise description $\theta_{ij} = \mathbb{P}[\xi_t = i, \xi_{t+1} = j]$
- Models considered (irreducible)

$$\Theta = \left\{\theta \in \mathbb{R}^{d\times d}_{++} : \sum_{i,j\in\Xi} \theta_{ij} = 1,\ \sum_{j\in\Xi} \theta_{ij} = \sum_{j\in\Xi} \theta_{ji}\ \ \forall i \in \Xi\right\}$$

- Estimator: $(\widehat{\theta}_T)_{ij} = \frac{1}{T}(1_{\sigma=i}1_{\xi_1=j} + \sum_{t=1}^{T-1} 1_{\xi_t=i}1_{\xi_{t+1}=j})$
  sufficient statistic for θ (Fisher-Neyman)
- Estimator state space

$$\Theta' = \left\{\theta \in \mathbb{R}^{d\times d}_{+} : \sum_{i,j\in\Xi} \theta_{ij} = 1\right\}$$

Correlated setting: finite state Markov chain

- $\{\xi_t\}_{t\in\mathbb{N}}$ stationary Markov chain, initial state σ
- pairwise description $\theta_{ij} = \mathbb{P}[\xi_t = i, \xi_{t+1} = j]$
- Models considered (irreducible)

$$\Theta = \left\{\theta \in \mathbb{R}^{d\times d}_{++} : \sum_{i,j\in\Xi} \theta_{ij} = 1,\ \sum_{j\in\Xi} \theta_{ij} = \sum_{j\in\Xi} \theta_{ji} \ \ \forall i \in \Xi\right\}$$

- Estimator: $(\widehat{\theta}_T)_{ij} = \frac{1}{T}(1_{\sigma=i}1_{\xi_1=j} + \sum_{t=1}^{T-1} 1_{\xi_t=i}1_{\xi_{t+1}=j})$ sufficient statistic for θ (Fisher-Neyman)
- Estimator state space

$$\Theta' = \left\{\theta \in \mathbb{R}^{d\times d}_{+} : \sum_{i,j\in\Xi} \theta_{ij} = 1\right\}$$

- "Distance measure" between estimator and underlying model $\rightarrow$ conditional relative entropy

Correlated setting: finite state Markov chain (cont'd)

Conditional relative entropy. For any $\theta \in \Theta, \theta' \in \Theta'$

$$\mathrm{D_c}(\theta' \| \theta) = \sum_{i,j \in \Xi} \theta'_{ij} \left(\log\left(\frac{\theta'_{ij}}{\sum_{k \in \Xi} \theta'_{ik}} \right) - \log\left(\frac{\theta_{ij}}{\sum_{k \in \Xi} \theta_{ik}} \right) \right)$$

Correlated setting: finite state Markov chain (cont'd)

Conditional relative entropy. For any $\theta \in \Theta, \theta' \in \Theta'$

$$\mathrm{D_c}(\theta' \| \theta) = \sum_{i,j \in \Xi} \theta'_{ij} \left(\log\left(\frac{\theta'_{ij}}{\sum_{k \in \Xi} \theta'_{ik}} \right) - \log\left(\frac{\theta_{ij}}{\sum_{k \in \Xi} \theta_{ik}} \right) \right)$$

- Similar properties as the relative entropy
  - non-negative
  - $\mathrm{D_c}(\theta' \| \theta) = 0 \Leftrightarrow \theta' = \theta$
  - non-convex in $\theta \Rightarrow$ solving DRO problem [Li, S., Kuhn, ICML 2021]

Correlated setting: finite state Markov chain (cont'd)

Conditional relative entropy. For any $\theta \in \Theta, \theta' \in \Theta'$

$$\mathrm{D_c}(\theta' \| \theta) = \sum_{i,j \in \Xi} \theta'_{ij} \left(\log\left(\frac{\theta'_{ij}}{\sum_{k \in \Xi} \theta'_{ik}} \right) - \log\left(\frac{\theta_{ij}}{\sum_{k \in \Xi} \theta_{ik}} \right) \right)$$

- Similar properties as the relative entropy
  - non-negative
  - $\mathrm{D_c}(\theta' \| \theta) = 0 \Leftrightarrow \theta' = \theta$
  - non-convex in θ $\Rightarrow$ solving DRO problem [Li, S., Kuhn, ICML 2021]

Lemma. The estimator $\widehat{\theta}_T$ satisfies an LDP with (regular) rate function $I(\widehat{\theta}_T, \theta) = \mathrm{D_c}(\widehat{\theta}_T \| \theta)$

Correlated setting: finite state Markov chain (cont'd)

Conditional relative entropy. For any $\theta \in \Theta, \theta' \in \Theta'$

$$\mathrm{D_c}(\theta' \| \theta) = \sum_{i,j \in \Xi} \theta'_{ij} \left(\log \left(\frac{\theta'_{ij}}{\sum_{k \in \Xi} \theta'_{ik}} \right) - \log \left(\frac{\theta_{ij}}{\sum_{k \in \Xi} \theta_{ik}} \right) \right)$$

- Similar properties as the relative entropy
  - non-negative
  - $\mathrm{D_c}(\theta' \| \theta) = 0 \Leftrightarrow \theta' = \theta$
  - non-convex in θ $\Rightarrow$ solving DRO problem [Li, S., Kuhn, ICML 2021]

Lemma. The estimator $\widehat{\theta}_T$ satisfies an LDP with (regular) rate function $I(\widehat{\theta}_T, \theta) = \mathrm{D_c}(\widehat{\theta}_T \| \theta)$

- [Dembo & Zeitouni, Chapter 3]
- Separation Theorem holds

IID process with unknown mean

- i.i.d. process $\{\xi_t\}_{t\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\theta}$, unknown mean $\mathbb{E}_{\mathbb{P}_{\theta}}[\xi_1] = \theta$

IID process with unknown mean

- i.i.d. process $\{\xi_t\}_{t \in \mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\theta$, unknown mean $\mathbb{E}_{\mathbb{P}_\theta}[\xi_1] = \theta$
- F_θ distribution of ξ_1 under $\mathbb{P}_\theta$

IID process with unknown mean

- i.i.d. process $\{\xi_t\}_{t\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\theta$, unknown mean $\mathbb{E}_{\mathbb{P}_\theta}[\xi_1] = \theta$
- F_θ distribution of ξ_1 under $\mathbb{P}_\theta$
- log-moment generating function $\Lambda(\lambda, \theta) = \log\left(\int_\Xi e^{\lambda^\top \xi}\, \mathrm{d}F_\theta(\xi)\right)$

IID process with unknown mean

- i.i.d. process $\{\xi_t\}_{t\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\theta}$,   unknown mean $\mathbb{E}_{\mathbb{P}_{\theta}}[\xi_1] = \theta$
- F_{θ} distribution of ξ_1 under $\mathbb{P}_{\theta}$
- log-moment generating function $\Lambda(\lambda, \theta) = \log\left(\int_{\Xi} e^{\lambda^\top \xi}\, \mathrm{d}F_{\theta}(\xi)\right)$
- Estimator $\widehat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \xi_t$
  - consistent by Law of Large Numbers
  - for many distributions is a sufficient statistic

IID process with unknown mean

- i.i.d. process $\{\xi_t\}_{t\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\theta}$, unknown mean $\mathbb{E}_{\mathbb{P}_\theta}[\xi_1] = \theta$
- F_θ distribution of ξ_1 under $\mathbb{P}_\theta$
- log-moment generating function $\Lambda(\lambda, \theta) = \log\left(\int_\Xi e^{\lambda^\top \xi}\, \mathrm{d}F_\theta(\xi)\right)$
- Estimator $\widehat{\theta}_T = \frac{1}{T}\sum_{t=1}^T \xi_t$
  - consistent by Law of Large Numbers
  - for many distributions is a sufficient statistic
- Cramér function $\Lambda^*(s, \theta) = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, s\rangle - \Lambda(\lambda, \theta)\}$

IID process with unknown mean

- i.i.d. process $\{\xi_t\}_{t\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\theta$, unknown mean $\mathbb{E}_{\mathbb{P}_\theta}[\xi_1] = \theta$
- F_θ distribution of ξ_1 under $\mathbb{P}_\theta$
- log-moment generating function $\Lambda(\lambda, \theta) = \log\left(\int_\Xi e^{\lambda^\top \xi}\, \mathrm{d}F_\theta(\xi)\right)$
- Estimator $\widehat{\theta}_T = \frac{1}{T}\sum_{t=1}^T \xi_t$
  - consistent by Law of Large Numbers
  - for many distributions is a sufficient statistic
- Cramér function $\Lambda^*(s, \theta) = \sup_{\lambda\in\mathbb{R}^d}\{\langle\lambda, s\rangle - \Lambda(\lambda, \theta)\}$

Lemma. The estimator $\widehat{\theta}_T$ satisfies an LDP with (regular) rate function $I(\theta', \theta) = \Lambda^*(\theta', \theta)$

IID process with unknown mean

- i.i.d. process $\{\xi_t\}_{t\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\theta$, unknown mean $\mathbb{E}_{\mathbb{P}_\theta}[\xi_1] = \theta$
- F_θ distribution of ξ_1 under $\mathbb{P}_\theta$
- log-moment generating function $\Lambda(\lambda, \theta) = \log\left(\int_\Xi e^{\lambda^\top \xi}\, \mathrm{d}F_\theta(\xi)\right)$
- Estimator $\widehat{\theta}_T = \frac{1}{T}\sum_{t=1}^T \xi_t$
  - consistent by Law of Large Numbers
  - for many distributions is a sufficient statistic
- Cramér function $\Lambda^*(s, \theta) = \sup_{\lambda\in\mathbb{R}^d}\{\langle \lambda, s\rangle - \Lambda(\lambda, \theta)\}$

Lemma. The estimator $\widehat{\theta}_T$ satisfies an LDP with (regular) rate function $I(\theta', \theta) = \Lambda^*(\theta', \theta)$

- Consequence of Cramér's Theorem
- Separation Theorem holds for many distributions

IID process with unknown mean (cont'd)

Each distribution induces an ambiguity set

$$\{\theta \in \Theta \;:\; \Lambda^*(\theta', \theta) \leq r\}$$

F_θ	$\Lambda^*(\theta', \theta)$	$\mathrm{dom}(\Lambda^*(\cdot, \theta))$
(a) Normal	$\frac{1}{2}(\theta' - \theta)^\top \Sigma^{-1}(\theta' - \theta)$	$\mathbb{R}^d$
(b) Exponential	$\frac{\theta' - \theta}{\theta} + \log(\theta/\theta')$	$\mathbb{R}_{++}$
(c) Poisson	$\theta' \log(\theta'/\theta) - \theta' + \theta$	$\mathbb{R}_{++}$
(d) Bernoulli	$\theta' \log(\frac{\theta'(1-\theta)}{\theta(1-\theta')}) - \log(\frac{1-\theta}{1-\theta'})$	$(0, 1)$

Many more possible, e.g., Gamma, Geometric, Binomial, ...

Summary

Meta-optimization problem

- optimizes over surrogate optimization models
- balances in-sample risk vs. out-of-sample disappointment
- pushes down the out-of-sample risk

Separation of estimation and optimization

- holds if $\widehat{\theta}_T$ is a sufficient statistic that obeys an LDP
- reminiscent of Rao-Blackwell theorem

Pareto-dominant solution is a DRO model

- ambiguity set is a rate-ball around $\widehat{\theta}_T$
- radius = decay rate of the out-of-sample disappointment
- invariant under homeomorphic transformations

Conclusion

1. **DRO is optimal:** Optimal data-driven predictor is given as a DRO problem centred around an estimator

2. **Ambiguity set**
   - **Structure** induced by underlying stochastic process via LDP
   - **Size** has operational meaning as the decay rate of the disappointment probability

3. Data-driven DRO framework for **non-i.i.d. data**

Outlook

- The proposed prescriptor is not consistent

$$\widehat{c}_T(x) \not\to c(x, \theta) \quad \text{as } T \to \infty$$

- **Idea:** Can we trade speed in the decay of

$$\mathbb{P}_\theta \left(c(\widehat{x}_T, \theta) > \widehat{c}_T(\widehat{x}_T) \right)$$

  to achieve consistency?

- Are there other statistical criteria for optimality?

$\Longrightarrow$ A. Ganguly and T. Sutter, *Optimal learning via Moderate Deviations Theory*, *arXiv:2305.14496*, 2023

Reference

This talk

- T. Sutter, B.P. Van Parys, and D. Kuhn, *A General Framework for Optimal Data-Driven Optimization*, arXiv:2010.06606, 2020
- A. Ganguly and T. Sutter, *Optimal learning via Moderate Deviations Theory* arXiv:2305.14496, 2023
- M. Li, T. Sutter, and D. Kuhn, *Distributionally Robust Optimization with Markovian Data*, ICML, 2021

Appendix

Interval estimation

Goal: Estimate cost $c(\theta)$ via a confidence interval $\widehat{\mathcal{I}}^{\star}_T = \mathcal{I}^{\star}_T(\widehat{\theta}_T)$ where $\mathcal{I}^{\star}_T(\theta) = [\underline{c}^{\star}_{T,r}(\theta), \bar{c}^{\star}_{T,r}(\theta)]$ with the following properties

① **Exponential accuracy:**

$$\mathbb{P}_{\theta}(c(\theta) \notin \widehat{\mathcal{I}}^{\star}_T) \leq e^{-rb_T}, \qquad 1 \ll b_T \ll T$$

Interval estimation

Goal: Estimate cost $c(\theta)$ via a confidence interval $\widehat{\mathcal{I}}_T^\star = \mathcal{I}_T^\star(\widehat{\theta}_T)$ where $\mathcal{I}_T^\star(\theta) = [\underline{c}_{T,r}^\star(\theta), \bar{c}_{T,r}^\star(\theta)]$ with the following properties

1. **Exponential accuracy:**

$$\mathbb{P}_\theta(c(\theta) \notin \widehat{\mathcal{I}}_T^\star) \leq e^{-rb_T}, \qquad 1 \ll b_T \ll T$$

2. **Minimality:** Any interval $\mathcal{I}_T(\theta) = [\underline{c}_{T,r}(\theta), \bar{c}_{T,r}(\theta)]$ satisfying 1 is eventually larger than $\mathcal{I}_T^\star(\theta)$

Interval estimation

Goal: Estimate cost $c(\theta)$ via a confidence interval $\widehat{\mathcal{I}}^\star_T = \mathcal{I}^\star_T(\widehat{\theta}_T)$ where $\mathcal{I}^\star_T(\theta) = [\underline{c}^\star_{T,r}(\theta), \bar{c}^\star_{T,r}(\theta)]$ with the following properties

① **Exponential accuracy:**

$$\mathbb{P}_\theta(c(\theta) \notin \widehat{\mathcal{I}}^\star_T) \leq e^{-rb_T}, \qquad 1 \ll b_T \ll T$$

② **Minimality:** Any interval $\mathcal{I}_T(\theta) = [\underline{c}_{T,r}(\theta), \bar{c}_{T,r}(\theta)]$ satisfying ① is eventually larger than $\mathcal{I}^\star_T(\theta)$

③ **Consistency:** $\widehat{\mathcal{I}}^\star_T \to \{c(\theta)\}$ as $T \to \infty$

Interval estimation

Goal: Estimate cost $c(\theta)$ via a confidence interval $\widehat{\mathcal{I}}_T^\star = \mathcal{I}_T^\star(\widehat{\theta}_T)$ where $\mathcal{I}_T^\star(\theta) = [\underline{c}_{T,r}^\star(\theta), \bar{c}_{T,r}^\star(\theta)]$ with the following properties

① **Exponential accuracy:**

$$\mathbb{P}_\theta(c(\theta) \notin \widehat{\mathcal{I}}_T^\star) \leq e^{-rb_T}, \qquad 1 \ll b_T \ll T$$

② **Minimality:** Any interval $\mathcal{I}_T(\theta) = [\underline{c}_{T,r}(\theta), \bar{c}_{T,r}(\theta)]$ satisfying ① is eventually larger than $\mathcal{I}_T^\star(\theta)$

③ **Consistency:** $\widehat{\mathcal{I}}_T^\star \to \{c(\theta)\}$ as $T \to \infty$

④ **Mischaracterization probability:**

$$\mathbb{P}_\theta(c(\theta') \notin \widehat{\mathcal{I}}_T^\star) > e^{-rb_T}, \qquad \forall \theta' \ : \ c(\theta') \neq c(\theta)$$

Interval estimation

Goal: Estimate cost $c(\theta)$ via a confidence interval $\widehat{\mathcal{I}}^\star_T = \mathcal{I}^\star_T(\widehat{\theta}_T)$ where $\mathcal{I}^\star_T(\theta) = [\underline{c}^\star_{T,r}(\theta), \bar{c}^\star_{T,r}(\theta)]$ with the following properties

1. **Exponential accuracy:**

$$\mathbb{P}_\theta(c(\theta) \notin \widehat{\mathcal{I}}^\star_T) \leq e^{-rb_T}, \qquad 1 \ll b_T \ll T$$

2. **Minimality:** Any interval $\mathcal{I}_T(\theta) = [\underline{c}_{T,r}(\theta), \bar{c}_{T,r}(\theta)]$ satisfying (1) is eventually larger than $\mathcal{I}^\star_T(\theta)$
3. **Consistency:** $\widehat{\mathcal{I}}^\star_T \to \{c(\theta)\}$ as $T \to \infty$
4. **Mischaracterization probability:**

$$\mathbb{P}_\theta(c(\theta') \notin \widehat{\mathcal{I}}^\star_T) > e^{-rb_T}, \qquad \forall \theta' \ : \ c(\theta') \neq c(\theta)$$

5. **Uniformly most accurate (UMA):** Any interval $\widehat{\mathcal{I}}_T$ satisfying (1) is such that

$$\mathbb{P}_\theta(c(\theta') \in \widehat{\mathcal{I}}^\star_T) \leq \mathbb{P}_\theta(c(\theta') \in \widehat{\mathcal{I}}_T), \qquad \forall \theta' \ : \ c(\theta') \neq c(\theta)$$

Heuristic CLT based confidence intervals

- Given a fixed α, CLT guarantees

$$\lim_{T\to\infty} \mathbb{P}_\theta(c(\theta) \notin \mathcal{I}^{\mathsf{CLT}}_{T,\alpha}(\widehat{\theta}_T)) \leq \alpha$$

- for the CLT-based interval

$$\mathcal{I}^{\mathsf{CLT}}_{T,\alpha}(\widehat{\theta}_T) = \left[c(\widehat{\theta}_T) - \kappa^{\mathsf{CLT}}_T(\alpha), c(\widehat{\theta}_T) + \kappa^{\mathsf{CLT}}_T(\alpha)\right]$$

$$\kappa^{\mathsf{CLT}}_T(\alpha) = \Phi^{-1}(1-\alpha/2)\sqrt{\nabla c(\widehat{\theta}_T)^\top S(\widehat{\theta}_T)\nabla c(\widehat{\theta}_T)}/\sqrt{T}$$

- Heuristic choice $\alpha = e^{-rb_T}$

Question: Does the CI $\mathcal{I}^{\mathsf{CLT}}_{T,\alpha}(\widehat{\theta}_T)$ for $\alpha = e^{-rT}$ satisfy any of the properties ① – ⑤?

Optimal confidence interval

- Interval $\widehat{\mathcal{I}}_T^\star = \mathcal{I}_T^\star(\widehat{\theta}_T)$ for $\mathcal{I}_T^\star(\theta) = [\underline{c}_{T,r}^\star(\theta), \overline{c}_{T,r}^\star(\theta)]$
  - $\underline{c}_{T,r}^\star(\theta') = \inf_{\theta\in\Theta}\{c(\theta) \ : \ I^M(a_T(\theta'-\theta),\theta) \le r\}$
  - $\overline{c}_{T,r}^\star(\theta') = \sup_{\theta\in\Theta}\{c(\theta) \ : \ I^M(a_T(\theta'-\theta),\theta) \le r\}$
- $I^M(\cdot,\theta)$: Moderate deviation rate function of $\widehat{\theta}_T$
- $a_T = \sqrt{T/b_T}, \quad 1 \ll b_T \ll T$

The confidence interval $\widehat{\mathcal{I}}_T^\star$ satisfies the properties ① – ⑤

- Θ can be infinite dimensional
- mild assumptions on $I^M(\cdot,\theta)$

Example: Asymptotic variance of OU process

- Ornstein-Uhlenbeck process

$$\mathrm{d}X_t = -\theta X_t \mathrm{d}t + \mathrm{d}W_t, \quad X_0 = 0$$

- Asymptotic variance

$$c(\theta) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \left(\mathbb{E}_\theta[X_t^2] - \mathbb{E}_\theta[X_t]^2\right) = \frac{1}{2\theta}$$

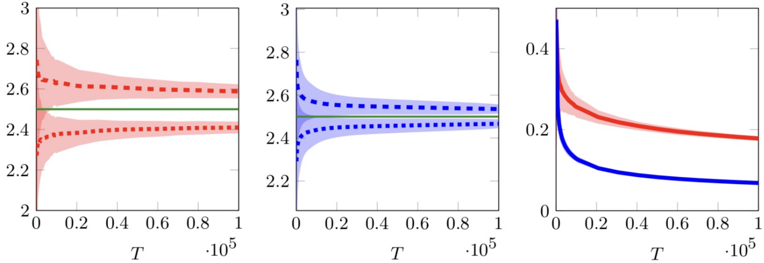
- Maximum likelihood estimator

$$\widehat{\theta}_T = -\frac{X_T^2 - T}{2\int_0^T X_t^2 \mathrm{d}t}, \qquad I^M(\vartheta, \theta) = \frac{\vartheta^2}{2\theta}$$

- Optimal CI $\mathcal{I}_T^\star(\widehat{\theta}_T) = [c(\widehat{\theta}_T) + \kappa_T^-, c(\widehat{\theta}_T) + \kappa_T^+]$

$$\kappa_T^\pm = \frac{1}{2\widehat{\theta}_T^2}\left(r_T \pm \sqrt{r_T^2 + 2\widehat{\theta}_T r_T}\right), \quad r_T = rb_T/T$$

Example: Asymptotic variance of OU process



(a) Upper and lower confidence bound (b) Upper and lower confidence bound (c) Interval length

- Optimal value $c(\theta)$
- Optimal interval $\mathcal{I}_T^\star(\widehat{\theta}_T)$
- CLT interval $\mathcal{I}_{T,\alpha}^{\mathsf{CLT}}(\widehat{\theta}_T)$, $\alpha = e^{-rb_T}$