A General Framework For Optimal Data-Driven Optimization

Tobias Sutter,[1] Bart Van Parys,[2] Daniel Kuhn [1]

[1] Risk Analytics and Optimization Chair, EPFL
www.epfl.ch/labs/rao/

[2] MIT Sloan School of Management
web.mit.edu/vanparys/www/

FNSNF Swiss National Science Foundation

NCCR Automation

EPFL

Data-Driven Decision-Making

Stochastic opti-
mization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Examples:

Data-Driven Decision-Making

Stochastic optimization problem	Family of probability measures	Data-generating process
$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$	$\{\mathbb{P}_\theta : \theta \in \Theta\}$	$\{\xi_t\}_{t \in \mathbb{N}}$

Examples:

- Expected loss $\quad c(x, \theta) = \mathbb{E}_\theta[\ell(x, \xi)]$

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Examples:

- Expected loss $\qquad c(x, \theta) = \mathbb{E}_\theta[\ell(x, \xi)]$
- Risk of loss $\qquad c(x, \theta) = \rho_\theta[\ell(x, \xi)]$

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Examples:

- Expected loss $\quad c(x, \theta) = \mathbb{E}_\theta[\ell(x, \xi)]$
- Risk of loss $\quad c(x, \theta) = \rho_\theta[\ell(x, \xi)]$
- Covariate information $\quad c(x, \theta) = \mathbb{E}_\theta[\ell(x, \xi) | C\xi \in B]$

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Examples:

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

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Assumptions:

Data-Driven Decision-Making

Stochastic optimization problem

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

Family of probability measures

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

Data-generating process

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

Assumptions:

- All measures defined on $(\Omega, \mathcal{F})$

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Assumptions:

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

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Examples:

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Examples:

- Finite-state i.i.d. processes

Data-Driven Decision-Making

Stochastic optimization problem	Family of probability measures	Data-generating process
$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$	$\{\mathbb{P}_\theta : \theta \in \Theta\}$	$\{\xi_t\}_{t \in \mathbb{N}}$

Examples:

- Finite-state i.i.d. processes
- Finite-state Markov chains

Data-Driven Decision-Making

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Family of probability measures

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

Data-generating process

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

Examples:

- Finite-state i.i.d. processes
- Finite-state Markov chains
- Vector-autoregressive processes

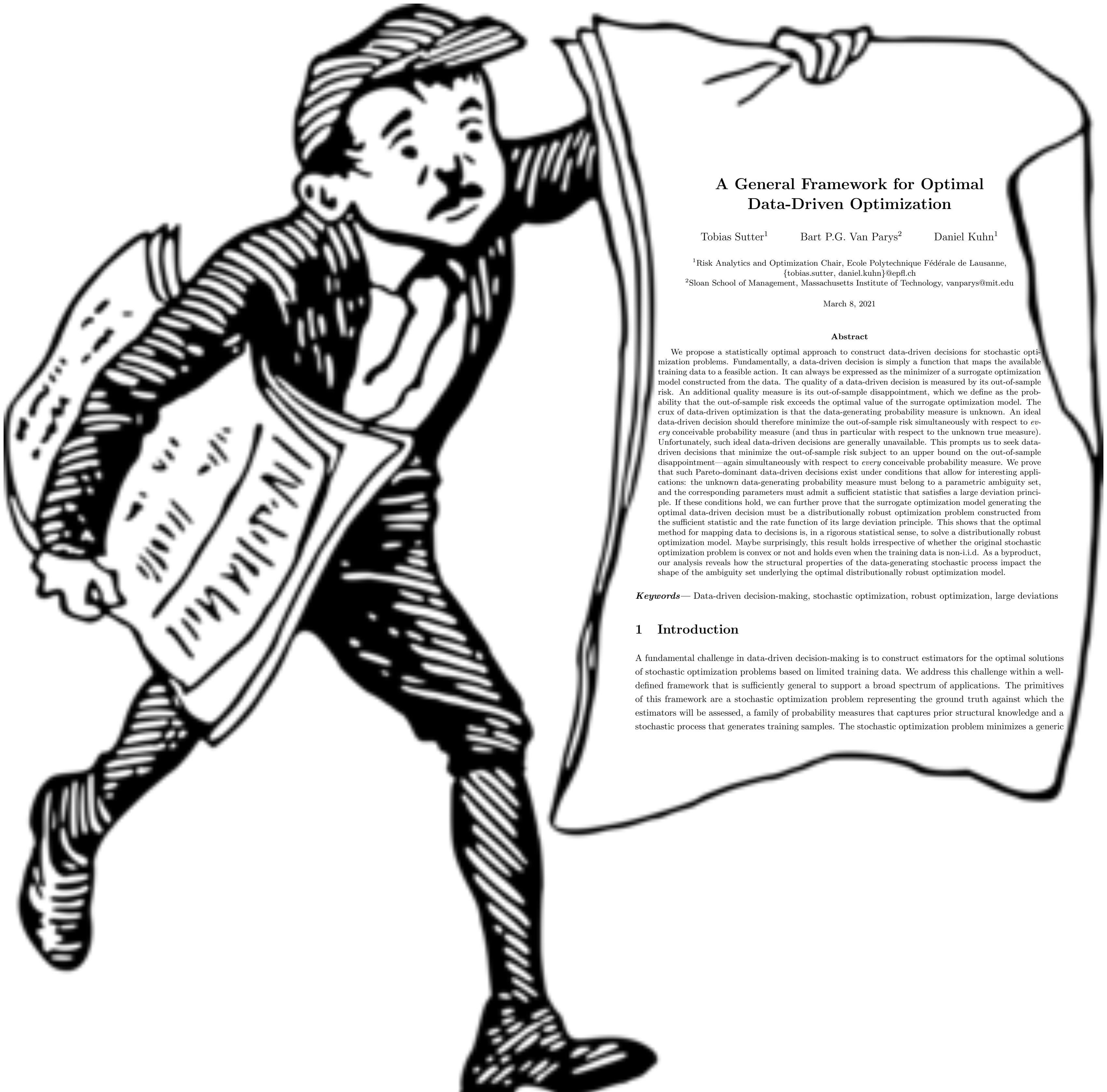
Data-Driven Decision-Making

Stochastic optimization problem	Family of probability measures	Data-generating process
$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$	$\{\mathbb{P}_\theta : \theta \in \Theta\}$	$\{\xi_t\}_{t \in \mathbb{N}}$

Examples:

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

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

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

- Order quantities $x \in X = \{1, \ldots, d\}$

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

- Order quantities $x \in X = \{1, \ldots, d\}$
- Demand $\xi \in \Xi = \{1, \ldots, d\}$

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

- Order quantities $x \in X = \{1, \ldots, d\}$
- Demand $\xi \in \Xi = \{1, \ldots, d\}$
- Expected cost $c(x, \theta) = \mathbb{E}_\theta \left[k\,x - p \min\{x, \xi\}\right]$

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

- Order quantities $x \in X = \{1, \ldots, d\}$
- Demand $\xi \in \Xi = \{1, \ldots, d\}$
- Expected cost $c(x, \theta) = \mathbb{E}_\theta \left[k\, x - p \min\{x, \xi\} \right]$

wholesale price

retail price

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

- Order quantities $x \in X = \{1, \ldots, d\}$
- Demand $\xi \in \Xi = \{1, \ldots, d\}$
- Expected cost $c(x, \theta) = \mathbb{E}_\theta \left[k\, x - p\, \underbrace{\min\{x, \xi\}}_{\text{sales}}\right]$

order quantity

sales

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Data-generating process

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

- Order quantities $x \in X = \{1, \ldots, d\}$
- Demand $\xi \in \Xi = \{1, \ldots, d\}$
- Expected cost $c(x, \theta) = \mathbb{E}_\theta \left[k\,x - p \min\{x, \xi\} \right]$

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Data-generating process

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

- Order quantities $x \in X = \{1, \ldots, d\}$
- Demand $\xi \in \Xi = \{1, \ldots, d\}$
- Expected cost $c(x, \theta) = \mathbb{E}_\theta \left[k\, x - p \min\{x, \xi\}\right]$

- Historical demands $\xi_t \in \Xi$

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Data-generating process

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

Family of probability measures

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

- Order quantities $x \in X = \{1, \ldots, d\}$
- Demand $\xi \in \Xi = \{1, \ldots, d\}$
- Expected cost $c(x, \theta) = \mathbb{E}_\theta \left[k\, x - p \min\{x, \xi\}\right]$
- Historical demands $\xi_t \in \Xi$

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

- Order quantities $x \in X = \{1, \ldots, d\}$
- Demand $\xi \in \Xi = \{1, \ldots, d\}$
- Expected cost $c(x, \theta) = \mathbb{E}_\theta\left[k\,x - p\min\{x, \xi\}\right]$

Data-generating process

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

- Historical demands $\xi_t \in \Xi$

Family of probability measures

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

- $\{\xi_t\}_{T \in \mathbb{N}}$ i.i.d. process under $\mathbb{P}_\theta$

Example: Newsvendor Problem

Stochastic optimization problem

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

- Order quantities $x \in X = \{1, \ldots, d\}$
- Demand $\xi \in \Xi = \{1, \ldots, d\}$
- Expected cost $c(x, \theta) = \mathbb{E}_\theta \left[k\, x - p \min\{x, \xi\}\right]$

Data-generating process

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

- Historical demands $\xi_t \in \Xi$

Family of probability 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

Surrogate Optimization Models

Original optimization problem:

$$\underset{x \in X}{\text{minimize}} \quad c(x, \theta)$$

Surrogate Optimization Models

Surrogate optimization problem:

$$\underset{x \in X}{\text{minimize}} \quad c(x, \widehat{\theta}_T)$$

$\xi_1, \xi_2, \ldots, \xi_T \longrightarrow$

$\longrightarrow \widehat{\theta}_T$

Surrogate Optimization Models

Surrogate optimization problem:

$$\underset{x \in X}{\text{minimize}} \quad \widehat{c}_T(x)$$

$\xi_1, \xi_2, \ldots, \xi_T \longrightarrow$

$\longrightarrow \widehat{c}_T$

Surrogate Optimization Models

Surrogate optimization problem:

$$\underset{x \in X}{\text{minimize}} \quad \widehat{c}_T(x)$$

Construction of $\widehat{c}_T$:

- Sample average approximation[1)]
- Regularized nominal model[2)]
- Predict-then-optimize approach[3)]
- Neural network model[4)]
- Distributionally robust optimization model[5)]
- etc.

[1)] Shapiro, *Annals of Statistics*, 1989; [2)] Hoerl & Kennard, *Technometrics*, 1970; [3)] Elmachtoub & Grigas, *Management Science*, 2021; [4)] Donti et al., *NIPS*, 2017; [5)] Delage & Ye, *Operations Research*, 2010; Mohajerin Esfahani & Kuhn, *Mathematical Programming*, 2018.

Terminology

Definitions:

- Data-driven predictor $\widehat{c}_T$
- Data-driven prescriptor $\widehat{x}_T \in \operatorname*{argmin}_{x \in X} \widehat{c}_T(x)$

Terminology

Definitions:

- Data-driven predictor $\widehat{c}_T$
- Data-driven prescriptor $\widehat{x}_T \in \underset{x \in X}{\operatorname{argmin}}\ \widehat{c}_T(x)$

determines the surrogate optimization model

Terminology

Definitions:

- Data-driven predictor $\widehat{c}_T$
- Data-driven prescriptor $\widehat{x}_T \in \underset{x \in X}{\operatorname{argmin}}\ \widehat{c}_T(x)$

any function that maps
$\xi_1, \xi_2, \ldots, \xi_T$ to X

Terminology

Definitions:

- Data-driven predictor $\widehat{c}_T$
- Data-driven prescriptor $\widehat{x}_T \in \operatorname*{argmin}_{x \in X} \widehat{c}_T(x)$

Performance measures:

Terminology

Definitions:

- Data-driven predictor $\widehat{c}_T$
- Data-driven prescriptor $\widehat{x}_T \in \underset{x \in X}{\operatorname{argmin}} \ \widehat{c}_T(x)$

Performance measures:

In-sample risk $\widehat{c}_T(\widehat{x}_T)$

Terminology

Definitions:

- Data-driven predictor $\widehat{c}_T$
- Data-driven prescriptor $\widehat{x}_T \in \operatorname*{argmin}_{x \in X} \widehat{c}_T(x)$

Performance measures:

In-sample risk $\widehat{c}_T(\widehat{x}_T)$

Out-of-sample risk $c(\widehat{x}_T, \theta)$

Terminology

Definitions:

- Data-driven predictor $\widehat{c}_T$
- Data-driven prescriptor $\widehat{x}_T \in \underset{x \in X}{\operatorname{argmin}}\ \widehat{c}_T(x)$

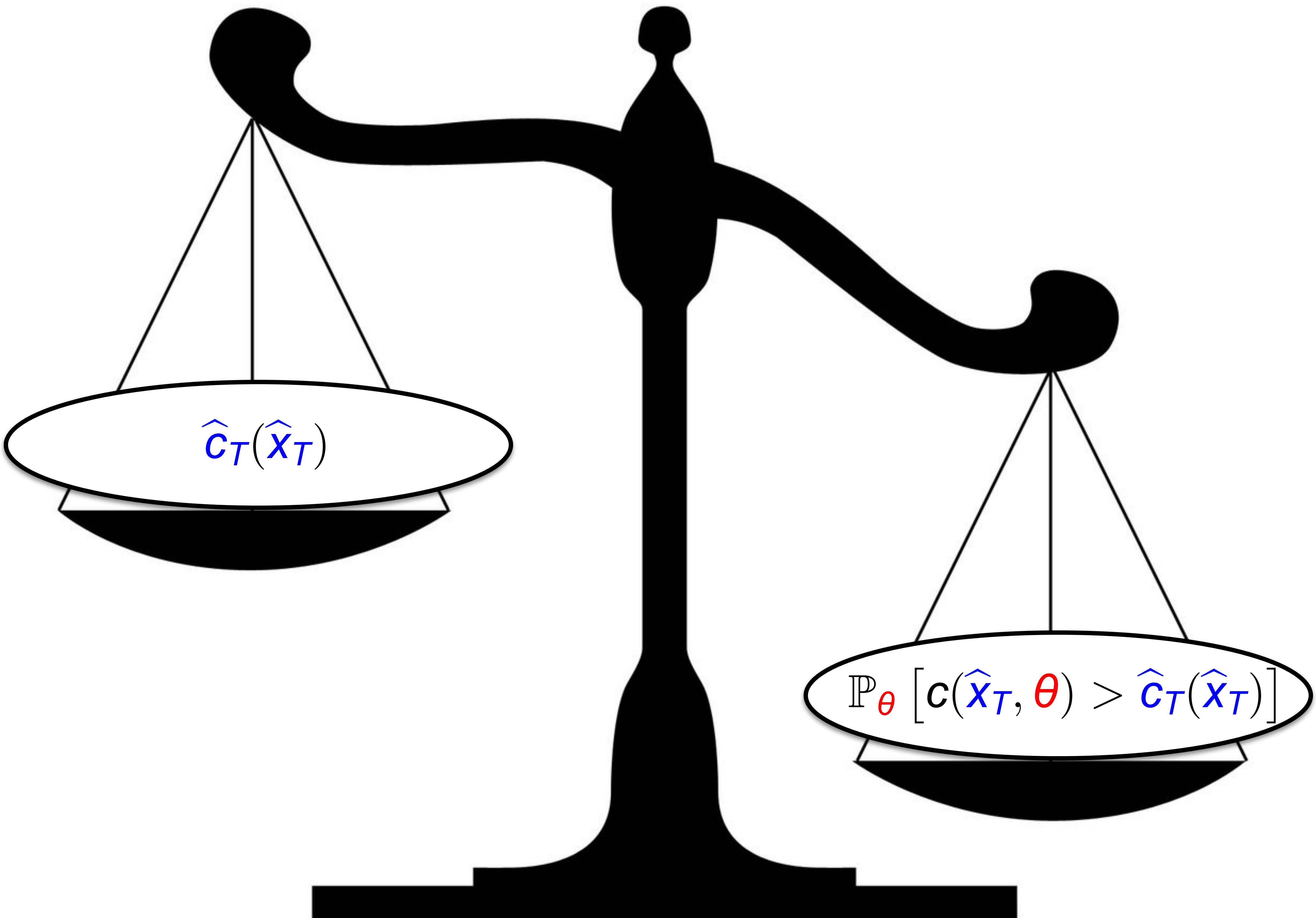
Performance measures:

In-sample risk $\widehat{c}_T(\widehat{x}_T)$

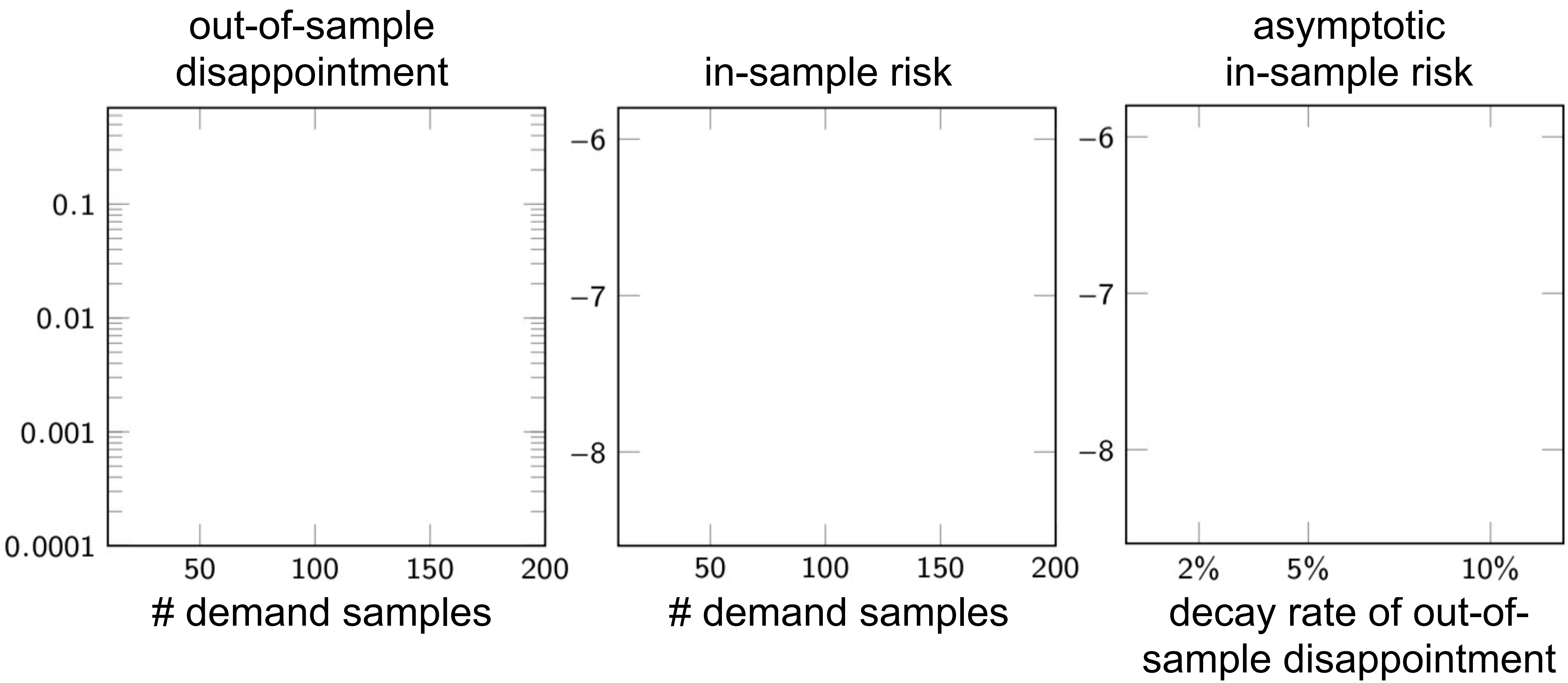
Out-of-sample risk $c(\widehat{x}_T, \theta)$

Out-of-sample disappointment $\mathbb{P}_\theta\left[c(\widehat{x}_T, \theta) > \widehat{c}_T(\widehat{x}_T)\right]$

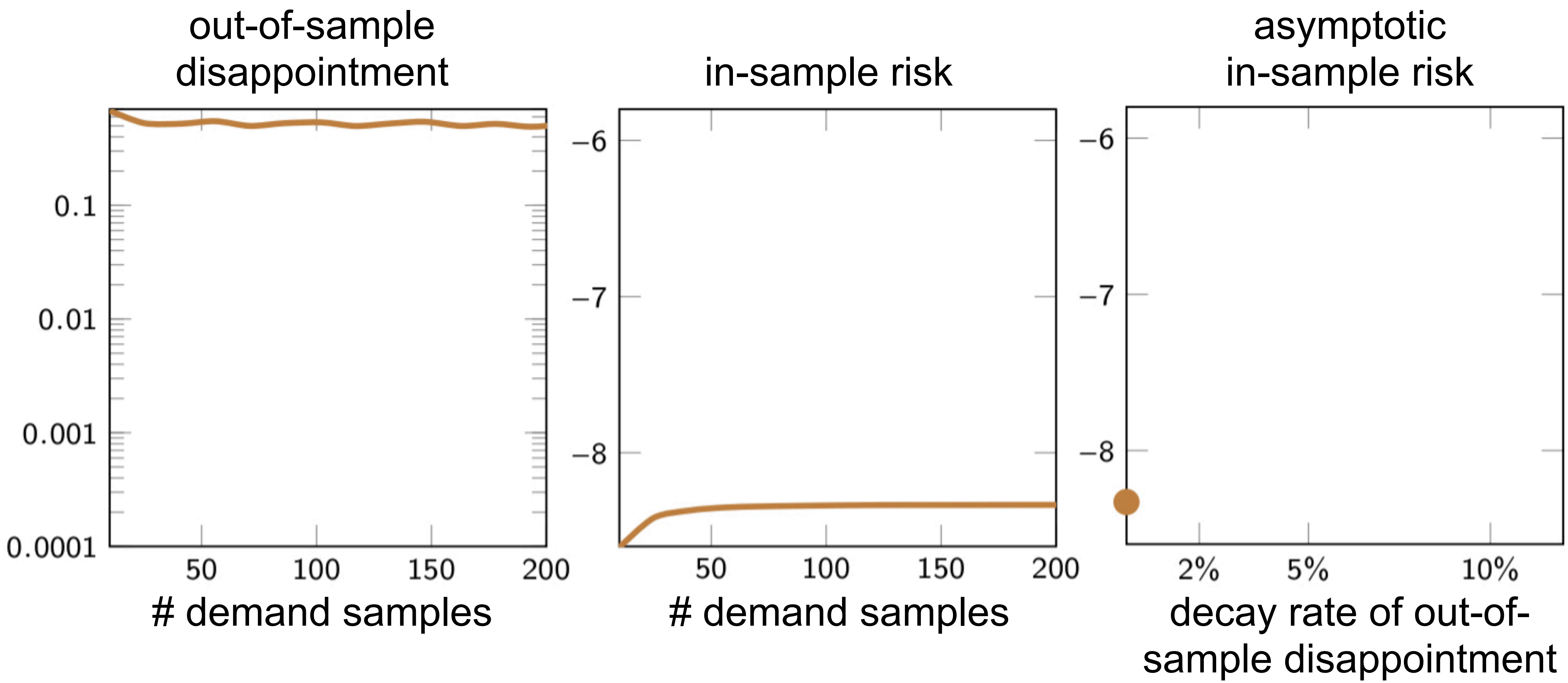
A Basic Trade-Off



Data-Driven Newsvendor Problem



Data-Driven Newsvendor Problem

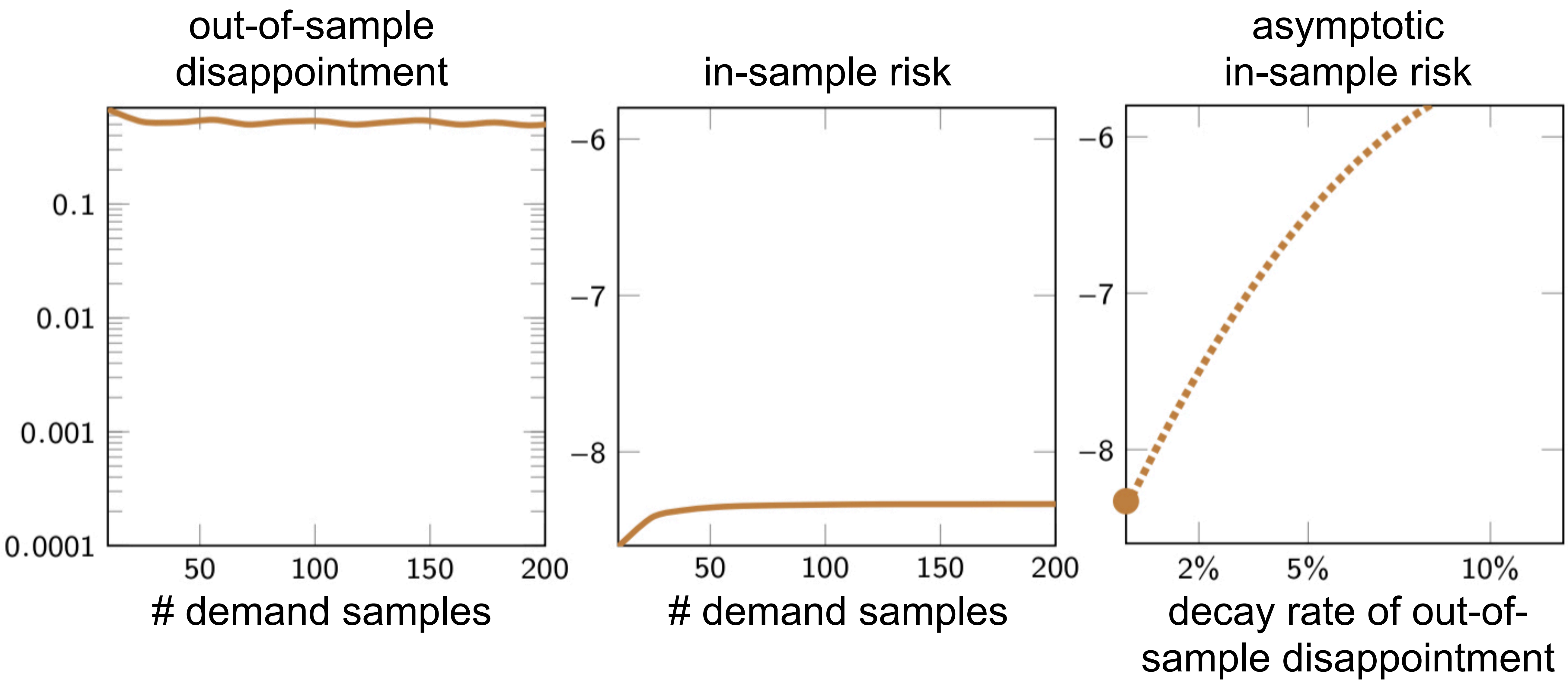


Model 1: SAA model[1)]

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

1) Shapiro, *Annals of Statistics*, 1989.

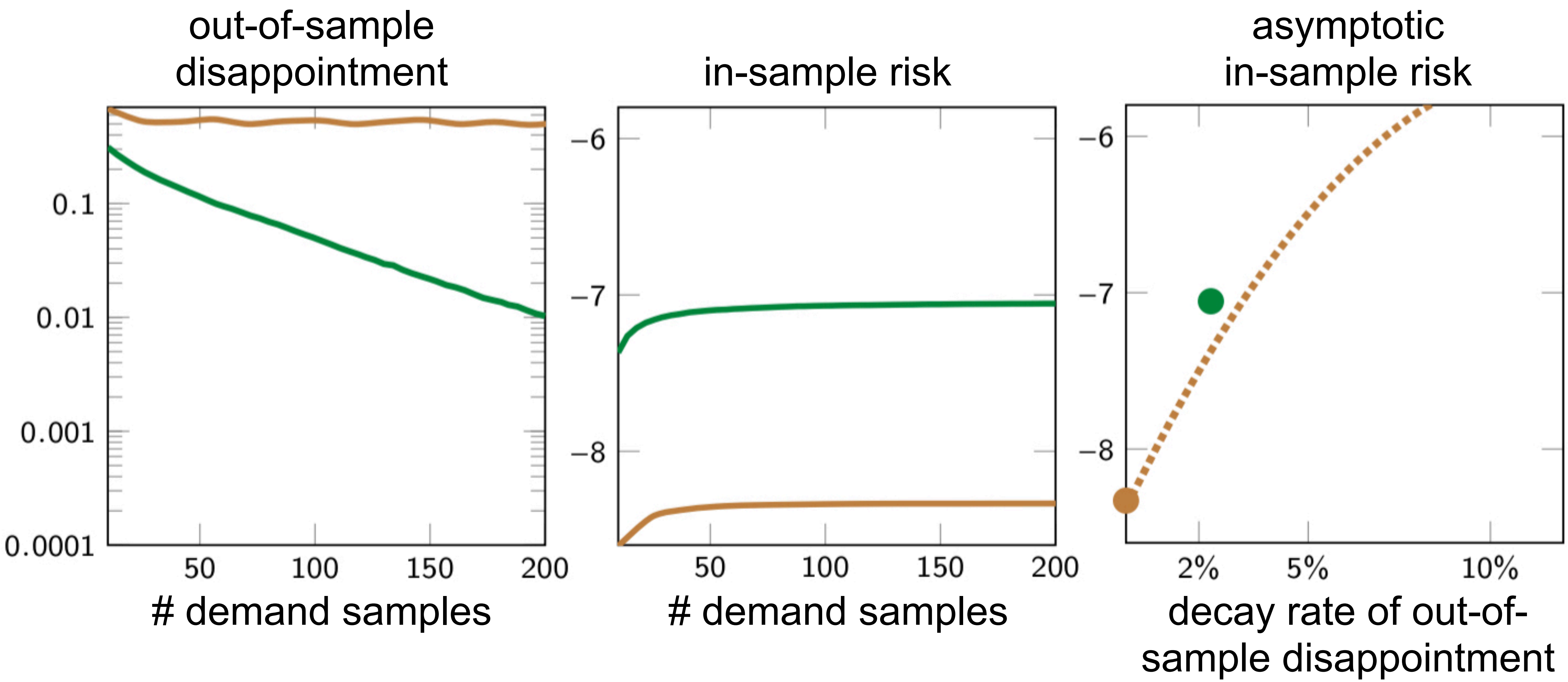
Data-Driven Newsvendor Problem



Model 2: SAA model with offset

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

Data-Driven Newsvendor Problem

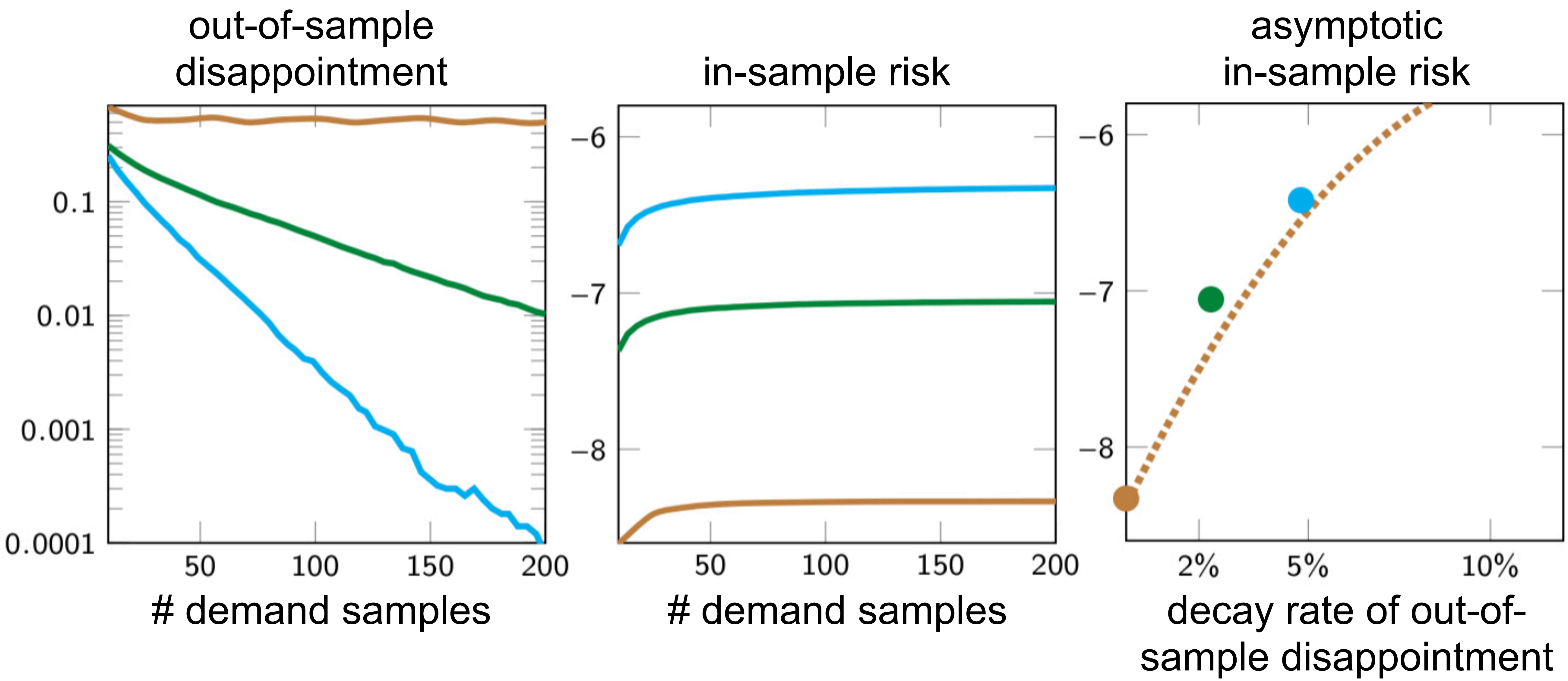


Model 3: DRO model with moment ambiguity set[1)]

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

[1)] Delage & Ye, *Operations Research*, 2010.

Data-Driven Newsvendor Problem



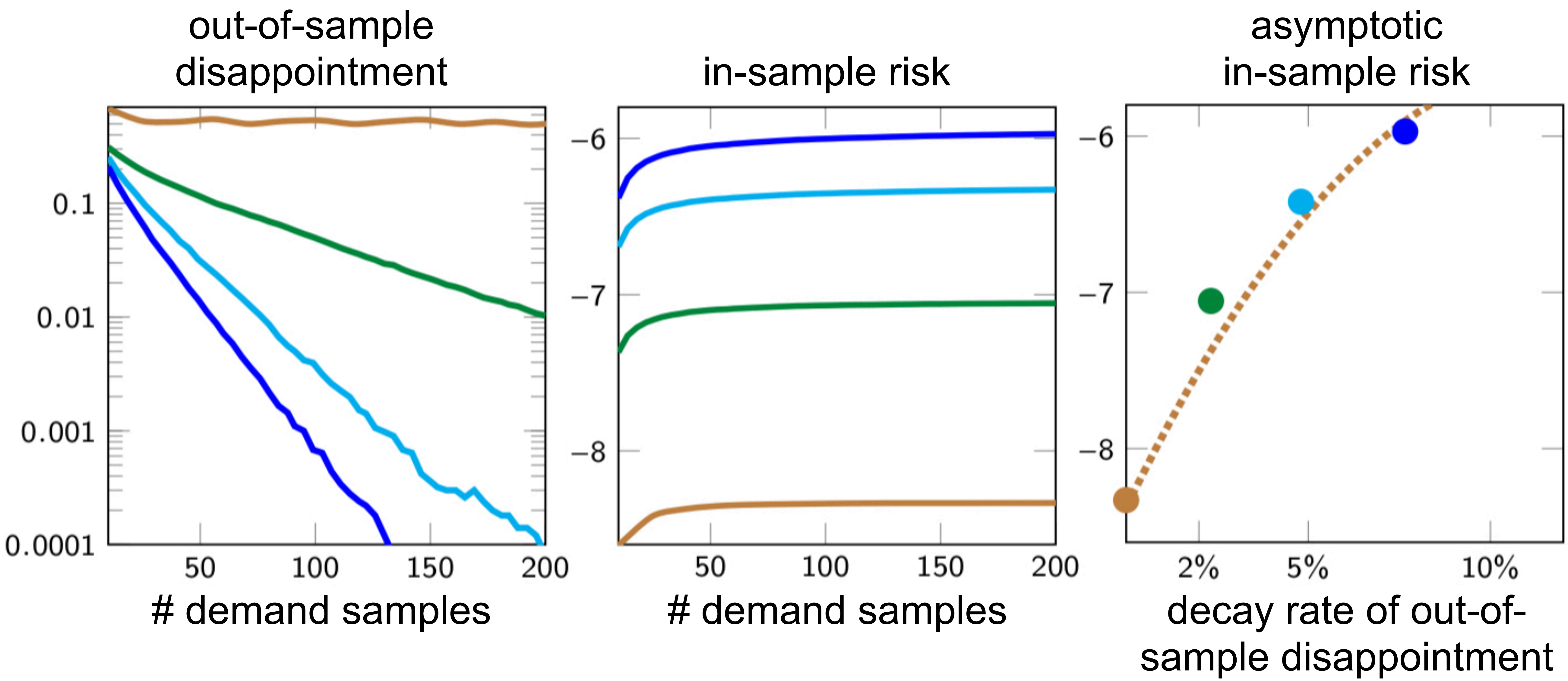
Model 3: DRO model with moment ambiguity set[1)]

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

---

1) Delage & Ye, *Operations Research*, 2010.

Data-Driven Newsvendor Problem

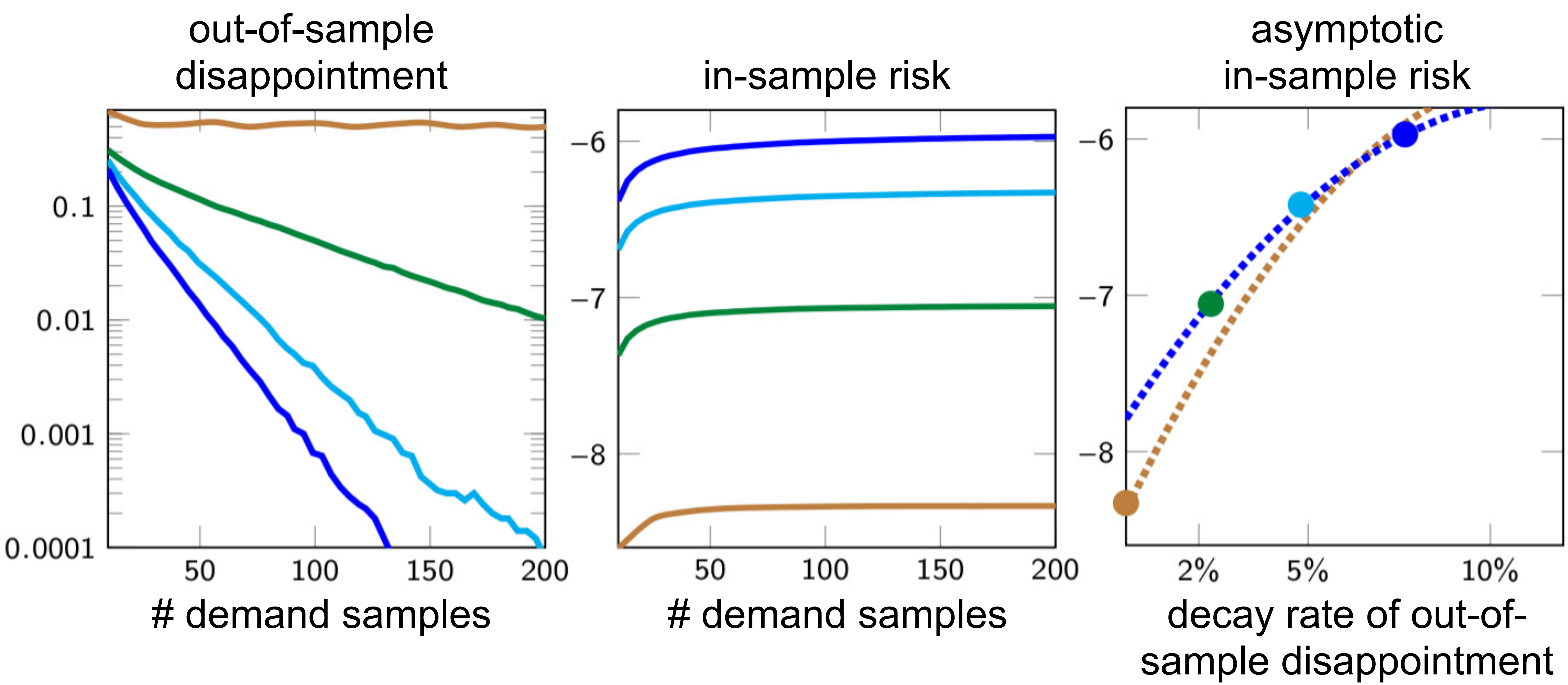


Model 3: DRO model with moment ambiguity set[1)]

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

[1)] Delage & Ye, *Operations Research*, 2010.

Data-Driven Newsvendor Problem

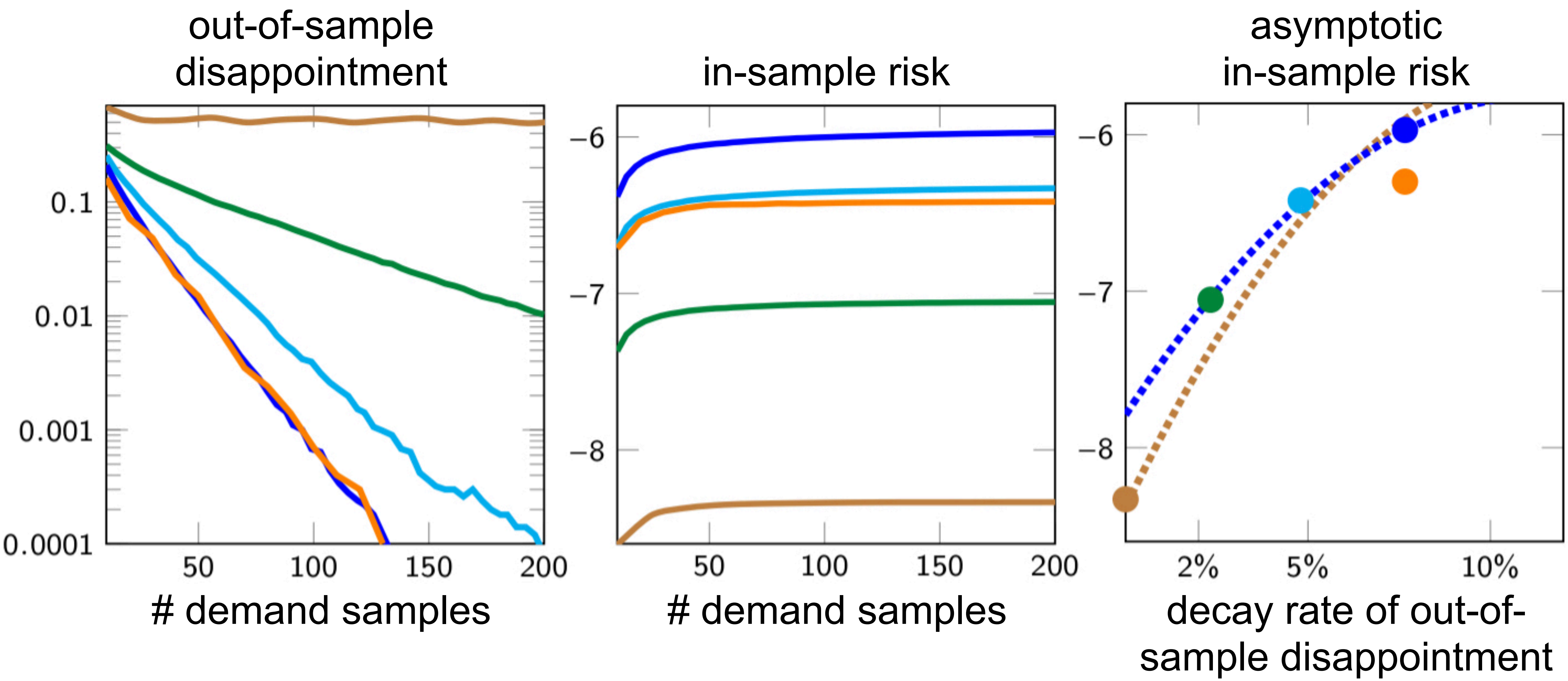


Model 3: DRO model with moment ambiguity set[1)]

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

[1)] Delage & Ye, *Operations Research*, 2010.

Data-Driven Newsvendor Problem

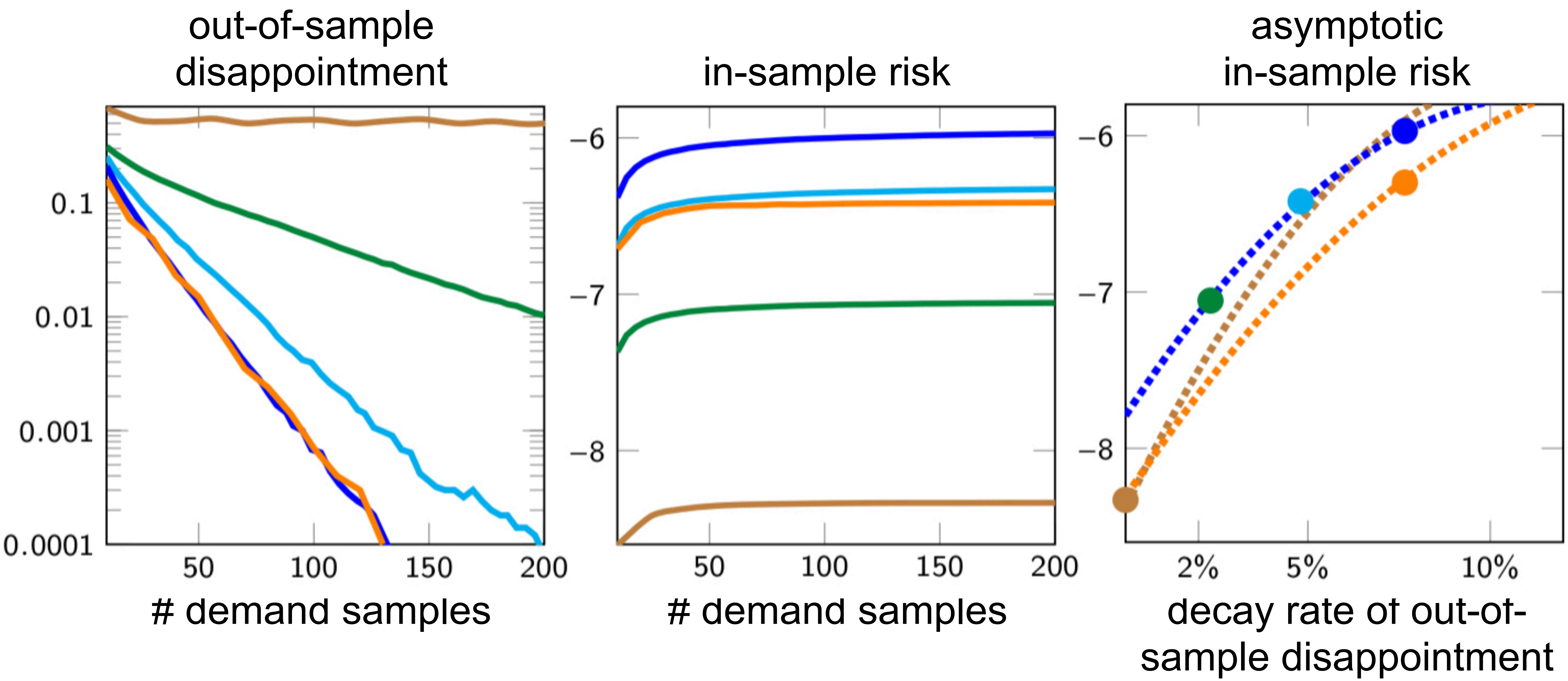


Model 4: DRO model with Wasserstein ambiguity set[1)]

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

1) Mohajerin Esfahani & Kuhn, *Mathematical Programming*, 2018.

Data-Driven Newsvendor Problem

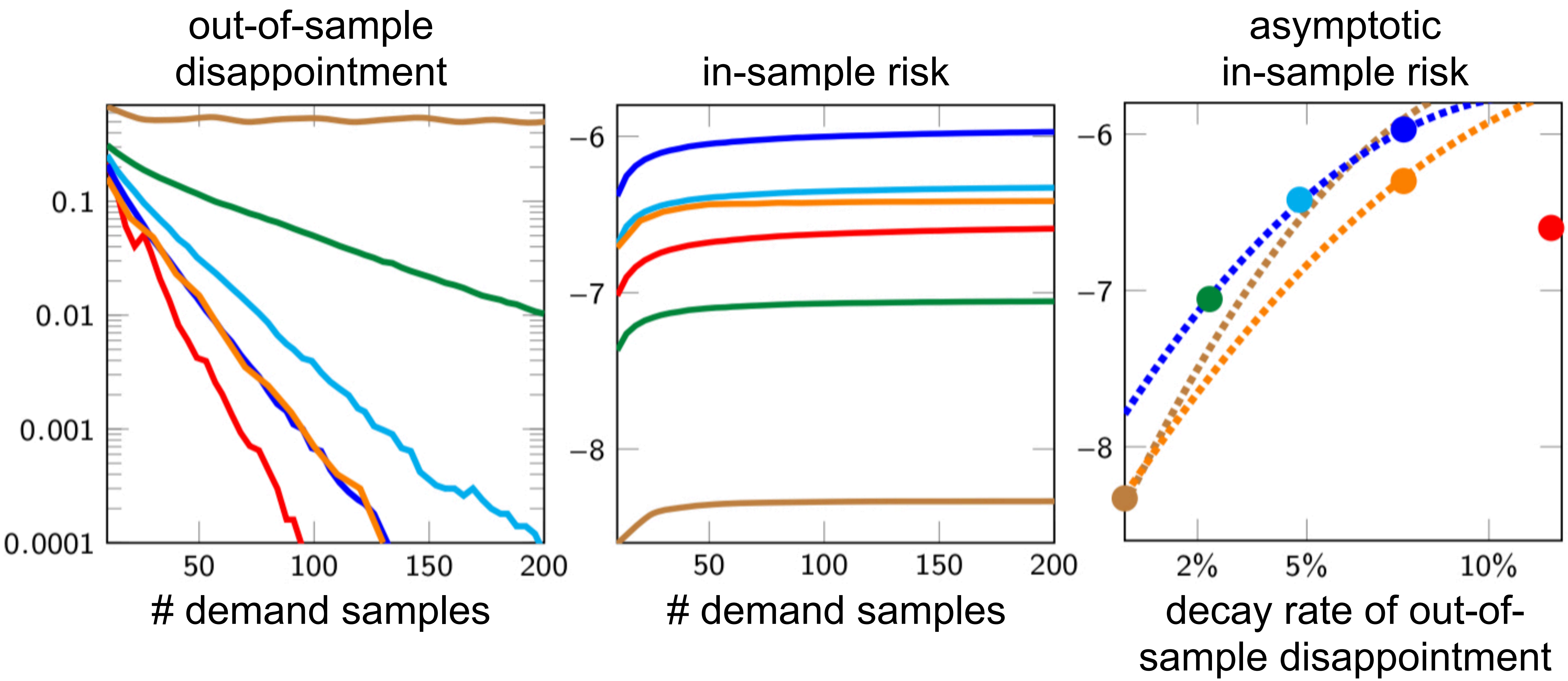


Model 4: DRO model with Wasserstein ambiguity set[1)]

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

[1)] Mohajerin Esfahani & Kuhn, *Mathematical Programming*, 2018.

Data-Driven Newsvendor Problem



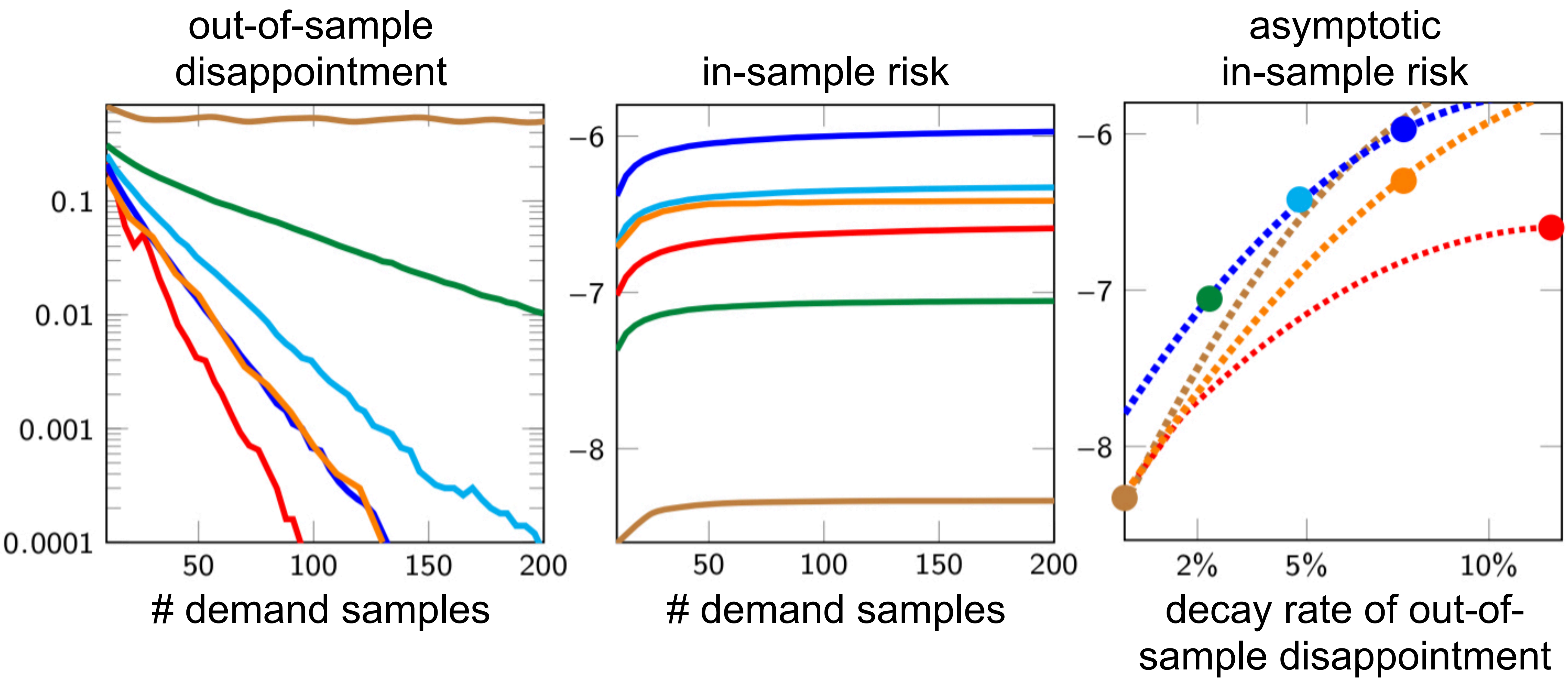
Model 5: DRO model with KL ambiguity set[1)]

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

---

1) Ben-Tal et al., *Management Science*, 2013.

Data-Driven Newsvendor Problem



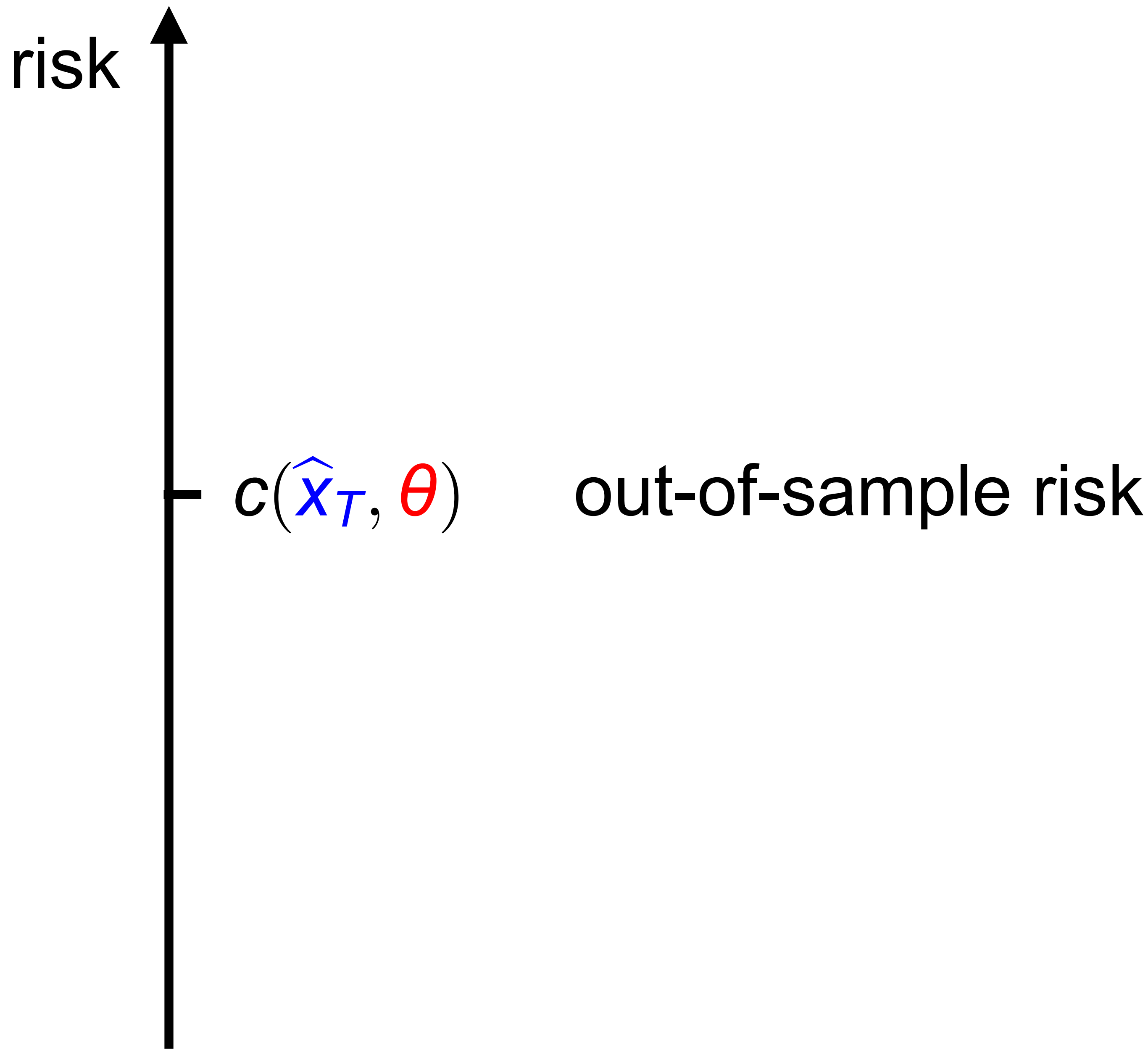
Model 5: DRO model with KL ambiguity set[1)]

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

1) Ben-Tal et al., *Management Science*, 2013.

Constructing “Optimal” Surrogate Optimization Models

Optimal Data-Driven Optimization



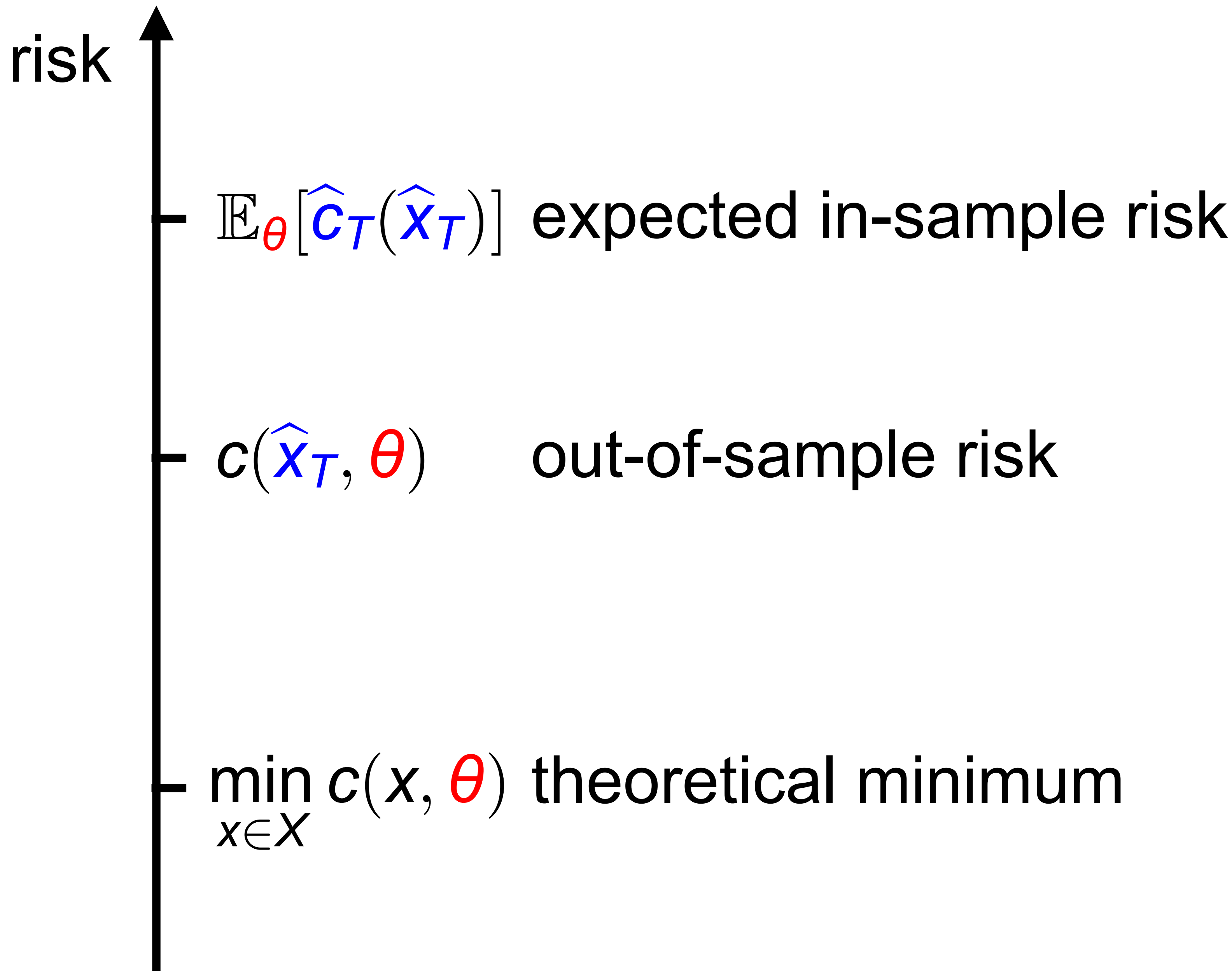
Optimal Data-Driven Optimization

risk

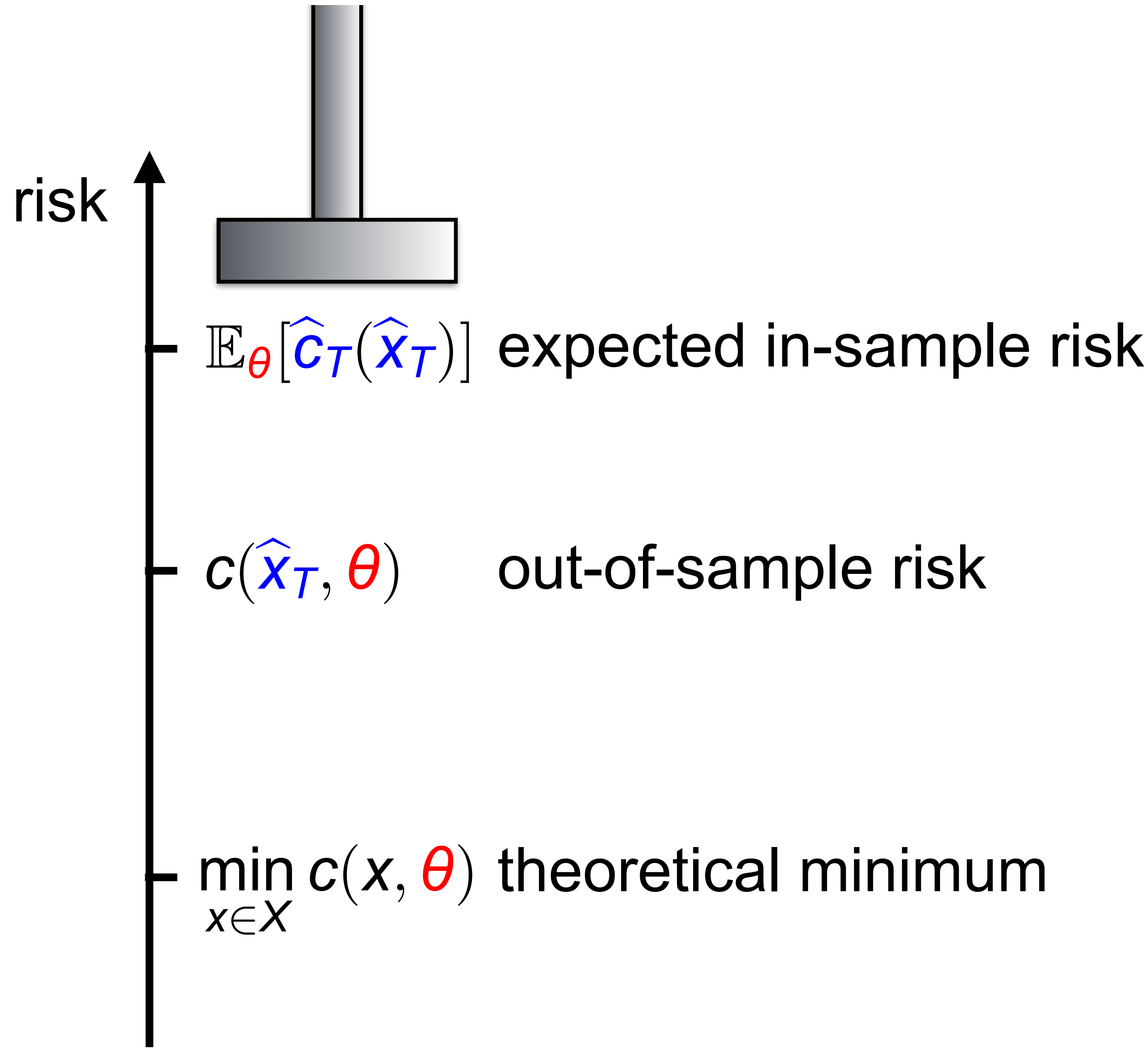
$c(\widehat{x}_T, \theta)$ out-of-sample risk

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

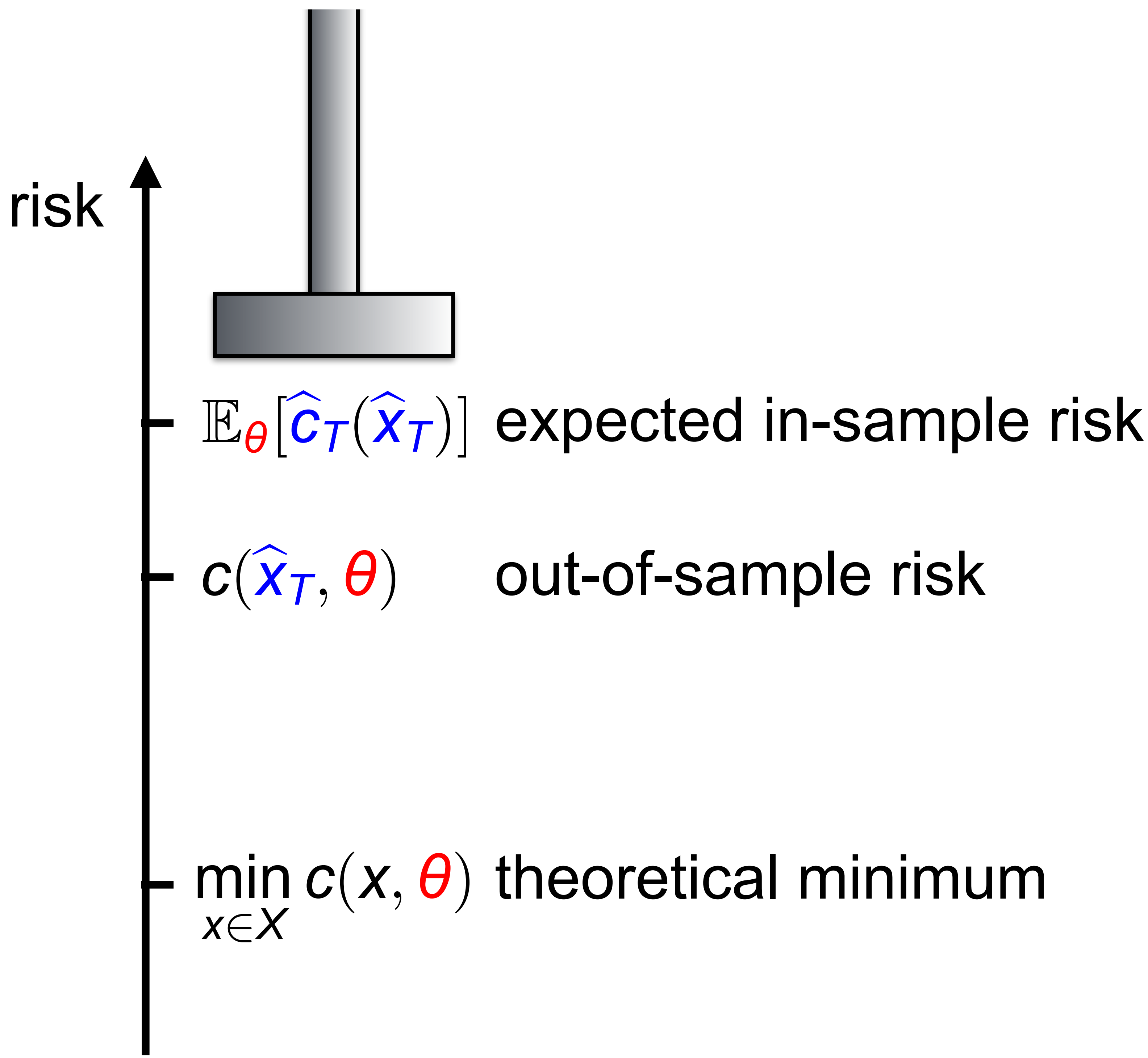
Optimal Data-Driven Optimization



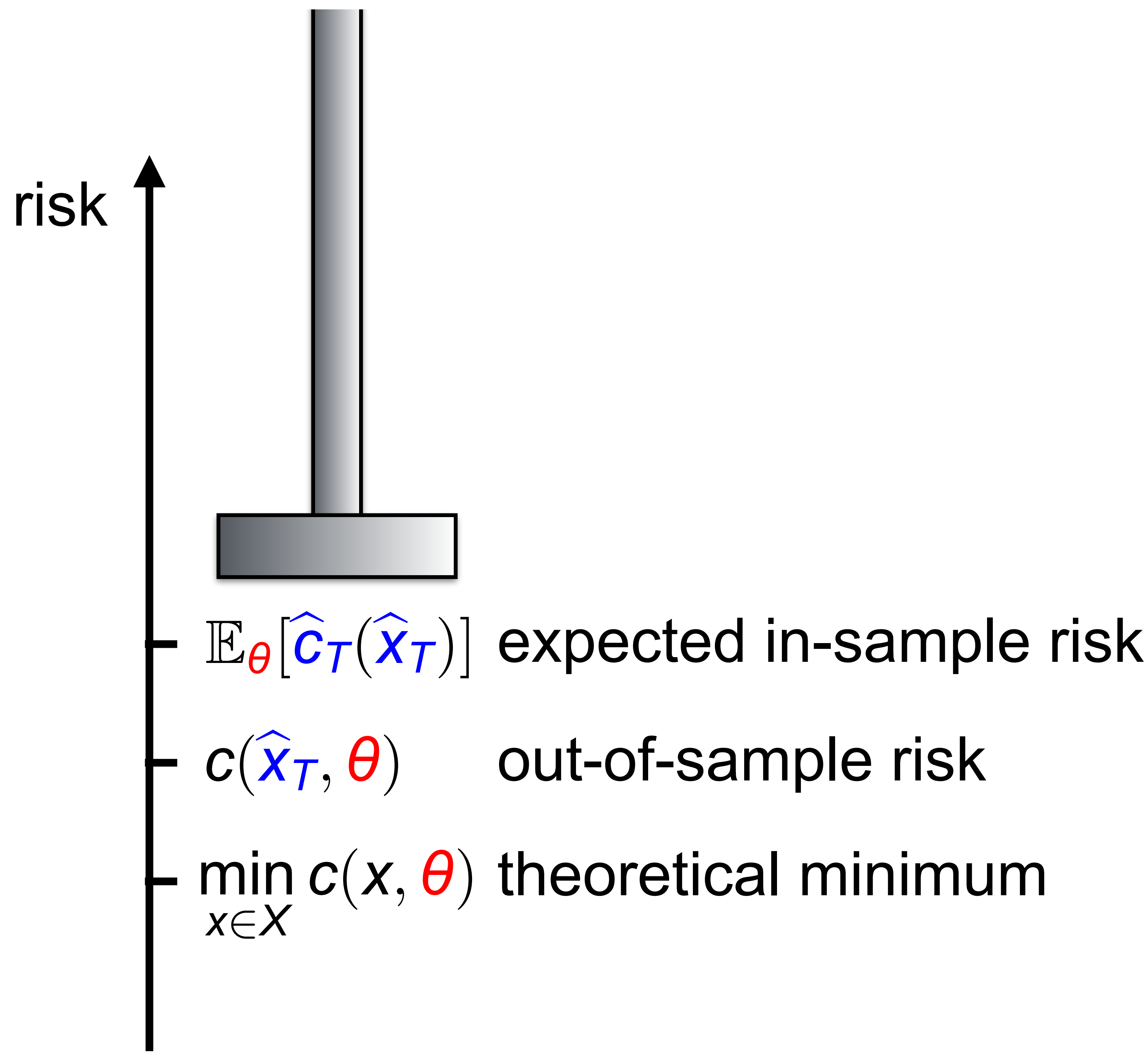
Optimal Data-Driven Optimization



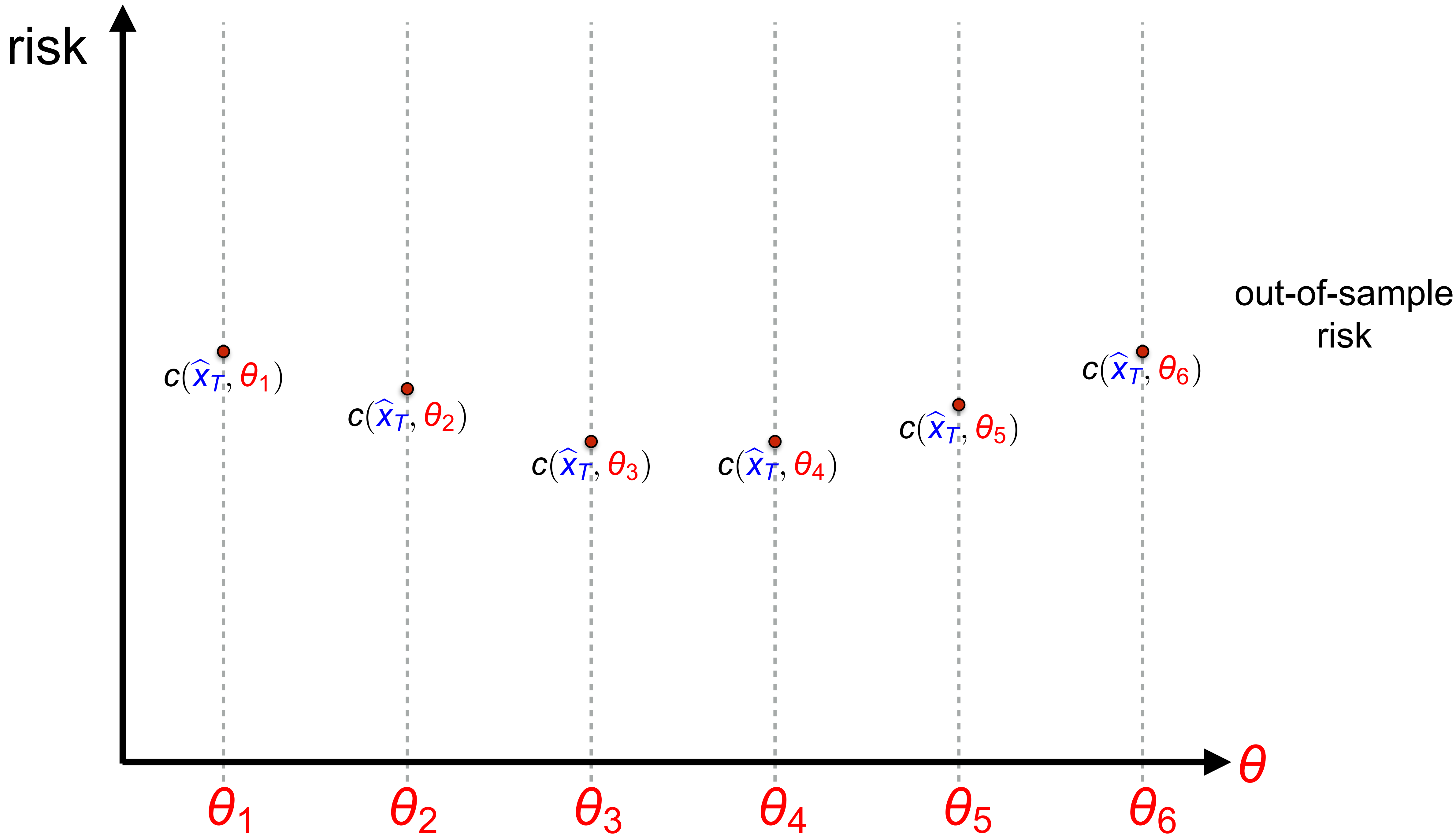
Optimal Data-Driven Optimization



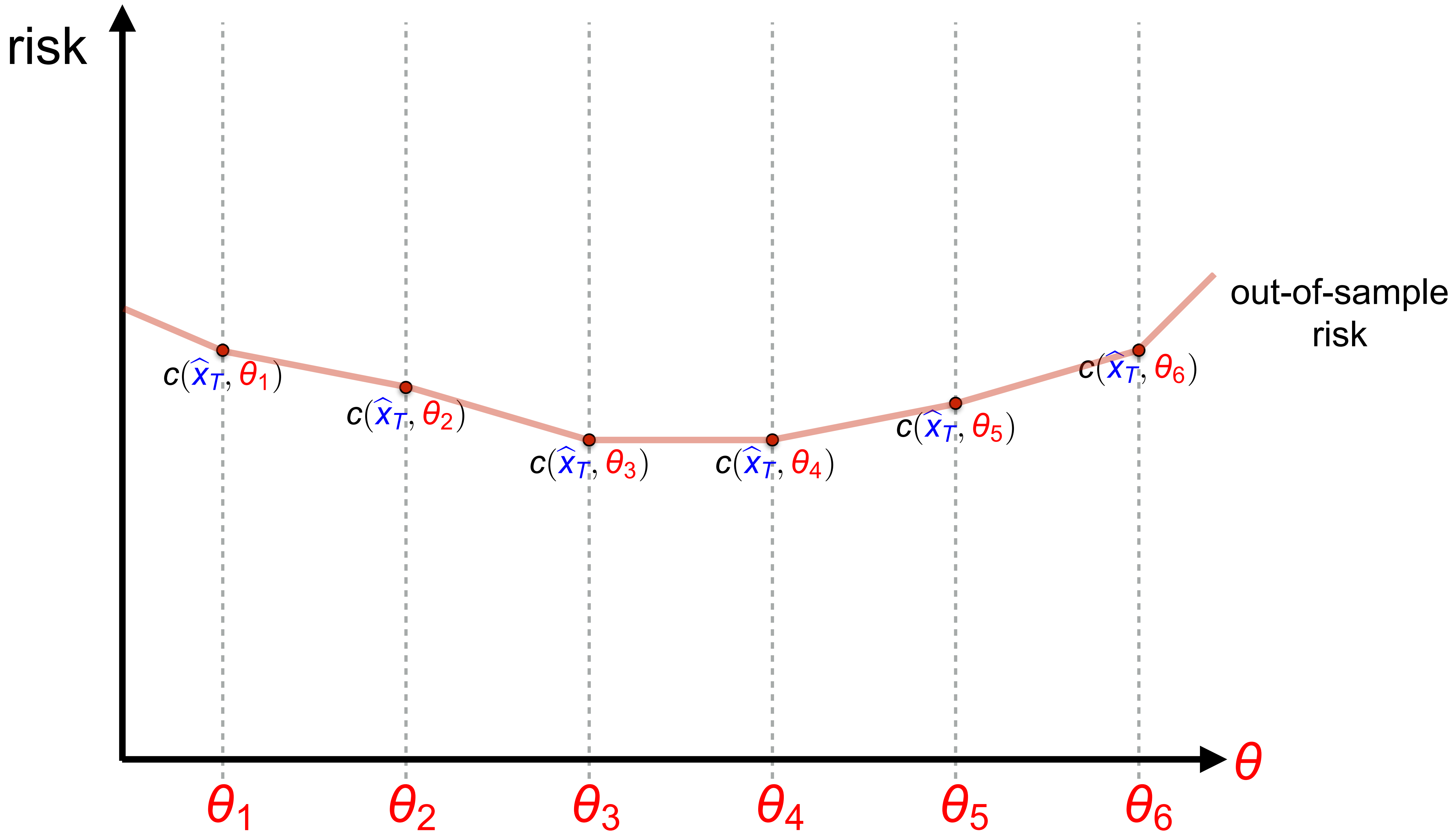
Optimal Data-Driven Optimization



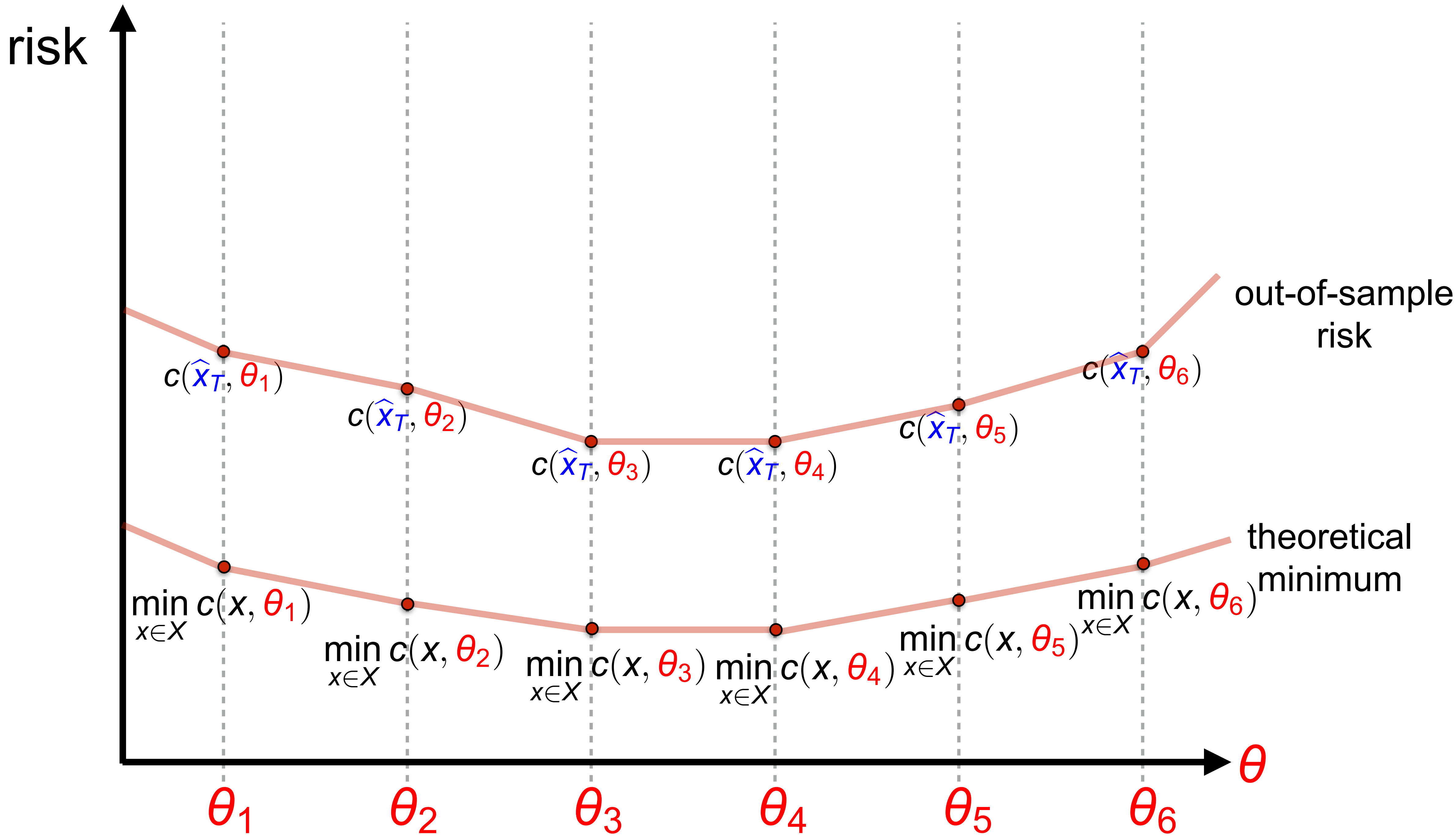
Optimal Data-Driven Optimization



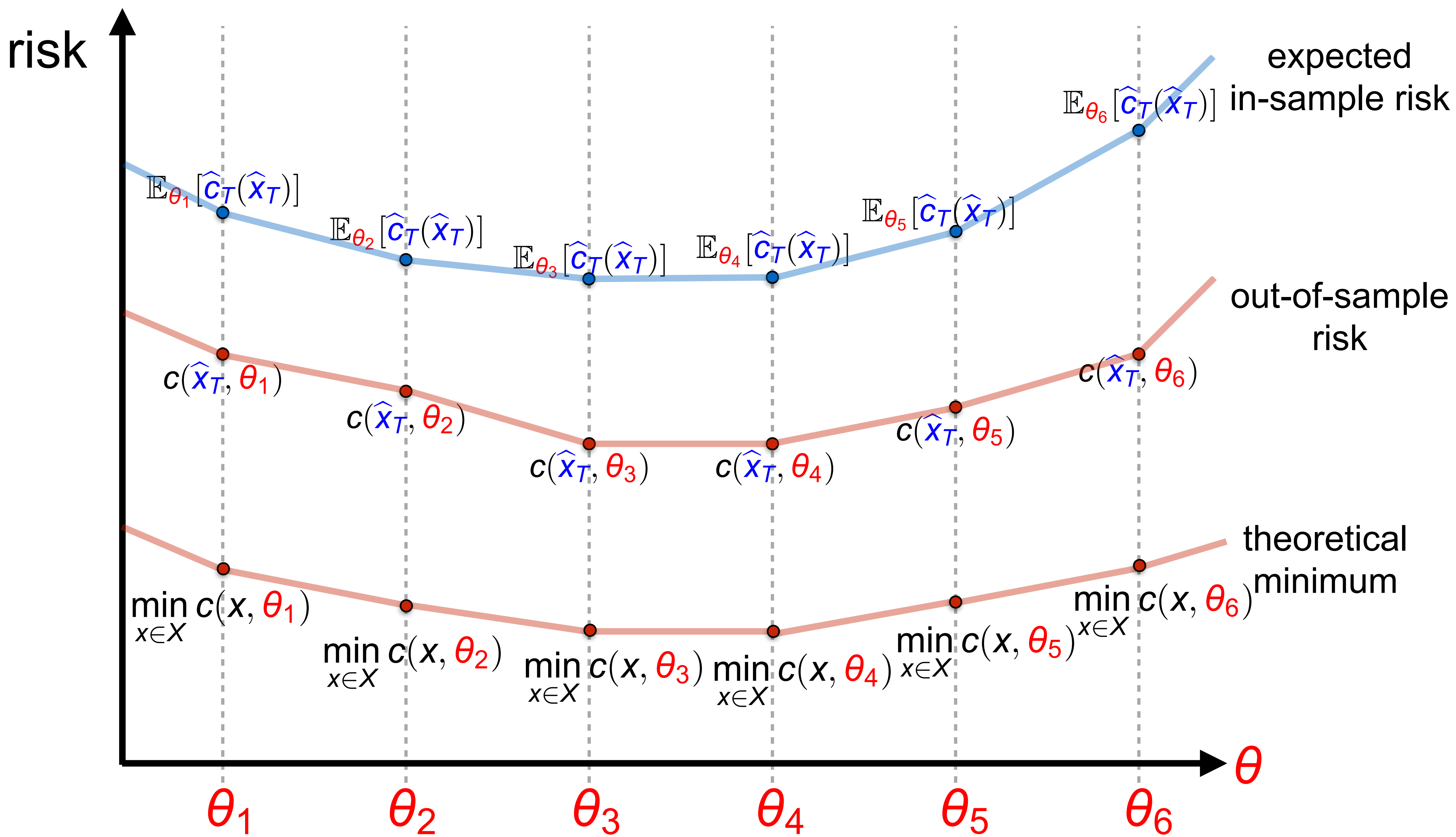
Optimal Data-Driven Optimization



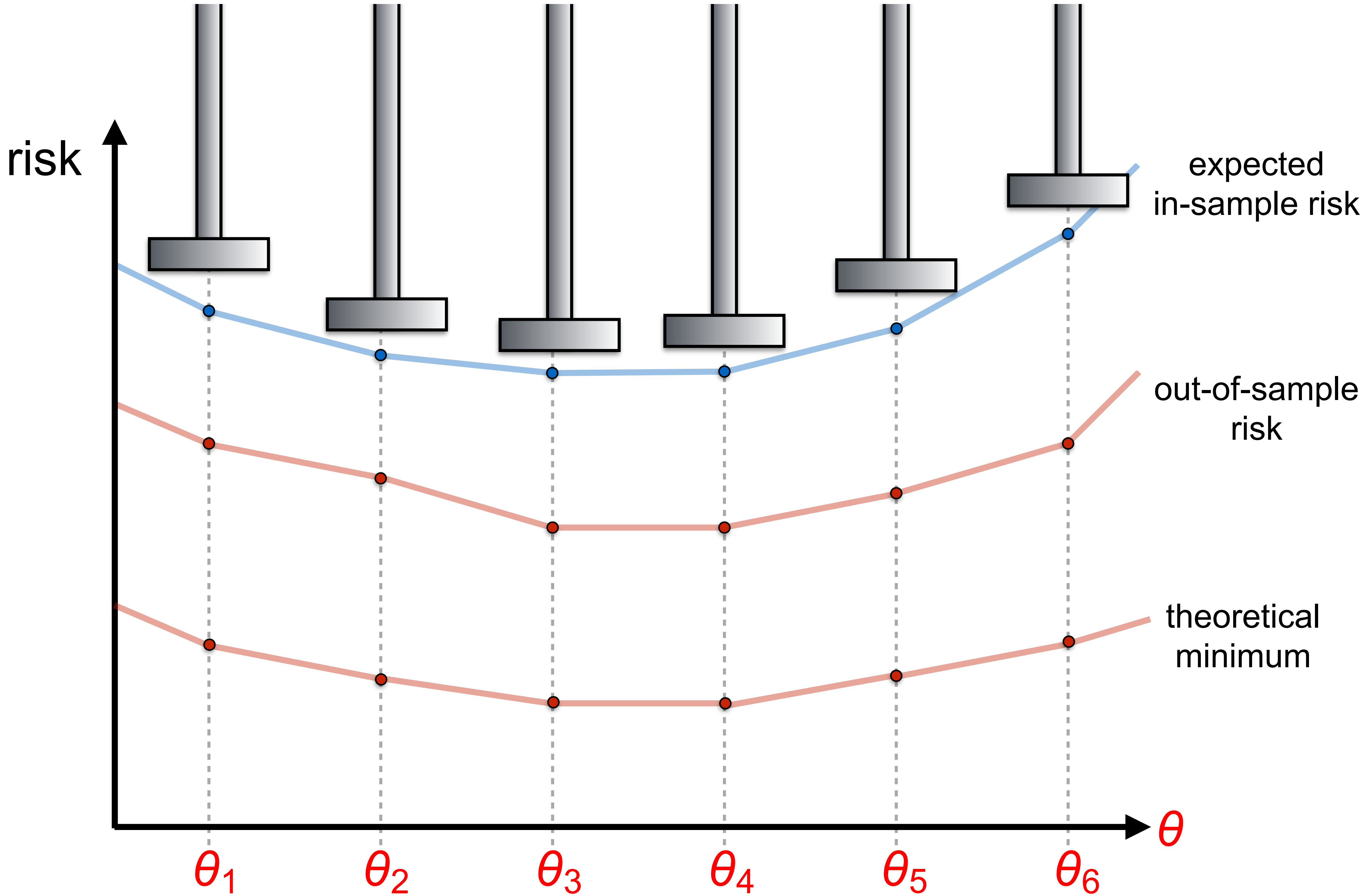
Optimal Data-Driven Optimization



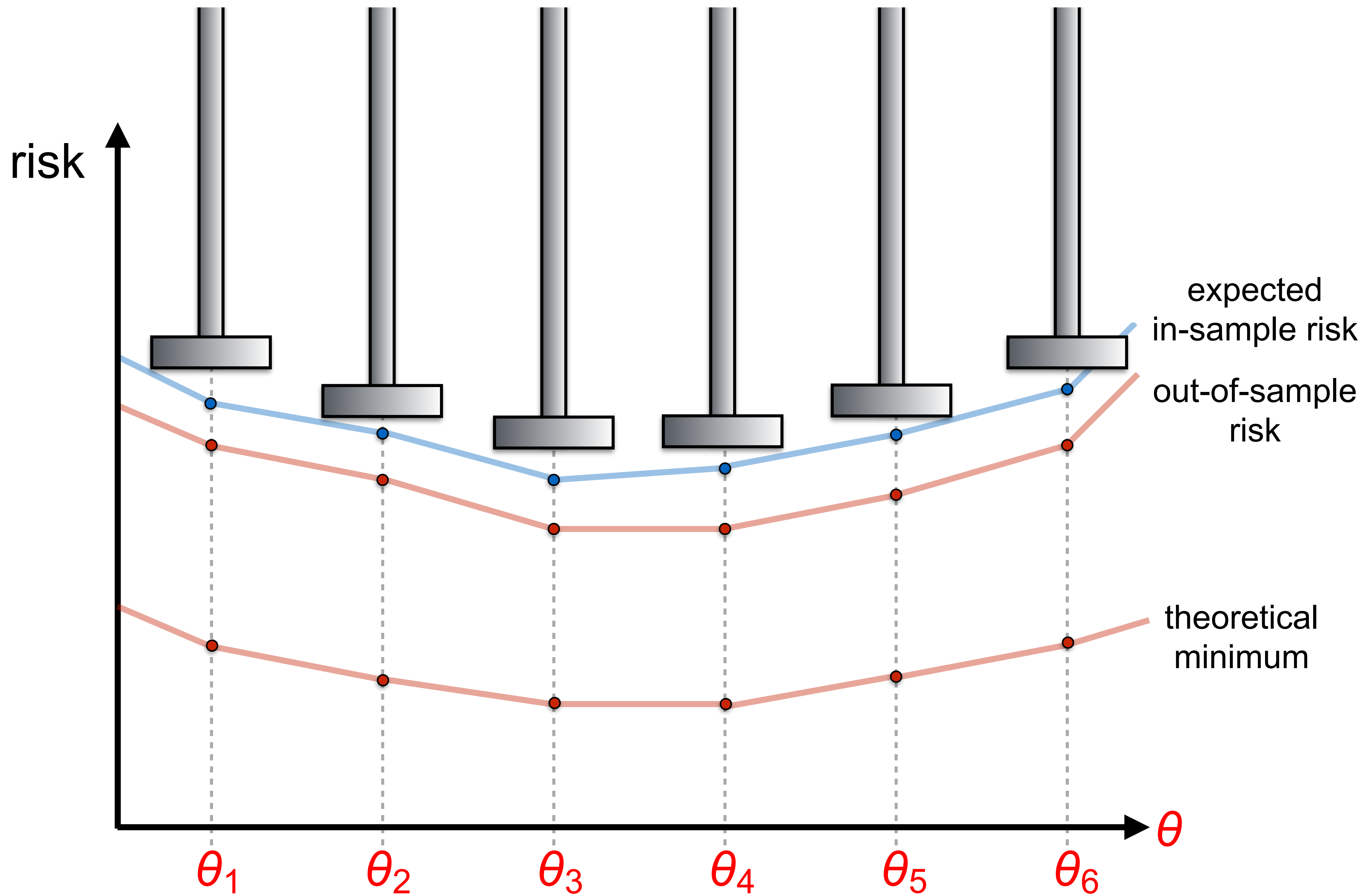
Optimal Data-Driven Optimization



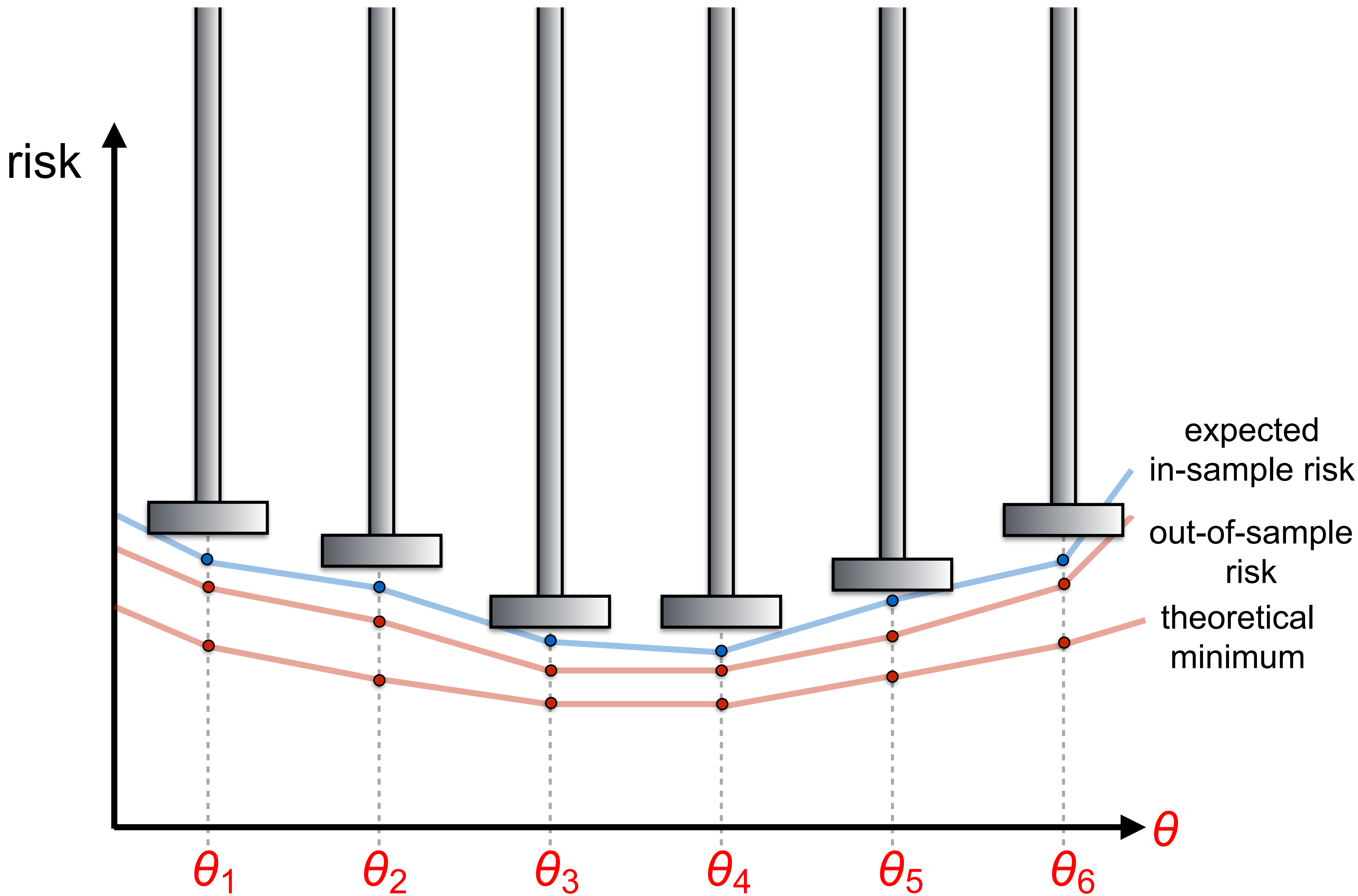
Optimal Data-Driven Optimization



Optimal Data-Driven Optimization



Optimal Data-Driven Optimization



Optimal Data-Driven Optimization

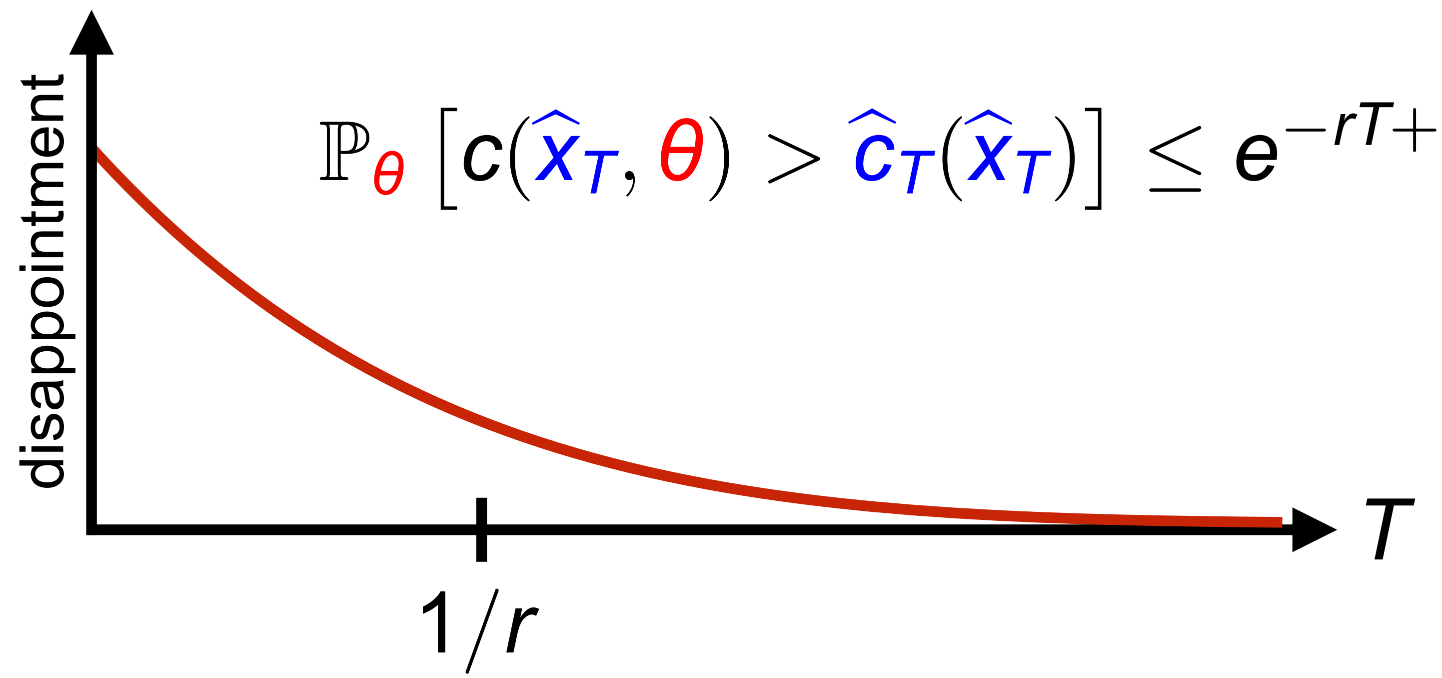
Minimize the in-sample risk and require that the out-of-sample disappointment decays exponentially at rate *r*

$$\begin{aligned}
\underset{\widehat{c}_T,\widehat{x}_T}{\text{minimize}} \quad & \left\{ \lim_{T\to\infty} \mathbb{E}_{\theta}\left[\widehat{c}_T(\widehat{x}_T)\right] \right\}_{\theta\in\Theta} \\
\text{subject to} \quad & \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\theta}\left[c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)\right] \leq -r \quad \forall \theta \in \Theta
\end{aligned}$$

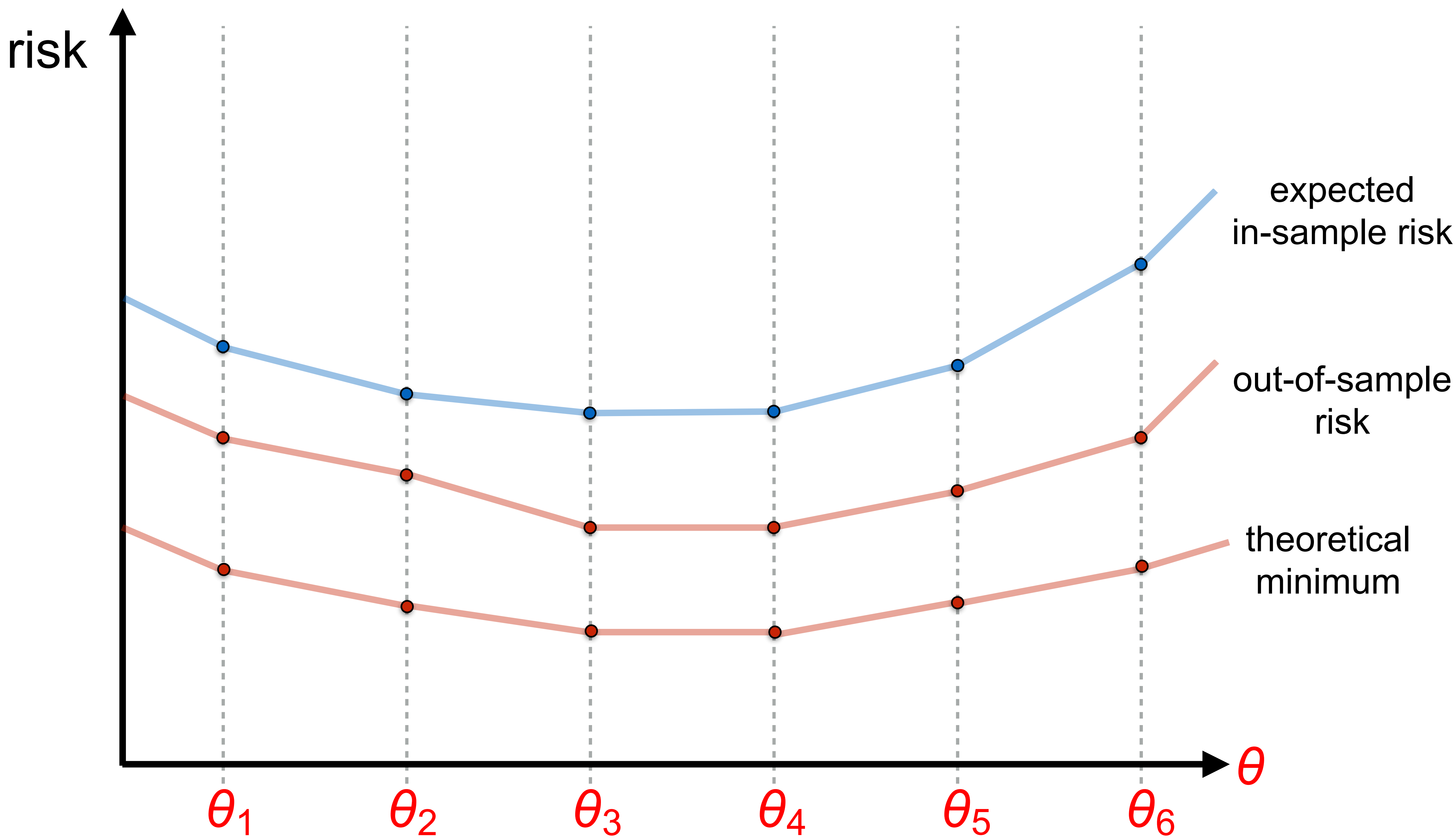
Optimal Data-Driven Optimization

Minimize the in-sample risk and require that the out-of-sample disappointment decays exponentially at rate r

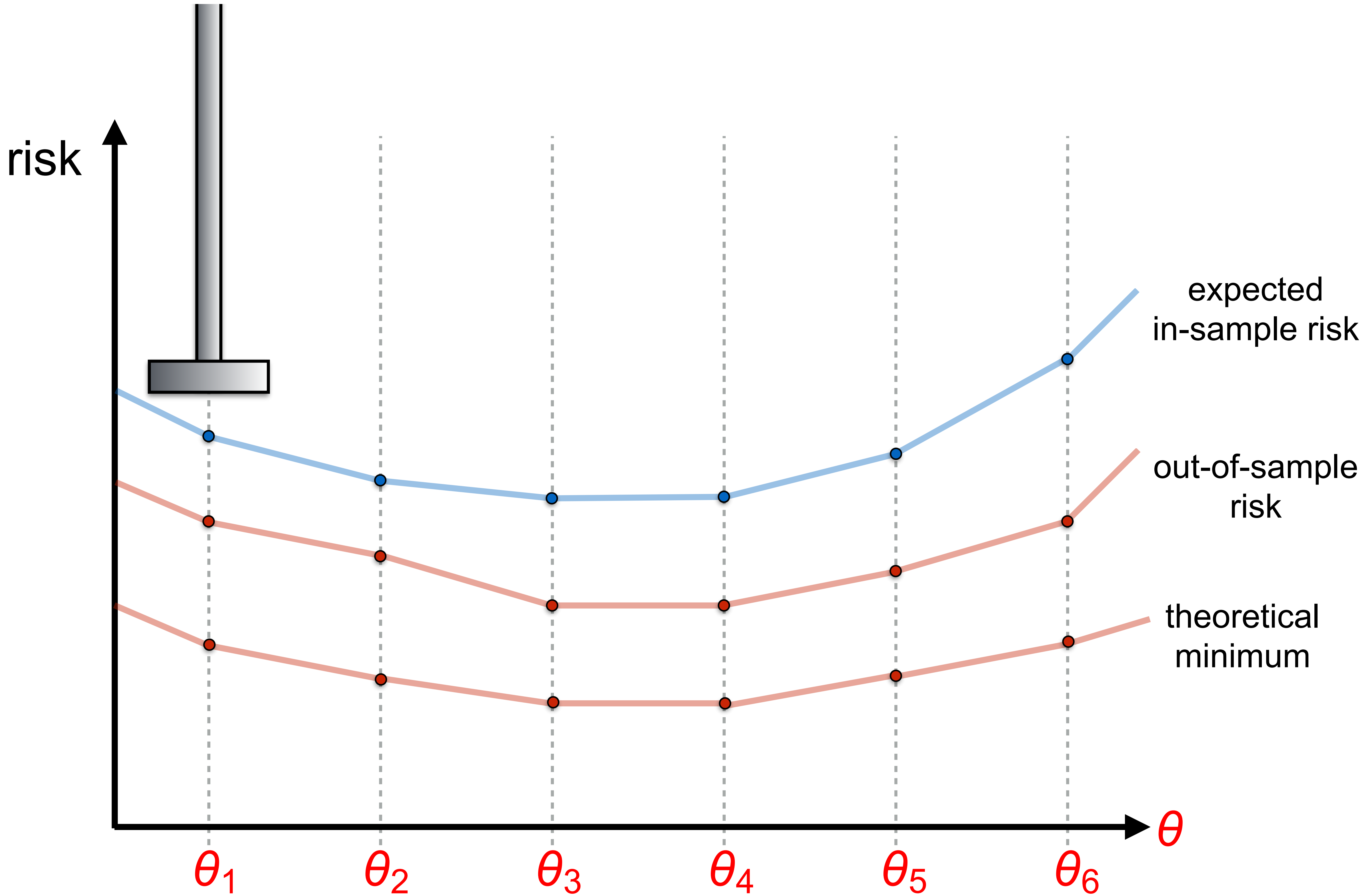
$$\begin{aligned} &\underset{\widehat{c}_T,\widehat{x}_T}{\text{minimize}} \quad \left\{ \lim_{T\to\infty} \mathbb{E}_{\theta}\left[\widehat{c}_T(\widehat{x}_T)\right] \right\}_{\theta\in\Theta} \\ &\text{subject to} \quad \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\theta}\left[c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)\right] \le -r \quad \forall \theta \in \Theta \end{aligned}$$



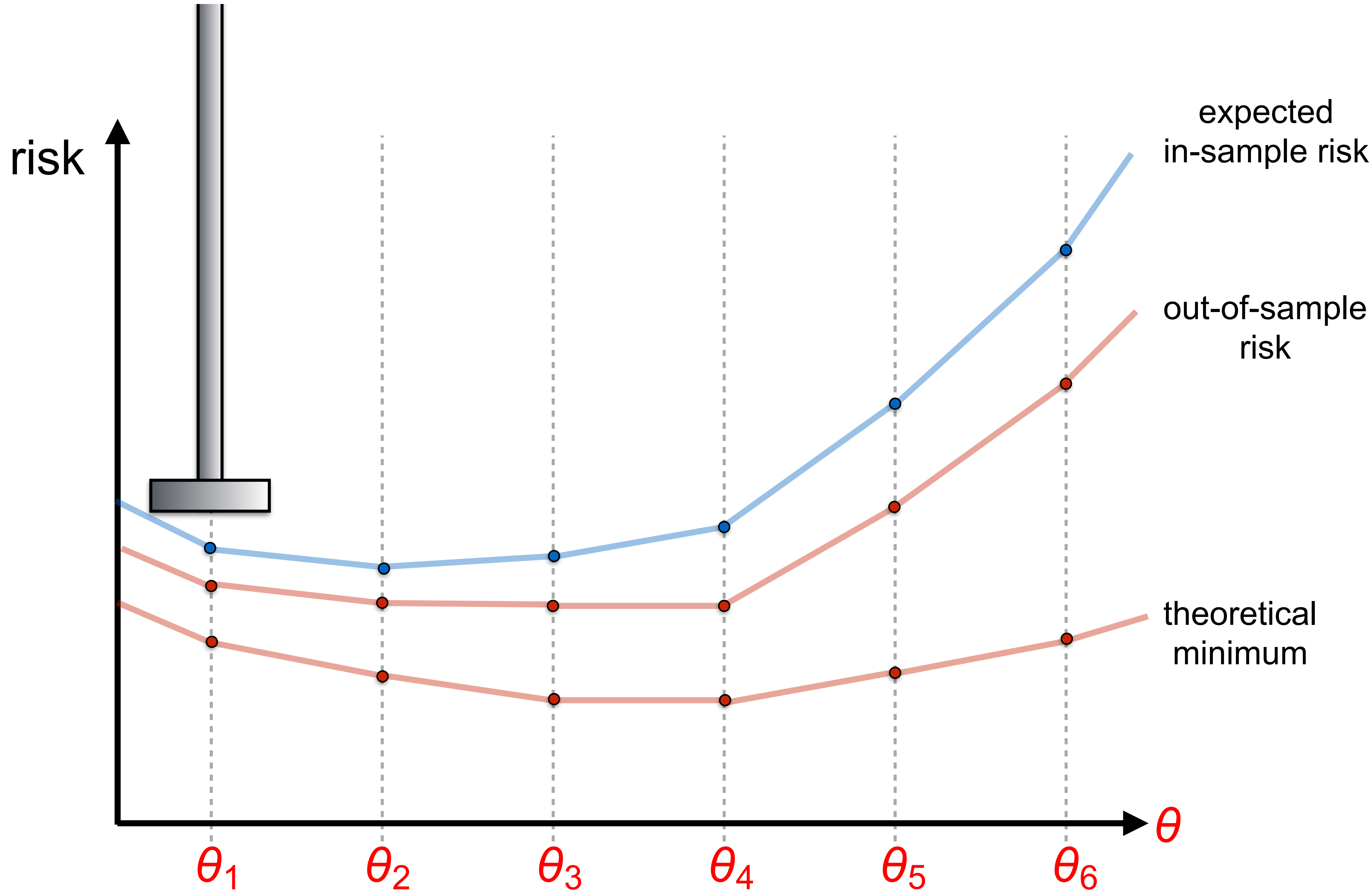
Optimal Data-Driven Optimization



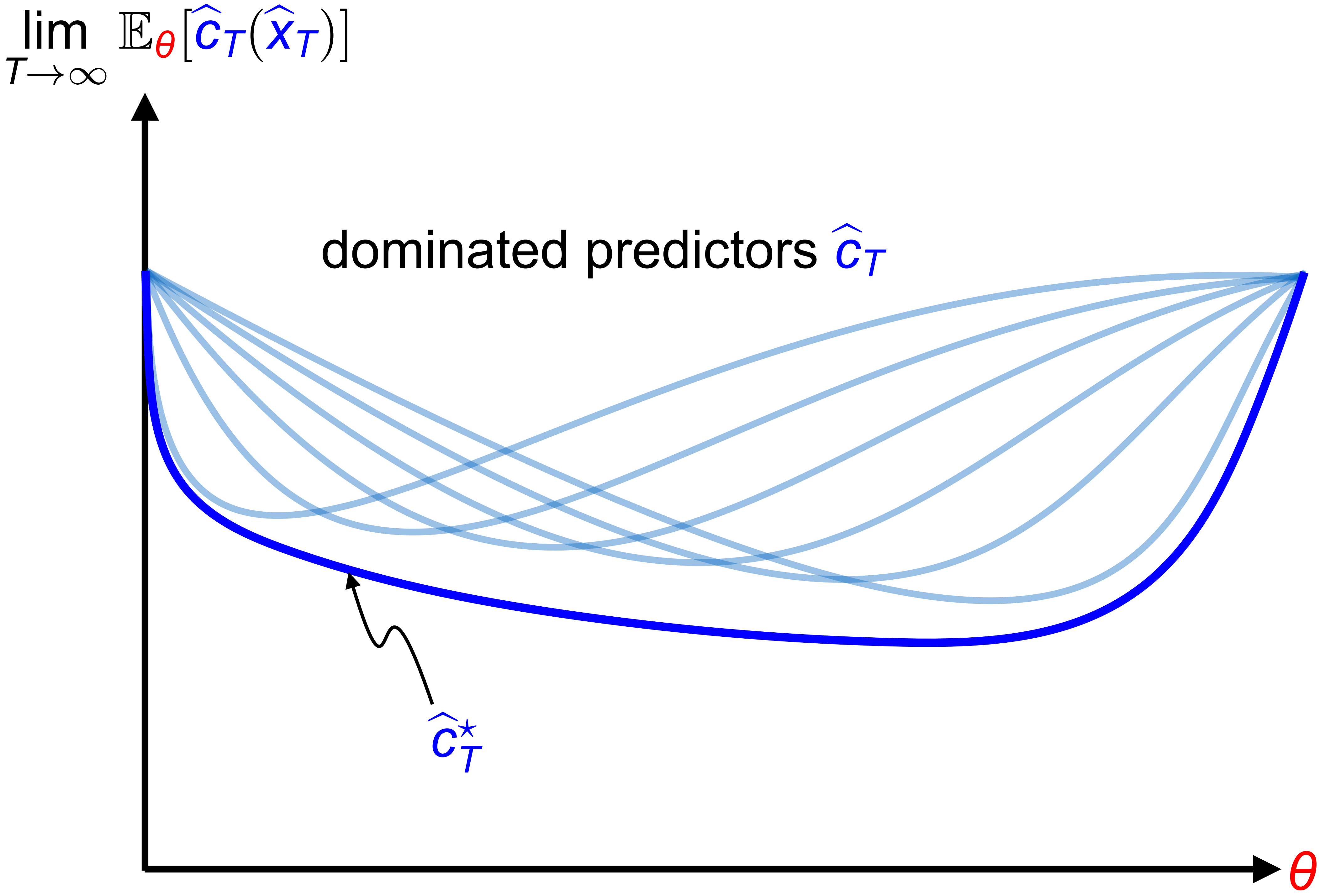
Optimal Data-Driven Optimization



Optimal Data-Driven Optimization



Pareto-Dominant Solutions



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

Meta-Optimization Problem (MOP)

MOP optimizes over *__all surrogate optimization models__*

$$
\begin{aligned}
&\underset{\widehat{c}_T,\widehat{x}_T}{\text{minimize}} \quad \left\{ \lim_{T\to\infty} \mathbb{E}_\theta \left[\widehat{c}_T(\widehat{x}_T) \right] \right\}_{\theta\in\Theta} \\
&\text{subject to} \quad \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta \left[c(\widehat{x}_T, \theta) > \widehat{c}_T(\widehat{x}_T) \right] \leq -r \quad \forall \theta \in \Theta
\end{aligned}
$$

Strengths:

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

Meta-Optimization Problem (MOP)

MOP optimizes over *all surrogate optimization models*

$$\begin{aligned}
&\underset{\widehat{c}_T, \widehat{x}_T}{\text{minimize}} \quad \left\{ \lim_{T\to\infty} \mathbb{E}_\theta \left[\widehat{c}_T(\widehat{x}_T) \right] \right\}_{\theta\in\Theta} \\
&\text{subject to} \quad \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta \left[c(\widehat{x}_T, \theta) > \widehat{c}_T(\widehat{x}_T) \right] \le -r \quad \forall \theta \in \Theta
\end{aligned}$$

Weaknesses:

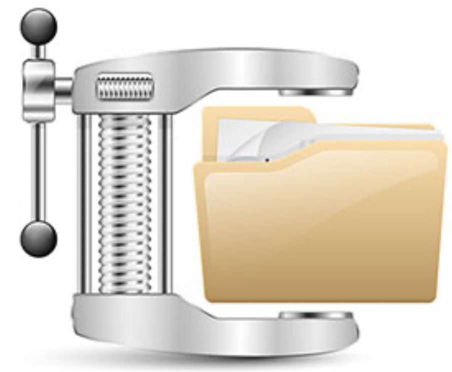
- performance criteria are asymptotic
- choice of r is subjective
- why insist on exponential decay?
- feasible/optimal models are biased

Restricted Meta-Optimization Problems

Data Compression

Compress the raw data to an estimator of θ

$$\xi_1, \xi_2, \ldots, \xi_T \in \mathbb{R}^{T \cdot d} \longrightarrow$$



$$\longrightarrow \widehat{\theta}_T \in \mathbb{R}^d$$

$\Longrightarrow$ compressed predictors depend on $\xi_1, \xi_2, \ldots, \xi_T$ and on T only indirectly through the summary statistic $\widehat{\theta}_T$

Compressed Predictors and Prescriptors

- Set $\widehat{c}_T(x) = \tilde{c}(x, \widehat{\theta}_T)$ for some continuous function $\tilde{c}$
- Set $\widehat{x}_T = \tilde{x}(\widehat{\theta}_T)$ for some quasi-continuous function $\tilde{x}$ with

$$\tilde{x}(\widehat{\theta}_T) \in \operatorname*{argmin}_{x \in X} \; \tilde{c}(x, \widehat{\theta}_T)$$

Restricted MOP

Restricted MOP over compressed predictors/prescriptors:

$$\underset{\tilde{c},\tilde{x}}{\text{minimize}} \quad \{\tilde{c}(\tilde{x}(\theta),\theta)\}_{\theta\in\Theta}$$

$$\text{subject to} \ \limsup_{T\to\infty} \frac{1}{T}\log \mathbb{P}_{\theta}\left[c(\tilde{x}(\widehat{\theta}_T),\theta) > \tilde{c}(\tilde{x}(\widehat{\theta}_T),\widehat{\theta}_T)\right] \leq -r \quad \forall \theta \in \Theta$$

Large Deviation Principle (LDP)

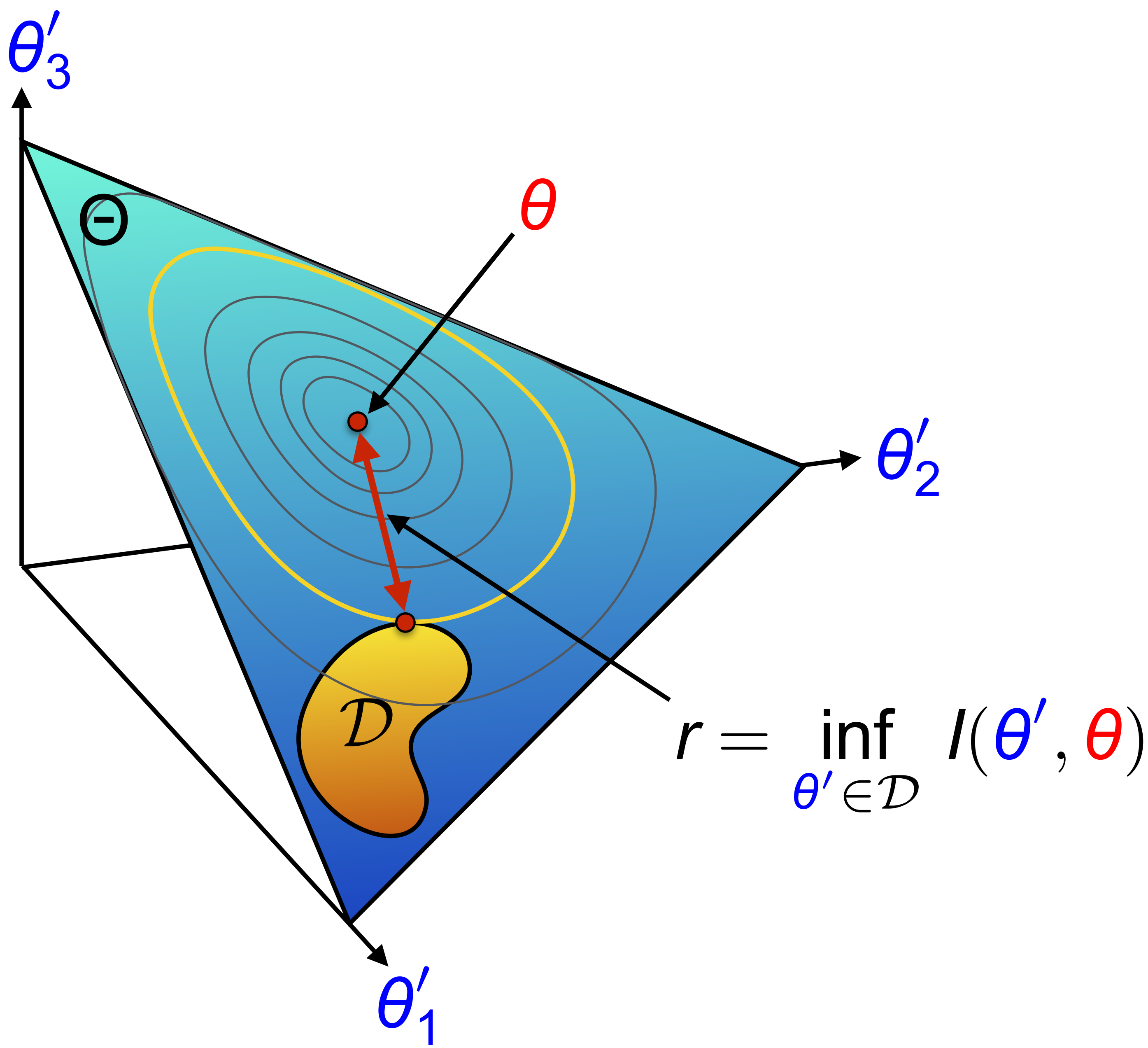
Definition:[1] The estimators $\widehat{\theta}_T$, $T \in \mathbb{N}$, satisfy an LDP if there exists a rate function $I(\theta', \theta)$ such that for all Borel sets $\mathcal{D} \subseteq \Theta$

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

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

[1] den Hollander, *American Mathematical Society*, 2008; Dembo & Zeitouni, *Springer*, 2009.

Large Deviation Principle (LDP)



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

DRO is Optimal

Assumption:

- $\widehat{\theta}_T$ satisfies an LDP with a “regular” rate function

Theorem 1 (DRO is optimal): The following distributionally robust compressed predictor is a Pareto-dominant solution for the <u>restricted</u> MOP.

$$\tilde{c}^{\star}(x, \widehat{\theta}_T) = \left\{ \begin{array}{ll} \sup\limits_{\theta \in \Theta} & c(x, \theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \leq r \end{array} \right.$$

DRO is Optimal

Assumption:

- $\widehat{\theta}_T$ satisfies an LDP with a "regular" rate function

Theorem 1 (DRO is optimal): The following distributionally robust compressed predictor is a Pareto-dominant solution for the <u>restricted</u> MOP.

$$\tilde{c}^\star(x, \widehat{\theta}_T) = \left\{ \begin{array}{ll} \sup_{\theta \in \Theta} & c(x, \theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \le r \end{array} \right.$$

Note:

- The shape of the ambiguity set is determined by $\widehat{\theta}_T$
- The "radius" of the ambiguity set is given by the decay rate r

Separation of Estimation and Optimization

Sufficient Statistic

Definition: $\widehat{\theta}_T$ is a sufficient statistic for θ if the distribution of $\xi_1, \xi_2, \ldots, \xi_T$ conditional on $\widehat{\theta}_T = \theta'$ is independent of $\theta \in \Theta$.

$\Longrightarrow$ Lossless compression

$\xi_1, \xi_2, \ldots, \xi_T \in \mathbb{R}^{T \cdot d}$ $\longrightarrow$  $\longrightarrow$ $\widehat{\theta}_T \in \mathbb{R}^d$

DRO is Optimal

Assumptions:

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

Theorem 2 (DRO is optimal)**:** The following distributionally robust surrogate optimization model is a Pareto-dominant solution for the original MOP.

$$\widehat{c}^{\star}_T(x) = \left\{ \begin{array}{ll} \sup\limits_{\theta \in \Theta} & c(x, \theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \leq r \end{array} \right.$$

DRO is Optimal

Assumptions:

- $\widehat{\theta}_T$ is a sufficient statistic for θ
- $\widehat{\theta}_T$ satisfies an LDP ...

Separation of estimation and optimization:

1) Evaluate the estimator

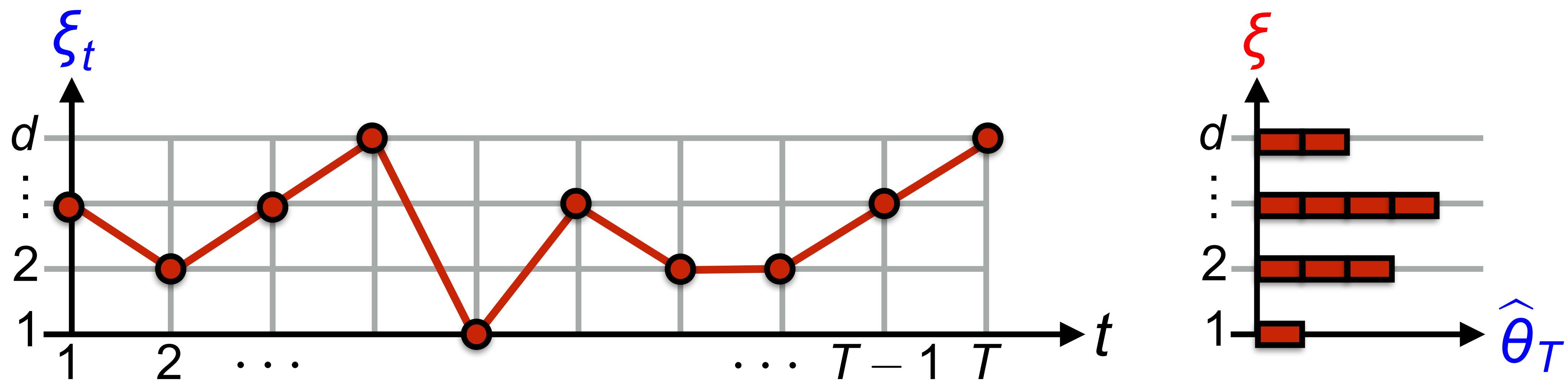
2) Solve the DRO problem

Data-Generating Processes

Newsvendor Problem Revisited

"Compressing" the raw data:

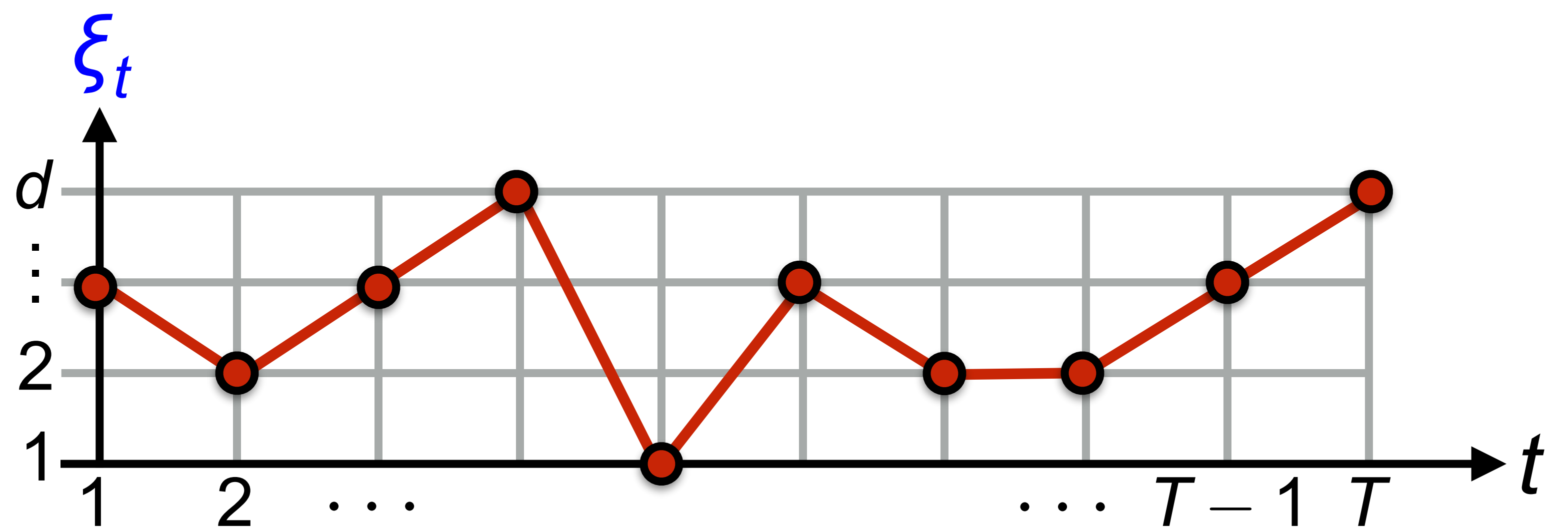
demand observations $(\xi_1, \xi_2, \ldots, \xi_T)$ $\leftrightsquigarrow$ empirical distribution $(\widehat{\theta}_T)_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\xi_t = i}$

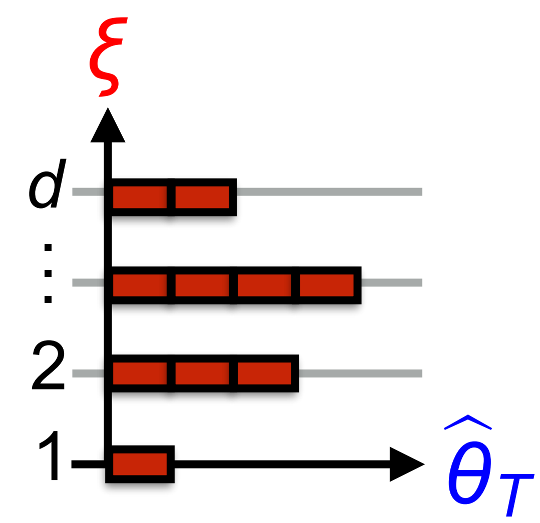


Newsvendor Problem Revisited

"Compressing" the raw data:

demand observations $(\xi_1, \xi_2, \ldots, \xi_T)$ $\Big\} \longleftrightarrow \Big\{$ empirical distribution $(\widehat{\theta}_T)_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\xi_t = i}$





- Fisher-Neyman:[1] $\widehat{\theta}_T$ is a sufficient statistic for θ
- Sanov:[2] $\widehat{\theta}_T$ satisfies an LDP with $I(\widehat{\theta}_T, \theta) = D_{\mathrm{KL}}(\widehat{\theta}_T \,\|\, \theta)$

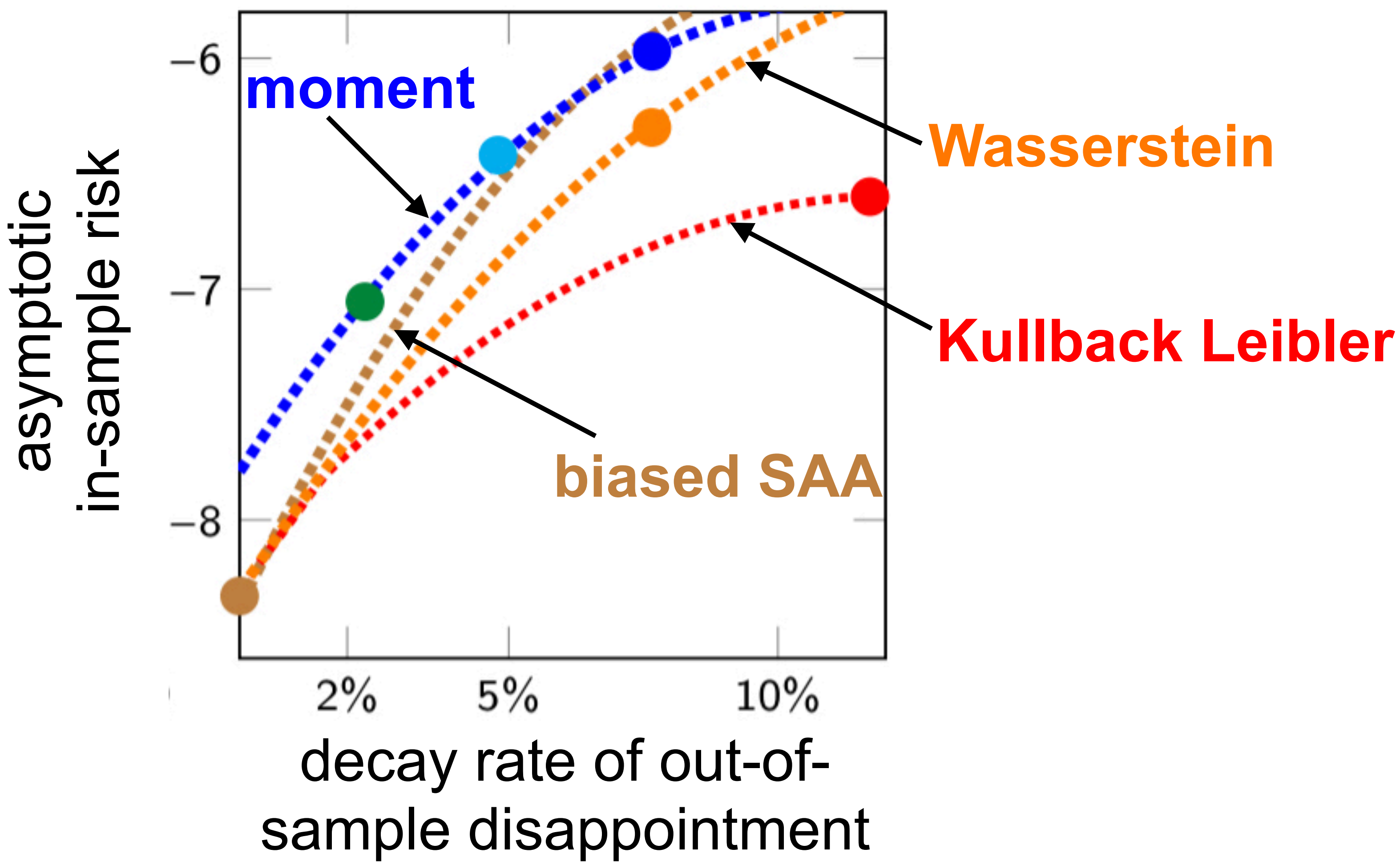
---

1) Lehmann & Casella, *Springer*, 1998;
2) Sanov, *Matematicheskii Sbornik*, 1957.

Newsvendor Problem Revisited

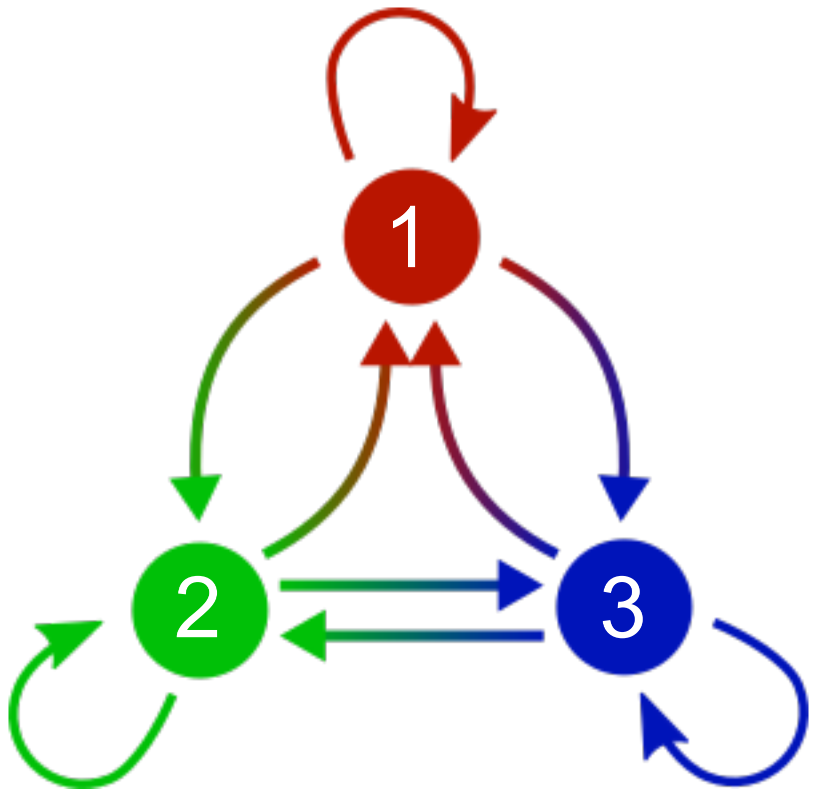
- The separation principle holds
- The optimal data-driven predictor is

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



Finite-State Markov Chains

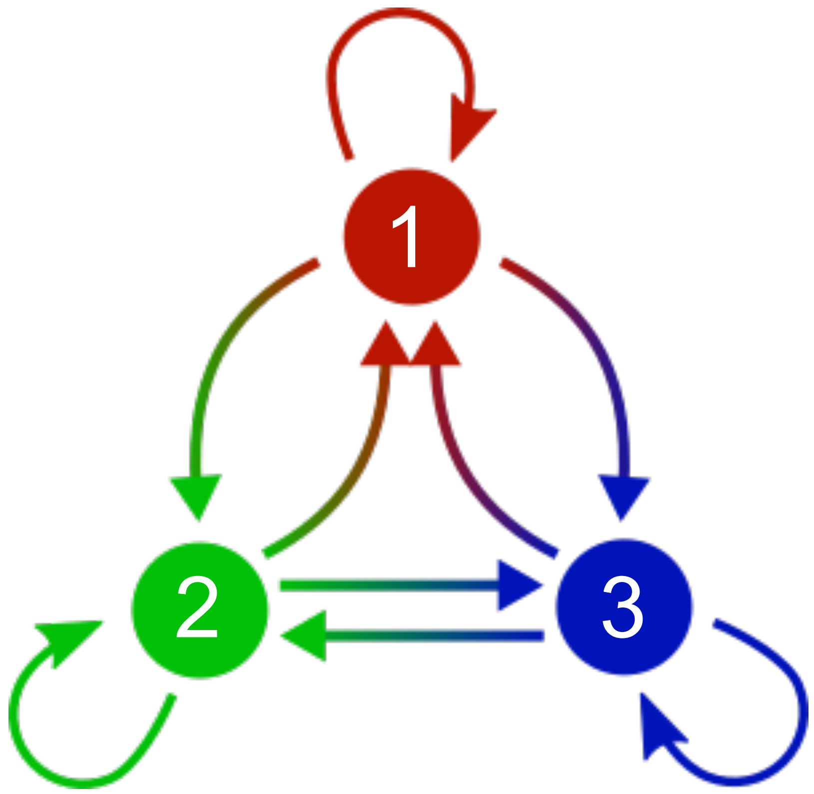
Assume that $\{\xi_t\}_{t\in\mathbb{N}}$ is a Markov chain on $\{1, 2, \ldots, d\}$



Finite-State Markov Chains

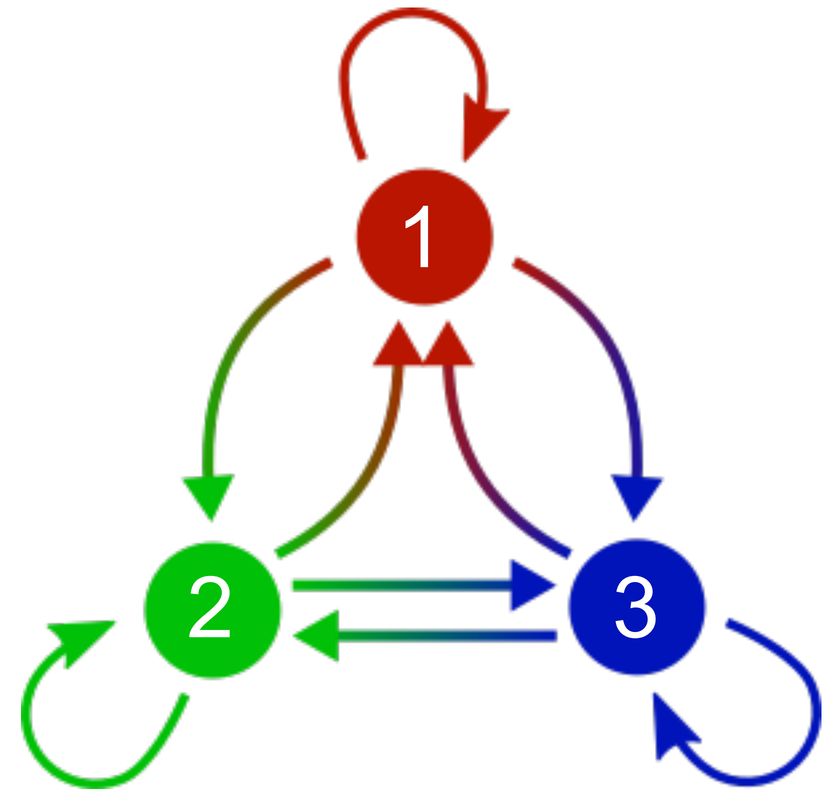
Assume that $\{\xi_t\}_{t\in\mathbb{N}}$ is a Markov chain on $\{1, 2, \ldots, d\}$

- $\theta_{ij} = \lim_{T\to\infty} \mathbb{P}_\theta(\xi_t = i,\ \xi_{t+1} = j)$
- all one-step transitions possible



Finite-State Markov Chains

Assume that $\{\xi_t\}_{t\in\mathbb{N}}$ is a Markov chain on $\{1, 2, \ldots, d\}$

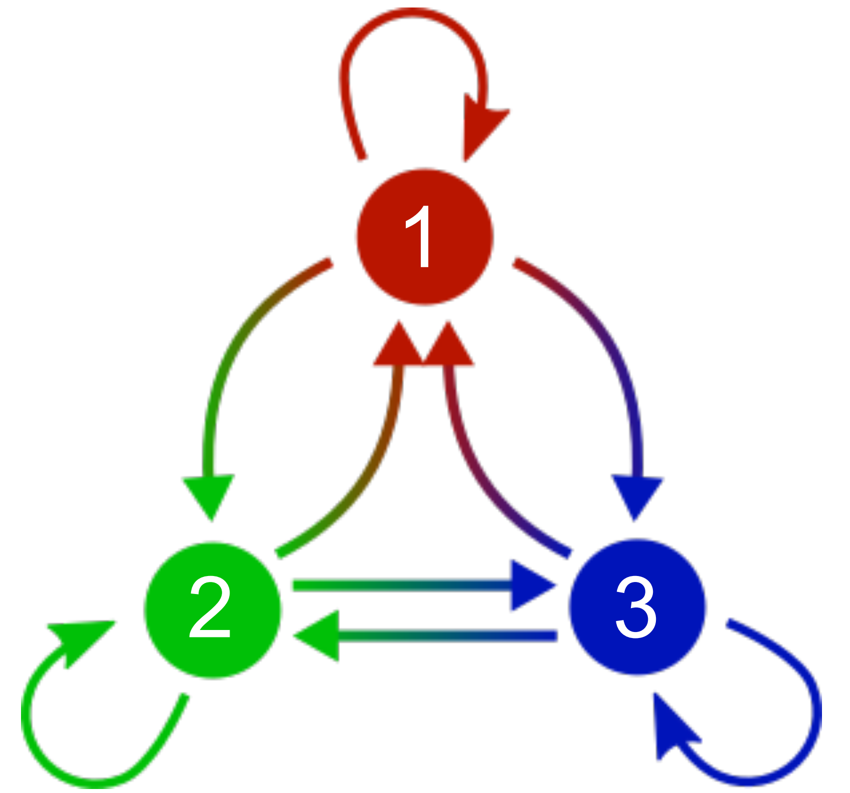


- $\theta_{ij} = \lim_{T\to\infty} \mathbb{P}_\theta(\xi_t = i,\ \xi_{t+1} = j)$
- all one-step transitions possible

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

Finite-State Markov Chains

Assume that $\{\xi_t\}_{t\in\mathbb{N}}$ is a Markov chain on $\{1, 2, \ldots, d\}$



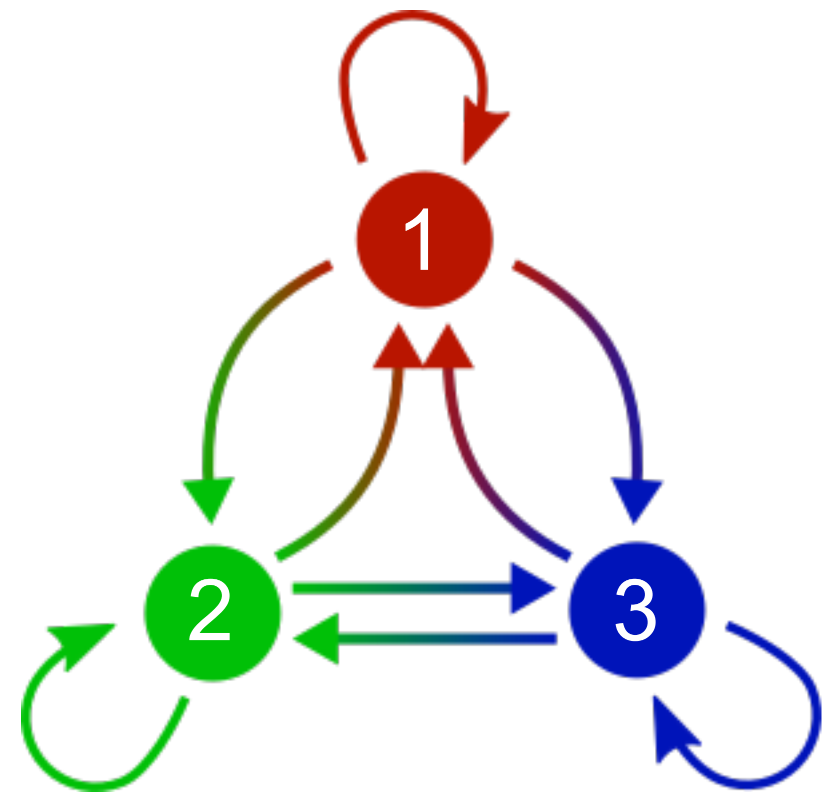
- $\theta_{ij} = \lim_{T\to\infty} \mathbb{P}_\theta(\xi_t = i,\ \xi_{t+1} = j)$
- all one-step transitions possible

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

all transitions have probability > 0

Finite-State Markov Chains

Assume that $\{\xi_t\}_{t\in\mathbb{N}}$ is a Markov chain on $\{1, 2, \ldots, d\}$



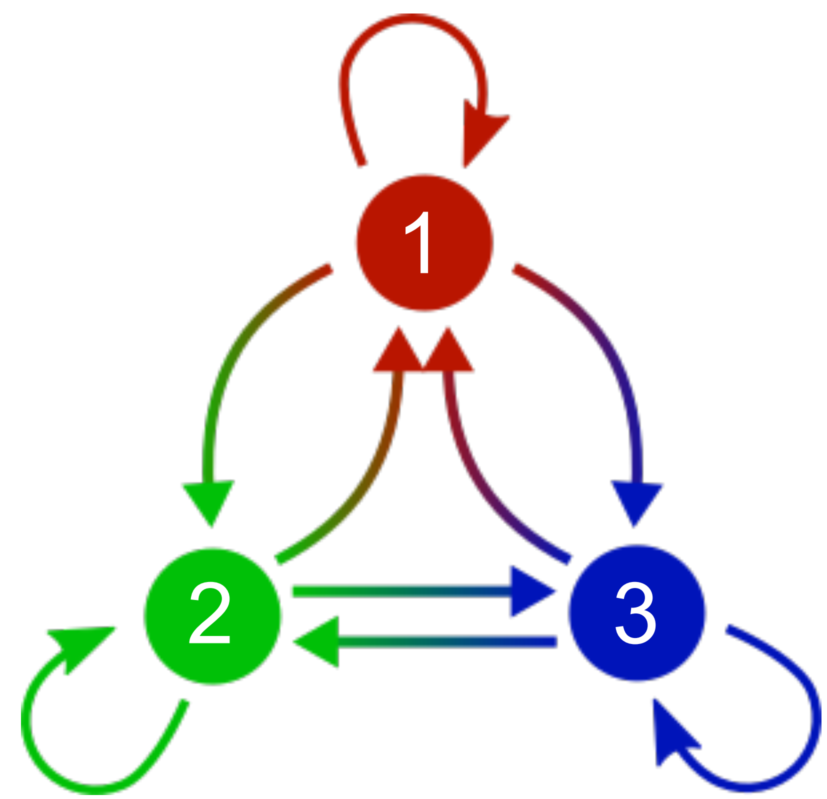
- $\theta_{ij} = \lim_{T\to\infty} \mathbb{P}_\theta(\xi_t = i,\ \xi_{t+1} = j)$
- all one-step transitions possible

sum of all entries $=1$,

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

Finite-State Markov Chains

Assume that $\{\xi_t\}_{t\in\mathbb{N}}$ is a Markov chain on $\{1,2,\ldots,d\}$



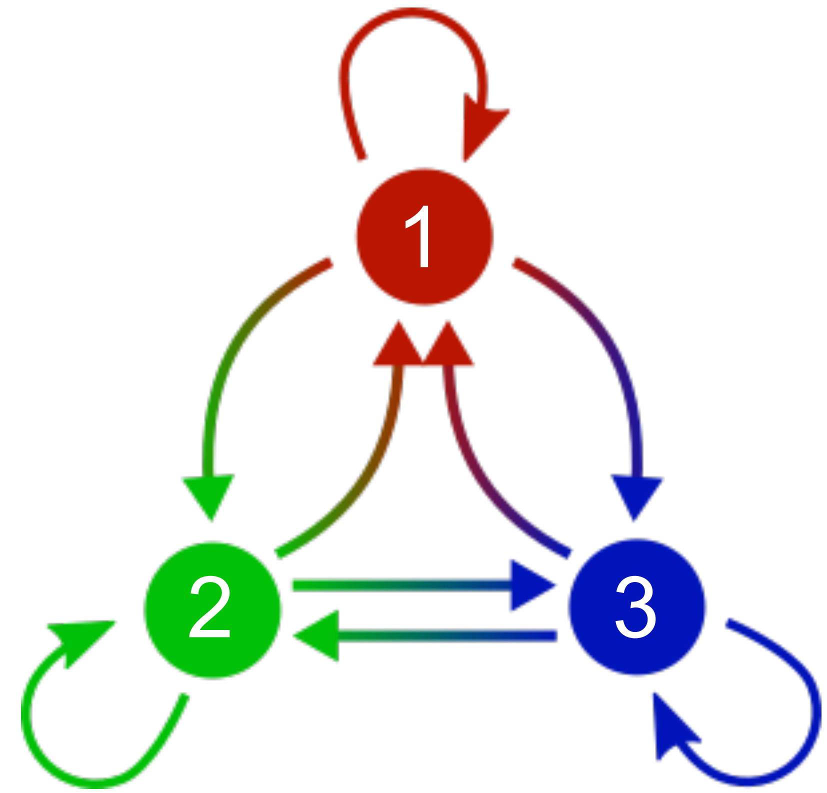
- $\theta_{ij} = \lim_{T\to\infty} \mathbb{P}_\theta(\xi_t = i,\ \xi_{t+1} = j)$
- all one-step transitions possible

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

invariant probability of state i

Finite-State Markov Chains

Assume that $\{\xi_t\}_{t\in\mathbb{N}}$ is a Markov chain on $\{1, 2, \ldots, d\}$



- $\theta_{ij} = \lim_{T\to\infty} \mathbb{P}_\theta(\xi_t = i,\ \xi_{t+1} = j)$
- all one-step transitions possible

$$\Theta = \left\{\theta \in \mathbb{R}^{d\times d}_{++} : \sum_{i,j} \theta_{ij} = 1,\ \ \sum_j \theta_{ij} = \underbrace{\sum_j \theta_{ji}}_{\text{invariant probability of state } i} \forall i\right\}$$

Finite-State Markov Chains

"Compressing" the raw data:

$$\left.\begin{array}{l}\text{available observations}\\ (\xi_1, \xi_2, \ldots, \xi_T)\end{array}\right\} \leftrightsquigarrow \left\{\begin{array}{l}\text{empirical doublet distribution}\\ (\widehat{\theta}_T)_{ij} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{(\xi_{t-1},\xi_t)=(i,j)}\end{array}\right.$$

Finite-State Markov Chains

"Compressing" the raw data:

$$\left.\begin{array}{l}\text{available observations}\\ (\xi_1, \xi_2, \ldots, \xi_T)\end{array}\right\} \leftrightsquigarrow \left\{\begin{array}{l}\text{empirical doublet distribution}\\ (\widehat{\theta}_T)_{ij} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{(\xi_{t-1},\xi_t)=(i,j)}\end{array}\right.$$

- Fisher-Neyman:[1] $\widehat{\theta}_T$ is a sufficient statistic for θ
- Dembo & Zeitouni:[2] $\widehat{\theta}_T$ satisfies an LDP with $I(\widehat{\theta}_T, \theta) = D_\mathrm{c}(\widehat{\theta}_T \,\|\, \theta)$

Definition: Conditional relative entropy

$$D_\mathrm{c}(\theta' \,\|\, \theta) = \sum_{i,j} \theta'_{ij} \left(\log\left(\frac{\theta'_{ij}}{\sum_k \theta'_{ik}}\right) - \log\left(\frac{\theta_{ij}}{\sum_k \theta_{ik}}\right)\right)$$

---

[1] Lehmann & Casella, *Springer*, 1998;
[2] Dembo & Zeitouni, *Springer*, 1998.

Autoregressive Gaussian Processes

Vector autoregressive processes with unknown drift:

- $\xi_{t+1} = \theta + A\xi_t + \varepsilon_{t+1}$ stationary, driven by Gaussian noise
- $\widehat{\theta}_T = (\mathbb{I}_d - A)\frac{1}{T}\sum_{t=1}^{T} \xi_t$ satisfies LDP but is *not* sufficient[1)]

Scalar autoregressive processes with unknown coefficient:

- $\xi_{t+1} = \theta\xi_t + \varepsilon_{t+1}$ stationary, driven by Gaussian noise
- Least squares and Yule-Walker estimators satisfy LDPs but are *not* sufficient[2)]

[1)] Dembo & Zeitouni, *Springer*, 1998;
[2)] Bercu et al., *Stochastic Processes and their Applications*, 1997.

I.I.D. Processes with Parametric CDFs

Assume that the $\{\xi_t\}_{t\in\mathbb{N}}$ are i.i.d. with any of the following CDFs:

- normal distribution with mean θ
- exponential distribution with rate parameter θ
- gamma distribution with scale parameter θ
- Poisson distribution with rate parameter θ
- Bernoulli distribution with success probability θ
- geometric distribution with success probability θ
- binomial distribution with success probability θ

I.I.D. Processes with Parametric CDFs

Then, $\widehat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \xi_t$ is sufficient[1)] and satisfies an LDP,[2)] where

- $\Lambda(\lambda, \theta) = \log \mathbb{E}_\theta \left[\exp(\lambda^\top \xi_t)\right]$ is the log-MGF, and
- $I(\theta', \theta) = \sup_\lambda \ \theta^\top \lambda - \Lambda(\lambda, \theta)$ is a “regular” rate function.

---

1) Lehmann & Casella, *Springer*, 1998;
2) Cramér, *Actualités scientifiques et industrielles*, 1938.

I.I.D. Processes with Parametric CDFs

Then, $\widehat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \xi_t$ is sufficient[1] and satisfies an LDP,[2] where

- $\Lambda(\lambda, \theta) = \log \mathbb{E}_\theta \left[\exp(\lambda^\top \xi_t) \right]$ is the log

Many convex uncertainty sets can be constructed in this way and are thus optimal for some i.i.d. process!

---

[1] Lehmann & Casella, *Springer*, 1998;
[2] Cramér, *Actualités scientifiques et industrielles*, 1938.

Summary & Conclusions

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

Conclusions

- **Data efficiency**
  - Pareto dominance reminiscent of Bahadur efficiency
- **Generality of results**
  - hold even for non-convex decision problems
  - hold even for non-i.i.d. data processes
- **Theoretical justification of DRO**
  - shape of ambiguity set depends on the data process
  - radius of ambiguity set has physical interpretation
- **Computation**
  - customized algorithms for new DRO models[1)]

1) Li et al., *ICML*, 2021.

This Talk is Based on…

[1] M. Li, T. Sutter and D. Kuhn. **Distributionally Robust Optimization Based on Markovian Data**. ICML, 2021.

[2] T. Sutter, W. Jongeneel, S. Shafieezadeh Abadeh and D. Kuhn. **From Moderate Deviations Theory to Distributionally Robust Optimization: Learning from Correlated Data**. *Working paper*, 2021.

[3] T. Sutter, B. Van Parys and D. Kuhn. **A General Framework for Optimal Data-Driven Optimization**. *arXiv:2010.06606*, 2020

[4] B. Van Parys, P. Mohajerin Esfahani and D. Kuhn. **From Data to Decisions: Distributionally Robust Optimization is Optimal**. *Management Science*, 2020.

https://www.epfl.ch/labs/rao/

Appendix: Proof Ideas

Optimizing over Optimization Problems

Restricted MOP for a fixed decision:

$$\begin{array}{ll} \underset{\tilde{c}}{\text{minimize}} & \{\tilde{c}(\theta)\}_{\theta \in \Theta} \\ \text{subject to} & \limsup_{T \to \infty} \frac{1}{T} \log \mathbb{P}_{\theta} \left[c(\theta) > \tilde{c}(\widehat{\theta}_T) \right] \leq -r \ \ \forall \theta \in \Theta \end{array}$$

Optimizing over Optimization Problems

Restricted MOP for a fixed decision:

$$\begin{aligned} \underset{\tilde{c}}{\text{minimize}} \quad & \{\tilde{c}(\theta)\}_{\theta \in \Theta} \\ \text{subject to} \quad & \limsup_{T \to \infty} \frac{1}{T} \log \mathbb{P}_\theta \left[c(\theta) > \tilde{c}(\widehat{\theta}_T) \right] \leq -r \;\; \forall \theta \in \Theta \end{aligned}$$

Pareto-dominant solution:

$$\tilde{c}^\star(\widehat{\theta}_T) = \left\{ \begin{array}{ll} \sup\limits_{\theta \in \Theta} & c(\theta) \\ \text{s.t.} & I(\widehat{\theta}_T, \theta) \leq r \end{array} \right.$$

Feasibility

Theorem: $\tilde{c}^\star$ is feasible in the MOP.

Feasibility

Theorem: $\tilde{c}^\star$ is feasible in the MOP.

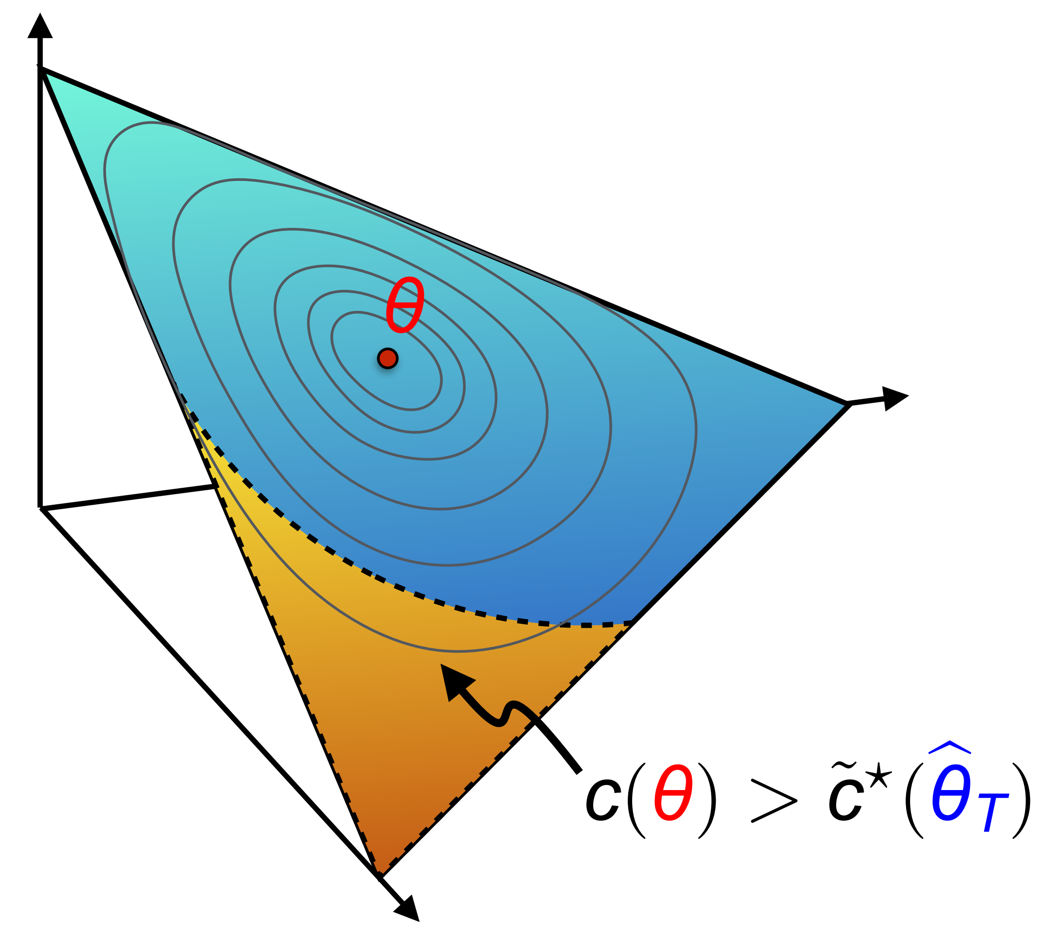
$$
\begin{aligned}
c(\theta) > \tilde{c}^\star(\widehat{\theta}_T) \quad & \Longrightarrow \quad c(\theta) > \sup_{\theta' \in \Theta} \left\{ c(\theta') : I(\widehat{\theta}_T, \theta') \le r \right\} \\
& \Longrightarrow \quad I(\widehat{\theta}_T, \theta) > r
\end{aligned}
$$

Feasibility

Theorem: $\tilde{c}^\star$ is feasible in the MOP.

$$c(\theta) > \tilde{c}^\star(\widehat{\theta}_T) \quad \Longrightarrow \quad c(\theta) > \sup_{\theta' \in \Theta} \left\{ c(\theta') : I(\widehat{\theta}_T, \theta') \leq r \right\}$$

$$\Longrightarrow \quad I(\widehat{\theta}_T, \theta) > r$$

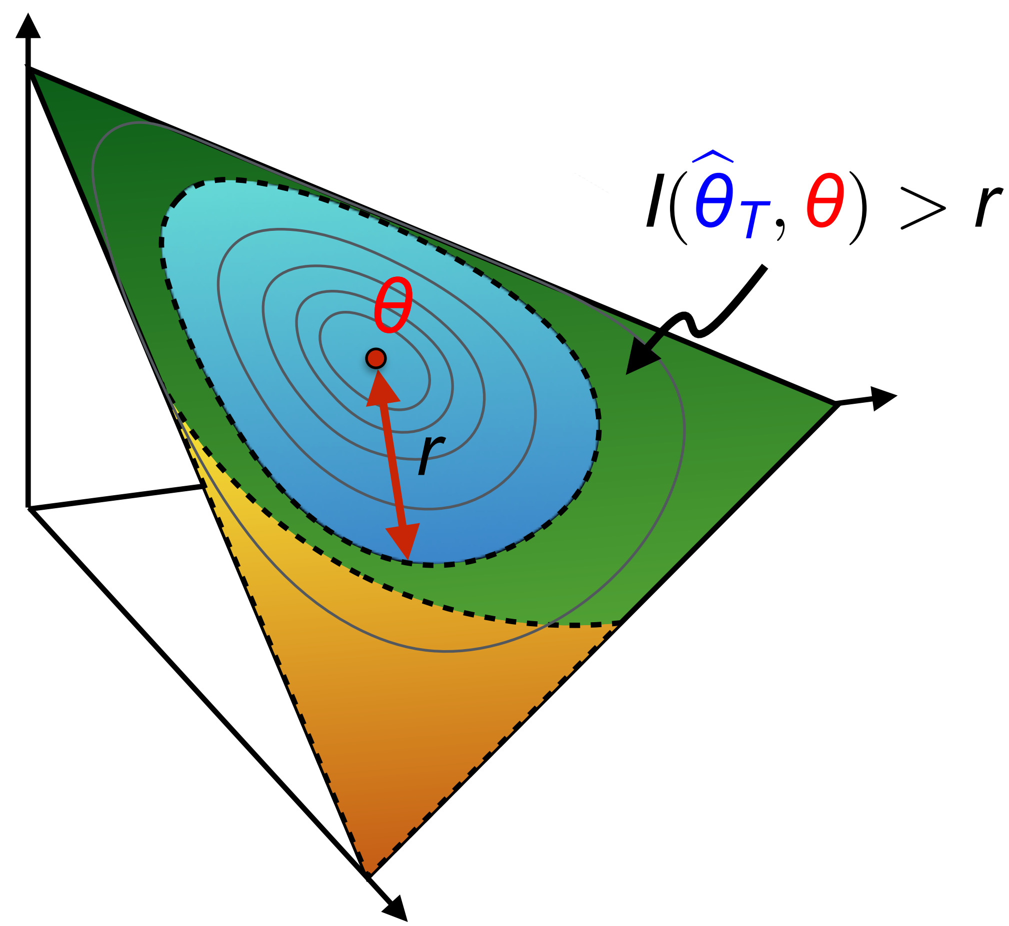


Feasibility

Theorem: $\tilde{c}^\star$ is feasible in the MOP.

$$c(\theta) > \tilde{c}^\star(\widehat{\theta}_T) \quad \Longrightarrow \quad c(\theta) > \sup_{\theta' \in \Theta} \left\{ c(\theta') : I(\widehat{\theta}_T, \theta') \leq r \right\}$$

$$\Longrightarrow \quad I(\widehat{\theta}_T, \theta) > r$$



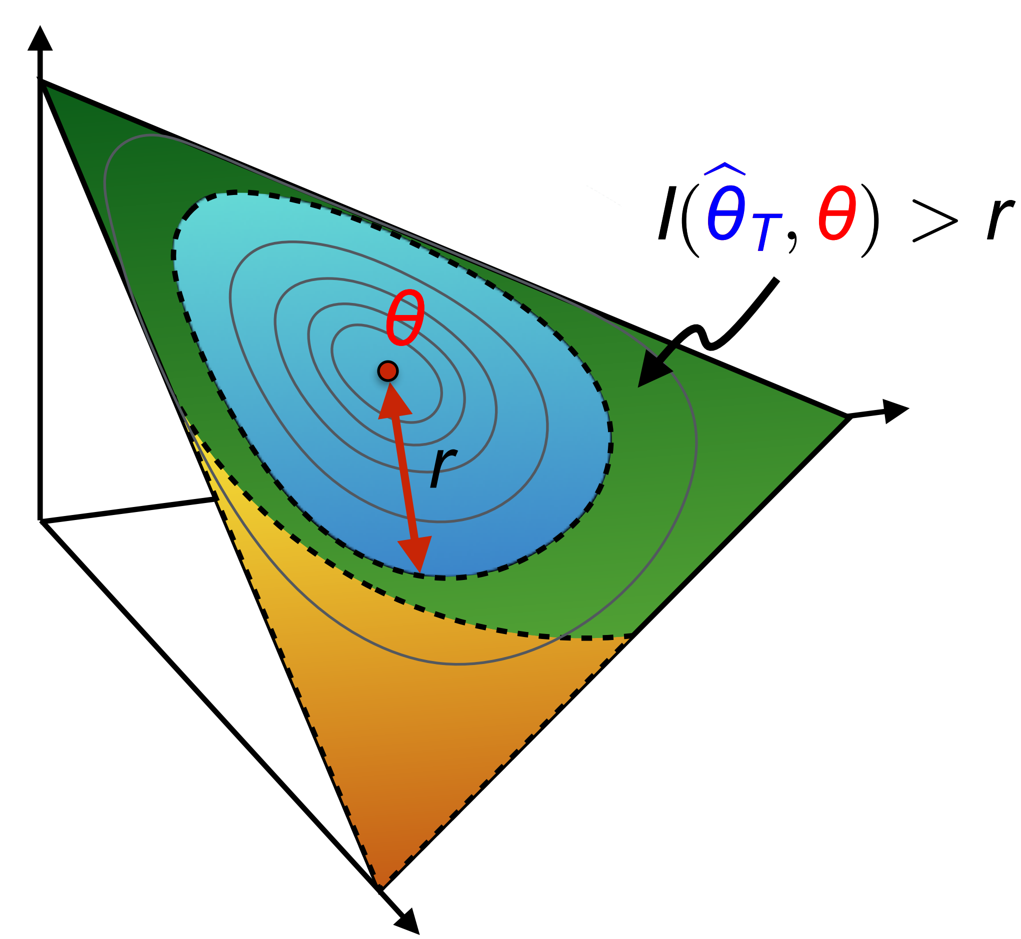
Feasibility

Theorem: $\tilde{c}^\star$ is feasible in the MOP.

$$c(\theta) > \tilde{c}^\star(\widehat{\theta}_T) \quad \Longrightarrow \quad c(\theta) > \sup_{\theta' \in \Theta} \left\{ c(\theta') : I(\widehat{\theta}_T, \theta') \leq r \right\}$$

$$\Longrightarrow \quad I(\widehat{\theta}_T, \theta) > r$$

$$\Longrightarrow \quad \mathbb{P}_\theta\left[c(\theta) > \tilde{c}^\star(\widehat{\theta}_T)\right] \leq \underbrace{\mathbb{P}_\theta\left[I(\widehat{\theta}_T, \theta) > r\right]}_{\leq e^{-rT+o(T)}\ \forall \theta \in \Theta}$$



Optimality

Theorem: If $r > 0$, then $\tilde{c}^{\star}$ is Pareto-dominant in the MOP.

Optimality

Theorem: If $r > 0$, then $\tilde{c}^\star$ is Pareto-dominant in the MOP.

Assume that there is $\tilde{c}$ feasible in MOP and θ_1 with $\tilde{c}(\theta_1) < \tilde{c}^\star(\theta_1)$

Optimality

Theorem: If $r > 0$, then $\tilde{c}^\star$ is Pareto-dominant in the MOP.

Assume that there is $\tilde{c}$ feasible in MOP and θ_1 with $\tilde{c}(\theta_1) < \tilde{c}^\star(\theta_1)$

$$\implies \quad \tilde{c}(\theta_1) < \sup_{\theta \in \Theta} \{c(\theta) : I(\theta_1, \theta) \leq r\}$$

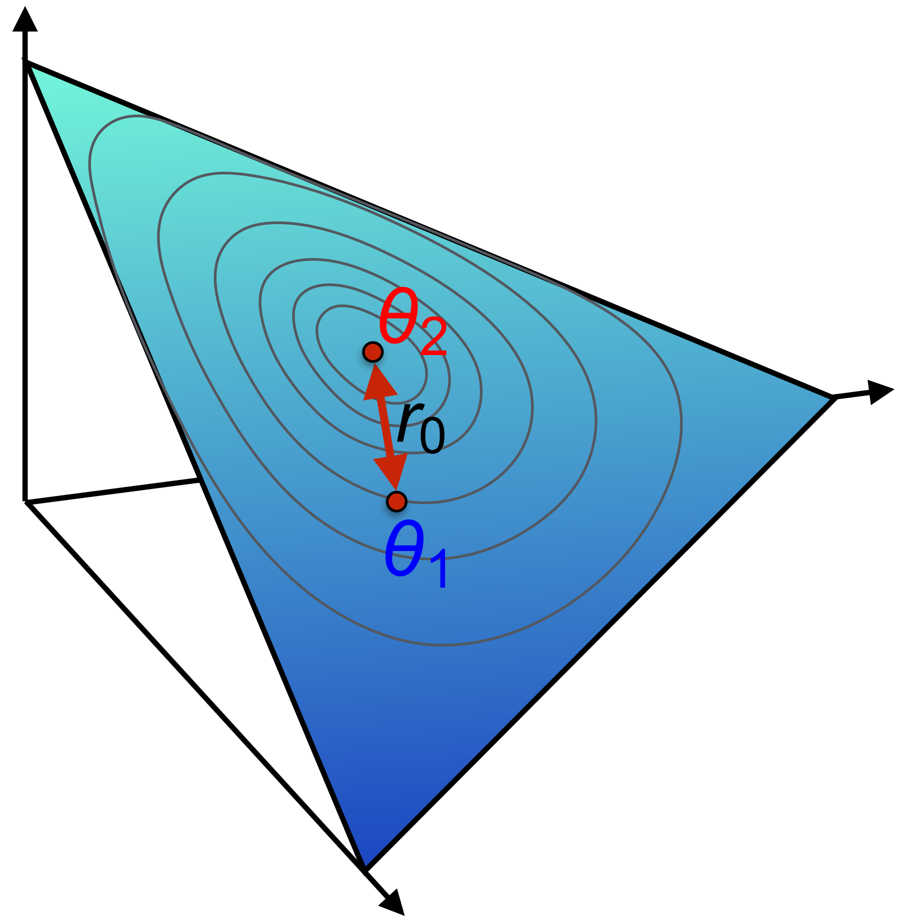
Optimality

Theorem: If $r > 0$, then $\tilde{c}^\star$ is Pareto-dominant in the MOP.

Assume that there is $\tilde{c}$ feasible in MOP and θ_1 with $\tilde{c}(\theta_1) < \tilde{c}^\star(\theta_1)$

$$\implies \quad \tilde{c}(\theta_1) < \sup_{\theta \in \Theta} \{c(\theta) : I(\theta_1, \theta) \le r\}$$

$$\implies \quad \exists \theta_2 : \quad \tilde{c}(\theta_1) < c(\theta_2) \ \text{ and } \ I(\theta_1, \theta_2) = r_0 < r$$



Optimality

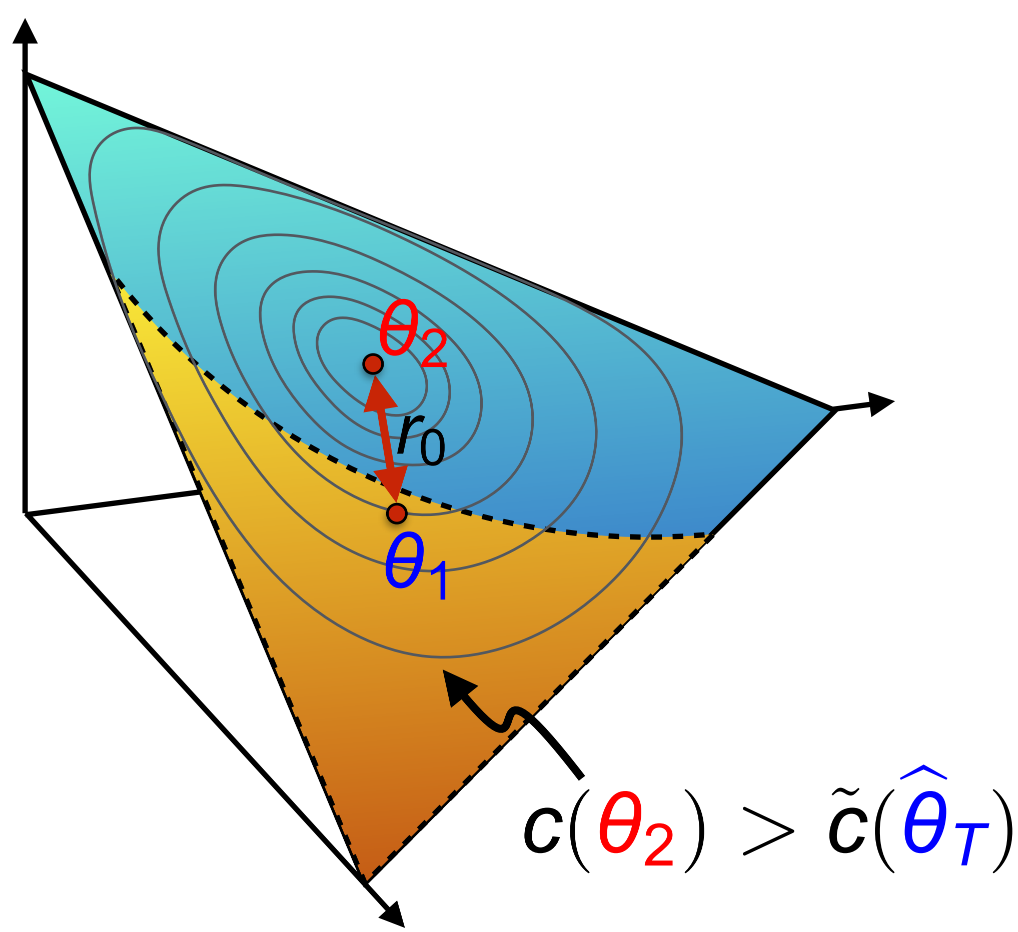
Theorem: If $r > 0$, then $\tilde{c}^\star$ is Pareto-dominant in the MOP.

Assume that there is $\tilde{c}$ feasible in MOP and θ_1 with $\tilde{c}(\theta_1) < \tilde{c}^\star(\theta_1)$

$$\implies \quad \tilde{c}(\theta_1) < \sup_{\theta \in \Theta} \{c(\theta) : I(\theta_1, \theta) \leq r\}$$

$$\implies \quad \exists \theta_2 : \quad \tilde{c}(\theta_1) < c(\theta_2) \text{ and } I(\theta_1, \theta_2) = r_0 < r$$

$$\implies \quad \mathbb{P}_{\theta_2}\left[c(\theta_2) > \tilde{c}(\widehat{\theta}_T)\right] \geq e^{-r_1 T + o(T)}$$



Optimality

Theorem: If $r > 0$, then $\tilde{c}^\star$ is Pareto-dominant in the MOP.

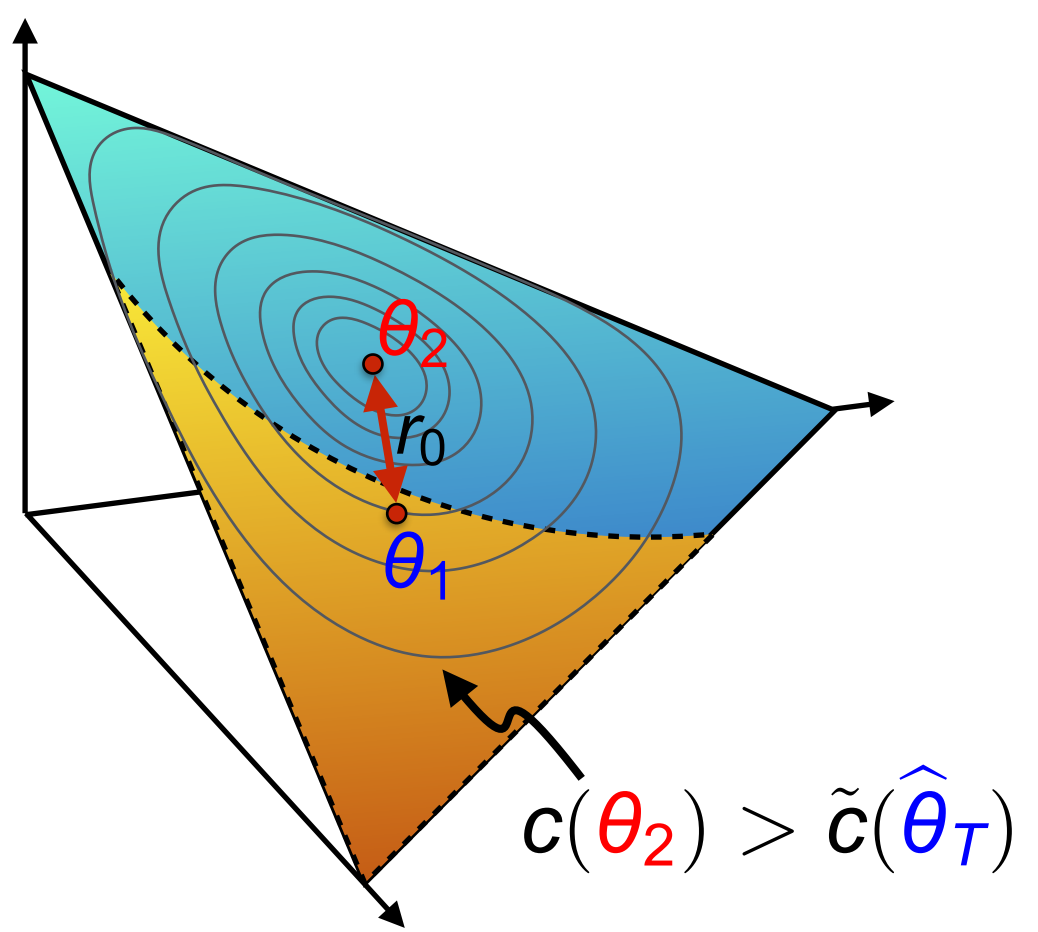
Assume that there is $\tilde{c}$ feasible in MOP and θ_1 with $\tilde{c}(\theta_1) < \tilde{c}^\star(\theta_1)$

$$\implies \quad \tilde{c}(\theta_1) < \sup_{\theta \in \Theta} \{c(\theta) : I(\theta_1, \theta) \leq r\}$$

$$\implies \quad \exists \theta_2 : \quad \tilde{c}(\theta_1) < c(\theta_2) \ \text{ and } \ I(\theta_1, \theta_2) = r_0 < r$$

$$\implies \quad \mathbb{P}_{\theta_2}\left[c(\theta_2) > \tilde{c}(\widehat{\theta}_T)\right] \geq e^{-r_1 T + o(T)}$$

$$\implies \quad \tilde{c} \text{ infeasible in MOP} \quad \lightning$$



Appendix: Data-Driven Control

From Data to Controllers?

Closed-loop LTI system: $x_{t+1} = \theta x_t + w_t$

Least squares estimator: $$\widehat{\theta}_T = \left(\sum_{t=1}^{T} \widehat{x}_t \, \widehat{x}_{t-1}^\top \right) \left(\sum_{t=1}^{T} \widehat{x}_{t-1} \, \widehat{x}_{t-1}^\top \right)^{-1}$$

Theorem: The modified least squares estimator $\theta + \sqrt[4]{T}(\widehat{\theta}_T - \theta)$ satisfies a moderate deviations principle with rate function

$$I(\theta', \theta) = \frac{1}{2} \mathrm{tr} \left({S_w}^{-1} (\theta' - \theta) S_\theta (\theta' - \theta)^\top \right),$$

where S_θ solves the Lyapunov equation $S_\theta = \theta S_\theta \theta^\top + S_w$.

$\Longrightarrow$ DRO bounds on $$J(\theta) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_\theta \left[x_t^\top (Q + K^\top R K) x_t \right]$$