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

February 21, 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 objective function that depends on the probability measure governing the uncertain problem parameters. Examples of such objective functions include the expected value or some risk measure of an uncertain loss function, the conditional expectation of an uncertain loss function given contextual covariates or the long-run average expected cost of a parametric control policy etc. The crux of data-driven decision-making is that the probability measure underlying the stochastic optimization problem is unknown. Throughout this paper we assume, however, that this probability measure is known to belong to a parametric family of the form $\{\mathbb{P}_\theta : \theta \in \Theta\}$. In addition, we assume that we have access to a finite trajectory of an exogenous stochastic process, which generates training samples that provide statistical information about θ. Examples of stochastic processes to be studied in this paper include processes of independent and identically distributed (i.i.d.) random variables on a finite state space, finite-state Markov chains, different classes of vector autoregressive processes or i.i.d. processes with parametric distribution functions, but many other examples are conceivable. These examples highlight that we actually allow the training samples to display serial dependence.

It is convenient to embed the original stochastic optimization problem into a parametric family of problems that are obtained by replacing the unknown true probability measure with any $\mathbb{P}_\theta$, $\theta \in \Theta$. The resulting stochastic optimization problems can be concisely represented as $\min_{x\in X} c(x, \theta)$, $\theta \in \Theta$, where X denotes the feasible set and $c(x, \theta)$ stands for the risk or cost of the decision x under the probability measure $\mathbb{P}_\theta$. As the parameter θ corresponding to the true probability measure is unknown, however, it is unclear which problem instance should be solved. We thus have no choice but to solve a data-driven surrogate optimization problem $\min_{x\in X} \widehat{c}_T(x)$, whose objective function $\widehat{c}_T$ is constructed independently of θ from T training samples. In the following, we denote by $\widehat{x}_T$ an optimal solution of the surrogate optimization problem, which is necessarily a function of the T training samples, too. For the sake of a succinct terminology, we henceforth refer to $\widehat{c}_T$ as a *data-driven predictor* because it predicts the risk of any decision x in view of the available data. Similarly, we refer to $\widehat{x}_T$ as a *data-driven prescriptor* because it prescribes a feasible decision in view of the available data. We emphasize that a data-driven prescriptor could be essentially *any* function that maps the available training data to a feasible decision. Indeed, it is easy to convince oneself that any such function can be expressed as the minimizer of a carefully constructed surrogate optimization problem. The main goal of this paper is to design—in a rigorous statistical sense—an 'optimal' surrogate optimization problem, which is equivalent to finding 'optimal' data-driven predictors and prescriptors.

The quality of a data-driven prescriptor $\widehat{x}_T$ under $\mathbb{P}_\theta$ is unequivocally measured by its out-of-sample risk $c(\widehat{x}_T, \theta)$. As the true θ is unknown, an ideal prescriptor would have to minimize the out-of-sample risk simultaneously for all $\theta \in \Theta$ and thus necessarily also for the unknown true θ. Unfortunately, such ideal data-driven prescriptors are unavailable for non-trivial stochastic optimization problems. To circumvent this difficulty, we recall that any data-driven prescriptor $\widehat{x}_T$ is induced by some data-driven predictor $\widehat{c}_T$, and we define $\widehat{c}_T(\widehat{x}_T)$ as the in-sample risk of $\widehat{x}_T$, which is a function of the training samples alone and therefore accessible to the decision-maker. Note, however, that $\widehat{x}_T$ may be induced by many different data-driven predictors $\widehat{c}_T$ and that our definition of the in-sample risk depends on the particular choice of $\widehat{c}_T$. In particular, $\widehat{c}_T$ could be shifted by a constant without affecting $\widehat{x}_T$. Minimizing the in-sample risk instead of the out-of-sample risk is therefore nonsensical unless we restrict the choice of $\widehat{c}_T$. To this end, we define the *out-of-sample disappointment* of $\widehat{x}_T$ under $\mathbb{P}_\theta$ as the probability that the in-sample risk strictly exceeds the out-of-sample risk of $\widehat{x}_T$. This means that if the out-of-sample disappointment is high, then the predicted risk of $\widehat{x}_T$ is likely to underestimate its true risk, which lulls the decision-maker into a false sense of security and invariably leads to disappointment in out-of-sample tests. Note that the true out-of-sample disappointment is again *in*accessible to the decision-maker because it depends on the unknown parameter θ.

By construction, however, the out-of-sample disappointment decreases as $\widehat{c}_T$ increases. This reasoning motivates us to formulate an optimization problem that finds data-driven predictor-prescriptor pairs with an optimal trade-off between in-sample risk and out-of-sample disappointment. As each data-driven predictor encodes itself a surrogate optimization problem, any optimization problem over $\widehat{c}_T$ and $\widehat{x}_T$ constitutes indeed a *meta-optimization problem*, that is, an optimization problem over surrogate optimization problems.

To describe the envisaged meta-optimization problem more precisely, we define the asymptotic in-sample risk of a data-driven predictor $\widehat{c}_T$ and the corresponding data-driven prescriptor $\widehat{x}_T$ under $\mathbb{P}_\theta$ as

$$\lim_{T\to\infty} \mathbb{E}_\theta \left[\widehat{c}_T(\widehat{x}_T)\right],$$

and we define the asymptotic decay rate of the out-of-sample disappointment of $\widehat{c}_T$ and $\widehat{x}_T$ under $\mathbb{P}_\theta$ as

$$\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta[c(\widehat{x}_T, \theta) > \widehat{c}_T(\widehat{x}_T)].$$

Both of these statistical performance indicators, which are well-defined under mild regularity conditions, depend on the unknown parameter θ. We therefore intend to optimize them simultaneously for all $\theta \in \Theta$, which leads to a multi-objective optimization problem. This problem minimizes the asymptotic in-sample risk simultaneously for all $\theta \in \Theta$ under the condition that the asymptotic decay rate of out-of-sample disappointment is smaller than $r \geq 0$ for every $\theta \in \Theta$. Multi-objective optimization problems typically only admit Pareto optimal solutions, that is, feasible solutions that are not Pareto dominated by any other feasible solution. Maybe surprisingly, however, we will see that the proposed meta-optimization problem sometimes admits Pareto dominant solutions, that is, feasible solutions that Pareto dominate all other feasible solutions. Moreover, if they exist, these solutions are available in closed-form and admit an intuitive interpretation.

Data-driven predictors and prescriptors are essentially arbitrary functions of the available T training samples. Processing or even storing such functions might easily become impractical for large T. A natural approach to simplify the proposed meta-optimization problem is to compress the observation history of the training samples into a statistic $\widehat{S}_T$ of constant dimension and to restrict attention to *compressed* data-driven predictors and prescriptors that depend on the training samples only indirectly through $\widehat{S}_T$. The resulting restricted meta-optimization problem is often easier to handle than the original meta-optimization problem.

We are now ready to summarize the main contributions of this work.

1. We prove that if the statistic $\widehat{S}_T$ satisfies a large deviation principle, then the restricted meta-optimization problem over all compressed data-driven predictors and prescriptors admits a Pareto dominant solution. Moreover, the optimal data-driven predictor evaluates, for every fixed decision x, the worst case of the risk $c(x, \theta)$ across all θ in a ball of radius r around $\widehat{S}_T$, where the discrepancy between $\widehat{S}_T$ and θ is measured via the rate function of the large deviation principle at hand. The surrogate optimization problem induced by this optimal predictor thus represents a distributionally robust optimization problem, and the radius r of the underlying ambiguity set coincides with the upper bound on the decay rate of the out-of-sample disappointment enforced by the restricted meta-optimization problem.

2. We demonstrate that the restricted meta-optimization problem and its Pareto dominant solution are invariant under homeomorphic coordinate transformations of the statistic $\widehat{S}_T$ as well as the distribution family $\{\mathbb{P}_\theta : \theta \in \Theta\}$. This implies that the chosen parametrizations, which are invariably somewhat arbitrary, have no impact on how the optimal data-driven prescriptor maps the raw data to decisions.

3. We prove that if the set $\{\mathbb{P}_\theta : \theta \in \Theta\}$ represents an exponential family with sufficient statistic $\widehat{S}_T$ and if $\widehat{S}_T$ satisfies a large deviation principle, then compressing the training samples into $\widehat{S}_T$ destroys no useful statistical information, and the original meta-optimization problem is indeed equivalent to the restricted meta-optimization problem. Thus, the original meta-optimization also admits a Pareto dominant solution that has a distributionally robust interpretation. This result establishes a separation principle that enables a decoupling of estimation and optimization, and it can be viewed as a non-trivial extension of the celebrated Rao-Blackwell theorem [16, 44] to data-driven decision problems.

4. We explicitly derive the optimal data-driven predictors corresponding to different data-generating stochastic processes including finite-state i.i.d. processes, finite-state Markov chains, two different classes of autoregressive processes as well as i.i.d. processes with parametric distribution functions.

The results of this paper suggest that the optimal method for mapping data to decisions is, in a rigorous statistical sense, to solve a distributionally robust optimization model. As we will see, this conclusion persists irrespective of whether the original stochastic optimization problem is convex or not, and it persists 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) surrogate optimization problem. This paper therefore generalizes the preliminary results for i.i.d. training samples on a finite state space reported in [55]. In fact, we will demonstrate through a running example that these results emerge as a special case of a significantly more general theory of data-driven decision-making. In addition, we will argue that the Pareto dominance properties of data-driven predictors are reminiscent of the Bahadur efficiency properties of classical statistical estimators [5].

The existing literature on data-driven stochastic optimization is vast. Arguably the most popular approach is the sample average approximation (SAA), which replaces the unknown true probability distribution of the uncertain parameters in the problem's objective function with the empirical distribution corresponding to the training samples. The asymptotic properties of the resulting SAA problem are well understood if the training samples are i.i.d.; see, *e.g.*, [32, 33, 48–51, 53]. In particular, the optimal value of the SAA problem is known to be strongly consistent and asymptotically normal [53, Sections 5.1.1–5.1.2], which facilitates a rigorous probabilistic error analysis that yields increasingly accurate confidence bounds as the sample size grows. If the sample size is small relative to the number of uncertain problem parameters, however, then the optimal solution of the SAA problem tends to display an excellent in-sample performance alongside a poor out-of-sample performance. This phenomenon can be interpreted as an overfitting effect, which is sometimes referred to as the optimization bias [52] or the optimizer's curse [54]. Data-driven distributionally robust optimization (DRO) has been widely championed as an effective means to combat this phenomenon. It seeks a decision that minimizes the worst-case risk with respect to all probability distributions in an ambiguity set constructed from the training samples. If one can guarantee that the unknown true distribution belongs to the ambiguity set with high probability, then the optimal value of the DRO problem provides an upper confidence bound on the out-of-sample performance of its optimal solution. Out-of-sample guarantees of this kind were first obtained for a Chebyshev ambiguity set that contains all probability distributions whose mean vectors and covariance matrices are close to the empirical mean and the empirical covariance matrix [21]. As the sample size grows, the moment estimates become increasingly accurate, in which case this Chebyshev ambiguity set reduces to the family of all probability distributions that share the same first- and second-order moments as the unknown true distribution. Since this family contains distributions with strikingly different shapes (and not only the true distribution), the optimal value of the corresponding DRO problem fails to be asymptotically consistent. Pertinent out-of-sample guarantees have also been established for ambiguity sets containing all probability distributions that are close to the empirical distribution with respect to some information divergence [9], for ambiguity sets containing all distributions that pass a statistical goodness-of-fit test against the observed training data [13] or for ambiguity sets containing all distributions that are close to the empirical distribution with respect to some Wasserstein distance [36, 42]. If these ambiguity sets are shrunk sufficiently slowly, then the corresponding DRO problems can be rendered asymptotically consistent without compromising their out-of-sample guarantees. By leveraging ideas from empirical likelihood theory, it has recently been shown that significantly tighter out-of-sample bounds can be obtained by relaxing the requirement that the ambiguity set must contain the unknown true distribution with high probability [25,37].

In view of the many ambiguity sets permeating the extant literature, it is natural to wonder which ones of them offer optimal statistical guarantees. For example, given an ambiguity set with a prescribed 'shape' determined by the choice of a specific information divergence or probability metric, it is natural to seek the smallest radius for which the corresponding DRO problem offers an upper confidence bound on the original stochastic optimization problem with a desired significance level. The scaling of the optimal radius with respect to the sample size T is indeed known both for divergence ambiguity sets [25, 37] as well as for Wasserstein ambiguity sets [17, 27]. A more challenging task than merely tuning the size would be to tune the size and the shape of the ambiguity set simultaneously. The study of optimal ambiguity sets was pioneered in [28], where the smallest convex ambiguity sets that satisfy a Bayesian robustness guarantee are identified under certain convexity assumptions about the stochastic optimization problem. In addition, ambiguity sets that offer optimal statistical guarantees in view of the central limit theorem are investigated in [37]. In this case the optimal ambiguity sets constitute carefully scaled Burg-entropy divergence balls centered at the empirical distribution. Recently it has been shown that if the training samples are i.i.d., then among *all* data-driven decisions whose out-of-sample risk is dominated by their in-sample risk with high confidence, the decision with the lowest in-sample risk can be computed by solving a DRO problem with a relative entropy ambiguity set centered at the empirical distribution [55]. This result indicates that, at least in simple stylized settings, data-driven DRO provides an optimal approach for mapping data to decisions. In this paper we endeavor to extend the main results of [55] to more general stochastic optimization problems, parametric distribution families and data-generating stochastic processes.

All statistical guarantees reviewed so far critically rely on the assumption that the training samples are i.i.d. Moreover, the literature on data-driven DRO with non-i.i.d. data is remarkably scarce. We are only aware of three recent papers addressing this topic. First, if the training samples are generated by a fast mixing process, then asymptotic confidence intervals for the optimal value of a stochastic optimization problem can be obtained by solving DRO problems with divergence ambiguity sets [25]. However, the resulting confidence bounds depend on the unknown probability distribution and are therefore primarily of theoretical interest. In addition, data-driven DRO models with Wasserstein ambiguity sets constructed from training samples following an autoregressive process are proposed in [24]. While these ambiguity sets offer rigorous out-of-sample guarantees, their shapes are chosen ad hoc. Finally, distributionally robust Markov decision processes with Wasserstein ambiguity sets for the uncertain transition kernel are developed in [23]. In this case the training dataset consists of multiple i.i.d. trajectories of serially correlated states, which may be difficult to acquire in practice. In contrast to all of these approaches, we devise here a principled approach to generate statistically optimal data-driven decisions based on a *single* trajectory of the data-generating process.

This paper is organized as follows. Section 2 formally introduces our framework for data-driven decision-making and constructs the meta-optimization problems that will be used for identifying optimal data-driven predictors and prescriptors. Sections 3 and 4 establish sufficient conditions under which the restricted and original meta-optimization problems admit Pareto dominant solutions, respectively, and Section 5 showcases practically relevant examples in which these conditions are satisfied. Section 6 concludes.

Notation. A multi-objective optimization problem $\min_{x\in\mathcal{X}}\{f_\alpha(x)\}_{\alpha\in\mathcal{A}}$ is determined by its feasible set $\mathcal{X}$ and its objective functions $f_\alpha : \mathcal{X} \to \mathbb{R}$ indexed by $\alpha \in \mathcal{A}$. A strong solution is a feasible solution $x^\star \in \mathcal{X}$ that Pareto dominates every other feasible solution in the sense that $f_\alpha(x^\star) \leq f_\alpha(x)$ for all $x \in \mathcal{X}$ and $\alpha \in \mathcal{A}$. A weak solution is a feasible solution $x^\star \in \mathcal{X}$ that is not Pareto dominated by any other feasible solution in the sense that there exists no $x \in \mathcal{X}$ such that $f_\alpha(x) \leq f_\alpha(x^\star)$ for all $\alpha \in \mathcal{A}$. A function $f : \mathcal{X} \to \mathcal{Y}$ from $\mathcal{X} \subseteq \mathbb{R}^n$ to $\mathcal{Y} \subseteq \mathbb{R}^m$ is called quasi-continuous at $x \in \mathcal{X}$ if for every neighbourhood $\mathcal{U} \subseteq \mathcal{X}$ of x and every neighborhood $\mathcal{V} \subseteq \mathcal{Y}$ of $f(x)$ there exists a non-empty open set $\mathcal{W} \subseteq \mathcal{U}$ with $f(x') \in \mathcal{V}$ for all $x' \in \mathcal{W}$. Note that $\mathcal{W}$ may not contain x. The n-dimensional probability simplex is denoted by $\Delta_n = \{x \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$. For any logical expression $\mathcal{E}$, the indicator function $\mathbf{1}_\mathcal{E}$ evaluates to 1 if $\mathcal{E}$ is true and to 0 otherwise, and for any $A, B \in \mathbb{R}^{n\times m}$ the trace inner product is denoted by $\langle A, B\rangle = \mathsf{tr}(A^\top B)$.

2 Data-driven optimization

Throughout this paper we assume that all random objects are defined on the same abstract probability space $(\Omega, \mathcal{F}, \mathbb{P}_\star)$, and we study a general stochastic optimization problem of the form

$$\min_{x\in X} \; c(x, \mathbb{P}_\star), \tag{2.1}$$

where the goal is to find a decision $x \in X \subseteq \mathbb{R}^n$ that minimizes a real-valued objective or 'cost' function $c(x, \mathbb{P}_\star)$ depending on the probability measure $\mathbb{P}_\star$. We assume throughout the paper that X is compact and that $c(x, \mathbb{P}_\star)$ is continuous in x. These assumptions guarantee that the minimum in (2.1) is attained.

Example 2.1 (Objective functions)**.** *A popular objective arising in operations research and statistics is to minimize the expected value of a loss function $\ell(x, \xi)$ that depends both on the decision x and an exogenous random vector $\xi \in \mathbb{R}^m$. Denoting the expectation operator with respect to $\mathbb{P}_\star$ by $\mathbb{E}_{\mathbb{P}_\star}[\cdot]$, we thus set*

$$c(x, \mathbb{P}_\star) = \mathbb{E}_{\mathbb{P}_\star}[\ell(x, \xi)]. \tag{2.2a}$$

In risk averse optimization [53, Chapter 6] the expectation is replaced with a risk measure $\varrho_{\mathbb{P}_\star}[\cdot]$. We thus set

$$c(x, \mathbb{P}_\star) = \varrho_{\mathbb{P}_\star}[\ell(x, \xi)]. \tag{2.2b}$$

Examples of risk measures include the variance, the value-at-risk or the conditional value-at-risk of the loss as well as their convex combinations with the expected loss. Decision makers sometimes have access to contextual covariates, that is, observable random variables that are correlated with the unobservable random variables impacting the loss function. In such situations it is beneficial to solve a conditional stochastic optimization problem that minimizes the conditional expectation of the loss given the contextual covariates [6, 14]. If the matrix $C \in \mathbb{R}^{m\times m_C}$ filters out m_C observable covariates from ξ and if these covariates are known to fall within a Borel set $B \subseteq \mathbb{R}^{m_C}$ (note that B could represent a singleton), then we set the objective function to

$$c(x, \mathbb{P}_\star) = \mathbb{E}_{\mathbb{P}_\star}[\ell(x, \xi)|C\xi \in B]. \tag{2.2c}$$

Contextual information may include weather forecasts, Twitter feeds or Google Trends data. Stochastic control, as a last example, aims to guide a dynamical system to a desirable state, assuming that the system's state obeys a recursion $s_{t+1} = f(s_t, u_t, \xi_t)$ that depends on some control inputs u_t and exogenous random disturbances ξ_t at time $t \in \mathbb{N}$. If the inputs are set to $u_t = \pi_x(s_t)$ for some control policy π_x parameterized by $x \in X$ and if $\ell(u_t, s_t)$ represents the cost at time t, then one may minimize the long-run average cost

$$c(x, \mathbb{P}_\star) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathbb{P}_\star}[\ell(\pi_x(s_t), s_t)]. \tag{2.2d}$$

Note that x impacts the objective function (2.2d) *both directly through the policy π_x as well as indirectly through the states s_t, $t \in \mathbb{N}$, which are defined by a recursion that depends on x. We also emphasize that some mild technical assumptions are needed for the objective functions* (2.2a)–(2.2d) *to be well-defined. But the above examples show that the abstract stochastic optimization problem* (2.1) *is remarkably general.*

When reasoning about the stochastic optimization problem (2.1), it is expedient to distinguish the *prediction problem*, which aims to evaluate the cost $c(x, \mathbb{P}_\star)$ associated with a fixed decision x, from the *prescription problem*, which aims to find a decision $x^\star$ that minimizes the cost $c(x, \mathbb{P}_\star)$ across all $x \in X$. We emphasize that any procedure for solving the prescription problem invariably necessitates a procedure for solving the corresponding prediction problem. As the prediction problem is reminiscent of an uncertainty quantification problem [38], however, it is of interest in its own right. Unfortunately, already the prediction problem poses two formidable challenges. On the one hand, the probability measure $\mathbb{P}_\star$, which is needed to evaluate the objective function, is usually unobservable. On the other hand, even if one had access to $\mathbb{P}_\star$, computing the objective function $c(x, \mathbb{P}_\star)$ for a fixed decision x might be difficult. For example, evaluating the expectation in (2.2a) is #P-hard even if $\ell(x, \xi)$ is defined as the non-negative part of an affine function of ξ and if ξ is uniformly distributed on the standard hypercube in $\mathbb{R}^m$ under the probability measure $\mathbb{P}_\star$ [29, Corollary 1].

In the following we develop a systematic approach for addressing the prediction and prescription problems when $\mathbb{P}_\star$ is only indirectly observable through finitely many training samples. We endeavor to keep the proposed framework as general as possible. In particular, we will forgo any restrictive independence assumptions and explicitly account for the possibility that the training data are serially dependent.

2.1 The data-driven newsvendor problem

We first exemplify several popular approaches to data-driven decision-making in the context of the classical newsvendor problem, which captures the fundamental dilemma faced by the seller of a perishable good. The textbook example of such a seller is a newsvendor who sells a daily newspaper that becomes worthless at the end of the day. At the beginning of each day, the newsvendor orders $x \in X$ newspapers from the publisher at the wholesale price $k \geq 0$, where $X = \{1, \ldots, d\}$. Then, the uncertain demand $\xi \in \Xi$ is revealed, where $\Xi = X$, and the newsvendor sells newspapers at the retail price $p > k$ until either the inventory or the demand is exhausted. The number of newspapers sold is thus given by $\min\{x, \xi\}$, and the total cost amounts to $\ell(x, \xi) = kx - p\min\{x, \xi\}$. If the probability measure $\mathbb{P}_\star$ governing the demand is known, then the problem of minimizing the expected cost can be formulated as a stochastic optimization problem of the form (2.1) with objective function $\mathbb{E}_{\mathbb{P}_\star}[\ell(x, \xi)] = \sum_{i\in\Xi} \ell(x, i)\,(\theta_\star)_i$, where the probability vector $\theta_\star \in \Delta_d$ is defined through $(\theta_\star)_i = \mathbb{P}_\star[\xi = i]$ for all $i \in \Xi$. Note that $\theta_\star$ captures all information about $\mathbb{P}_\star$ that is needed to solve the newsvendor's decision problem. By slight abuse of notation, we may thus identify $\mathbb{P}_\star$ with $\theta_\star$

and use $c(x,\theta_\star)$ as a shorthand for the expected cost of any fixed order quantity $x \in X$. If the demands on different days are i.i.d., then the law of large numbers guarantees that $\min_{x\in X} c(x,\theta_\star)$ represents the minimum cost attainable by the newsvendor on average in the long run.

In reality, the probability measure $\mathbb{P}_\star$ and the probability vector $\theta_\star$ are unobservable and must be estimated from historical demand realizations $\xi_t \in \Xi$, $t = 1,\dots,T$, which we refer to as training samples. We assume here for simplicity that the training samples are mutually independent, but the general methods developed in this paper do not rely on this assumption. Given a batch of only T training samples, the newsvendor now seeks to answer three intertwined questions: (i) What is the expected cost of a given ordering decision? (ii) How many newspapers should be ordered so as to minimize the expected cost? (iii) What is the probability that the unknown true expected cost of the chosen ordering decision exceeds the estimated cost?

In the following we designate all estimators (*i.e.*, all functions of the data) with a superscript '$\widehat{\ }$' as well as a subscript 'T' indicating the size of the underlying dataset. For example, we use $\widehat{c}_T(x)$ to denote an estimator of the expected cost $c(x,\theta_\star)$ constructed from T demand samples, where x is any feasible ordering decision. Similarly, we use $\widehat{x}_T$ to denote an estimator for the optimal ordering decision constructed from T demand samples. Below we assume that $\widehat{x}_T \in \arg\min_{x\in X} \widehat{c}_T(x)$, that is, we assume that any estimator for the optimal ordering decision is induced by some estimator for the expected cost function. Note that this assumption can be imposed without loss of generality. Indeed, any estimator $\widehat{x}_T$ for the optimal decision can be expressed as a minimizer of a cost function estimator; for example, we may set $\widehat{c}_T(x) = (x - \widehat{x}_T)^2$.

Questions (i) and (ii) above address the construction of the estimators $\widehat{c}(x)$ and $\widehat{x}_T$, respectively, while question (iii) asks for the probability of the event $c(\widehat{x}_T,\theta_\star) > \widehat{c}_T(\widehat{x}_T)$. In this event the true (out-of-sample) expected cost of the chosen decision $\widehat{x}_T$ exceeds the estimated (in-sample) expected cost, which might lead to a budget overrun and force the newsvendor into financial distress. In the following we refer to the probability of this event (with respect to the sampling of the training dataset) as the *out-of-sample disappointment*. Note that in the event $c(\widehat{x}_T,\theta_\star) < \widehat{c}_T(\widehat{x}_T)$ there is also a discrepancy between the estimated budget and the true expected cost. However, in this event the newsvendor faces no severe financial repercussions.

There are countless possibilities to construct cost estimators $\widehat{c}_T(x)$ and the corresponding decision estimators $\widehat{x}_T$ from the training data. Different estimators may offer different statistical guarantees and display different computational properties. However, the existing literature offers little guidance on how to choose among these many estimators. Moreover, there could exist yet undiscovered estimators that dominate all known estimators in terms of some meaningful statistical criteria. In the following we will compare different estimators in terms of the exponential decay rate of their out-of-sample disappointment, which is defined as

$$\lim_{T\to\infty} -\frac{1}{T}\log \mathbb{P}_\star[c(\widehat{x}_T,\theta_\star) > \widehat{c}_T(\widehat{x}_T)],$$

and in terms of their asymptotic in-sample cost, which is defined as $\lim_{T\to\infty} \mathbb{E}_{\mathbb{P}_\star}[\widehat{c}_T(\widehat{x}_T)]$. We will see later that these quantities are well-defined for a wide range of estimators. In the remainder we thus view a pair of cost and decision estimators as 'desirable' if the asymptotic in-sample cost is low (*i.e.*, the expected cost of $\widehat{x}_T$ is predicted to be low) and if the decay rate of the out-of-sample disappointment is high (*i.e.*, the probability that the true expected cost of $\widehat{x}_T$ exceeds the predicted cost decays quickly as T grows).

Arguably one of the simplest conceivable cost estimators is the empirical cost $\widehat{c}_T(x) = \frac{1}{T}\sum_{t=1}^T \ell(x,\xi_t)$. Thus, we have $\widehat{c}_T(x) = c(x,\widehat{S}_T)$, where $c(x,\theta) = \sum_{i\in\Xi} \ell(x,i)\theta_i$ represents the expected cost of the decision x when the demand uncertainty is described by the probability vector $\theta \in \Delta_d$, and the statistic $\widehat{S}_T \in \Delta_d$ stands for the empirical probability vector, whose i^{th} component $(\widehat{S}_T)_i = \frac{1}{T}\sum_{t=1}^T 1_{\xi_t = i}$ records the empirical frequency of the i^{th} demand scenario for each $i \in \Xi$. Using the central limit theorem, one can show that the expected in-sample cost of the empirical cost estimator and its induced decision estimator converges to the true optimum $\min_{x\in X} c(x, \theta_\star)$ and that the out-of-sample disappointment converges to 50% as T grows. Thus, the decay rate of the out-of-sample disappointment vanishes completely; see also [55, Example 2].

A naïve approach to force the out-of-sample disappointment to decay would be to add a constant positive penalty ε to the empirical cost estimator, thereby increasing its asymptotic in-sample cost and thus introducing a conservative bias. This reasoning suggests that the in-sample cost and the out-of-sample disappointment stand in direct competition. In order to provide a better intuition for the trade-off between these statistical performance criteria, we further investigate three distributionally robust cost estimators of the form $\widehat{c}_T(x) = \max_{\theta\in\widehat{\Theta}_T} c(x, \theta)$, which evaluate the worst-case expected cost of the decision x with respect to all probability vectors from within some ambiguity set $\widehat{\Theta}_T \subseteq \Delta_d$ constructed from the training data.

Traditionally, distributionally robust optimization has mostly studied moment ambiguity sets such as

$$\widehat{\Theta}_T = \left\{\theta \in \Delta_d : \left|\sum_{i\in\Xi} i^j \theta_i - \sum_{i\in\Xi} i^j (\widehat{S}_T)_i\right| \le \varepsilon \ \forall j = 1, \ldots, J\right\}.$$

All probability vectors in this ambiguity set share, to within an absolute error tolerance $\varepsilon \ge 0$, the same moments of all orders up to J as the empirical probability vector $\widehat{S}_T$. In the subsequent numerical experiments we set $J = 4$. The tolerance ε is usually tuned to ensure that $\widehat{\Theta}_T$ contains the unknown data-generating probability vector $\theta_\star$ with a prescribed high confidence; see [21, Section 3] for the first results of this kind. The recent literature has witnessed an increasing interest in Wasserstein ambiguity sets of the form

$$\widehat{\Theta}_T = \left\{\theta \in \Delta_d : \mathsf{d}_{\mathsf{W}}(\theta, \widehat{S}_T) \le \varepsilon\right\},$$

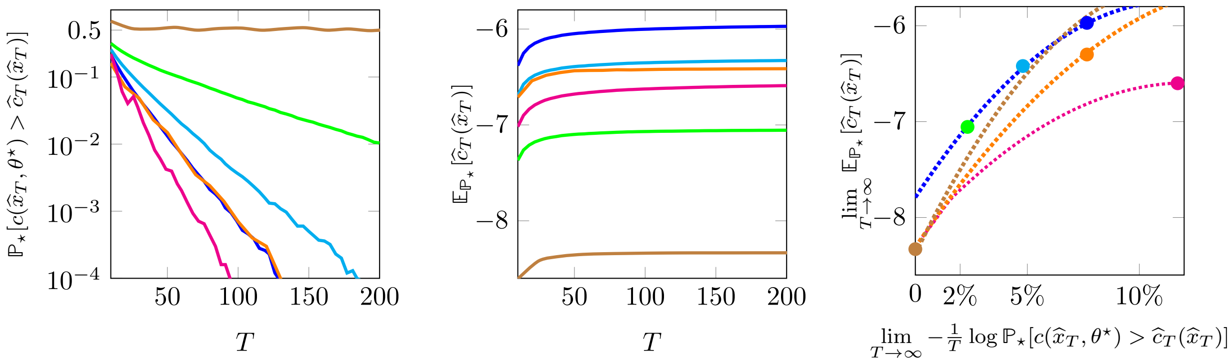
where $\mathsf{d}_{\mathsf{W}}(\theta, \widehat{S}_T)$ denotes the first Wasserstein distance between θ and $\widehat{S}_T$ [30]. This ambiguity set can be viewed as a Wasserstein ball of radius $\varepsilon \ge 0$ around $\widehat{S}_T$ in Δ_d. Unlike the moment ambiguity set, the Wasserstein ambiguity set shrinks to the singleton that contains merely the empirical probability vector if we set $\varepsilon = 0$. In general, ε can again be tuned to ensure that $\theta_\star \in \widehat{\Theta}_T$ with any prescribed high confidence [42, Section 3]. Finally, we also study relative entropy ambiguity sets of the form

$$\widehat{\Theta}_T = \left\{\theta \in \Delta_d : D(\widehat{S}_T \| \theta) \le \varepsilon\right\},$$

where $D(\widehat{S}_T\|\theta)$ stands for the relative entropy (or Kullback-Leibler divergence) of $\widehat{S}_T$ with respect to θ. This ambiguity set also shrinks to a singleton for $\varepsilon = 0$, and ε can again be tuned to guarantee that $\widehat{\Theta}_T$ covers $\theta_\star$ with a prescribed probability [9, Section 3]. In contrast to most of the existing literature on distributionally robust optimization, here we are *not* concerned about whether or not the ambiguity set covers $\theta_\star$. Instead, we view any distributionally robust optimization model simply as a vehicle for transforming data to decisions, and we are merely interested in the statistical properties of the resulting cost and decision estimators.

Figure 1 visualizes the out-of-sample disappointment and the expected in-sample cost as well as the trade-off between the asymptotic in-sample cost and the decay rate of the out-of-sample disappointment for different estimators. Figure 1a shows that, as a function of T, the out-of-sample disappointment always traces out an almost perfect straight line on a logarithmic scale. This observation suggests that the out-of-sample disappointment decays exponentially and is therefore faithfully represented by its decay rate.

The solid lines in Figures 1a and 1b correspond to the empirical cost estimator (light brown) and to different distributionally robust cost estimators with a moment ambiguity set (green: $\varepsilon = 0.05$, light blue: $\varepsilon = 0.15$,



(a) Out-of-sample disappointment versus training sample size T

(b) In-sample cost versus training sample size T

(c) Asymptotic in-sample cost versus decay rate of out-of-sample disappointment

Figure 1: Statistical properties of different cost and decision estimators for a data-driven newsvendor problem with ordering cost $k = 5$ and retail price $p = 7$, where the demand ξ follows a shifted binomial distribution with 10 trials, success probability 0.5 and shift 1. All probabilities and expectations involving random training data are evaluated empirically using 10^4 independent training sets.

dark blue: $\varepsilon = 0.20$), a Wasserstein ambiguity set (orange: $\varepsilon = 0.275$) and a relative entropy ambiguity set (magenta: $\varepsilon = 0.05$). As expected, the empirical cost estimator is the most optimistic one in the sense that it displays the lowest in-sample cost, but its out-of-sample disappointment fails to decay. Any distributionally robust cost estimator becomes increasingly pessimistic as the size parameter ε of the underlying ambiguity set increases. The dashed lines in Figure 1c visualize the trade-off between the asymptotic in-sample cost and the decay rate of the out-of-sample disappointment for the naïve penalized empirical cost estimator $\widehat{c}(x) = c(x, \widehat{S}_T) + \varepsilon$ (light brown) and for the distributionally robust cost estimators with a moment ambiguity set (dark blue), a Wasserstein ambiguity set (orange) and a relative entropy ambiguity set (magenta) as ε is swept. The six dots in Figure 1c correspond to the six estimators investigated in Figures 1a and 1b. As expected, the dashed lines corresponding to the distributionally robust cost estimators with a Wasserstein and a relative entropy ambiguity set intersect because both of these estimators reduce to the empirical cost estimator for $\varepsilon = 0$. Maybe surprisingly, the distributionally robust cost estimators associated with the relative entropy ambiguity set dominate those associated with the Wasserstein ambiguity set and even more so those associated with the moment ambiguity set, that is, their asymptotic in-sample cost is lowest for any fixed decay rate of the out-of-sample disappointment. They also dominate the penalized empirical cost estimators. It is now natural to ask whether there exists a globally *least conservative* cost estimator whose asymptotic in-sample risk is minimal across *all* conceivable cost estimators (not necessarily only distributionally robust ones) with the same decay rate of the out-of-sample disappointment. In the remainder of the paper we endeavor to answer this fundamental question under significantly more general conditions.

2.2 Data-driven predictors and prescriptors

We now return to the general stochastic optimization problem (2.1), and we assume that the unknown probability measure $\mathbb{P}_\star$ must be learned from a finite sample path of a stochastic process $\{\xi_t\}_{t\in\mathbb{N}}$ with state space $\Xi \subseteq \mathbb{R}^m$. Like any random object, this data-generating stochastic process is defined on the measurable space $(\Omega, \mathcal{F})$. From now on we assume that even though the probability measure $\mathbb{P}_\star$ is unknown, it belongs to a known finitely parametrized ambiguity set. This premise is formalized in the following assumption.

Assumption 2.1 (Finitely parametrized ambiguity set). *The probability measure $\mathbb{P}_\star$ belongs to a finitely parametrized ambiguity set $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, where Θ is a relatively open convex subset of $\mathbb{R}^d$, and $\mathbb{P}_\theta$ is a probability measure on $(\Omega, \mathcal{F})$ for every $\theta \in \Theta$.*

As each $\theta \in \Theta$ encodes a unique probabilistic model $\mathbb{P}_\theta$, for ease of terminology, we will henceforth refer to θ as a *model* and to Θ as the *model space*. The ambiguity set $\mathcal{P}$ is meant to capture all structural information on $\mathbb{P}_\star$ that is available before observing any statistical data. This justifies our assumption that $\mathcal{P}$ is known to contain the probability measure $\mathbb{P}_\star$ with certainty (and not only with high confidence). Assumption 2.1 thus implies that there exists a model $\theta_\star \in \Theta$ with $\mathbb{P}_{\theta_\star} = \mathbb{P}_\star$.

To provide some intuition for the abstract concepts introduced in this paper, we use the class of i.i.d. stochastic processes with a finite state space as a running example. This example will further show that the approach to data-driven decision-making developed in [55] emerges as a simple special case of a considerably more general framework. Several alternative data generation processes will be discussed in Section 5.

Example 2.2 (Ambiguity set for finite state i.i.d. processes). *Assume that $\Xi = \{1, \ldots, d\}$, the random variables ξ_t are serially independent under $\mathbb{P}_\star$ and $\mathbb{P}_\star[\xi_t = i] = (\theta_\star)_i > 0$ for all $i \in \Xi$ and $t \in \mathbb{N}$. The vector $\theta_\star$ thus encodes the unknown probability mass function of ξ_t, which is independent of t. These assumptions imply that $\mathbb{P}_\star$ belongs to an ambiguity set of the form $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, where $\Theta = \{\theta \in \mathbb{R}^d_{++} : \sum_{i=1}^d \theta_i = 1\}$ is the positive probability simplex, and each θ encodes a probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ satisfying*

$$\mathbb{P}_\theta[\xi_t = i_t \ \forall t = 1, \ldots, T] = \prod_{t=1}^{T} \theta_{i_t} \quad \forall i_t \in \Xi, \ t = 1, \ldots, T, \ T \in \mathbb{N}.$$

We now embed the original stochastic optimization problem (2.1) into a family of problems corresponding to the probability measures $\mathbb{P}_\theta$, $\theta \in \Theta$. Therefore, by slightly abusing notation with the goal to avoid clutter, we henceforth parametrize the objective function of problem (2.1) by θ instead of $\mathbb{P}_\theta$.

Definition 2.1 (Model-based predictors and prescriptors). *For any fixed model $\theta \in \Theta$, we define the model-based predictor $c(x, \theta)$ as the objective function of problem (2.1) when $\mathbb{P}_\star$ is replaced with $\mathbb{P}_\theta$ and the corresponding model-based prescriptor $x^\star(\theta) \in \arg\min_{x \in X} c(x, \theta)$ as a decision that minimizes $c(x, \theta)$ over $x \in X$.*

The stochastic program (2.1) can now be identified with the *prescription problem* of computing $x^\star(\theta^\star)$. Similarly, the evaluation of the objective function of a given decision $x \in X$ in (2.1) can be identified with the *prediction problem* of computing $c(x, \theta^\star)$. In the remainder we impose the following regularity condition.

Assumption 2.2 (Uniform continuity and boundedness of the model-based predictor). *The model-based predictor $c(x, \theta)$ is uniformly continuous and bounded on $X \times \Theta$.*

Note that if $c(x, \theta)$ is uniformly continuous and bounded on $X \times \Theta$, then it admits a unique uniformly continuous and bounded extension to $X \times \mathrm{cl}\, \Theta$ [1, Theorem 5.15]. By slight abuse of notation, we will denote this extension by $c(x, \theta)$, too. Assumption 2.2 is trivially satisfied by the newsvendor problem of Section 2.1. As neither the model-based predictor $c(x, \theta_\star)$ nor the model-based prescriptor $x^\star(\theta_\star)$ can be evaluated for the unknown true model $\theta_\star$, we will now approximate them by functions of the available data. In order to formally define data-driven predictors and prescriptors, we denote by $\xi_{[T]} = (\xi_1, \ldots, \xi_T)$ the history of the data-generating process up to time T, and we let $\mathcal{F}_T \subseteq \mathcal{F}$ be the σ-algebra generated by $\xi_{[T]}$ for any $T \in \mathbb{N}$. We also use $\mathbb{E}_\theta[\cdot]$ to denote the expectation operator with respect to $\mathbb{P}_\theta$ for any model $\theta \in \Theta$.

Definition 2.2 (Data-driven predictors). *A decision-dependent stochastic process $\widehat{c} = \{\widehat{c}_T(x)\}_{T\in\mathbb{N},\, x\in X}$ valued in $\mathbb{R}$ is called a data-driven predictor if it satisfies the following conditions.*

(i) ***Continuity in the decisions.*** *The random variable $\widehat{c}_T(x)$ is continuous in $x \in X$ for all $T \in \mathbb{N}$.*

(ii) ***Non-anticipativity.*** *The process $\{\widehat{c}_T(x)\}_{T\in\mathbb{N}}$ is adapted to the filtration $\{\mathcal{F}_T\}_{T\in\mathbb{N}}$ for every $x \in X$.*

(iii) ***Uniform integrability.*** *There exists a non-negative random variable $\bar{c}$ such that $\mathbb{E}_\theta[\bar{c}] < \infty$ for all $\theta \in \Theta$ and $|\widehat{c}_T(x)| \leq \bar{c}$ $\mathbb{P}_\theta$-almost surely for all $T \in \mathbb{N}$, $x \in X$ and $\theta \in \Theta$.*

(iv) ***Convergence of objective.*** *There exists a deterministic Borel-measurable function $c_\infty : X \times \Theta \to \mathbb{R}$ such that, as T grows, $\widehat{c}_T(x)$ converges in probability under $\mathbb{P}_\theta$ to $c_\infty(x, \theta)$ for every $x \in X$ and $\theta \in \Theta$.*

(v) ***Convergence of optimal value.*** *There exists a deterministic Borel-measurable function $v_\infty : \Theta \to \mathbb{R}$ such that, as T grows, $\min_{x\in X} \widehat{c}_T(x)$ converges in probability under $\mathbb{P}_\theta$ to $v_\infty(\theta)$ for every $\theta \in \Theta$.*

If we use data-driven predictors as accessible proxies for inaccessible model-based predictors, then it is reasonable to assume that they share all known properties of the model-based predictors. The continuity condition (i) in Definition 2.2 is thus a natural consequence of Assumption 2.2. In addition, as X is compact, this condition guarantees that the data-driven decision problem $\min_{x\in X} \widehat{c}_T(x)$ is sovlable for every $T \in \mathbb{N}$. The non-anticipativity condition (ii) implies via [3, Theorem 5.4.2] that for any $T \in \mathbb{N}$ there exists a Borel-measurable function $f_T : X \times \Xi^T \to \mathbb{R}$ with $\widehat{c}_T(x) = f_T(x, \xi_{[T]})$. This means that $\widehat{c}_T(x)$ may depend only on the history $\xi_{[T]}$ of the data-generating process observed up to time T. The uniform integrability condition (iii) is of technical nature and non-restrictive in all examples studied in this paper. The convergence condition (iv) implies that for any fixed $\theta \in \Theta$, the predictor $\widehat{c}_T(x)$ represents a consistent estimator for $c_\infty(x, \theta)$ if the data is generated under $\mathbb{P}_\theta$. Note that we explicitly allow for the possibility that $c_\infty(x, \theta) \neq c(x, \theta)$, that is, $\widehat{c}_T(x)$ may in fact be a *biased* estimator for the model-based predictor $c(x, \theta)$. Similarly, the convergence condition (v) implies that for any fixed $\theta \in \Theta$ the optimal value $\widehat{c}_T(\widehat{x}_T)$ of the data-driven optimization problem $\min_{x\in X} \widehat{c}_T(x)$ represents a consistent estimator for $v_\infty(\theta)$ if the data is generated under $\mathbb{P}_\theta$. Thus, it may be a *biased* estimator for the optimal value of the stochastic optimization problem $\min_{x\in X} c(x, \theta)$. We emphasize that condition (v) is not implied by conditions (i)–(iv). To see this, set $X = [0, 1]$, and define a degenerate decision-dependent process $\widehat{c}$ through $\widehat{c}_T(x) = \min\{1, T|x - 1/T|\}$ if T is odd and $\widehat{c}_T(x) = 1$ if T is even. Thus, there is no randomness in $\widehat{c}$, and conditions (i)–(iii) are trivially satisfied. Condition (iv) is also satisfied if we set $c_\infty(x, \theta) = 1$. As $\min_{x\in X} \widehat{c}_T(x)$ evaluates deterministically to 0 if T is odd and to 1 if T is even, however, condition (v) is violated. This example proves that condition (v) is not redundant.[1] From now on we denote the set of all data-driven predictors in the sense of Definition 2.2 by $\widehat{\mathcal{C}}$.

Definition 2.3 (Data-driven prescriptors). *A stochastic process $\widehat{x} = \{\widehat{x}_T\}_{T\in\mathbb{N}}$ valued in X is called a data-driven prescriptor if it satisfies the following conditions.*

(i) ***Non-anticipativity.*** *The process $\{\widehat{x}_T\}_{T\in\mathbb{N}}$ is adapted to the filtration $\{\mathcal{F}_T\}_{T\in\mathbb{N}}$.*

[1] Similarly, one can show that if all conditions of Definition 2.2 hold, then $v_\infty(\theta)$ may differ from $\min_{x\in X} c_\infty(x, \theta)$.

(ii) ***Compatibility with a data-driven predictor.*** *There exists a data-driven predictor $\widehat{c}$ that induces the data-driven prescriptor $\widehat{x}$ in the sense that $\widehat{x}_T \in \arg\min_{x \in X} \widehat{c}_T(x)$ for all $T \in \mathbb{N}$.*

The non-anticipativity condition (i) implies via [3, Theorem 5.4.2] that for any $T \in \mathbb{N}$ there exists a Borel-measurable function $g_T : \Xi^T \to \mathbb{R}$ with $\widehat{x}_T = g_T(\xi_{[T]})$. The compatibility condition (ii) requires that any data-driven prescriptor is a minimizer of some data-driven predictor. From now on we use $\widehat{\mathcal{X}}$ to denote the set of all data-driven predictor-prescriptor-pairs of the form $(\widehat{c}, \widehat{x})$, where $\widehat{x}$ is induced by $\widehat{c}$.

One can show that *any* data-driven predictor $\widehat{c}$ induces a (not necessarily unique) data-driven prescriptor $\widehat{x}$. The reason for this is that since X is compact and since $\widehat{c}_T(x)$ depends continuously on $x \in X$ and represents an $\mathcal{F}_T$-measurable random variable for every fixed x, there exists an $\mathcal{F}_T$-measurable random vector $\widehat{x}_T \in \arg\min_{x \in X} \widehat{c}_T(x)$ thanks to [46, Theorem 14.37]. Combining $\widehat{x}_T$ for all $T \in \mathbb{N}$ yields the desired prescriptor.

We emphasize that essentially any procedure for mapping the available data to an 'asymptotically deterministic' feasible decision defines a data-driven prescriptor. Indeed, if a stochastic process $\widehat{x} = \{\widehat{x}_T\}_{T \in \mathbb{N}}$ with state space X is adapted to the filtration $\{\mathcal{F}_T\}_{T \in \mathbb{N}}$ and converges in probability under $\mathbb{P}_\theta$ to some deterministic Borel-measurable function $x_\infty(\theta)$ for every $\theta \in \Theta$, then one readily verifies that the decision-dependent stochastic process $\widehat{c} = \{\widehat{c}_T(x)\}_{T \in \mathbb{N}}$ defined through $\widehat{c}_T(x) = \min\{1, \|x - \widehat{x}_T\|_2\}$ for all $T \in \mathbb{N}$ is a data-driven predictor in the sense of Definition 2.2 that induces $\widehat{x}$. This example shows that the notion of a data-driven prescriptor is very general. Moreover, Definition 2.3 does not even require $\widehat{x}_T$ to converge.

Example 2.3 (Empirical predictor for finite state i.i.d. processes)**.** *In the context of Example 2.2, assume that the model-based predictor represents an expected loss, that is, set $c(x, \theta) = \mathbb{E}_\theta[\ell(x, \xi)]$ for some loss function $\ell(x, \xi)$ that is continuous in x and bounded on $X \times \Xi$, and assume that the random variable ξ has the same distribution as the i.i.d. training samples $\{\xi_t\}_{t \in \mathbb{N}}$. Note that the newsvendor problem of Section 2.1 satisfies all of these assumptions, for instance. We now define the empirical predictor $\widehat{c}$ through*

$$\widehat{c}_T(x) = \frac{1}{T} \sum_{t=1}^{T} \ell(x, \xi_t) \quad \forall T \in \mathbb{N}.$$

Note that the empirical predictor simply evaluates the sample average of the loss across the observed dataset and represents a data-driven predictor in the sense of Definition 2.2. While the continuity condition (i) and the non-anticipativity assumption (ii) hold by construction, the boundedness condition (iii) holds because $\widehat{c}_T(x)$ is trivially bounded by the finite constant $\overline{c} = \max_{x \in X,\, \xi \in \Xi} |\ell(x, \xi)|$ for every $T \in \mathbb{N}$. The convergence condition (iv) follows by setting $c_\infty(x, \theta) = c(x, \theta)$ and observing that $\lim_{T \to \infty} |\widehat{c}_T(x) - c(x, \theta)| = 0$ $\mathbb{P}_\theta$-almost surely for all $\theta \in \Theta$ thanks to the strong law of large numbers. Note also that $c_\infty(x, \theta)$ is continuous by Assumption 2.2. As ℓ is continuous and X is compact, the uniform law of large numbers [43, Lemma 2.4] further guarantees that $\lim_{T \to \infty} \sup_{x \in X} \|\frac{1}{T} \sum_{t=1}^{T} \ell(x, \xi_t) - c(x, \theta)\| = 0$ $\mathbb{P}_\theta$-almost surely for all $\theta \in \Theta$. Therefore, the convergence condition (v) is satisfied if we set $v_\infty(\theta) = \min_{x \in X} c(x, \theta)$.

We will now investigate sequences of surrogate decision problems of the form $\min_{x \in X} \widehat{c}_T(x)$ indexed by $T \in \mathbb{N}$, where $\widehat{c}$ is a data-driven predictor. As the set $\widehat{\mathcal{C}}$ of admissible predictors is vast, there are endless possibilities to design such surrogate decision problems. An ideal design should have the following property for *any* model $\theta \in \Theta$: If the observable data is generated by $\mathbb{P}_\theta$, then the surrogate decision problem $\min_{x \in X} \widehat{c}_T(x)$, which must be constructed without knowledge of θ, should provide a 'good' approximation for the stochastic optimization problem $\min_{x \in X} c(x, \theta)$ corresponding to model θ. If such an ideal design can be found, it will provide—in particular—a 'good' approximation for the actual decision problem corresponding to the unknown true model $\theta_\star$. Intuitively, a data-driven predictor $\widehat{c}$ and the corresponding predictor $\widehat{x}$ provide a 'good' design if the data-driven objective function $\widehat{c}_T(x)$ is close to the function $c(x,\theta)$ for large T and if the data-driven decision $\widehat{x}_T$ is near-optimal in the decision problem $\min_{x\in X} c(x,\theta)$ for large T whenever the data is generated by $\mathbb{P}_\theta$. In the following we will formalize these intuitions.

The key idea is to find the best possible data-driven predictor $\widehat{c}$ by solving an optimization problem over $\widehat{\mathcal{C}}$ and to find the best possible data-driven prescriptor $\widehat{x}$ by solving an optimization problem over $\widehat{\mathcal{X}}$. As any predictor $\widehat{c}$ encodes a procedure for transforming data to surrogate optimization problems, an optimization problem over $\widehat{\mathcal{C}}$ can be viewed as an optimization problem over optimization problems. We will therefore refer to it as a *meta-optimization model.* As any data-driven prescriptor is induced by a data-driven predictor, an optimization problem over the set $\widehat{\mathcal{X}}$ of predictor-prescriptor pairs can also be viewed as a meta-optimization problem. In the special case when the data is generated by a simple i.i.d. processes, such meta-optimization problems were already studied in [55]. Here, we will show that these ideas have a much wider scope.

To formulate the desired meta-optimization problems, we first need to introduce some terminology. For any fixed model $\theta \in \Theta$ and data-driven predictor $\widehat{c}$, we will henceforth refer to $\widehat{c}_T(x)$ as the *in-sample risk* and to $c(x,\theta)$ as the *out-of-sample risk* of the decision $x \in X$. Specifically, if $\widehat{x}$ is a data-driven prescriptor induced by $\widehat{c}$, then $\widehat{c}_T(\widehat{x}_T)$ and $c(\widehat{x}_T,\theta)$ represent the in-sample and out-of-sample risk of $\widehat{x}_T$, respectively. We emphasize that the out-of-sample risk under the true model $\theta_\star$ is the actual quantity of interest as it represents the objective function value of a given candidate decision in the true stochastic optimization problem (2.2a). If the data-generating process is ergodic (which is the case for all examples studied Section 5), then the out-of-sample risk also coincides almost surely with the average cost incurred of the given candidate decision along an infinitely long sample path. Unfortunately, only the in-sample risk is observable at the time when the decision problem needs to be solved. Of course, the out-of-sample risk can in principle be computed for any model $\theta \in \Theta$. But the benefits of this capability remain limited as long as $\theta_\star$ is unknown.

The ideal meta-optimization problem over all data-driven prescriptors would be tailored to the length T of the available observation history and would minimize the out-of-sample risk $c(\widehat{x}_T,\theta_\star)$ of $\widehat{x}_T$ over all $(\widehat{c},\widehat{x}) \in \widehat{\mathcal{X}}$. As the true model $\theta_\star$ is unknown, however, such an approach would only be successful if there existed a Pareto dominant prescriptor that minimizes the out-of-sample risk of $\widehat{x}_T$ simultaneously for all models $\theta \in \Theta$ (and thus in particular for $\theta_\star$). Unfortunately, finding such a Pareto dominant prescriptor seems too ambitious and is probably impossible. This prompts us to work with an alternative notion of optimality. The key idea is to minimize the in-sample risk subject to a constraint that forces the out-of-sample risk to be smaller than or equal to the in-sample risk. As both the in-sample and the out-of-sample risk are random objects, we impose this constraint probabilistically. To this end, we define a notion of *out-of-sample disappointment.*

Definition 2.4 (Out-of-sample disappointment)**.** *For any data-driven predictor $\widehat{c}$ the probability*

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

is referred to as the out-of-sample prediction disappointment of $x \in X$ at time T under model $\theta \in \Theta$. Similarly, for any data-driven prescriptor $\widehat{x}$ induced by a data-driven predictor $\widehat{c}$ the probability

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

is termed the out-of-sample prescription disappointment at time T under model $\theta \in \Theta$.

Note that the out-of-sample disappointment represents the probability that the out-of-sample risk strictly exceeds the in-sample risk. Intuitively, a smaller out-of-sample disappointment should be preferred over a large out-of-sample disappointment. For example, in the context of the newsvendor problem studied in Section 2.1, a high out-of-sample disappointment entailed a high probability of budget overruns.

The meta-optimization problem to be developed below aims to minimize the in-sample risk. As $\widehat{c}_T(x)$ for $x \in X$ as well as $\widehat{c}_T(\widehat{x}_T)$ are random variables, however, this informal objective is not well-defined. The properties of a data-driven predictor laid out in Definition 2.2 further imply that even the *expected* in-sample risk is not well-defined. Indeed, if the data is generated under $\mathbb{P}_\theta$, then $\mathbb{E}_\theta[\widehat{c}_T(x)]$ converges to $c_\infty(x,\theta)$ as T grows, where c_∞ is the Borel-measurable function whose existence is postulated in Definition 2.2(iv). This follows directly from Lemma B.1, which applies because of conditions (iii) and (iv) in Definition 2.2. The same lemma implies that $\mathbb{E}_\theta[\widehat{c}_T(\widehat{x}_T)]$ converges to $v_\infty(\theta)$ as T grows, where v_∞ is the Borel-measurable function whose existence is postulated in Definition 2.2(v).

The above reasoning indicates that both the out-of-sample disappointment as well as the expected in-sample risk depend on the data-generating model θ and the length T of the available observation history. As θ is unobservable, however, the meta-optimization problem to be developed may *not* depend on θ for otherwise its solution would not be implementable. Even though T is known to the decision-maker, we did not manage to construct a meta-optimization problem that adapts to T and can still be solved. To eliminate the dependence on both θ as well as T, we thus propose to minimize the *asymptotic* expected in-sample performance of *every* $\theta \in \Theta$ subject to an upper bound on the *asymptotic* exponential decay rate of the out-of-sample disappointment for *every* $\theta \in \Theta$. Note that since the proposed meta-optimization problem accommodates multiple objective functions (one for each $\theta \in \Theta$) and a constraint that must hold for all realizations of the uncertain parameter $\theta \in \Theta$, it constitutes a *robust multi-objective optimization problem.*

The meta-optimization problem for finding the best data-driven predictor can thus be formulated as

$$\begin{array}{ll} \underset{\widehat{c}\in\widehat{\mathcal{C}}}{\text{minimize}} & \left\{\lim\limits_{T\to\infty} \mathbb{E}_\theta\left[\widehat{c}_T(x)\right]\right\}_{x\in X,\,\theta\in\Theta} \\ \text{subject to} & \limsup\limits_{T\to\infty} \dfrac{1}{T}\log \mathbb{P}_\theta[c(x,\theta) > \widehat{c}_T(x)] \leq -r \quad \forall x\in X,\ \theta\in\Theta. \end{array} \tag{2.4a}$$

Recall that the asymptotic expected in-sample risk $\lim_{T\to\infty}\mathbb{E}_\theta[\widehat{c}_T(x)]$ under model θ is well-defined and coincides with the limit function $c_\infty(x,\theta)$ of Definition 2.2(iv) for every $x \in X$ and $\theta \in \Theta$. The constraint requires that the out-of-sample disappointment under model θ satisfies $\mathbb{P}_\theta[c(x,\theta) > \widehat{c}_T(x)] \leq e^{-rT+o(T)}$ for every $x \in X$ and $\theta \in \Theta$, where $r > 0$ is a risk aversion parameter chosen by the decision-maker.

Similarly, the meta-optimization problem for finding the best predictor-prescriptor-pair can be formulated as

$$\begin{array}{ll} \underset{(\widehat{c},\widehat{x})\in\widehat{\mathcal{X}}}{\text{minimize}} & \left\{\lim\limits_{T\to\infty} \mathbb{E}_\theta\left[\widehat{c}_T(\widehat{x}_T)\right]\right\}_{\theta\in\Theta} \\ \text{subject to} & \limsup\limits_{T\to\infty} \dfrac{1}{T}\log \mathbb{P}_\theta[c(\widehat{x}_T,\theta) > \widehat{c}_T(\widehat{x}_T)] \leq -r \quad \forall \theta\in\Theta. \end{array} \tag{2.4b}$$

As above, $\lim_{T\to\infty}\mathbb{E}_\theta[\widehat{c}_T(\widehat{x}_T)]$ is well-defined and coincides with the limit function $v_\infty(\theta)$ of Definition 2.2(v) for every $\theta \in \Theta$, and the constraint requires that $\mathbb{P}_\theta[c(\widehat{x}_T,\theta) > \widehat{c}_T(x)] \leq e^{-rT+o(T)}$ for every $\theta \in \Theta$.

To gain some intuition for the rate constraint in (2.4a), recall from Definition 2.2(iv) that $\widehat{c}_T(x)$ converges in probability to $c_\infty(x,\theta)$. As convergence in probability implies convergence in distribution, we thus have

$$\lim_{T\to\infty}\mathbb{P}_\theta[c(x,\theta) \geq \widehat{c}_T(x)] = 1_{c(x,\theta)\geq c_\infty(x,\theta)} \quad \forall x\in X,\ \theta\in\Theta \text{ with } c(x,\theta)\neq c_\infty(x,\theta).$$

The rate constraint in (2.4a) requires the out-of-sample disappointment to converge to 0 as T grows. The above reasoning thus implies that the rate constraint is *not* satisfiable if there exists a decision $x \in X$ and a model $\theta \in \Theta$ with $c(x, \theta) > c_\infty(x, \theta)$. In other words, any feasible data-driven predictor $\widehat{c}$ must asymptotically exceed (or match) the model-based predictor $c(x, \theta)$ for all $x \in X$ and $\theta \in \Theta$. This conclusion is consistent with the reasoning that led to the meta-optimization problem (2.4a). In the remainder of the paper we will show that an exponentially decaying out-of-sample disappointment necessitates indeed a biased data-driven predictor that *strictly* overestimates $c(x, \theta)$. In addition, the bias increases with the desired decay rate r.

Both meta-optimization problems (2.4a) and (2.4b) seek an optimal trade-off between in-sample performance and out-of-sample disappointment. A similar trade-off arises in statistical hypothesis testing, where one often seeks to optimally balance the exponential decay rates of the type I and type II errors.

Multi-objective optimization problems such as (2.4a) and (2.4b) typically only admit Pareto optimal solutions, *i.e.*, feasible solutions that are not Pareto dominated by any other feasible solution. Perhaps surprisingly, in the remainder of this paper we will show that under some regularity conditions both (2.4a) and (2.4b) admit Pareto dominant solutions, *i.e.*, feasible solutions that Pareto dominate all other feasible solutions. Moreover, these solutions admit intuitive closed-form expressions.

2.3 Data compression

A defining property of data-driven predictors and prescriptors is that they are adapted to the filtration generated by the data. Thus, they can be seen as sequences of functions that map the increasingly high-dimensional observation history $\xi_{[T]} \in \mathbb{R}^{dT}$ to a cost estimate or a decision, respectively. Processing or even storing such functions might easily become impractical for large T. As a remedy, we will try to compress the observation history $\xi_{[T]}$ into a statistic $\widehat{S}_T$ of constant dimension d without sacrificing useful information.

Definition 2.5 (Statistic)**.** *A stochastic process $\widehat{S} = \{\widehat{S}_T\}_{T \in \mathbb{N}}$ with a closed state space $\mathbb{S} \subseteq \mathbb{R}^d$ is called a statistic if it is adapted to the filtration $\{\mathcal{F}_T\}_{T \in \mathbb{N}}$ and if there exists a local homeomorphism $S_\infty : \Theta \to \mathbb{S}$ such that, as T grows, $\widehat{S}_T$ converges in probability under $\mathbb{P}_\theta$ to $S_\infty(\theta)$ for every $\theta \in \Theta$. If $S_\infty(\theta) = \theta$ for all $\theta \in \Theta$, then the statistic $\widehat{S}$ is called a consistent model estimator.*

As $\widehat{S}$ is adapted to $\{\mathcal{F}_T\}_{T \in \mathbb{N}}$, we know from [3, Theorem 5.4.2] that for any $T \in \mathbb{N}$ there exists a Borel-measurable function $h_T : \Xi^T \to \mathbb{R}^d$ with $\widehat{S}_T = h_T(\xi_{[T]})$. In the following we will always assume that the state space $\mathbb{S} \subseteq \mathbb{R}^d$ is defined as the smallest closed set that satisfies $\mathbb{P}_\theta[\widehat{S}_T \in \mathbb{S}] = 1$ for all $\theta \in \Theta$ and $T \in \mathbb{N}$. It is also useful to define $\mathbb{S}_\infty = \{S_\infty(\theta) : \theta \in \Theta\} \subseteq \mathbb{S}$ as the set of all asymptotic realizations of the statistic $\widehat{S}$. As Θ is open with respect to the subspace topology on Θ and as S_∞ constitutes a local homeomorphism, $\mathbb{S}_\infty$ is an open subset of $\mathbb{S}$ with respect to the subspace topology on $\mathbb{S}$.

Example 2.4 (Empirical distribution for finite state i.i.d. processes)**.** *In the context of the finite state i.i.d. processes described in Example 2.2, we define the empirical distribution $\widehat{S}_T \in \mathbb{R}^d$ through*

$$(\widehat{S}_T)_i = \frac{1}{T} \sum_{t=1}^{T} 1_{\xi_t = i} \quad \forall i \in \Xi,\ T \in \mathbb{N}. \tag{2.5}$$

Thus, the i^{th} component of $\widehat{S}_T$ records the empirical frequency of observing state i over the first T time periods. By construction, $\widehat{S} = \{\widehat{S}_T\}_{T \in \mathbb{N}}$ constitutes a consistent model estimator in the sense of Definition 2.5. Indeed, the strong law of large numbers guarantees that, under $\mathbb{P}_\theta$, the empirical distribution $\widehat{S}_T$ converges almost surely (and thus in probability) to $S_\infty(\theta) = \theta$ for every $\theta \in \Theta$. Hence, the set $\mathbb{S}_\infty$ coincides with the *open probability simplex Θ. As the support of $\widehat{S}_T$ is given by $\Delta_d \cap (\mathbb{Z}^d/T)$ for each $T \in \mathbb{N}$, we also have*

$$\mathbb{S} = \mathrm{cl}\left(\cup_{T\in\mathbb{N}} \Delta_d \cap (\mathbb{Z}^d/T)\right) = \mathrm{cl}\left(\Delta_d \cap \mathbb{Q}^d\right) = \Delta_d = \mathrm{cl}\,\Theta.$$

We are now ready to introduce families of data-driven predictors and prescriptors that depend on the data only indirectly through a statistic, which may or may not be a consistent model estimator. To our best knowledge, all predictors and prescriptors studied in the existing literature can be represented in this form.

Definition 2.6 (Compressed data-driven predictors and prescriptors)**.** *If $\mathbb{S}$ and $\mathbb{S}_\infty$ represent the state space and the set of asymptotic realizations of a statistic $\widehat{S}$, then $\tilde{c} : X \times \mathbb{S} \to \mathbb{R}$ is called a compressed data-driven predictor if it is bounded and continuous in x on $X \times \mathbb{S}$ and continuous in (x, s) on $X \times \mathbb{S}_\infty$. In addition, $\tilde{x} : \mathbb{S} \to X$ is called a compressed data-driven prescriptor if it is quasi-continuous on $\mathbb{S}_\infty$ and there exists a compressed data-driven predictor $\tilde{c}$ that induces $\tilde{x}$ in the sense that $\tilde{x}(s) \in \arg\min_{x\in X} \tilde{c}(x, s)$ for all $s \in \mathbb{S}$.*

One can show that *any* compressed data-driven predictor $\tilde{c}$ induces a (not necessarily unique) compressed data-driven prescriptor $\tilde{x}$. To see this, note first that the multifunction $\arg\min_{x\in X} \tilde{c}(x, s)$ is non-empty-valued because X is compact and $\tilde{c}(x, s)$ is continuous in $x \in X$ for every fixed $x \in \mathbb{S}$. Moreover, the restriction of this multifunction to $\mathbb{S}_\infty$ admits a quasi-continuous selector. This follows from the reasoning after Definition 3 in [55], which applies here because $\tilde{c}$ is continuous on $X \times \mathbb{S}_\infty$ and X is compact. Note also that any compressed data-driven predictor $\tilde{c}$ and the underlying statistic $\widehat{S}$ induce a data-driven predictor $\widehat{c}$ defined through $\widehat{c}_T(x) = \tilde{c}(x, \widehat{S}_T)$ for all $x \in X$ and $T \in \mathbb{N}$. One readily verifies that $\widehat{c}$ satisfies all conditions of Definition 2.2. Indeed, conditions (i)–(iii) follow directly from the definitions of the statistic $\widehat{S}$ and the compressed data-driven predictor $\tilde{c}$. To check condition (iv), fix a probability measure $\mathbb{P}_\theta$, and recall that $\widehat{S}_T$ converges in probability to $S_\infty(\theta)$. By the continuous mapping theorem [26, Theorem 3.2.4], which applies because $S_\infty(\theta) \in \mathbb{S}_\infty$ and because $\tilde{c}(x, s)$ is continuous in $s \in \mathbb{S}_\infty$ for every fixed $x \in X$, we may then conclude that $\widehat{c}_T(x)$ converges in probability to $\tilde{c}(x, S_\infty(\theta))$. As this reasoning applies to every model $\theta \in \Theta$, condition (iv) holds with $c_\infty(x, \theta) = \tilde{c}(x, S_\infty(\theta))$. To check condition (v), fix again a probability measure $\mathbb{P}_\theta$, and introduce a real-valued function $\tilde{v}(s) = \min_{x\in X} \tilde{c}(x, s)$, which is continuous in $s \in \mathbb{S}$ by Berge's maximum theorem [11, pp. 115–116]. Invoking the continuous mapping theorem as above, it then follows that $\tilde{v}(\widehat{S}_T) = \min_{x\in X} \widehat{c}_T(x)$ converges in probability to $\tilde{v}(S_\infty(\theta))$. As this reasoning applies to every model $\theta \in \Theta$, condition (v) holds with $v_\infty(\theta) = \tilde{v}(S_\infty(\theta))$, which is a continuous function by construction. Finally, any compressed data-driven prescriptor $\tilde{x}$ and the corresponding statistic $\widehat{S}$ induce a data-driven prescriptor $\widehat{x}$ defined trough $\widehat{x}_T = \tilde{x}(\widehat{S}_T)$ for all $T \in \mathbb{N}$. One readily verifies that $\widehat{x}$ satisfies all conditions of Definition 2.3. Indeed, condition (i) follows directly from the defining properties of a statistic. To check condition (ii), recall that any compressed data-driven prescriptor $\tilde{x}$ is induced by some compressed data-driven predictor $\tilde{c}$. Next, define an ordinary data-driven predictor $\widehat{c}$ through $\widehat{c}_T(x) = \tilde{c}(x, \widehat{S}_T)$ for all $x \in X$ and $T \in \mathbb{N}$. By our earlier reasoning, $\widehat{c}$ satisfies indeed all conditions of Definition 2.2. Then, we have

$$\widehat{x}_T = \tilde{x}(\widehat{S}_T) \in \arg\min_{x\in X} \tilde{c}(x, \widehat{S}_T) = \arg\min_{x\in X} \widehat{c}_T(x) \quad \forall T \in \mathbb{N},$$

where the two equalities follow from the definitions of $\widehat{x}_T$ and $\widehat{c}_T$, respectively, while the membership relation holds by assumption. Hence, $\widehat{x}$ is induced by the data-driven predictor $\widehat{c}$, and thus condition (ii) holds.

In analogy to our conventions of Section 2.2, from now on we denote the set of all compressed data-driven predictors by $\tilde{\mathcal{C}}$ and the set of all compressed data-driven predictor-prescriptor-pairs by $\tilde{\mathcal{X}}$.

Example 2.5 (Empirical predictor for finite state i.i.d. processes revisited). *If $\widehat{S} = \{\widehat{S}_T\}_{T\in\mathbb{N}}$ is any statistic whose state space $\mathbb{S}$ is a subset of* $\mathrm{cl}(\Theta)$*, then the model-based predictor c of Definition 2.1 constitutes a trivial compressed data-driven predictor with respect to $\widehat{S}$. Indeed, recall that c admits a continuous extension to $X \times$* $\mathrm{cl}(\Theta)$ *thanks to Assumption 2.2. In the context of the finite state i.i.d. processes described in Example 2.2, it is natural to set $\widehat{S}_T$ to the empirical distribution over the first T observations as in Example 2.4. In this case, we have $\mathbb{S} = \mathrm{cl}(\Theta)$, which ensures that $\widehat{c}_T(x) = c(x, \widehat{S}_T)$ is well-defined for every $x \in X$ and $T \in \mathbb{N}$. In the special case when $c(x, \theta) = \mathbb{E}_\theta[\ell(x, \xi)]$, a direct calculation shows that $\widehat{c}_T(x) = \frac{1}{T}\sum_{t=1}^T \ell(x, \xi_t)$. Thus, the data-driven predictor $\widehat{c} = \{\widehat{c}_T\}_{T\in\mathbb{N}}$ induced by c and $\widehat{S}$ coincides with the empirical predictor of Example 2.3.*

We now consider a restriction of the meta-optimization problem (2.4a) that optimizes only over *compressed* data-driven predictors $\tilde{c} \in \tilde{\mathcal{C}}$. As in (2.4a), we minimize the asymptotic expected in-sample performance of *every* $\theta \in \Theta$ subject to an upper bound on the asymptotic exponential decay rate of the out-of-sample disappointment for *every* $\theta \in \Theta$. Identifying each compressed data-driven predictor $\tilde{c}$ with an ordinary data-driven predictor $\widehat{c}$ defined through $\widehat{c}_T(x) = \tilde{c}(x, \widehat{S}_T)$, $T \in \mathbb{N}$, and observing that $\lim_{T\to\infty} \mathbb{E}_\theta[\tilde{c}(x, \widehat{S}_T)] = \tilde{c}(x, S_\infty(\theta))$ for all $x \in X$ and $\theta \in \Theta$ thanks to the continuous mapping theorem [26, Theorem 3.2.4] and Lemma B.1, the restricted meta-optimization problem can be formulated as follows.

$$\begin{array}{ll} \underset{\tilde{c}\in\tilde{\mathcal{C}}}{\text{minimize}} & \{\tilde{c}(x, S_\infty(\theta))\}_{x\in X,\, \theta\in\Theta} \\ \text{subject to} & \limsup\limits_{T\to\infty} \dfrac{1}{T} \log \mathbb{P}_\theta[c(x, \theta) > \tilde{c}(x, \widehat{S}_T)] \leq -r \quad \forall x \in X,\ \theta \in \Theta \end{array} \tag{2.6a}$$

Likewise, identifying each compressed data-driven predictor-prescriptor pair $(\tilde{c}, \tilde{x})$ with an ordinary data-driven predictor-prescriptor pair $(\widehat{c}, \widehat{x})$ defined through $\widehat{c}_T(x) = \tilde{c}(x, \widehat{S}_T)$ and $\widehat{x}_T = \tilde{x}(\widehat{S}_T)$, $T \in \mathbb{N}$, and observing that $\lim_{T\to\infty} \mathbb{E}_\theta[\tilde{c}(\tilde{x}(\widehat{S}_T), \widehat{S}_T)] = \tilde{c}(\tilde{x}(S_\infty(\theta)), S_\infty(\theta))$ for all $\theta \in \Theta$ thanks to the continuous mapping theorem and Lemma B.1, we obtain the following restriction of the meta-optimization problem (2.4b).

$$\begin{array}{ll} \underset{(\tilde{c},\tilde{x})\in\tilde{\mathcal{X}}}{\text{minimize}} & \{\tilde{c}(\tilde{x}(S_\infty(\theta)), S_\infty(\theta))\}_{\theta\in\Theta} \\ \text{subject to} & \limsup\limits_{T\to\infty} \dfrac{1}{T} \log \mathbb{P}_\theta[c(\tilde{x}(\widehat{S}_T), \theta) > \tilde{c}(\tilde{x}(\widehat{S}_T), \widehat{S}_T)] \leq -r \quad \forall \theta \in \Theta \end{array} \tag{2.6b}$$

Focusing on compressed data-driven predictors and prescriptors seems natural and is indeed the *de facto* standard. On the one hand, one would expect that the corresponding restricted meta-optimization problems (2.6) are easier to solve than the original meta-optimization problems (2.4). On the other hand, it is unclear how much performance is sacrificed by this restriction. In the following we will first show that the restricted meta-optimization problems (2.6) admit Pareto dominant solutions whenever the underlying statistic $\widehat{S}$ satisfies a large deviation principle. Later we will show that the compressed and original meta-optimization problems are equivalent whenever $\widehat{S}$ represents a sufficient statistic.

3 Pareto dominant predictors and prescriptors

In order to construct Pareto dominant solutions for the restricted meta-optimization problems (2.6), if they exist, we first review and extend some fundamental definitions and concepts from large deviations theory.

3.1 Large deviation principles

Large deviations theory provides bounds on the exponential rate at which the probabilities of atypical realizations of a given statistic $\widehat{S}$ decay as the length T of the observation history grows. These bounds are expressed in terms of a rate function, which depends on a realization of $\widehat{S}$ and the data-generating model θ.

Definition 3.1 (Rate function [22, Section 2.1])**.** *A function $I : \mathbb{S} \times \operatorname{cl}\Theta \to [0,\infty]$ is called a rate function if $I(s,\theta)$ is lower semi-continuous in s throughout $\mathbb{S} \times \operatorname{cl}\Theta$.*

Definition 3.2 (Large deviation principle)**.** *The statistic $\widehat{S} = \{\widehat{S}_T\}_{T\in\mathbb{N}}$ with state space $\mathbb{S}$ satisfies a large deviation principle (LDP) with rate function I, if for all $\theta \in \Theta$ and Borel sets $\mathcal{D} \subseteq \mathbb{S}$ we have*

$$-\inf_{s\in\operatorname{int}\mathcal{D}} I(s,\theta) \le \liminf_{T\to\infty} \frac{1}{T}\log \mathbb{P}_\theta[\widehat{S}_T \in \mathcal{D}] \tag{3.1a}$$

$$\le \limsup_{T\to\infty} \frac{1}{T}\log \mathbb{P}_\theta[\widehat{S}_T \in \mathcal{D}] \le -\inf_{s\in\operatorname{cl}\mathcal{D}} I(s,\theta). \tag{3.1b}$$

As the subspace topology on $\mathbb{S}$ induced by the Euclidean topology on $\mathbb{R}^d$ is Hausdorff, we know from [22, Lemma 4.1.4] that if $\widehat{S}$ satisfies an LDP, then the inequalities (3.1) uniquely determine the rate function on $\mathbb{S}\times\Theta$. However, Definition 3.1 requires the rate function to be defined on $\mathbb{S}\times\operatorname{cl}\Theta$. Even though its values on the boundary of Θ are immaterial, extending the rate function to $\mathbb{S}\times\operatorname{cl}\Theta$ will simplify some of the derivations in Section 3, provided the extension preserves the regularity conditions of Definition 3.3 below.

Before defining regular rate functions, we discuss a few immediate consequences of the inequalities (3.1). First, as $\widehat{S}_T$ converges in probability to $S_\infty(\theta)$ under $\mathbb{P}_\theta$, the LDP bound (3.1b) implies that

$$0 = \lim_{T\to\infty}\frac{1}{T}\log\mathbb{P}_\theta\left[\|\widehat{S}_T - S_\infty(\theta)\|_2 \le \frac{1}{k}\right] \le -\inf_{s\in\mathbb{S}}\left\{I(s,\theta) : \|s - S_\infty(\theta)\|_2 \le \frac{1}{k}\right\} \quad \forall k\in\mathbb{N}.$$

Thus, there is a sequence $\{s_k\}_{k\in\mathbb{N}}$ in $\mathbb{S}$ that converges to $S_\infty(\theta)$ and satisfies $\liminf_{k\in\mathbb{N}} I(s_k,\theta)\le 0$, which implies via the non-negativity and lower semi-continuity of the rate function that $I(S_\infty(\theta),\theta)=0$. This in turn implies that the minima of the optimization problems in (3.1a) and (3.1b) evaluate to 0 whenever $S_\infty(\theta)$ falls within the interior of $\mathcal{D}$. In this case the LDP inequalities reduce to the trivial statement that $\mathbb{P}_\theta[\widehat{S}_T\in\mathcal{D}]$ converges to 1 as T grows. In general, the LDP inequalities (3.1) imply that the probability $\mathbb{P}_\theta[\widehat{S}_T\in\mathcal{D}]$ is bounded below by $e^{-\overline{r}\,T+o(T)}$, where $\overline{r}=\inf_{s\in\operatorname{int}\mathcal{D}} I(s,\theta))$ represents the I-distance between θ and the interior of $\mathcal{D}$, and bounded above by $e^{-\underline{r}\,T+o(T)}$, where $\underline{r}=\inf_{s\in\operatorname{cl}\mathcal{D}} I(s,\theta)$ represents the I-distance between θ and the closure of $\mathcal{D}$. For non-pathological sets $\mathcal{D}$ that emerge in applications, one often has $\underline{r}=\overline{r}$. In this case the exponential decay rate of $\mathbb{P}_\theta[\widehat{S}_T\in\mathcal{D}]$ is uniquely determined by (3.1).

In the following we will study the class of compressed data-driven predictors and prescriptors corresponding to a statistic $\widehat{S}$ that satisfies an LDP. We will see that the *optimal* predictors and prescriptors within this class (that is, the Pareto dominant solutions of the meta-optimization problems (2.6a) and (2.6b)) can be constructed in closed form if this LDP's rate function is regular in the sense of the following definition.

Definition 3.3 (Regular rate function)**.** *We call a rate function I regular if the following conditions hold.*

(i) **Radial monotonicity in θ.** $\operatorname{cl}\{\theta\in\Theta : I(s,\theta) < r\} = \{\theta\in\operatorname{cl}\Theta : I(s,\theta)\le r\}$ *for all $s\in\mathbb{S}_\infty$, $r>0$.*

(ii) **Continuity.** *$I(s,\theta)$ is continuous on $\mathbb{S}\times\Theta$.*

(iii) **Level-compactness.** $\{(s,\theta)\in\mathbb{S}\times\operatorname{cl}\Theta : I(s,\theta)\le r\}$ *is compact for every $r\ge 0$.*

Definition 3.3 strengthens the more common notion of a 'good' rate function. Recall that a rate function I is called good if $\{s\in\mathbb{S} : I(s,\theta)\le r\}$ is compact for every $r\ge 0$ and $\theta\in\Theta$ [22, Section 1.2]. As $\mathbb{S}_\infty$ is a subset of $\mathbb{S}$ and as $\mathbb{S}$ is closed, condition (iii) of Definition 3.3 implies indeed that every regular rate function is good. Note also that the radial monotonicity condition (i) may be difficult to check in practice. A more easily checkable sufficient condition for radial monotonicity is that for any $s \in \mathbb{S}_\infty$, $\theta \in \operatorname{cl}\Theta$ and $\theta_s \in \Theta$ with $S_\infty(\theta_s) = s$ (note that θ_s exists because $s \in \mathbb{S}_\infty$) we have

$$I(s, (1-\lambda)\theta_s + \lambda\theta) \leq I(s, \theta) \quad \forall \lambda \in [0, 1), \tag{3.2}$$

where the inequality is strict if $I(s, \theta) > 0$. The inequality (3.2) actually inspired the name 'radial monotonicity' for condition (i). We will now prove that (3.2) together with condition (iii) implies condition (i). To this end, fix any $s \in \mathbb{S}_\infty$, and set $A(s) = \{\theta \in \Theta : I(s, \theta) < r\}$ and $B(s) = \{\theta \in \operatorname{cl}\Theta : I(s, \theta) \leq r\}$. By construction, we have $A(s) \subseteq B(s)$. As $B(s)$ is compact thanks to condition (iii), this even implies that $\operatorname{cl} A(s) \subseteq B(s)$. It remains to be shown that (3.2) implies the converse inclusion $B(s) \subseteq \operatorname{cl} A(s)$. To this end, fix any $\theta \in B(s)$, and choose any $\theta_s \in \Theta$ with $s = S_\infty(\theta_s)$, which exists because $s \in \mathbb{S}_\infty$. Next, define $\theta(\lambda) = (1-\lambda)\theta_s + \lambda\theta$ for all $\lambda \in [0, 1)$. As $\theta_s \in \Theta$ and $\theta \in \operatorname{cl}\Theta$, and as Θ is open and convex, the line segment principle [12, Proposition 1.3.1] implies that $\theta(\lambda) \in \Theta$ for all $\lambda \in [0, 1)$. By (3.2), we also have $I(s, \theta(\lambda)) \leq I(s, \theta) \leq r$ for all $\lambda \in [0, 1)$, where at least one of the two inequalities is strict. This reasoning shows that $\theta(\lambda) \in A(s)$ for all $\lambda \in [0, 1)$. As $\theta(\lambda)$ approaches θ arbitrarily closely when λ increases towards 1, we may finally conclude that $\theta \in \operatorname{cl} A(s)$. We have therefore shown that condition (i) holds.

Example 3.1 (An LDP for finite state i.i.d. processes)**.** *Consider the class of finite state i.i.d. processes introduced in Example 2.2, and let $\widehat{S}_T$ be the empirical distribution defined in Example 2.4. The classical Sanov theorem [22, Theorem 2.1.10] asserts that $\widehat{S}$ satisfies an LPD with good rate function $I(s, \theta) = D(s\|\theta)$, where $D(s\|\theta) = \sum_{i=1}^d s_i \log(s_i/\theta_i)$ denotes the relative entropy of s with respect to θ, and where we use the standard conventions that $0 \log(0/p) = 0$ for any $p \geq 0$ and $p \log(p/0) = \infty$ for any $p > 0$. The relative entropy is also referred to as Kullback-Leibler divergence. By the information inequality [20, Theorem 2.6.3], it is non-negative and vanishes if and only if $\theta = s$. Moreover, the relative entropy is a regular rate function in the sense of Definition 3.3. To see this, note that $D(s\|\theta)$ is continuous on $\mathbb{S} \times \Theta$ and level-compact [22, pp. 13–18]. In addition, $D(s\|\theta)$ is jointly convex in s and θ thanks to [20, Theorem 2.7.2]. Recalling from Example 2.4 that S_∞ is the identity function, for any $s \in \mathbb{S}_\infty = \Theta$ and $\theta \in \operatorname{cl}\Theta$ we thus have*

$$D(s\|(1-\lambda)s + \lambda\theta) \leq (1-\lambda)D(s\|s) + \lambda D(s\|\theta) \leq D(s\|\theta) \quad \forall \lambda \in [0, 1),$$

where the second inequality holds because $D(s\|s) = 0$ and $D(s\|\theta) \geq 0$. Note that this inequality is strict for $D(s\|\theta) > 0$, and thus $D(s\|\theta)$ is radially monotonic in θ due to (3.2)*. Hence, the relative entropy is indeed a regular rate function. In Section 5 we will present a broad spectrum of additional data-generating stochastic processes for which there exists a statistic that satisfies an LDP with a regular rate function.*

We are now ready to demonstrate that if the statistic $\widehat{S}$ satisfies an LDP with a regular rate function, then the restricted meta-optimization problems (2.6) admit Pareto dominant solutions. In Section 3.2, we first construct a compressed data-driven predictor that is strongly optimal in (2.6a). In Section 3.3 we then construct a compressed data-driven predictor-prescriptor pair that is strongly optimal in (2.6b).

3.2 Distributionally robust predictors

In order to solve the meta-optimization problems (2.6) over compressed data-driven predictors and prescriptors, we assume that the statistic $\widehat{S}$ satisfies an LDP and that the underlying rate function is regular.

Assumption 3.1 (LDP)**.** *The statistic $\widehat{S}$ satisfies an LDP with a regular rate function.*

We will impose Assumption 3.1 throughout the rest of this section. We can now construct a compressed data-driven predictor, which will later be shown to represent a Pareto dominant solution for (2.6a).

Definition 3.4 (Distributionally robust predictor)**.** *The function $\tilde{c}^\star : X \times \mathbb{S} \to \mathbb{R}$ defined through*

$$\tilde{c}^\star(x,s) = \begin{cases} \max\limits_{\theta \in \operatorname{cl}\Theta} \ \{c(x,\theta) : I(s,\theta) \le r\} & \text{if } \exists\, \theta \in \operatorname{cl}\Theta \text{ with } I(s,\theta) \le r, \\ \sup\limits_{\theta \in \operatorname{cl}\Theta} \ \ c(x,\theta) & \text{if } \nexists\, \theta \in \operatorname{cl}\Theta \text{ with } I(s,\theta) \le r, \end{cases} \tag{3.3}$$

is the distributionally robust predictor induced by the rate function I and the risk aversion parameter r.

Note that the maximum of the first optimization problem in (3.3) is indeed attained because the feasible set is compact due to the level-compactness of the regular rate function $I(s,\theta)$ and because the objective function $c(x,\theta)$ is continuous in θ on $X \times \operatorname{cl}\Theta$ thanks to the discussion after Assumption 2.2. In addition, the supremum of the second optimization problem in (3.3) is finite because $c(x,\theta)$ is bounded on $X \times \operatorname{cl}\Theta$. The following proposition confirms that $\tilde{c}^\star$ is a compressed data-driven predictor in the sense of Definition 2.6.

Proposition 3.1 (Continuity of $\tilde{c}^\star$)**.** *If the rate function I is regular and $r > 0$, then the distributionally robust predictor $\tilde{c}^\star(x,s)$ is bounded and continuous in x on $X \times \mathbb{S}$ and continuous in (x,s) on $X \times \mathbb{S}_\infty$.*

The proof of Proposition 3.1 is technical and therefore relegated to Appendix A. Intuitively, the compressed data-driven predictor $\tilde{c}^\star(x,s)$ evaluates the worst-case objective function of the stochastic optimization problem (2.2) over all probability measures $\mathbb{P}_\theta$ corresponding to models $\theta \in \operatorname{cl}\Theta$ that reside in an I-ball of radius r around s. Thus, $\tilde{c}^\star(x,s)$ admits a distributionally robust interpretation, which justifies our terminology.

Example 3.2 (Distributionally robust predictors for finite state i.i.d. processes)**.** *Consider the class of finite state i.i.d. processes introduced in Example 2.2, and let $\widehat{S}_T$ be the empirical distribution defined in Example 2.4. From Example 3.1 we know that $\widehat{S}$ satisfies an LDP and that the underlying regular rate function coincides with the relative entropy. Thus, the distributionally robust predictor* (3.4) *simplifies to*

$$\tilde{c}^\star(x,s) = \max_{\theta \in \Delta_d} \ \{c(x,\theta) : D(s\|\theta) \le r\}\,.$$

This problem is feasible for every possible estimator realization because $\mathbb{S} = \Delta_d$ (see Example 2.4), and if $c(x,\theta) = \mathbb{E}_\theta[\ell(x,\xi)]$, then it is equivalent to the one-dimensional convex minimization problem

$$\tilde{c}^\star(x,s) = \min_{\alpha \ge \bar{\ell}(x)} \alpha - e^{-r} \prod_{i=1}^{d} (\alpha - \ell(x,i))^{s_i}$$

with $\bar{\ell}(x) = \max_{i \in \Xi} \ell(x,i)$, which can be solved efficiently via line search methods (see [55, Proposition 2]).

The following theorem establishes that the distributionally robust predictor (3.3) strikes indeed an optimal balance between expected in-sample performance and out-of-sample disappointment.

Theorem 3.1 (Optimality of $\tilde{c}^\star$)**.** *If Assumptions 2.1, 2.2 and 3.1 hold and if $r > 0$, then $\tilde{c}^\star$ is a Pareto dominant solution of the meta-optimization problem* (2.6a).

Proof. We first show (Step 1) that $\tilde{c}^\star$ is feasible in problem (2.6a), and subsequently (Step 2) we demonstrate that $\tilde{c}^\star$ Pareto dominates any other feasible solution of problem (2.6a).

Step 1. Proposition 3.1 readily implies that $\tilde{c}^\star \in \tilde{\mathcal{C}}$. It remains to be shown that the out-of-sample disappointment of $\tilde{c}^\star$ decays at a rate of at least r. To this end, fix any $x \in X$ and $\theta \in \Theta$, and define the sets

$$A(x,\theta) = \{s \in \mathbb{S} : c(x,\theta) > \tilde{c}^\star(x,s)\} \quad \text{and} \quad B(x,\theta) = \{s \in \mathbb{S} : I(s,\theta) > r\}.$$

We may assume without loss of generality that $A(x,\theta) \neq \emptyset$ for otherwise the out-of-sample disappointment $\mathbb{P}_\Theta[\widehat{S}_T \in A(x,\theta)]$ vanishes for all $T \in \mathbb{N}$ and thus decays at any exponential rate. We will now show that $A(x,\theta) \subseteq B(x,\theta)$. To this end, choose any $s \in A(x,\theta)$, and assume that $I(s,\theta) \leq r$. Thus, we have

$$c(x,\theta) > \tilde{c}^\star(x,s) = \max_{\theta' \in \operatorname{cl}\Theta} \{c(x,\theta') : I(s,\theta') \leq r\} \geq c(x,\theta),$$

where the strict inequality holds because $s \in A(x,\theta)$, the equality follows from the definition of $\tilde{c}^\star(x,s)$ and the assumption that $A(x,\theta) \neq \emptyset$, and the weak inequality exploits the assumptions that $\theta \in \Theta$ and $I(s,\theta) \leq r$. The resulting conclusion is manifestly false, which implies that $I(s,\theta) > r$, that is, $s \in B(x,\theta)$. As $s \in A(x,\theta)$ was chosen arbitrarily, we have thus shown that $A(x,\theta) \subseteq B(x,\theta)$. This result further implies that

$$\operatorname{cl} A(x,\theta) \subseteq \operatorname{cl} B(x,\theta) \subseteq \{s \in \mathbb{S} : I(s,\theta) \geq r\}, \tag{3.4}$$

where the second inclusion holds because the set on the right hand side covers $B(x,\theta)$ and is closed thanks to the continuity of the regular rate function $I(s,\theta)$ in s on $\mathbb{S}$. The above reasoning implies that

$$\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta[c(x,\theta) > \tilde{c}^\star(x,\widehat{S}_T)] = \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta[\widehat{S}_T \in A(x,\theta)] \leq -\inf_{s \in \operatorname{cl} A(x,\theta)} I(s,\theta) \leq -r,$$

where the first inequality follows from (3.1b), which applies because $\widehat{S}$ satisfies an LDP with rate function $I(s,\theta)$, and the second inequality is a direct consequence of (3.4). As $x \in X$ and $\theta \in \Theta$ were chosen arbitrarily, we may thus conclude that $\tilde{c}^\star$ is feasible in problem (2.6a).

Step 2. We now prove that $\tilde{c}^\star$ Pareto dominates every other feasible solution of problem (2.6a). Assume to the contrary that there exists a compressed data-driven predictor $\tilde{c} \in \tilde{\mathcal{C}}$ that is feasible in (2.6a) but is *not* Pareto dominated by $\tilde{c}^\star$. Hence, there exist a decision $x_0 \in X$ and a model $\theta_0 \in \Theta$ with

$$\varepsilon = \tilde{c}^\star(x_0, S_\infty(\theta_0)) - \tilde{c}(x_0, S_\infty(\theta_0)) > 0.$$

As we will see, the above inequality implies that $\tilde{c}$ is infeasible in (2.6a). This contradiction will reveal that our initial assumption must have been false and that there cannot be any feasible $\tilde{c}$ that Pareto dominates $\tilde{c}^\star$.

Define $s_0 = S_\infty(\theta_0)$, and recall from the discussion below Definition 3.2 that $I(s_0,\theta_0) = 0$. Thus, we have

$$\tilde{c}^\star(x_0,s_0) = \max_{\theta \in \operatorname{cl}\Theta} \{c(x_0,\theta) : I(s_0,\theta) \leq r\}. \tag{3.5}$$

By construction, θ_0 is feasible in (3.5). From the discussion after Definition 3.4 we further know that the maximization problem (3.5) is solvable. In the following we denote by $\theta^\star \in \operatorname{cl}\Theta$ an arbitrary maximizer. Feasibility of $\theta^\star$ then guarantees that $I(s_0,\theta^\star) \leq r$, and optimality implies the identity

$$\tilde{c}^\star(x_0,s_0) = c(x_0,\theta^\star). \tag{3.6}$$

Recall now that $s_0 \in \mathbb{S}_\infty$ and that $r > 0$. Recall also that the rate function $I(s,\theta)$ is regular and thus radially monotonic thanks to Assumption 3.1. This implies that there exist $\theta_k^\star \in \Theta$, $k \in \mathbb{N}$, such that $I(s_0,\theta_k^\star) < r$ for all $k \in \mathbb{N}$ and $\lim_{k\to\infty} \theta_k^\star = \theta^\star$. In addition, as $c(x,\theta)$ is continuous on $X \times \operatorname{cl}\Theta$ thanks to the discussion after Assumption 2.2, this further implies that there exists $\theta_0^\star \in \Theta$ with $I(s_0,\theta_0^\star) = r_0 < r$ and

$$c(x_0,\theta^\star) < c(x_0,\theta_0^\star) + \varepsilon.$$

Using this inequality, we then find

$$\tilde{c}(x_0, s_0) = \tilde{c}^\star(x_0, s_0) - \varepsilon = c(x_0, \theta^\star) - \varepsilon < c(x_0, \theta_0^\star), \tag{3.7}$$

where the first equality follows from the definitions of ε and s_0, and the second equality holds due to (3.6). In the following we will use (3.7) to show that the prediction disappointment $\mathbb{P}_{\theta_0^\star}[c(x, \theta_0^\star) > \tilde{c}(x, \widehat{S}_T)]$ of $\tilde{c}$ under decision x_0 and model $\theta_0^\star$ decays no faster than $e^{r_0 T}$ for large sample sizes T. As $r_0 < r$, this will imply that $\tilde{c}$ is *in*feasible in (2.6a). To this end, we define the set of disappointing realizations of $\widehat{S}$ as

$$\mathcal{D}(x_0, \theta_0^\star) = \{s \in \mathbb{S} : c(x_0, \theta_0^\star) > \tilde{c}(x_0, s)\}.$$

By (3.7), this disappointment set contains s_0. As $s_0 \in \mathbb{S}_\infty \subseteq \mathbb{S}$ and as $\mathbb{S}_\infty$ is open, we further know that s_0 resides in the interior of $\mathbb{S}$. In addition, $\tilde{c}(x, s)$ is continuous in s on $X \times \mathbb{S}_\infty$. We may thus conclude that s_0 belongs in fact to the interior of $\mathcal{D}(x_0, \theta_0^\star)$ and that

$$\inf_{s \in \operatorname{int} \mathcal{D}(x_0, \theta_0^\star)} I(s, \theta_0^\star) \leq I(s_0, \theta_0^\star) = r_0,$$

where the equality holds by the definition of r_0. As the statistic $\widehat{S}$ satisfies an LDP with rate function $I(s, \theta)$, the above inequality in conjunction with (3.1a) finally implies that

$$-r < -r_0 \leq - \inf_{s \in \operatorname{int} \mathcal{D}(x_0, \theta_0^\star)} I(s, \theta_0^\star) \leq \liminf_{T \to \infty} \frac{1}{T} \log \mathbb{P}_{\theta_0^\star}[\widehat{S}_T \in \mathcal{D}(x_0, \theta_0^\star)].$$

By the definition of $\mathcal{D}(x_0, \theta_0^\star)$, this means that the out-of-sample disappointment of $\tilde{c}$ corresponding to x_0 and $\theta_0^\star$ decays strictly slower than e^{rT} as T grows. This in turn contradicts our assumption that $\tilde{c}$ is feasible in (2.6a) and thus implies that $\tilde{c}^\star$ is a Pareto dominant solution of problem (2.6a). □

3.3 Distributionally robust prescriptors

We will now demonstrate that if the statistic $\widehat{S}$ satisfies an LDP with a regular rate function, then the distributionally robust predictor $\tilde{c}^\star$ of Definition 3.4 and any compressed data-driven prescriptor $\tilde{x}^\star$ induced by $\tilde{c}^\star$ represent a Pareto dominant solution for the meta-optimization problem (2.6b).

Definition 3.5 (Distributionally robust prescriptor)**.** *If $\tilde{c}^\star$ is a distributionally robust predictor in the sense of Definition 3.4, then any function $\tilde{x}^\star : \mathbb{S} \to X$ that is quasi-continuous on $\mathbb{S}_\infty$ and satisfies*

$$\tilde{x}^\star(s) \in \arg\min_{x \in X} \tilde{c}^\star(x, s) \quad \forall s \in \mathbb{S} \tag{3.8}$$

is a distributionally robust prescriptor.

One can show that any distributionally robust predictor $\tilde{c}^\star$ induces at least one distributionally robust prescriptor $\tilde{x}^\star$. To see this, note first that the multifunction $\arg\min_{x \in X} \tilde{c}^\star(x, s)$ is non-empty-valued because X is compact and $\tilde{c}^\star(x, s)$ is continuous in x on $X \times \mathbb{S}$; see Proposition 3.1. Moreover, the restriction of this multifunction to $\mathbb{S}_\infty$ admits a quasi-continuous selector. This follows from the reasoning after Definition 3 in [55], which applies here because $\tilde{c}^\star$ is continuous on $X \times \mathbb{S}_\infty$ and X is compact. Therefore, $(\tilde{c}^\star, \tilde{x}^\star)$ belongs to the family $\mathcal{X}$ of all compressed data-driven predictor-prescriptor-pairs.

Theorem 3.2 (Optimality of $(\tilde{c}^\star, \tilde{x}^\star)$)**.** *If Assumptions 2.1, 2.2 and 3.1 hold and if $r > 0$, then $(\tilde{c}^\star, \tilde{x}^\star)$ is a Pareto dominant solution of the meta-optimization problem* (2.6b)*.*

Proof. We first show that $(\tilde{c}^\star, \tilde{x}^\star)$ is feasible in problem (2.6b) (Step 1), and subsequently we demonstrate that $(\tilde{c}^\star, \tilde{x}^\star)$ Pareto dominates any other feasible solution of problem (2.6b) (Step 2).

Step 1. From the discussion after Definition 3.5 we already know that $(\tilde{c}^\star, \tilde{x}^\star) \in \tilde{\mathcal{X}}$. It remains to be shown that the out-of-sample disappointment of $(\tilde{c}^\star, \tilde{x}^\star)$ decays at a rate of at least r. To this end, fix any $\theta \in \Theta$, and define the set of all estimator realizations that lead to disappointment for *some* decision $x \in X$ as

$$A(\theta) = \left\{ s \in \mathbb{S} : \max_{x \in X} \{c(x,\theta) - \tilde{c}^\star(x,s)\} > 0 \right\}.$$

The maximum in this definition is indeed attained because X is compact and because the model-based predictor $c(x,\theta)$ and the distributionally robust predictor $\tilde{c}^\star(x,s)$ are continuous in x thanks to Assumption 2.2 and Proposition 3.1, respectively. Note also that $A(\theta) = \cup_{x \in X} A(x,\theta)$, where $A(x,\theta)$ is defined as in Step 1 of the proof of Theorem 3.1. The inclusion (3.4) and the continuity of $I(s,\theta)$ in s thus ensure that

$$\mathrm{cl}\, A(\theta) \subseteq \mathrm{cl}\{s \in \mathbb{S} : I(s,\theta) \geq r\} = \{s \in \mathbb{S} : I(s,\theta) \geq r\}, \tag{3.9}$$

respectively. The above reasoning implies that

$$\begin{aligned} &\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta \left[c(\tilde{x}^\star(\widehat{S}_T),\theta) > \tilde{c}^\star(\tilde{x}^\star(\widehat{S}_T), \widehat{S}_T) \right] \\ &\qquad \leq \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta \left[\max_{x\in X} \left\{ c(x,\theta) - \tilde{c}^\star(x, \widehat{S}_T) \right\} > 0 \right] \leq -\inf_{s \in \mathrm{cl}\, A(\theta)} I(s,\theta) \leq -r. \end{aligned}$$

Indeed, the first inequality holds because $\tilde{x}^\star(\widehat{S}_T) \in X$ $\mathbb{P}_\theta$-almost surely, the second follows from the definition of $A(\theta)$ and the LDP bound (3.1b) and the third inequality is a direct consequence of (3.9). As $\theta \in \Theta$ was chosen arbitrarily, we may thus conclude that $(\tilde{c}^\star, \tilde{x}^\star)$ is feasible in problem (2.6b).

Step 2. We now prove that $(\tilde{c}^\star, \tilde{x}^\star)$ Pareto dominates every other feasible solution of problem (2.6b). Assume to the contrary that there exists a compressed data-driven predictor-prescriptor pair $(\tilde{c}, \tilde{x}) \in \tilde{\mathcal{X}}$ that is feasible in (2.6b) but is *not* Pareto dominated by $(\tilde{c}^\star, \tilde{x}^\star)$. Hence, there exists a model $\theta_0 \in \Theta$ with

$$\tilde{c}^\star(\tilde{x}^\star(S_\infty(\theta_0)), S_\infty(\theta_0)) - \tilde{c}(\tilde{x}(S_\infty(\theta_0)), S_\infty(\theta_0)) > 0. \tag{3.10}$$

As we will see, (3.10) implies that $(\tilde{c}, \tilde{x})$ is infeasible in (2.6b). This contradiction will reveal that our initial assumption must have been false and that there cannot be any feasible $(\tilde{c}, \tilde{x})$ that Pareto dominates $(\tilde{c}^\star, \tilde{x}^\star)$.

As X is compact and independent of s and as $\tilde{c}^\star(x,s)$ is continuous on $X \times \mathbb{S}_\infty$ by virtue of Proposition 3.1, Berge's maximum theorem [11, pp. 115–116]) implies that $\tilde{c}^\star(\tilde{x}^\star(s), s) = \min_{x\in X} \tilde{c}^\star(x,s)$ is continuous on $\mathbb{S}_\infty$. In addition, $S_\infty(\theta)$ is continuous on Θ by Definition 2.5, and thus the combination $\tilde{c}^\star(\tilde{x}^\star(S_\infty(\theta)), S_\infty(\theta))$ is also continuous on Θ. The exact same arguments can be used to show that $\tilde{c}(\tilde{x}(S_\infty(\theta)), S_\infty(\theta))$ is continuous on Θ as well. This implies that the strict inequality (3.10) remains valid under small perturbations of θ_0. Next, set $s_0 = S_\infty(\theta_0)$, and recall from Definition 2.5 that $S_\infty(\theta)$ is a local homeomorphism and thus locally surjective. Recall also from Definition 2.6 that $\tilde{x}(s)$ is quasi-continuous on $\mathbb{S}_\infty$ and therefore continuous on a dense subset of $\mathbb{S}_\infty$ [18]. By perturbing θ_0 if necessary, we may thus assume without loss of generality that $\tilde{x}(s)$ is continuous at $s_0 = S_\infty(\theta_0)$ while still maintaining the strict inequality (3.10).

Identifying x_0 with $\tilde{x}(s_0)$, we can now reuse the arguments from the proof of Theorem 3.1 that led to the inequality (3.7) to show that there exists a model $\theta_0^\star \in \Theta$ with $I(s_0, \theta_0^\star) = r_0 < r$ and

$$\tilde{c}(\tilde{x}(s_0), s_0) < c(\tilde{x}(s_0), \theta_0^\star). \tag{3.11}$$

Details are omitted to avoid redundancy. In the following we will use this inequality to show that the prediction disappointment $\mathbb{P}_{\theta_0^\star}[c(\tilde{x}(\widehat{S}_T), \theta_0^\star) > \tilde{c}(\tilde{x}(\widehat{S}_T), \widehat{S}_T)]$ of the predictor-prescriptor pair $(\tilde{c}, \tilde{x})$ under model $\theta_0^\star$ decays no faster than $e^{r_0 T}$ for large sample sizes T. As $r_0 < r$, this will imply that $(\tilde{c}, \tilde{x})$ is *in*feasible in (2.6b). To this end, we define the set of disappointing realizations of $\widehat{S}$ as

$$\mathcal{D}(\theta_0^\star) = \{s \in \mathbb{S} : c(\tilde{x}(s), \theta_0^\star) > \tilde{c}(\tilde{x}(s), s)\}.$$

By (3.11), this disappointment set contains s_0. As $s_0 \in \mathbb{S}_\infty \subseteq \mathbb{S}$ and as $\mathbb{S}_\infty$ is open, we further know that s_0 resides in the interior of $\mathbb{S}$. In addition, $\tilde{c}(x, s)$ is continuous in s on $X \times \mathbb{S}_\infty$ and $\tilde{x}(s)$ is continuous at $s_0 \in \mathbb{S}_\infty$ by the construction of s_0. We may thus conclude that s_0 belongs in fact to the interior of $\mathcal{D}(\theta_0^\star)$ and that

$$\inf_{s \in \operatorname{int} \mathcal{D}(\theta_0^\star)} I(s, \theta_0^\star) \leq I(s_0, \theta_0^\star) = r_0,$$

where the equality holds by the definition of r_0. Together with the LDP bound (3.1a), this implies that

$$-r < -r_0 \leq - \inf_{s \in \operatorname{int} \mathcal{D}(\theta_0^\star)} I(s, \theta_0^\star) \leq \liminf_{T \to \infty} \frac{1}{T} \log \mathbb{P}_{\theta_0^\star}[\widehat{S}_T \in \mathcal{D}(\theta_0^\star)].$$

By the definition of $\mathcal{D}(\theta_0^\star)$, this means that the out-of-sample disappointment of $(\tilde{c}, \tilde{x})$ corresponding to $\theta_0^\star$ decays strictly slower than e^{rT} as T grows. This in turn contradicts our assumption that $(\tilde{c}, \tilde{x})$ is feasible in (2.6b) and thus implies that $(\tilde{c}^\star, \tilde{x}^\star)$ is a Pareto dominant solution of problem (2.6b). □

3.4 Invariance under coordinate transformations

We now demonstrate that the restricted meta-optimization problems (2.6) are invariant under homeomorphic coordinate transformations of the state space $\mathbb{S}$ and the model space $\operatorname{cl} \Theta$. Note first that if $\widehat{S}$ is a statistic and $\psi : \mathbb{S} \to \mathbb{S}$ is a homeomorphism, then $\psi(\widehat{S}) = \{\psi(\widehat{S}_T)\}_{T \in \mathbb{N}}$ is also a statistic in the sense of Definition 2.5. Indeed, $\psi \circ S_\infty$ is a local homeomorphism because ψ is continuous and S_∞ is a local homeomorphism. In addition, for any fixed $\theta \in \Theta$, we know that $\widehat{S}_T$ converges in probability to $S_\infty(\theta)$ under $\mathbb{P}_\theta$. The continuous mapping theorem [26, Theorem 3.2.4] thus implies that $\psi(\widehat{S}_T)$ converges in probability to $\psi(S_\infty(\theta))$ under $\mathbb{P}_\theta$.

If the transformed statistic $\psi(\widehat{S})$ satisfies an LDP with a regular rate function, then Theorems 3.1 and 3.2 imply that the corresponding distributionally robust predictors and prescriptors must provide Pareto dominant solutions to the compressed meta-optimization problems (2.6). In the following we demonstrate that these Pareto dominant solutions corresponding to different homeomorphisms ψ are indeed all equivalent.

Proposition 3.2 (Invariance under coordinate transformations of $\mathbb{S}$)**.** *If Assumptions 2.1, 2.2 and 3.1 hold and $\psi : \mathbb{S} \to \mathbb{S}$ is a homeomorphism, then the statistic $\psi(\widehat{S})$ satisfies an LDP with regular rate function $I_\psi(s, \theta) = I(\psi^{-1}(s), \theta)$. In addition, $\tilde{c}_\psi^\star(x, s) = \tilde{c}^\star(x, \psi^{-1}(s))$ is the distributionally robust predictor induced by I_ψ, and $\tilde{x}_\psi^\star(s) = \tilde{x}^\star(\psi^{-1}(s))$ is a corresponding distributionally robust prescriptor.*

Proof. By the contraction principle [22, Theorem 4.2.1], which applies because ψ is continuous, the transformed statistic $\psi(\widehat{S})$ satisfies an LDP with rate function $I_\psi(s, \theta) = I(\psi^{-1}(s), \theta)$. Since the homeomorphism ψ has a continuous inverse and preserves compactness, one readily verifies that I_ψ inherits the radial monotonicity in θ, the continuity on $\mathbb{S} \times \Theta$, and the level-compactness from I. Thus, I_ψ is regular in the sense of Definition 3.3. By Definition 3.4, the distributionally robust predictor induced by I_ψ is given by

$$\tilde{c}_\psi^\star(x, s) = \begin{cases} \max\limits_{\theta \in \operatorname{cl} \Theta} \{c(x, \theta) : I_\psi(s, \theta) \leq r\} & \text{if } \exists\, \theta \in \operatorname{cl} \Theta \text{ with } I_\psi(s, \theta) \leq r, \\ \sup\limits_{\theta \in \operatorname{cl} \Theta} c(x, \theta) & \text{if } \nexists\, \theta \in \operatorname{cl} \Theta \text{ with } I_\psi(s, \theta) \leq r. \end{cases}$$

Clearly, we have $\tilde{c}^\star_\psi(x,s) = \tilde{c}^\star(x, \psi^{-1}(s))$ by the definition of I_ψ. Next, define $\tilde{x}^\star_\psi(s) = \tilde{x}^\star(\psi^{-1}(s))$, and note that $\tilde{x}^\star_\psi$ inherits quasi-continuity from $\tilde{x}^\star$ because ψ is continuous. As $\tilde{x}^\star$ satisfies (3.8), we further have

$$\tilde{x}^\star_\psi(s) = \tilde{x}^\star(\psi^{-1}(s)) \in \arg\min_{x\in X} \tilde{c}^\star(x, \psi^{-1}(s)) = \arg\min_{x\in X} \tilde{c}^\star_\psi(x,s),$$

and thus $\tilde{x}^\star(\psi^{-1}(s))$ is a distributionally robust prescriptor corresponding to $\tilde{c}^\star_\psi$. $\square$

Proposition 3.2 implies that the data-driven predictor induced by $\tilde{c}^\star_\psi$ and the transformed statistic $\psi(\widehat{S})$ coincides with that induced by $\tilde{c}^\star$ and the original statistic $\widehat{S}$ because $\tilde{c}^\star_\psi(x, \psi(\widehat{S}_T)) = \tilde{c}^\star(x, \widehat{S}_T)$ for all $T \in \mathbb{N}$. Similarly, we have $\tilde{x}^\star_\psi(\psi(\widehat{S}_T)) = \tilde{x}^\star(\widehat{S}_T)$ for all $T \in \mathbb{N}$. Thus, homeomorphic transformations of the estimator $\widehat{S}$ have no impact on how we map the raw data $\xi_{[T]}$ to a prediction of the cost or to a decision.

Similar invariance properties hold under coordinate transformations of the model space. To see this, note that if $\varphi : \mathrm{cl}\,\Theta \to \mathrm{cl}\,\Theta$ is a homeomorphism, then φ maps Θ onto Θ thanks to a simple generalization of [41, Exercise 5.4] and because $\mathrm{cl}\,\Theta$ is convex. This implies that the transformed ambiguity set $\mathcal{P}_\varphi = \{\mathbb{P}_{\varphi(\theta)} : \theta \in \Theta\}$ coincides with the original ambiguity set $\mathcal{P}$. We now show that the key properties of ambiguity sets, model-based predictors and regular rate functions are preserved and that the distributionally robust predictors and prescriptors are invariant under homeomorphic coordinate transformations of $\mathrm{cl}\,\Theta$.

Proposition 3.3 (Invariance under coordinate transformations of $\mathrm{cl}\,\Theta$)**.** *If Assumptions 2.1, 2.2 and 3.1 hold and $\varphi : \mathrm{cl}\,\Theta \to \mathrm{cl}\,\Theta$ is a homeomorphism, then $\mathcal{P}_\varphi = \{\mathbb{P}_{\varphi(\theta)} : \theta \in \Theta\}$ is a finitely parametrized ambiguity set in the sense of Assumption 2.1, the model-based predictor $c_\varphi(x,\theta) = c(x, \varphi^{-1}(\theta))$ satisfies Assumption 2.2, and the rate function $I_\varphi(s,\theta) = I(s, \varphi^{-1}(\theta))$ is regular. In addition, the distributionally robust predictor and any corresponding distributionally robust prescriptor are invariant under this coordinate transformation.*

Proof. The assertions concerning $\mathcal{P}_\varphi$ and $c_\varphi(x,\theta)$ follow immediately from the defining properties of a homeomorphism. In addition, the transformed rate function $I_\varphi(x,\theta)$ is non-negative and lower semi-continuous in s on $\mathbb{S} \times \mathrm{cl}\,\Theta$, and it satisfies the continuity and level-compactness conditions of Definition 3.3. All these properties are inherited from the original rate function $I(x,\theta)$ because φ^{-1} is continuous. To show that $I_\varphi(x,\theta)$ satisfies the radial monotonicity condition of Definition 3.3, we introduce the sets

$$A = \{\theta \in \Theta : I(s,\theta) < r\} \quad \text{and} \quad B = \{\theta \in \mathrm{cl}\,\Theta : I(s,\theta) \le r\}$$

and note that $\mathrm{cl}\,A = B$ because the original rate function is radially monotonic. Similarly, we introduce

$$A_\varphi = \{\theta \in \Theta : I_\varphi(s,\theta) < r\} \quad \text{and} \quad B_\varphi = \{\theta \in \mathrm{cl}\,\Theta : I_\varphi(s,\theta) \le r\}.$$

By the definition of $I_\varphi(s,\theta)$ and because φ maps Θ onto Θ, we have $A_\varphi = \varphi(A)$. Similarly, as φ maps $\mathrm{cl}\,\Theta$ onto $\mathrm{cl}\,\Theta$, we have $B_\varphi = \varphi(B)$. To prove that the new rate function is radially monotonic, we need to show that $\mathrm{cl}\,A_\varphi = B_\varphi$. As $A_\varphi \subseteq B_\varphi$ and B_φ is closed thanks to the level-compactness of $I_\varphi(s,\theta)$, we have $\mathrm{cl}\,A_\varphi \subseteq B_\varphi$. To prove the converse inclusion, select any $\theta \in B_\varphi$, and note that $\varphi^{-1}(\theta) \in B = \mathrm{cl}\,A$. Thus, there exist $\theta_k \in A$, $k \in \mathbb{N}$, such that $\lim_{k\to\infty} \theta_k = \varphi^{-1}(\theta)$. As $\varphi(A) = A_\varphi$, we then have $\varphi(\theta_k) \in A_\varphi$ for all $k \in \mathbb{N}$, and as φ is continuous, we have $\lim_{k\to\infty} \varphi(\theta_k) = \theta$. This implies that $\theta \in \mathrm{cl}\,A_\varphi$. As $\theta \in B_\varphi$ was chosen arbitrarily, we have thus shown that $B_\varphi \subseteq \mathrm{cl}\,A_\varphi$ and consequently that $I_\varphi(x,\theta)$ is regular.

The distributionally robust predictor of Definition 3.4 and the distributionally robust prescriptor of Definition 3.5 are thus manifestly invariant under homeomorphic coordinate transformations of $\mathrm{cl}\,\Theta$. $\square$

Propositions 3.2 and 3.3 testify to the reasonableness of Assumptions 2.1, 2.2 and 3.1.

4 Separation of estimation and optimization

We are now ready to tackle a fundamental question in data-driven decision-making that is of theoretical as well as practical interest: Under what conditions on the statistic $\widehat{S}$ can we restrict the class of *all* data-driven predictors and prescriptors to the subclass of all *compressed* data-driven predictors and prescriptors induced by $\widehat{S}$ without incurring any loss of optimality? In other words, we aim to identify conditions under which any decision-relevant information contained in the raw data $\xi_{[T]}$ is also contained in the summary statistic $\widehat{S}_T$ for every $T \in \mathbb{N}$, such that the meta-optimization problems (2.4a) and (2.4b) become equivalent to (2.6a) and (2.6b), respectively. In statistical estimation it is well known that the possibility of lossless compression is intimately related to the existence of a sufficient statistic; see, *e.g.*, [39]. In the following, we will argue that such a result also holds in the context of data-driven decision-making. Although this result has intuitive appeal, it seems not to have been established before, and we find it surprisingly difficult to prove.

Definition 4.1 (Sufficient statistic)**.** *A statistic $\widehat{S}$ with state space $\mathbb{S}$ is called sufficient for θ if the conditional distribution of $\xi_{[T]}$ given $\widehat{S}_T = s$ under $\mathbb{P}_\theta$ is independent of $\theta \in \Theta$ for all $s \in \mathbb{S}$ and $T \in \mathbb{N}$.*

Intuitively, $\widehat{S}$ is a sufficient statistic for θ if knowing the full observation history $\xi_{[T]}$ provides no advantage for estimating θ over only knowing $\widehat{S}_T$. In other words, compressing $\xi_{[T]}$ into $\widehat{S}_T$, which is equivalent to a Borel-measurable function of $\xi_{[T]}$, does not destroy any information that could be useful for estimating θ. The Pitman-Koopman-Darmois theorem [34] implies that if the observed data is i.i.d. over time, then there exists a sufficient statistic if and only if the data generation process belongs to an exponential family. Even though we do not restrict attention to i.i.d. processes, this result prompts us to require that the ambiguity set $\mathcal{P}$ represents an exponential family of stochastic processes. To formalize this requirement, we henceforth denote by $\mathbb{P}_\theta^T$ the restriction of the probability measure $\mathbb{P}_\theta$ to the σ-algebra $\mathcal{F}_T$ generated by $\xi_{[T]}$ for all $T \in \mathbb{N}$ and $\theta \in \Theta$. In addition, for any $T \in \mathbb{N}$ we define the log-moment generating function $\Lambda_T : \mathbb{R}^d \times \Theta \to (-\infty, +\infty]$ of $\widehat{S}_T$ through $\Lambda_T(\lambda, \theta) = \log \mathbb{E}_\theta[\exp(\langle \lambda, \widehat{S}_T \rangle)]$ if the expectation is finite and $\Lambda_T(\lambda, \theta) = +\infty$ otherwise. As $\Lambda_T(0, \theta) = 0$ by construction, the function $\Lambda_T(\lambda, \theta)$ is proper in λ. Moreover, $\Lambda_T(\lambda, \theta)$ is convex and lower semi-continuous in λ thanks to [7, Theorem 7.1].

Assumption 4.1 (Exponential family of stochastic processes)**.** *The ambiguity set $\mathcal{P}$ represents a time-homogeneous exponential family of stochastic processes. This means that there exist a baseline model $\bar{\theta} \in \Theta$, a continuous parametrization function $g : \Theta \to \mathbb{R}^d$ and a sequence of log-partition functions $A_T : \mathbb{R}^d \to (-\infty, +\infty]$ for $T \in \mathbb{N}$ defined through $A_T(\lambda) = \Lambda_T(\lambda, \bar{\theta})$ such that $Tg(\theta) \in \mathrm{dom}(A_T)$ for all $\theta \in \Theta$ and*

$$\frac{\mathrm{d}\mathbb{P}_\theta^T}{\mathrm{d}\mathbb{P}_{\bar{\theta}}^T} = \exp\Big(\langle Tg(\theta), \widehat{S}_T \rangle - A_T(Tg(\theta)) \Big) \quad \forall T \in \mathbb{N},\ \theta \in \Theta. \tag{4.1}$$

Exponential families that obey Assumption 4.1 are called time-homogeneous because the parametrization function g is independent of T [35, Section 3.1]. As the Radon-Nikodym derivative (4.1) is strictly positive, all probability measures within a given exponential family are mutually equivalent. For every $T \in \mathbb{N}$ and $\theta \in \Theta$, the log-partition function A_T ensures that the probability measure $\mathbb{P}_\theta^T$ is normalized, and it inherits properness, convexity and lower semi-continuity from the log-moment generating function Λ_T. Even though the log-partition function was defined as the log-moment generating function corresponding to the baseline model $\bar{\theta}$, any other log-moment generating function corresponding to an arbitrary model $\theta \in \Theta$ can be recovered from A_T. This follows from the change of measure formula (4.1) and the observation that expectations of $\mathcal{F}_T$-measurable functions with respect to $\mathbb{P}_\theta$ depend only on the restriction of $\mathbb{P}_\theta$ to $\mathcal{F}_T$, *i.e.*,

$$\begin{aligned}\Lambda_T(\lambda,\theta) &= \log \mathbb{E}_{\bar\theta}\left[\exp\left(\langle \lambda, \widehat{S}_T\rangle + \langle Tg(\theta), \widehat{S}_T\rangle - A_T(Tg(\theta))\right)\right] \\ &= A_T(\lambda + Tg(\theta)) - A_T(Tg(\theta)).\end{aligned} \tag{4.2}$$

Assumption 4.1 guarantees via the Fisher-Neyman factorization theorem [39, Theorem 6.5] that the statistic $\widehat{S}$ is sufficient. This can also be verified directly. Indeed, if it is known that $\widehat{S}_T = s$ for some $s \in \mathbb{S}$, then the Radon-Nikodym derivative (4.1) reduces to a deterministic function, and therefore the conditional distribution of $\xi_{[T]}$ given $\widehat{S}_T = s$ is identical under $\mathbb{P}^T_\theta$ and $\mathbb{P}^T_{\bar\theta}$, that is, it does not depend on $\theta \in \Theta$. As this argument holds for every $s \in \mathbb{S}$ and $T \in \mathbb{N}$, we may conclude that $\widehat{S}$ is indeed a sufficient statistic.

The next assumption will ensure via the celebrated Gärtner-Ellis theorem that $\widehat{S}$ also satisfies an LDP.

Assumption 4.2 (Log-moment generating functions)**.** *The log-moment generating functions corresponding to the statistic $\widehat{S}$ display the following properties. First, we have $\Lambda_T(\lambda,\theta) < \infty$ for all $\lambda \in \mathbb{R}^d$ and $T \in \mathbb{N}$, and the limiting log-moment generating function $\Lambda : \mathbb{R}^d \times \Theta \to (-\infty, \infty]$ defined as the limit*

$$\Lambda(\lambda,\theta) = \lim_{T\to\infty} \frac{1}{T}\Lambda_T(T\lambda,\theta) \tag{4.3}$$

exists as an extended real number for all $\lambda \in \mathbb{R}^d$ and $\theta \in \Theta$. In addition, the origin belongs to the interior of $\operatorname{dom}\Lambda(\cdot,\theta)$ *for all $\theta \in \Theta$. Finally, the gradient $\nabla_\lambda \Lambda(\lambda,\theta)$ exists on the interior of* $\operatorname{dom}\Lambda(\cdot,\theta)$*, and its norm tends to infinity when λ approaches the boundary of* $\operatorname{dom}\Lambda(\cdot,\theta)$ *for all $\theta \in \Theta$.*

As $\Lambda_T(0,\theta) = 0$ for all $T \in \mathbb{N}$, it is clear that $\Lambda(0,\theta) = 0$, that is, the origin belongs to $\operatorname{dom}\Lambda(\cdot,\theta)$. Note that Assumption 4.2 imposes the stronger condition that the origin belongs to the interior of $\operatorname{dom}\Lambda(\cdot,\theta)$. Recall next that the log-moment generating functions $\Lambda_T(\lambda,\theta)$ are convex in λ for all $T \in \mathbb{N}$. By [22, Lemma 2.3.9], their asymptotic counterpart $\Lambda(\lambda,\theta)$ inherits convexity in λ. Assumption 4.2 further stipulates that $\operatorname{dom}\Lambda_T(\cdot,\theta) = \mathbb{R}^d$, which implies via [7, Theorem 7.2] that $\Lambda_T(\lambda,\theta)$ is analytical in λ, throughout all of $\mathbb{R}^d$, for all $T \in \mathbb{N}$. By leveraging the dominated convergence theorem, it is then easy to prove that $\nabla_\lambda \Lambda_T(0,\theta) = \mathbb{E}_\theta[\widehat{S}_T]$ for all $T \in \mathbb{N}$. The following lemma extends this result to the gradient $\nabla_\lambda\Lambda(0,\theta)$ of the limiting log-moment generating function, which exists thanks to Assumption 4.2.

Lemma 4.1. *If Assumption 4.2 holds, then we have $\nabla_\lambda\Lambda(0,\theta) = \lim_{T\to\infty}\mathbb{E}_\theta[\widehat{S}_T]$ for all $\theta \in \Theta$.*

Proof. Fix any $\theta \in \Theta$, and recall that $\mathbb{E}_\theta[\widehat{S}_T] = \nabla_\lambda\Lambda_T(0,\theta) = \frac{1}{T}\nabla_\lambda[\Lambda_T(T\lambda,\theta)]_{\lambda=0}$ for all $T \in \mathbb{N}$. Driving T to infinity, the claim follows if we can interchange the limit and the gradient on the right hand side of this identity to obtain $\lim_{T\to\infty}\nabla_\lambda[\frac{1}{T}\Lambda_T(T\lambda,\theta)]_{\lambda=0} = \nabla_\lambda\Lambda(0,\theta)$. To this end, select a tolerance $\varepsilon > 0$ and a direction $b \in \mathbb{R}^d$. By the definition of the directional derivative of $\Lambda(\lambda,\theta)$ at $\lambda = 0$ along the direction b, there exists a step size $h > 0$ such that $(\Lambda(hb,\theta) - \Lambda(0,\theta))/h < \langle\nabla_\lambda\Lambda(0,\theta), b\rangle + \varepsilon$. In addition, by the definition (4.3) of the limiting log-moment generating function Λ, there exists $T_0 \in \mathbb{N}$ such that

$$(\Lambda_T(hbT,\theta) - \Lambda_T(0,\theta))/(Th) < \langle\nabla_\lambda\Lambda(0,\theta), b\rangle + \varepsilon \quad \forall T \geq T_0.$$

Next, the first-order condition of convexity for $\Lambda_T(T\lambda,\theta)/T$ guarantees that

$$\langle\nabla_\lambda\Lambda_T(0,\theta), b\rangle \leq (\Lambda_T(hbT,\theta) - \Lambda_T(0,\theta))/(Th) \quad \forall T \in \mathbb{N}.$$

Combining the last two inequalities then yields the estimate $\lim_{T\to\infty}\langle\nabla_\lambda\Lambda_T(0,\theta), b\rangle < \langle\nabla_\lambda\Lambda(0,\theta), b\rangle + \varepsilon$. As $\varepsilon > 0$ was chosen arbitrarily, this implies that $\lim_{T\to\infty}\langle\nabla_\lambda\Lambda_T(0,\theta), b\rangle \leq \langle\nabla_\lambda\Lambda(0,\theta), b\rangle$, and as $b \in \mathbb{R}^d$ was also chosen arbitrarily, we may in fact conclude that $\lim_{T\to\infty}\nabla_\lambda\Lambda_T(0,\theta) = \nabla_\lambda\Lambda(0,\theta)$. $\square$

Remark 4.1. *Lemma 4.1 admits the following generalization. If Assumptions 4.1 and 4.2 hold and $\eta = g(\theta') - g(\theta)$ for some $\theta, \theta' \in \Theta$, then one can proceed as in the proof of Lemma 4.1 to show that*

$$\nabla_\lambda[\Lambda(\lambda, \theta)]_{\lambda=\eta} = \lim_{T\to\infty} \mathbb{E}_\theta[\widehat{S}_T \cdot \exp(\langle \eta, T\widehat{S}_T\rangle - \Lambda_T(T\eta, \theta))].$$

The following example shows that Assumptions 4.1 and 4.2 are satisfied if the observable data is governed by an i.i.d. process with a finite state space and if $\widehat{S}_T$ denotes the empirical distribution.

Example 4.1 (Exponential families of finite state i.i.d. processes)**.** *Consider the class of finite state i.i.d. processes introduced in Example 2.2, and let $\widehat{S}_T$ be the empirical distribution defined in Example 2.4. In this case, the Assumptions 4.1 and 4.2 are satisfied. To see this, set the baseline model θ_0 to the uniform probability vector, that is, set $(\theta_0)_i = 1/d$ for all $i = 1, \dots d$. Recalling that the probability of observing $\xi_{[T]}$ is given by $\prod_{t=1}^T \theta_{\xi_t}$ under an arbitrary model $\theta \in \Theta$ and by d^{-T} under the baseline model $\bar{\theta}$, we then find*

$$\frac{\mathrm{d}\mathbb{P}_\theta^T}{\mathrm{d}\mathbb{P}_{\bar{\theta}}^T} = d^T \prod_{t=1}^T \theta_{\xi_t} = d^T \prod_{j=1}^d \theta_j^{\sum_{t=1}^T 1_{\xi_t=j}} = \exp(\langle T\log\theta, \widehat{S}_T\rangle + T\log d),$$

where $\log\theta$ is evaluated component-wise. In addition, the T^{th} log-moment generating function is given by

$$\begin{aligned}\Lambda_T(\lambda, \theta) &= \log \mathbb{E}_\theta\left[\exp(\tfrac{1}{T}\textstyle\sum_{t=1}^T \sum_{j=1}^d \lambda_j 1_{\xi_t=j})\right]\\ &= \log \mathbb{E}_\theta\left[\textstyle\prod_{t=1}^T \exp(\tfrac{1}{T}\sum_{j=1}^d \lambda_j 1_{\xi_t=j})\right]\\ &= T\log \mathbb{E}_\theta\left[\exp(\tfrac{1}{T}\textstyle\sum_{j=1}^d \lambda_j 1_{\xi_1=j})\right]\\ &= T\log \textstyle\sum_{i=1}^d \theta_i \exp(\tfrac{1}{T}\sum_{j=1}^d \lambda_j 1_{i=j}) = T\log\sum_{i=1}^d \theta_i e^{\lambda_i/T},\end{aligned}$$

where the second equality follows from the serial independence of the observations, and the third inequality holds because all observations have the same marginal distribution as ξ_1. Thus, the family of all finite state i.i.d. processes corresponding to the models $\theta \in \Theta$ form a time-homogeneous exponential family with parametrization function $g(\theta) = \log\theta$ and log-partition function $A_T(\lambda) = \Lambda_T(\lambda, \theta_0) = T\log\frac{1}{d}\sum_{i=1}^d e^{\lambda_i/T}$, which ensures that $A_T(Tg(\theta)) = -T\log d$. This confirms Assumption 4.1 and consequently shows that the empirical distribution is a sufficient statistic for θ. Next, observe that $\Lambda_T(\lambda, \theta) < \infty$ for all $T \in \mathbb{N}$ and $\lambda \in \mathbb{R}^d$,

$$\Lambda(\lambda, \theta) = \lim_{T\to\infty} \frac{1}{T}\Lambda_T(T\lambda, \theta) = \log\sum_{i=1}^d \theta_i e^{\lambda_i} < \infty \quad \forall \lambda \in \mathbb{R}^d$$

and $\partial_{\lambda_i}\Lambda(\lambda, \theta) = \theta_i e^{\lambda_i}/(\sum_{j=1}^d \theta_j e^{\lambda_j})$ for all $i = 1, \dots, d$ and $\lambda \in \mathbb{R}^d$. Therefore, $\Lambda(\lambda, \theta)$ is smooth and convex in λ and continuous in θ on $\mathbb{R}^d \times \Theta$. These findings imply that Assumption 4.2 holds.

Assumption 4.2 guarantees via the celebrated Gärtner-Ellis theorem that $\widehat{S}$ satisfies an LDP.

Theorem 4.1 (Gärtner-Ellis theorem [22, Theorem 2.3.6])**.** *If the limiting log-moment generating function Λ satisfies Assumption 4.2, then the statistic $\widehat{S}$ satisfies an LDP with good rate function*

$$I(s, \theta) = \sup_{\lambda\in\mathbb{R}^d} \langle \lambda, s\rangle - \Lambda(\lambda, \theta). \tag{4.4}$$

Note that the limiting log-moment generating function $\Lambda(\lambda, \theta)$ and the rate function $I(s, \theta)$ of Theorem 4.1 are only defined on $\mathbb{R}^d \times \Theta$. However, it is usually easy to extend $I(s, \theta)$ to $\mathbb{R}^d \times \mathrm{cl}\,\Theta$ so that it becomes a rate function in the sense of Definition 3.1. In Section 5 we will provide several examples where $I(s,\theta)$ can even be extended to a *regular* rate function on $\mathbb{R}^d \times \mathrm{cl}\,\Theta$. Note that $I(s,\theta)$ displays the following properties for every fixed $\theta \in \Theta$. First, it coincides with the convex conjugate of the limiting log-moment generating function $\Lambda(\lambda,\theta)$ with respect to λ. Consequently, $I(s,\theta)$ represents a pointwise supremum of affine functions and is thus convex and lower semi-continuous in s. By Assumption 4.2, $\Lambda(\lambda,\theta)$ is essentially smooth in λ, that is, the gradient $\nabla_\lambda \Lambda(\lambda,\theta)$ exists on the interior of $\mathrm{dom}\,\Lambda(\cdot,\theta)$, and its norm tends to infinity when λ approaches the boundary of $\mathrm{dom}\,\Lambda(\cdot,\theta)$. This implies via [45, Theorem 26.3] that the rate function $I(s,\theta)$ is strictly convex in s on the relative interior of $\mathrm{dom}\, I(\cdot,\theta)$. Conversely, if $I(s,\theta)$ is strictly convex in s, then the same theorem guarantees that $\Lambda(\lambda,\theta)$ is essentially smooth in λ. This implication is sometimes useful to verify Assumption 4.2. As $\Lambda(0,\theta)=0$, we further have $I(s,\theta)\geq 0$ for all $s\in\mathbb{S}$. Finally, as we will show in the following lemma, Assumption 4.2 implies via the Gärtner-Ellis theorem that $S_\infty(\theta) = \nabla_\lambda\Lambda(0,\theta)$.

Lemma 4.2 (Asymptotic consistency of $\widehat{S}$)**.** *If Assumption 4.2 holds, then, as T grows, $\widehat{S}_T$ converges in probability under $\mathbb{P}_\theta$ to $\nabla_\lambda\Lambda(0,\theta)$ for every $\theta\in\Theta$. This implies that $S_\infty(\theta) = \nabla_\lambda\Lambda(0,\theta)$.*

Proof. Fix any $\theta\in\Theta$ and $\varepsilon>0$, and set $s_0 = \nabla_\lambda\Lambda(0,\theta)$. By Theorem 4.1 we have

$$\limsup_{T\to\infty} \frac{1}{T}\log \mathbb{P}_\theta[\|\widehat{S}_T - s_0\| \geq \varepsilon] \leq -\inf_{s\in\mathbb{R}^d}\{I(s,\theta) : \|s-s_0\|\geq\varepsilon\} = -I(s^\star,\theta) \tag{4.5}$$

for some $s^\star \neq s_0$, where the equality holds because the good rate function $I(s,\theta)$ has compact sublevel sets. Next, we show that $I(s^\star,\theta)>0$. Suppose for the sake of argument that $I(s^\star,\theta)=0$. This implies that

$$\langle \lambda, s^\star\rangle = \langle \lambda, s^\star\rangle - I(s^\star,\theta) \leq \Lambda(\lambda,\theta) = \Lambda(\lambda,\theta) - \Lambda(0,\theta) \quad \forall \lambda\in\mathbb{R}^d,$$

where the inequality exploits the definition of $I(s^\star,\theta)$. Setting $\lambda = \delta v$ for $\delta>0$ and $v\in\mathbb{R}^d$, we then find

$$\langle v, s^\star\rangle \leq \lim_{\delta\downarrow 0}\frac{1}{\delta}\left(\Lambda(\delta v,\theta)-\Lambda(0,\theta)\right) = \langle v, \nabla_\lambda\Lambda(0,\theta)\rangle \quad \forall v\in\mathbb{R}^d,$$

which in turn implies that $s^\star = \nabla_\lambda\Lambda(0,\theta) = s_0$ and thus contradicts the construction of $s^\star$. We therefore conclude that $I(x^\star,\theta)>0$, which ensures via (4.5) that $\mathbb{P}_\theta[\|\widehat{S}_T - s_0\|\geq\varepsilon] \leq \exp(-T\cdot I(x^\star,\theta)+o(T))$ tends to 0 as T grows. The claim then follows because $\theta\in\Theta$ and $\varepsilon>0$ were chosen arbitrarily. $\square$

The following example shows that the Gärtner-Ellis theorem subsumes Sanov's theorem as a special case.

Example 4.2 (An LDP for finite state i.i.d. processes revisited)**.** *Consider the class of finite state i.i.d. processes of Example 2.2, and let $\widehat{S}_T$ be the empirical distribution defined in Example 2.4. From Example 4.1 we know that the limiting log-moment generating function is given by $\Lambda(\lambda,\theta) = \log\sum_{i=1}^d \theta_i e^{\lambda_i}$ and that Assumptions 4.1 and 4.2 are satisfied. By Theorem 4.1, $\widehat{S}$ thus satisfies an LPD with good rate function*

$$I(s,\theta) = \sup_{\eta\in\mathbb{R}^d} \langle \eta, s\rangle - \Lambda(\eta,\theta) = D(s\|\theta)\,,$$

where the second equality follows from an elementary but tedious calculation. This reasoning reveals that Sanov's theorem [22, Theorem 2.1.10], which describes an LDP for the empirical distributions on i.i.d. data, emerges as a special case of the Gärtner-Ellis theorem. Recall also from Example 3.1 that the relative entropy admits a lower semi-continuous extension to $\mathbb{S}\times\mathrm{cl}\,\Theta = \Delta_d\times\Delta_d$ and constitutes a regular rate function.

We will now demonstrate that if the statistic $\widehat{S}$ not only satisfies an LDP with a regular rate function but is also sufficient, then even the original meta-optimization problems (2.4) admit Pareto dominant solutions that are available in closed form. To this end, denote as usual by $\tilde{c}^\star$ the distributionally robust predictor of Definition 3.4, and introduce a data-driven predictor $\widehat{c}^\star$ defined through $\widehat{c}^\star_T(x) = \tilde{c}^\star(x, \widehat{S}_T)$ for all $T \in \mathbb{N}$.

Theorem 4.2 (Optimality of $\widehat{c}^\star$)**.** *If the Assumptions 2.1, 2.2, 4.1 and 4.2 hold, the rate function* (4.4) *is regular and $r > 0$, then $\widehat{c}^\star$ is a Pareto dominant solution of the meta-optimization problem* (2.4a)*.*

The assumptions of Theorem 4.2 ensure via the Gärtner-Ellis theorem that $\widehat{S}$ satisfies an LDP, and thus they imply the assumptions of Theorem 3.1. From Theorem 3.1 we further know that $\tilde{c}^\star$ represents a Pareto dominant solution to the restricted meta-optimization problem (2.6a) over compressed data-driven predictors. The discussion after Example 2.5 finally implies that the objective function value of $\widehat{c}^\star$ in (2.4a) coincides with that of $\tilde{c}^\star$ in (2.6a) for every fixed decision $x \in X$ and model $\theta \in \Theta$, that is, we have

$$\lim_{T\to\infty} \mathbb{E}_\theta[\widehat{c}^\star_T(x)] = \tilde{c}^\star(x, S_\infty(\theta)).$$

As Theorem 4.2 identifies $\widehat{c}^\star$ as a Pareto dominant solution to (2.4a), the above identity thus implies that the original meta-optimization problem (2.4a) is indeed equivalent to the restricted meta-optimization problem (2.6a). In other words, compressing the raw data $\xi_{[T]}$ into $\widehat{S}_T$ incurs no loss of optimality.

Theorem 4.2 can be interpreted as establishing a separation principle that enables a decoupling of estimation and optimization. Instead of directly solving a data-driven optimization problem of the form $\min_{x\in X} \widehat{c}_T(x)$ constructed from the raw data $\xi_{[T]}$, which may become increasingly difficult as T grows, we can first solve an estimation problem that evaluates the statistic $\widehat{S}_T$ and subsequently solve an optimization problem $\min_{x\in X} \tilde{c}(x, \widehat{S}_T)$ constructed merely from $\widehat{S}_T$. Theorem 4.2 guarantees that if these two data-driven optimization problems are designed optimally, then no optimality is sacrificed by this separation.

Proof of Theorem 4.2. As $\tilde{c}^\star \in \tilde{\mathcal{C}}$ is a compressed data-driven predictor in the sense of Definition 2.6, $\widehat{c}^\star$ constitutes a data-driven predictor in the sense of Definition 2.2, that is, $\widehat{c}^\star \in \widehat{\mathcal{C}}$. This follows from the discussion after Definition 2.6. Similarly, as $\tilde{c}^\star$ satisfies the rate constraint in (2.6a), one readily verifies that $\widehat{c}^\star$ satisfies the rate constraint in (2.4a). We may thus conclude that $\widehat{c}^\star$ is feasible in (2.4a). In the remainder of the proof we will show that $\widehat{c}^\star$ Pareto dominates every other feasible solution of problem (2.4a).

Assume for the sake of contradiction that there exists a data-driven predictor $\widehat{c}$ that is feasible in (2.4a) but not dominated by $\widehat{c}^\star$. Hence, there exist a decision $x_0 \in X$ and a model $\theta_0 \in \Theta$ with

$$\lim_{T\to\infty} \mathbb{E}_{\theta_0}[\widehat{c}^\star_T(x_0)] - \lim_{T\to\infty} \mathbb{E}_{\theta_0}[\widehat{c}_T(x_0)] > 0.$$

As we will see, the above inequality implies that $\widehat{c}$ is infeasible in (2.4a). This contradiction will reveal that our initial assumption must have been false and that there cannot be any feasible $\widehat{c}$ that Pareto dominates $\widehat{c}^\star$.

Since the data-driven predictor $\widehat{c}$ must satisfy the conditions (iii) and (iv) of Definition 2.2, we may conclude via Lemma B.1 that $\lim_{T\to\infty} \mathbb{E}_{\theta_0}[\widehat{c}_T(x_0)] = c_\infty(x_0, \theta_0)$, where c_∞ is the Borel-measurable function whose existence is postulated in Definition 2.2(iv). Similarly, since $\widehat{S}_T$ converges in probability under $\mathbb{P}_{\theta_0}$ to $S_\infty(\theta_0)$ and since Proposition 3.1 ensures that $\tilde{c}^\star(x_0, s)$ is bounded and continuous in s on $\mathbb{S}_\infty$, we may invoke the continuous mapping theorem [26, Theorem 3.2.4] and the fact that convergence in probability implies convergence in distribution to conclude that $\lim_{T\to\infty} \mathbb{E}_{\theta_0}[\widehat{c}^\star_T(x_0)] = \lim_{T\to\infty} \mathbb{E}_{\theta_0}[\tilde{c}^\star(x_0, \widehat{S}_T)] = \tilde{c}^\star(x_0, S_\infty(\theta_0))$. In summary, we have thus shown that $\tilde{c}^\star(x_0, S_\infty(\theta_0)) - c_\infty(x_0, \theta_0) > 0$.

Defining $s_0 = S_\infty(\theta_0)$, we may reuse the reasoning at the beginning of Step 2 in the proof of Theorem 3.1 to show that there exists $\theta_0^\star \in \Theta$ with $I(s_0, \theta_0^\star) = r_0 < r$ and $c(x_0, \theta_0^\star) - c_\infty(x_0, \theta_0) > 0$. In the following, we select any $\varepsilon > 0$ that is strictly smaller than $c(x_0, \theta_0^\star) - c_\infty(x_0, \theta_0)$ and any $\delta > 0$. Thus, we have

$$\begin{aligned} \mathbb{P}_{\theta_0^\star}\left[\widehat{c}_T(x_0) < c(x_0, \theta_0^\star)\right] &\geq \mathbb{P}_{\theta_0^\star}\left[\widehat{c}_T(x_0) < c(x_0, \theta_0^\star) \ \wedge \ \widehat{S}_T \in \mathbb{B}_\delta(s_0)\right] \\ &\geq \mathbb{P}_{\theta_0^\star}\left[\widehat{c}_T(x_0) < c_\infty(x_0, \theta_0) + \varepsilon \ \wedge \ \widehat{S}_T \in \mathbb{B}_\delta(s_0)\right] \\ &\geq \mathbb{P}_{\theta_0^\star}\left[\left|\widehat{c}_T(x_0) - c_\infty(x_0, \theta_0)\right| < \varepsilon \ \wedge \ \widehat{S}_T \in \mathbb{B}_\delta(s_0)\right], \end{aligned} \tag{4.6}$$

where $\mathbb{B}_\delta(s_0)$ denotes the Euclidean ball of radius δ around s_0. Here, the second inequality holds because $c_\infty(x_0, \theta_0) + \varepsilon < c(x_0, \theta_0^\star)$ thanks to the choice of ε. The other two inequalities are elementary. By Assumption 4.1, the probability measures $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_{\theta_0^\star}$ both belong to an exponential family of the form (4.1) and are therefore equivalent. The chain rule for Radon-Nikodym derivatives thus implies that

$$\frac{\mathrm{d}\mathbb{P}^T_{\theta_0}}{\mathrm{d}\mathbb{P}^T_{\theta_0^\star}} = \frac{\mathrm{d}\mathbb{P}^T_{\theta_0}}{\mathrm{d}\mathbb{P}^T_{\bar\theta}} \left(\frac{\mathrm{d}\mathbb{P}^T_{\theta_0^\star}}{\mathrm{d}\mathbb{P}^T_{\bar\theta}}\right)^{-1} = \frac{\exp\left(\langle Tg(\theta_0), \widehat{S}_T\rangle - A_T(Tg(\theta_0))\right)}{\exp\left(\langle Tg(\theta_0^\star), \widehat{S}_T\rangle - A_T(Tg(\theta_0^\star))\right)} = \exp\left(\langle \eta, T\widehat{S}_T\rangle - \Lambda_T(T\eta, \theta_0^\star)\right), \tag{4.7}$$

where $\eta = g(\theta_0) - g(\theta_0^\star) \in \mathbb{R}^d$ characterizes an exponential tilting between $\mathbb{P}_{\theta_0}$ and $\mathbb{P}_{\theta_0^\star}$. Here, the second equality follows from (4.1), while the third equality exploits the relation (4.2), which ensures that

$$A_T(Tg(\theta_0)) - A_T(Tg(\theta_0^\star)) = A_T(T\eta + g(T\theta_0^\star)) - A_T(Tg(\theta_0^\star)) = \Lambda_T(T\eta, \theta_0^\star).$$

In order to simplify the subsequent arguments, we introduce the $\mathcal{F}_T$-measurable Bernoulli random variable

$$\widehat{\zeta}_T = \begin{cases} 1 & \text{if } |\widehat{c}_T(x_0) - c_\infty(x_0, \theta_0)| < \varepsilon \text{ and } \widehat{S}_T \in \mathbb{B}_\delta(s_0), \\ 0 & \text{otherwise.} \end{cases}$$

Combining all preparatory results derived so far, we then obtain

$$\begin{aligned} &\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\theta_0^\star}\left[\widehat{c}_T(x_0) < c(x_0, \theta_0^\star)\right] \\ &\quad \geq \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\theta_0^\star}\left[\left|\widehat{c}_T(x_0) - c_\infty(x_0, \theta_0)\right| < \varepsilon \ \wedge \ \widehat{S}_T \in \mathbb{B}_\delta(s_0)\right] && \text{(4.8a)} \\ &\quad = \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{E}_{\theta_0^\star}\left[\widehat{\zeta}_T\right] && \text{(4.8b)} \\ &\quad = \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{E}_{\theta_0^\star}\left[\widehat{\zeta}_T \cdot \exp(-\langle \eta, T\widehat{S}_T\rangle + \Lambda_T(T\eta, \theta_0^\star)) \cdot \exp(\langle \eta, T\widehat{S}_T\rangle - \Lambda_T(T\eta, \theta_0^\star))\right] && \text{(4.8c)} \\ &\quad = \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{E}_{\theta_0}\left[\widehat{\zeta}_T \cdot \exp(-\langle \eta, T\widehat{S}_T\rangle + \Lambda_T(T\eta, \theta_0^\star))\right] && \text{(4.8d)} \\ &\quad = \Lambda(\eta, \theta_0^\star) - \langle \eta, s_0\rangle + \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{E}_{\theta_0}\left[\widehat{\zeta}_T \cdot \exp(\langle T\eta, s_0 - \widehat{S}_T\rangle)\right], && \text{(4.8e)} \end{aligned}$$

where (4.8a) and (4.8b) follow from (4.6) and the definition of $\widehat{\zeta}_T$, respectively, (4.8c) is obtained by multiplying $\widehat{\zeta}_T$ by 1, and (4.8d) follows from (4.7) and the Radon-Nikodym theorem. Equation (4.8e), finally, is obtained by extracting the deterministic factor $\exp(\Lambda_T(T\eta, \theta_0^\star) - \langle T\eta, s_0\rangle)$ from the expectation and recalling that $\lim_{T\to\infty} \frac{1}{T}\Lambda_T(T\eta, \theta_0^\star) = \Lambda(\eta, \theta_0^\star)$. As $\widehat{\zeta}_T = 1$ only if $\|\widehat{S}_T - s_0\| \leq \delta$, the Cauchy-Schwartz inequality implies that $\langle \eta, s_0 - \widehat{S}_T \rangle \geq -\|\eta\|\delta$ whenever $\widehat{\zeta}_T = 1$. Thus, (4.8) implies that

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\theta_0^\star} \left[\widehat{c}_T(x_0) < c(x_0, \theta_0^\star)\right] \\
&\quad \geq \Lambda(\eta, \theta_0^\star) - \big\langle \eta, s_0 \big\rangle - \|\eta\|\delta + \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{E}_{\theta_0} \left[\widehat{\zeta}\right] \\
&\quad \geq \Lambda(\eta, \theta_0^\star) - \big\langle \eta, s_0 \big\rangle - \|\eta\|\delta \\
&\qquad\quad + \limsup_{T\to\infty} \frac{1}{T} \log \left(\mathbb{P}_{\theta_0} \left[|\widehat{c}_T(x_0) - c_\infty(x_0, \theta_0)| < \varepsilon\right] + \mathbb{P}_{\theta_0} \left[\widehat{S}_T \in \mathbb{B}_\delta(s_0)\right] - 1 \right) \\
&\quad \geq \Lambda(\eta, \theta_0^\star) - \big\langle \eta, s_0 \big\rangle - \|\eta\|\delta,
\end{aligned} \tag{4.9}$$

where the second inequality follows from the definition of the random variable $\widehat{\zeta}_T$ and the elementary insight that $\mathbb{P}_{\theta_0}[A \cap B] \geq \mathbb{P}_{\theta_0}[A] + \mathbb{P}_{\theta_0}[B] - 1$ for all events $A, B \in \mathcal{F}$. The last inequality holds because

$$\lim_{T\to\infty} \mathbb{P}_{\theta_0} \left[|\widehat{c}_T(x_0) - c_\infty(x_0, \theta_0)| < \varepsilon\right] = 1$$

thanks to the definition of a data-driven predictor (see Definition 2.2(iv)) and because

$$\lim_{T\to\infty} \mathbb{P}_{\theta_0} \left[\widehat{S}_T \in \mathbb{B}_\delta(s_0)\right] = 1 \quad \forall \delta > 0$$

thanks to the definition of a statistic (see Definition 2.5) and the definition of $s_0 = S_\infty(\theta_0)$. As (4.9) holds for every $\delta > 0$, we have effectively shown that

$$\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_{\theta_0^\star} \left[\widehat{c}_T(x_0) < c(x_0, \theta_0^\star)\right] \geq \Lambda(\eta, \theta_0^\star) - \big\langle \eta, s_0 \big\rangle = -I(s_0, \theta_0^\star) = -r_0 > -r,$$

which contradicts the feasibility of $\widehat{c}_T$ in (2.4a). Here, the first equality holds because $\eta = g(\theta_0) - g(\theta_0^\star)$ is a maximizer of the unconstrained convex optimization problem on the right hand side of (4.4) at $s = s_0$ and $\theta = \theta_0^\star$, which defines the rate function of the Gärtner-Ellis theorem. To see this, note that

$$\begin{aligned}
\nabla_\lambda \left[\big\langle \lambda, s_0 \big\rangle - \Lambda(\lambda, \theta_0^\star)\right]_{\lambda=\eta} &= s_0 - \nabla_\lambda \left[\Lambda(\lambda, \theta_0^\star)\right]_{\lambda=\eta} \\
&= s_0 - \lim_{T\to\infty} \mathbb{E}_{\theta_0^\star}[\widehat{S}_T \cdot \exp(\langle \eta, T\widehat{S}_T \rangle - \Lambda_T(T\eta, \theta_0^\star))] \\
&= s_0 - \lim_{T\to\infty} \mathbb{E}_{\theta_0}[\widehat{S}_T] = s_0 - \nabla_\lambda \left[\Lambda(\lambda, \theta_0)\right]_{\lambda=0} = 0,
\end{aligned}$$

where the second equality holds due to Remark 4.1, the third equality follows from (4.7) and the Radon-Nikodym theorem, the fourth equality exploits Lemma 4.1, and the fifth equality follows from Lemma 4.2 and the definition of s_0. In summary, we may conclude that our initial assumption was false and that $\widehat{c}^\star$ indeed Pareto dominates every other feasible solution of problem (2.4a). $\square$

Next, we show that the meta-optimization problem (2.4b) over data-driven predictor-prescriptor pairs also admits a Pareto dominant solution. To this end, define the distributionally robust predictor $\tilde{c}^\star$ and the corresponding data-driven predictor $\widehat{c}^\star$ as before, and let $\tilde{x}^\star$ be a distributionally robust prescriptor as in Definition 3.5. Then, introduce a data-driven prescriptor $\widehat{x}^\star$ defined through $\widehat{x}^\star_T = \tilde{x}^\star(\widehat{S}_T)$ for all $T \in \mathbb{N}$.

Theorem 4.3 (Optimality of $(\widehat{c}^\star, \widehat{x}^\star)$)**.** *If the Assumptions 2.1, 2.2, 4.1 and 4.2 hold, the rate function* (4.4) *is regular and $r > 0$, then $(\widehat{c}^\star, \widehat{x}^\star)$ is a Pareto dominant solution of the meta-optimization problem* (2.4b).

The assumptions of Theorem 4.3 imply the assumptions of Theorem 3.2, which in turn implies that $(\tilde{c}^\star, \tilde{x}^\star)$ represents a Pareto dominant solution to the restricted meta-optimization problem (2.6b). The discussion after Definition 2.6 further implies that the objective function value of $(\widehat{c}^\star, \widehat{x}^\star)$ in (2.4b) coincides with that of $(\tilde{c}^\star, \tilde{x}^\star)$ in (2.6b) for every fixed model $\theta \in \Theta$, that is, we have

$$\lim_{T\to\infty} \mathbb{E}_\theta[\widehat{c}_T^\star(\widehat{x}_T^\star)] = \tilde{c}^\star(\tilde{x}^\star(S_\infty(\theta)), S_\infty(\theta)).$$

As Theorem 4.3 identifies $(\widehat{c}^\star, \widehat{x}^\star)$ as a Pareto dominant solution to (2.4b), the original meta-optimization problem (2.4b) is thus equivalent to the restricted meta-optimization problem (2.6b). Therefore, Theorem 4.3 establishes another separation principle that enables a decoupling of estimation and optimization.

Proof of Theorem 4.3. As $(\tilde{c}^\star, \tilde{x}^\star) \in \tilde{\mathcal{X}}$ comprises a compressed data-driven predictor and a compressed data-driven prescriptor in the sense of Definition 2.6, $(\widehat{c}^\star, \widehat{x}^\star)$ comprises a data-driven predictor in the sense of Definition 2.2, and a data-driven prescriptor in the sense of Definition 2.3, that is, $(\widehat{c}^\star, \widehat{x}^\star) \in \widehat{\mathcal{X}}$. This follows from the discussion after Definition 2.6. Similarly, as $(\tilde{c}^\star, \tilde{x}^\star)$ satisfies the rate constraint in (2.6b), one readily verifies that $(\widehat{c}^\star, \widehat{x}^\star)$ satisfies the rate constraint in (2.4b). Hence, $(\widehat{c}^\star, \widehat{x}^\star)$ is feasible in (2.4b). Below we will further show that $(\widehat{c}^\star, \widehat{x}^\star)$ Pareto dominates every other feasible solution of problem (2.4b).

Assume for the sake of contradiction that there exists a data-driven predictor-prescriptor pair $(\widehat{c}, \widehat{x})$ that is feasible in (2.4b) but not dominated by $(\widehat{c}^\star, \widehat{x}^\star)$. Hence, there exist a model $\theta_0 \in \Theta$ with

$$\lim_{T\to\infty} \mathbb{E}_{\theta_0}[\widehat{c}_T^\star(\widehat{x}_T^\star)] - \lim_{T\to\infty} \mathbb{E}_{\theta_0}[\widehat{c}_T(\widehat{x}_T)] > 0. \tag{4.10}$$

In the following we will show that this inequality contradicts our assumption that $(\widehat{c}, \widehat{x})$ is feasible in (2.4a).

By the defining properties of data-driven predictors and prescriptors, $\widehat{c}_T(\widehat{x}_T) = \min_{x\in X} \widehat{c}_T(x)$ converges in probability under $\mathbb{P}_{\theta_0}$ to $v_\infty(\theta_0)$, where v_∞ is the Borel-measurable function whose existence is postulated in Definition 2.2(v). As there exists a random variable $\overline{c}$ with $\mathbb{E}_{\theta_0}[\overline{c}] < \infty$ and $|\widehat{c}_T(\widehat{x}_T)| \le \overline{c}$ $\mathbb{P}_{\theta_0}$-almost surely for all $T \in \mathbb{N}$ (see Definition 2.2(iii)), Lemma B.1 implies that $\lim_{T\to\infty} \mathbb{E}_{\theta_0}[\widehat{c}_T(\widehat{x}_T)] = v_\infty(\theta_0)$. Similarly, since $\widehat{S}_T$ converges in probability under $\mathbb{P}_{\theta_0}$ to $S_\infty(\theta_0)$ and since $\tilde{c}^\star(x, s)$ is bounded and continuous in (x, s) on $\mathbb{S}_\infty$ thanks to Proposition 3.1, the continuous mapping theorem [26, Theorem 3.2.4] implies that $\widehat{c}_T^\star(\widehat{x}_T^\star) = \tilde{c}^\star(\tilde{x}^\star(\widehat{S}_T), \widehat{S}_T)$ converges in probability under $\mathbb{P}_{\theta_0}$ to $\tilde{c}^\star(\tilde{x}^\star(S_\infty(\theta_0)), S_\infty(\theta_0))$. This in turn ensures via Lemma B.1 that $\lim_{T\to\infty} \mathbb{E}_{\theta_0}[\widehat{c}_T^\star(\widehat{x}_T^\star)] = \tilde{c}^\star(\tilde{x}^\star(S_\infty(\theta_0)), S_\infty(\theta_0))$.

We now introduce the optimal value function $\tilde{v}^\star : \mathbb{S} \to \mathbb{R}$ through $\tilde{v}^\star(s) = \tilde{c}^\star(\tilde{x}^\star(s), s) = \min_{x\in X} \tilde{c}^\star(x, s)$. Note that $\tilde{v}^\star$ inherits boundedness from $\tilde{c}^\star$ and is continuous in $s \in \mathbb{S}$ by Berge's maximum theorem [11, pp. 115–116]. The above arguments show that (4.10) is equivalent to $\varepsilon = \frac{1}{3}[\tilde{v}^\star(S_\infty(\theta_0)) - v_\infty(\theta_0)] > 0$.

By Lemma B.2, which applies because all realizations of the random variable $\widehat{x}_T$ fall into the compact set X for all $T \in \mathbb{N}$, there exists a deterministic function $x_\infty : \Theta \to X$ that satisfies

$$\limsup_{T\to\infty} \mathbb{P}_\theta \left[\|\widehat{x}_T - x_\infty(\theta)\| < \rho\right] > 0 \quad \forall \rho > 0 \quad \forall \theta \in \Theta. \tag{4.11}$$

Defining $x_0 = x_\infty(\theta_0)$ and $s_0 = S_\infty(\theta_0)$, we may reuse the reasoning at the beginning of Step 2 in the proof of Theorem 3.1 to show that there exists $\theta_0^\star \in \Theta$ with $I(s_0, \theta_0^\star) = r_0 < r$ and

$$\tilde{c}^\star(x_0, s_0) < c(x_0, \theta_0^\star) + \varepsilon.$$

As $x_0 \in X$ by construction, we further have

$$\tilde{c}^\star(x_0, s_0) \ge \min_{x\in X} \tilde{c}^\star(x, s_0) = \tilde{v}^\star(s_0) = v_\infty(\theta_0) + 3\varepsilon,$$

where the two equalities follow from the definitions of $\tilde{v}^\star$ and ε, respectively. Combining the two inequalities above then yields $c(x_0,\theta_0^\star) > v_\infty(\theta_0)+2\varepsilon$. As $c(x,\theta_0^\star)$ is continuous in $x \in X$ thanks to Assumption 2.2, we may finally conclude that there exists a tolerance $\rho>0$ such that

$$c(x,\theta_0^\star) > v_\infty(\theta_0)+\varepsilon \quad \forall x \in X:\ \|x-x_0\|<\rho. \tag{4.12}$$

Armed with these preliminary results, we are now ready to show that $(\widehat{c},\widehat{x})$ fails to be feasible in (2.4a). To this end, we may use a similar reasoning as in the proof of Theorem 4.2 to demonstrate that

$$\begin{aligned}
&\limsup_{T\to\infty}\frac{1}{T}\log \mathbb{P}_{\theta_0^\star}[\widehat{c}_T(\widehat{x}_T) < c(\widehat{x}_T,\theta_0^\star)]\\
&\quad\geq \limsup_{T\to\infty}\frac{1}{T}\log \mathbb{P}_{\theta_0^\star}\left[\widehat{c}_T(\widehat{x}_T) < v_\infty(\theta_0)+\varepsilon \ \wedge\ \widehat{S}_T \in \mathbb{B}_\delta(s_0)\ \wedge\ \|\widehat{x}_T-x_0\|<\rho\right]\\
&\quad\geq \Lambda(\eta,\theta_0^\star)-\langle\eta,s_0\rangle-\|\eta\|\delta\\
&\qquad\quad+\limsup_{T\to\infty}\frac{1}{T}\log \mathbb{P}_{\theta_0}\left[\widehat{c}_T(\widehat{x}_T) < v_\infty(\theta_0)+\varepsilon \ \wedge\ \widehat{S}_T \in \mathbb{B}_\delta(s_0)\ \wedge\ \|\widehat{x}_T-x_0\|<\rho\right],
\end{aligned} \tag{4.13}$$

where the two inequalities follow from (4.12) and from a change of measure argument akin to (4.8)–(4.9). Details are omitted for brevity of exposition. As $\widehat{c}_T(\widehat{x}_T)$ converges in probability to $v_\infty(\theta_0)$ under $\mathbb{P}_{\theta_0}$ (by the definition of a data-driven predictor) and as $\widehat{S}_T$ converges in probability to $S_\infty(\theta_0)$ under $\mathbb{P}_{\theta_0}$ (by the definition of a statistic and by the construction of $s_0=S_\infty(\theta_0)$), one readily verifies that

$$\lim_{T\to\infty}\mathbb{P}_{\theta_0}\left[\widehat{c}_T(\widehat{x}_T) < v_\infty(\theta_0)\ \wedge\ \widehat{S}_T\in\mathbb{B}_\delta(s_0)\right]=1.$$

As $x_0=x_\infty(\theta_0)$, (4.11) further implies that $\limsup_{T\to\infty}\mathbb{P}_\theta[\|\widehat{x}_T-x_0\|<\rho]>0$ for all $\delta>0$. Thus, we have

$$\begin{aligned}
&\limsup_{T\to\infty}\frac{1}{T}\log \mathbb{P}_{\theta_0}\left[\widehat{c}_T(\widehat{x}_T) < v_\infty(\theta_0)+\varepsilon \ \wedge\ \widehat{S}_T \in \mathbb{B}_\delta(s_0)\ \wedge\ \|\widehat{x}_T-x_0\|<\rho\right]\\
&\quad\geq \limsup_{T\to\infty}\frac{1}{T}\log\left(\mathbb{P}_{\theta_0}\left[\widehat{c}_T(\widehat{x}_T) < v_\infty(\theta_0)\ \wedge\ \widehat{S}_T\in\mathbb{B}_\delta(s_0)\right]+\mathbb{P}_\theta[\|\widehat{x}_T-x_0\|<\rho]-1\right)=0,
\end{aligned}$$

where the inequality follows using the elementary insight that $\mathbb{P}_{\theta_0}[A\cap B]\geq \mathbb{P}_{\theta_0}[A]+\mathbb{P}_{\theta_0}[B]-1$ for all events $A,B\in\mathcal{F}$. Combining this estimate with (4.13) then yields

$$\limsup_{T\to\infty}\frac{1}{T}\log \mathbb{P}_{\theta_0^\star}[\widehat{c}_T(\widehat{x}_T) < c(\widehat{x}_T,\theta_0^\star)] \geq \Lambda(\eta,\theta_0^\star)-\big\langle\eta,s_0\big\rangle-\|\eta\|\delta.$$

As the above inequality holds for every $\delta>0$, we have effectively shown that

$$\limsup_{T\to\infty}\frac{1}{T}\log \mathbb{P}_{\theta_0^\star}[\widehat{c}_T(\widehat{x}_T) < c(\widehat{x}_T,\theta_0^\star)] \geq \Lambda(\eta,\theta_0^\star)-\big\langle\eta,s_0\big\rangle = -I(s_0,\theta_0^\star) = -r_0 > -r,$$

where the two equalities follow from the relation $I(s_0,\theta_0^\star)=\langle\eta,s_0\rangle-\Lambda(\eta,\theta_0^\star)$ established in the proof of Theorem 4.2 and the definition of r_0, respectively. This contradicts the feasibility of $(\widehat{c},\widehat{x})$ in (2.4b). $\square$

Theorems 4.2 and 4.3 are reminiscent of the celebrated Rao-Blackwell theorem [16, 44], which asserts that any given estimator $\widehat{\theta}_T$ of the unknown parameter θ can be improved by conditioning it on a sufficient statistic $\widehat{S}_T$. The resulting estimator $\mathbb{E}_\theta[\widehat{\theta}_T|\widehat{S}_T]$ is non-inferior to $\widehat{\theta}_T$ with respect to the mean squared error criterion and depends on the available data only through $\widehat{S}_T$. The proof of the Rao-Blackwell theorem critically relies on Jensen's inequality, which is applicable because the mean squared error is convex in $\widehat{\theta}_T$. Unfortunately, it is not possible to improve a given data-driven predictor $\widehat{c}_T(x)$ by simply conditioning it on $\widehat{S}_T$. This approach fails because the out-of-sample disappointment is *non*-convex in $\widehat{c}_T(x)$. The proofs of Theorems 4.2 and 4.3 are therefore substantially more involved than that of the Rao-Blackwell theorem.

Example 4.3 (Optimal predictors and prescriptors for finite state i.i.d. processes)**.** *Consider the class of finite state i.i.d. processes of Example 2.2, and let $\widehat{S}_T$ be the empirical distribution defined in Example 2.2. We know from Example 3.1 that $\widehat{S}_T$ satisfies an LDP with regular rate function $D(s\|\theta)$. By Theorems 3.1 and 3.2, the distributionally robust predictor $\tilde{c}^\star$ with a relative entropy ambiguity set and the corresponding prescriptor $\tilde{x}^\star$ thus provide Pareto dominant solutions for the restricted meta-optimization problems* (2.6)*. From Example 4.1 we further know that Assumptions 4.1 and 4.2 hold. By Theorems 4.2 and 4.3 the data-driven predictor $\widehat{c}^\star$ and the corresponding prescriptor $\widehat{x}^\star$ induced by $\tilde{c}^\star$ and $\tilde{x}^\star$, respectively, thus provide Pareto dominant solutions for the original meta-optimization problems* (2.4)*.*

5 Data-generating processes

We now describe several data-generating processes for which the restricted meta-optimization problems (2.6) or even the original meta-optimization problems (2.4) admit Pareto dominant solutions.

5.1 Finite-state Markov chains

Assume that $\{\xi_t\}_{t=1}^T$ represents a time-homogeneous ergodic Markov chain with state space $\Xi = \{1, \ldots, m\}$ and dummy deterministic initial state $\xi_0 = i_0 \in \Xi$ satisfying $\lim_{t\to\infty} \mathbb{P}_\star[\xi_t = i,\, \xi_{t+1} = j] = (\theta_\star)_{ij} > 0$ for all $i, j \in \Xi$. The matrix $\theta_\star$ encodes the stationary probability mass function of the doublet (ξ_t, ξ_{t+1}), and thus

$$\sum_{j\in\Xi}(\theta_\star)_{ij} = \lim_{t\to\infty}\sum_{j\in\Xi}\mathbb{P}_\star[\xi_t = i,\, \xi_{t+1} = j] = \lim_{t\to\infty}\mathbb{P}_\star[\xi_t = i] = \lim_{t\to\infty}\sum_{j\in\Xi}\mathbb{P}_\star[\xi_{t-1} = j,\, \xi_t = i] = \sum_{j\in\Xi}(\theta_\star)_{ji},$$

i.e., the row sums of $\theta_\star$ coincide with the respective column sums. These properties of $\theta_\star$ prompt us to define

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

as the set of all strictly positive doublet probability mass functions with balanced marginals. Note that every $\theta \in \Theta$ induces a unique row vector $\pi_\theta \in \mathbb{R}^{1\times m}_{++}$ of stationary state probabilities and a unique transition probability matrix $P_\theta \in \mathbb{R}^{m\times m}_{++}$ defined through $(\pi_\theta)_i = \sum_{j\in\Xi}\theta_{ij}$ and $(P_\theta)_{ij} = \theta_{ij}/(\pi_\theta)_i$, respectively. By construction, P_θ is a stochastic matrix whose rows represent strictly positive probability vectors, and the stationary distribution π_θ satisfies $\pi_\theta P_\theta = \pi_\theta$; see Ross [47, Chapter 4] for further details on Markov chains. We conclude that $\mathbb{P}_\star$ belongs to a finitely parametrized ambiguity set of the form $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, where each model $\theta \in \Theta$ encodes a probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ with

$$\mathbb{P}_\theta[\xi_{[T]} = (i_1, \ldots, i_T)] = \textstyle\prod_{t=1}^T (P_\theta)_{i_{t-1}i_{t+1}} \quad \forall (i_1, \ldots, i_T) \in \Xi^T,\ T \in \mathbb{N}.$$

Note also that Θ is embedded in a Euclidean space of finite dimension $d = m^2$. In summary, we have thus shown that Assumption 2.1 holds. Next, we define the empirical doublet distribution $\widehat{S}_T \in \mathbb{R}^{m\times m}$ through

$$(\widehat{S}_T)_{ij} = \tfrac{1}{T}\textstyle\sum_{t=1}^T \mathbb{1}_{(\xi_{t-1},\xi_t)=(i,j)} \quad \forall i, j \in \Xi. \tag{5.1}$$

By construction, $\widehat{S} = \{\widehat{S}_T\}_{T\in\mathbb{N}}$ constitutes a statistic with state space

$$\mathbb{S} = \mathrm{cl}\left(\cup_{T\in\mathbb{N}}\Delta_{m\times m} \cap (\mathbb{Z}^{m\times m}/T)\right) = \mathrm{cl}\left(\Delta_{m\times m} \cap \mathbb{Q}^{m\times m}\right) = \Delta_{m\times m}.$$

We emphasize that $\mathbb{S}$ is a strict superset of the model space Θ. The ergodic theorem for Markov chains further ensures that the empirical doublet distribution $\widehat{S}_T$ converges $\mathbb{P}_\theta$-almost surely to the true doublet distribution θ as T grows; see [47, Theorem 4.1]. Consequently, we have $S_\infty(\theta) = \theta$ for all $\theta \in \Theta$, which implies that $\widehat{S}$ is a consistent model estimator in the sense of Definition 2.5 and that the set $\mathbb{S}_\infty$ of all asymptotic realizations of $\widehat{S}$ coincides with Θ. In addition, the function S_∞ is clearly a local homeomorphism.

We now follow the reasoning in [15] to show that the ambiguity set $\mathcal{P}$ represents a time-homogeneous exponential family. Specifically, we define the baseline model $\bar{\theta} \in \Theta$ through $\bar{\theta}_{ij} = 1/m^2$ for all $i, j \in \Xi$. The observations ξ_t, $t \in \mathbb{N}$, are thus serially independent and uniformly distributed under $\mathbb{P}_{\bar{\theta}}$, and the corresponding transition probability matrix satisfies $(P_{\bar{\theta}})_{ij} = 1/m$ for all $i, j \in \Xi$. In addition, the probability of observing $\xi_{[T]}$ under $\mathbb{P}_{\bar{\theta}}$ is given by $1/m^T$, and

$$\begin{aligned}
\frac{\mathrm{d}\mathbb{P}^T_\theta}{\mathrm{d}\mathbb{P}^T_{\bar{\theta}}} &= m^T \prod_{t=1}^{T} (P_\theta)_{\xi_{t-1}\xi_t} \\
&= m^T \prod_{i,j\in\Xi} (P_\theta)_{ij}^{\sum_{t=1}^T 1_{(\xi_{t-1},\xi_t)=(i,j)}} \\
&= m^T \prod_{i,j\in\Xi} (P_\theta)_{ij}^{T(\widehat{S}_T)_{ij}} = \exp\Big(\langle T\log(P_\theta), \widehat{S}_T\rangle + T\log m\Big),
\end{aligned}$$

where the logarithm of the matrix P_θ is evaluated element-wise. This reveals that $\mathcal{P}$ constitutes an exponential family in the sense of Assumption 4.1 with parametrization function $g(\theta) = \log(P_\theta)$ and that $\widehat{S}$ is a sufficient statistic. The T^{th} log-moment generating function $\Lambda_T(\lambda, \theta)$—and thus also the log-partition function $A_T(\lambda)$—admit no concise closed-form expression. However, the proof of [22, Theorem 3.1.2] implies that the limiting log-moment generating function $\Lambda(\lambda, \theta) = \lim_{T\to\infty} \frac{1}{T}\Lambda_T(T\lambda, \theta)$ is everywhere finite and differentiable in λ for all $\theta \in \Theta$. In addition, we have $\nabla_\lambda \Lambda(0, \theta) = \lim_{T\to\infty} \mathbb{E}_\theta[\widehat{S}_T] = \mathbb{E}_\theta[\lim_{T\to\infty} \widehat{S}_T] = \theta$, where the three equalities follow from Lemma 4.1, the dominated convergence theorem and our insight that $\widehat{S}$ converges $\mathbb{P}_\theta$-almost surely to θ, respectively. Hence, Assumption 4.2 holds, which ensures via the Gärtner-Ellis theorem that $\widehat{S}$ satisfies an LDP; see also [22, Theorem 3.1.13]. The corresponding rate function $I(s, \theta)$ is given by the convex conjugate of the limiting log-moment generating function $\Lambda(\lambda, \theta)$ with respect to λ, which coincides with conditional relative entropy of s with respect to θ [22, Section 3.1.3].

Definition 5.1 (Conditional relative entropy)**.** *Using the standard convention that* $0\log(0/p) = 0$ *for any* $p \geq 0$*, the conditional relative entropy of* $s \in \mathbb{S}$ *with respect to* $\theta \in \Theta$ *is defined as*

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

If we denote the i^{th} rows of the transition probability matrices P_s and P_θ by $(P_s)_{i\cdot}$ and $(P_\theta)_{i\cdot}$, respectively,[2] and if we denote the relative entropy as usual by $D(\cdot\|\cdot)$, then an elementary calculation reveals that

$$D_c(s\|\theta) = \sum_{i\in\Xi} (\pi_s)_i \, D((P_s)_{i\cdot}\|(P_\theta)_{i\cdot}).$$

Thus, $D_c(s\|\theta)$ can be viewed as the relative entropy distance between the transition probability vectors under s and θ emanating from a random state of the Markov chain, averaged by the invariant state distribution associated with s. This interpretation justifies the name 'conditional relative entropy.' Note also that Definition 5.1 specifies $D_c(s\|\theta)$ only on $\mathbb{S}\times\Theta$ and that $D_c(s\|\theta)$ is continuous on $\mathbb{S}\times\Theta$ thanks to our standard

[2] If $(\pi_s)_i = \sum_{j\in\Xi} s_{ij} = 0$, then we may define without loss of generality $(P_s)_{ij} = 1$ if $j = i$ and $(P_s)_{ij} = 0$ otherwise.

conventions for the logarithm. We emphasize that $D_c(s\|\theta)$ cannot be continuously extended beyond $\mathbb{S}\times\Theta$. However, $D_c(s\|\theta)$ admits a unique lower semi-continuous extension to $\mathbb{S}\times\operatorname{cl}\Theta$, which is obtained by setting

$$D_c(s\|\theta) = \lim_{\delta\downarrow 0}\inf_{(s',\theta')\in\mathbb{S}\times\Theta}\left\{D_c(s'\|\theta') : \|(s',\theta')-(s,\theta)\|\le\delta\right\} \quad \forall (s,\theta)\in\mathbb{S}\times(\operatorname{cl}\Theta\backslash\Theta);$$

see also [46, Definition 1.5]. In the following, we will always mean this lower semi-continuous extension to $\mathbb{S}\times\operatorname{cl}\Theta$ when referring to the conditional relative entropy $D_c(s\|\theta)$. The next proposition establishes that the conditional relative entropy represents a regular rate function in the sense of Definition 3.3.

Proposition 5.1 (Properties of the conditional relative entropy)**.** *The conditional relative entropy $D_c(s\|\theta)$ is a regular rate function in the sense of Definition 3.3. In addition, $D_c(s\|\theta)$ is convex in s.*

Proof. The continuity of $D_c(s\|\theta)$ on $\mathbb{S}\times\Theta$ follows directly from Definition 5.1 and our standard conventions for the logarithm. Next, we show that the level sets of the form $\{(s,\theta)\in\mathbb{S}\times\operatorname{cl}\Theta : D_c(s\|\theta)\le r\}$ are compact for all fixed thresholds $r\ge 0$. It is clear that all these level sets are bounded because both $\mathbb{S}$ and $\operatorname{cl}\Theta$ are bounded. In addition, they are closed because $D_c(s\|\theta)$ is lower semi-continuous on $\mathbb{S}\times\operatorname{cl}\Theta$ by construction. To prove that the conditional relative entropy $D_c(s\|\theta)$ is radially monotonic in θ, we first observe that the following equivalent inequalities hold for all vectors $v,w\in\mathbb{R}^d_{++}$ thanks to Jensen's inequality.

$$\sum_{j\in\Xi}\frac{w_j}{\sum_{k\in\Xi}w_k}\left(\frac{v_j}{w_j}\right)^2 \ge \left(\sum_{j\in\Xi}\frac{w_j}{\sum_{k\in\Xi}w_k}\frac{v_j}{w_j}\right)^2 \quad\iff\quad \sum_{j\in\Xi}\frac{v_j^2}{w_j}\ge\frac{\left(\sum_{j\in\Xi}v_j\right)^2}{\sum_{k\in\Xi}w_k} \tag{5.2}$$

Note further that these inequalities are strict unless v and w are parallel, in which case the fraction v_j/w_j is constant in j. Next, select any $\theta\in\operatorname{cl}\Theta$ and $s\in\mathbb{S}_\infty=\Theta$, and define $\theta(\lambda)=(1-\lambda)s+\lambda\theta$ for any $\lambda\in[0,1)$. By construction, $\theta(\lambda)\in\Theta$ for all $\lambda\in[0,1)$. Basic algebra further implies that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}D_c(s\|\theta(\lambda)) &= \sum_{i,j\in\Xi}s_{ij}\left(\frac{\sum_{k\in\Xi}\theta_{ik}-s_{ik}}{\sum_{k\in\Xi}\theta_{ik}(\lambda)}-\frac{\theta_{ij}-s_{ij}}{\theta_{ij}(\lambda)}\right)\\
&= \frac{1}{\lambda}\sum_{i,j\in\Xi}s_{ij}\left(\frac{\sum_{k\in\Xi}\theta_{ik}(\lambda)-s_{ik}}{\sum_{k\in\Xi}\theta_{ik}(\lambda)}-\frac{\theta_{ij}(\lambda)-s_{ij}}{\theta_{ij}(\lambda)}\right)\\
&= \frac{1}{\lambda}\sum_{i\in\Xi}\left(\sum_{j\in\Xi}\frac{(s_{ij})^2}{\theta_{ij}(\lambda)}-\frac{\left(\sum_{j\in\Xi}s_{ij}\right)^2}{\sum_{j\in\Xi}\theta_{ij}(\lambda)}\right)\ge 0 \quad \forall\lambda\in(0,1),
\end{aligned}$$

where the inequality follows from (5.2). Note that this inequality collapses to an equality if and only if each row of θ is parallel to the corresponding row of s, that is, if and only if the transition probability matrices P_θ and P_s induced by θ and s, respectively, are identical. In this special case we have $D_c(s\|\theta(\lambda))=0$ for all $\lambda\in[0,1)$. Otherwise, $D_c(s\|\theta(\lambda))$ is strictly monotonically increasing in λ. Thus, $D_c(s\|\theta)$ satisfies the inequality (3.2), which is sufficient for radial monotonicity; see the discussion after Definition 3.3. In summary, we have thus shown that the rate function $D_c(s\|\theta)$ is indeed regular. Finally, $D_c(s\|\theta)$ is convex in s because the perspective function $v\log(v/w)$ is convex in $(v,w)\in\mathbb{R}^2_+$ [19, Section 3.2.6] and because convexity is preserved under combinations with linear functions [19, Section 3.2.2]. $\square$

By Theorems 3.1 and 3.2, we may now conclude that the distributionally robust predictor $\tilde{c}^\star$ with a conditional relative entropy ambiguity set and the corresponding prescriptor $\tilde{x}^\star$ provide Pareto dominant solutions for the restricted meta-optimization problems (2.6). Moreover, by Theorems 4.2 and 4.3 the data-driven predictor $\widehat{c}^\star$ and the corresponding prescriptor $\widehat{x}^\star$ induced by $\tilde{c}^\star$ and $\tilde{x}^\star$, respectively, provide Pareto dominant solutions for the original meta-optimization problems (2.4). As $D_c(s\|\theta)$ fails to be convex in θ, computing $\tilde{c}^\star(x,s)$ for a fixed $x \in X$ and $s \in \mathbb{S}$ necessitates the solution of a challenging nonconvex optimization problem with $\mathcal{O}(m^2)$ decision variables. An efficient Frank-Wolfe algorithm that decomposes this problem into a sequence of convex programs with $\mathcal{O}(m)$ decision variables is developed in [40].

5.2 Vector autoregressive processes with unknown drift

Assume now that the observable data $\{\xi_t\}_{t=1}^T$ follows a vector autoregressive process of the form

$$\xi_{t+1} = \theta_\star + A\xi_t + \varepsilon_{t+1} \quad \forall t \in \mathbb{N} \tag{5.3}$$

with state space $\Xi = \mathbb{R}^d$, where the drift term $\theta_\star \in \mathbb{R}^d$ is deterministic but unknown. Assume further that the disturbances $\{\varepsilon_t\}_{t\in\mathbb{N}}$ are normally distributed with zero mean and known positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and that the initial state ξ_1 and all disturbances are mutually independent under $\mathbb{P}_\star$. Finally, assume that $A \in \mathbb{R}^{d\times d}$ is asymptotically stable in the sense that all of its eigenvalues reside strictly inside the complex unit circle. Hence, the process $\{\xi_t\}_{t\in\mathbb{N}}$ is ergodic and admits a unique stationary distribution [56, Example 3.43 and Proposition 3.44]. It is well known that the stationary distribution is Gaussian with mean vector $(\mathbb{1}_d - A)^{-1}\theta_\star$ and that its covariance matrix R_0 is the unique solution to the discrete Lyapunov equation $R_0 = AR_0A^\top + \Sigma$, see, *e.g.*, [2, Section 6.10 E]. In the remainder of this section we will assume that the process $\{\xi_t\}_{t\in\mathbb{N}}$ is stationary under $\mathbb{P}_\star$. This means that ξ_t follows the stationary distribution for every $t \in \mathbb{N}$. An elementary calculation further reveals that the cross-covariance matrix $R_\delta \in \mathbb{R}^{d\times d}$ of any ξ_t and ξ_s with $\delta = t - s$ is given by $R_\delta = A^\delta R_0$ if $\delta \geq 0$ and $R_\delta = R_0(A^{-\delta})^\top$ if $\delta < 0$.

Assume now that the drift $\theta_\star$ is known to belong to an open convex set $\Theta \subseteq \mathbb{R}^d$ that captures any available structural information. We then define an ambiguity set $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, where each $\theta \in \Theta$ encodes a probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ under which the observations $\{\xi_t\}_{t\in\mathbb{N}}$ are jointly normally distributed with mean vector $\mathbb{E}_\theta[\xi_t] = (\mathbb{1}_d - A)^{-1}\theta$ for all $t \in \mathbb{N}$ and cross-covariance matrix $\mathbb{E}_\theta[(\xi_t - \mathbb{E}_\theta[\xi_t])(\xi_s - \mathbb{E}_\theta[\xi_s])^\top] = R_{t-s}$ for all $s, t \in \mathbb{N}$. In this setting, a natural estimator for θ is the scaled sample mean

$$\widehat{S}_T = (\mathbb{1}_d - A)\frac{1}{T}\sum_{t=1}^T \xi_t \quad \forall T \in \mathbb{N} \tag{5.4}$$

with state space $\mathbb{S} = \mathbb{R}^d$. By [56, Theorem 3.34], which applies because the data process is ergodic, $\widehat{S}$ represents a consistent model estimator in the sense of Definition 2.5. Consequently, we have $S_\infty(\theta) = \theta$ for all $\theta \in \Theta$, and $\mathbb{S}_\infty = \Theta$. In addition, the function S_∞ is clearly a local homeomorphism. The next proposition asserts that the statistic $\widehat{S}$ also satisfies an LDP with a regular quadratic rate function.

Proposition 5.2 (LDP for stationary autoregressive processes with unknown drift)**.** *If $\{\xi_t\}_{t\in\mathbb{N}}$ follows a stationary autoregressive process of the form* (5.3) *with drift $\theta \in \Theta$, then the scaled sample mean* (5.4) *satisfies an LDP with regular convex quadratic rate function $I(s,\theta) = \frac{1}{2}(s-\theta)^\top\Sigma^{-1}(s-\theta)$.*

Proof. Fix any $\theta \in \Theta$, and define a probability measure $\mathbb{P}'_\theta$ on $(\Omega, \mathcal{F})$ under which the observations $\{\xi_t\}_{t\in\mathbb{N}}$ are jointly normally distributed with $\mathbb{E}'_\theta[\xi_t] = 0$ and $\mathbb{E}'_\theta[\xi_t\xi_s^\top] = R_\delta$ for $\delta = t - s$, where $\mathbb{E}'_\theta[\cdot]$ denotes the expectation under $\mathbb{P}'_\theta$. Note that $\mathbb{P}'_\theta$ and $\mathbb{P}_\theta$ assign different means to ξ_t but are otherwise indistinguishable.

The log-moment generating function of the sample mean $\widehat{\mu}_T = \frac{1}{T}\sum_{t=1}^T \xi_t$ under $\mathbb{P}'_\theta$ is then given by

$$\Lambda'_T(\lambda,\theta) = \log \mathbb{E}'_\theta\left[\exp(\lambda^\top \widehat{\mu}_T)\right] = \frac{1}{2}\lambda^\top \mathbb{E}'_\theta\left[\widehat{\mu}_T\widehat{\mu}_T^\top\right]\lambda = \frac{1}{2T^2}\sum_{s,t=1}^T \lambda^\top R_{t-s}\lambda = \frac{1}{2T}\sum_{\delta=-T}^T\left(1-\frac{|\delta|}{T}\right)\lambda^\top R_\delta\lambda,$$

where the second equality follows from the formula for the mean value of a lognormal random variable, while the third equality exploits the definitions of the sample mean $\widehat{\mu}_T$ and the cross-covariance matrix of the observations ξ_t and ξ_s. This implies that the limiting log-moment generating function is representable as

$$\begin{aligned}\Lambda'(\lambda,\theta) &= \lim_{T\to\infty}\frac{1}{T}\Lambda'_T(T\lambda,\theta) = \lim_{T\to\infty}\frac{1}{2}\sum_{\delta=-T}^T\left(1-\frac{|\delta|}{T}\right)\lambda^\top R_\delta\lambda\\ &= \frac{1}{2}\lambda^\top\left(R_0 + \lim_{T\to\infty}\sum_{\delta=1}^T\left(1-\frac{\delta}{T}\right)A^\delta R_0 + \lim_{T\to\infty}\sum_{\delta=1}^T\left(1-\frac{\delta}{T}\right)R_0(A^\delta)^\top\right)\lambda\\ &= \frac{1}{2}\lambda^\top\left((\mathbb{1}_d - A)^{-1}R_0 + R_0(\mathbb{1}_d - A^\top)^{-1} - R_0\right)\lambda,\end{aligned}$$

where the second equality holds because $R_\delta = A^\delta R_0$ for $\delta > 0$ and $R_\delta = R_0(A^{-\delta})^\top$ for $\delta < 0$, while the third equality follows from the asymptotic stability of A and the geometric series formulas

$$\lim_{T\to\infty}\sum_{\delta=1}^T A^\delta = (\mathbb{1}_d - A)^{-1} - \mathbb{1}_d \qquad\text{and}\qquad \lim_{T\to\infty}\sum_{\delta=1}^T\frac{\delta}{T}A^\delta = 0.$$

We have thus demonstrated that $\Lambda'(\lambda,\theta) = \frac{1}{2}\lambda^\top Q\lambda$ is independent of θ and quadratic in λ with Hessian matrix $Q = (\mathbb{1}_d - A)^{-1}R_0 + R_0(\mathbb{1}_d - A^\top)^{-1} - R_0$. By the definitions of Q and R_0 we have

$$(\mathbb{1}_d - A)Q(\mathbb{1}_d - A^\top) = R_0 - AR_0A^\top = \Sigma. \tag{5.5}$$

As A is asymptotically stable, the matrix $\mathbb{1}_m - A$ is invertible. The above relation thus implies that Q inherits positive definiteness from Σ. Consequently, the quadratic function $\Lambda'(\lambda,\theta)$ is strictly convex in λ and is easily seen to satisfy Assumption 4.2. The Gärtner-Ellis Theorem thus ensures that the sample mean satisfies an LDP under $\mathbb{P}'_\theta$ with good rate function $I'_{\widehat{\mu}}(\mu,\theta) = \sup_{\lambda\in\mathbb{R}^d}\langle\lambda,\mu\rangle - \Lambda'(\lambda,\theta) = \frac{1}{2}\mu^\top Q^{-1}\mu$; see Theorem 4.1. Note that under $\mathbb{P}_\theta$ the expected value of the sample mean $\widehat{\mu}_T$ is given by $(\mathbb{1}_d - A)^{-1}\theta$ instead of 0, but the distribution of $\widehat{\mu}_T$ has the same shape under $\mathbb{P}_\theta$ and $\mathbb{P}'_\theta$. Therefore, $\widehat{\mu}_T$ also satisfies an LDP under $\mathbb{P}_\theta$ with good rate function $I_{\widehat{\mu}}(\mu,\theta) = \frac{1}{2}(\mu - (\mathbb{1}_d - A)^{-1}\theta)^\top Q^{-1}(\mu - (\mathbb{1}_d - A)^{-1}\theta)$. As $\widehat{S}_T = (\mathbb{1}_d - A)\widehat{\mu}_T$ and as A is asymptotically stable, the contraction principle [22, Theorem 4.2.1] further implies that the scaled sample mean $\widehat{S}_T$ defined in (5.4) satisfies an LDP with good rate function

$$I(s,\theta) = I_{\widehat{\mu}}((\mathbb{1}_d - A)^{-1}s,\theta) = \frac{1}{2}(s-\theta)^\top(\mathbb{1}_m - A^\top)^{-1}Q^{-1}(\mathbb{1}_m - A)^{-1}(s-\theta) = \frac{1}{2}(s-\theta)^\top\Sigma^{-1}(s-\theta),$$

where the last equality follows from (5.5). The regularity of the rate function $I(s,\theta)$ is easy to check. □

Proposition 5.2 generalizes [22, Exercise 2.3.23], which focuses on scalar autoregressive processes without drift. The results of this section imply via Theorems 3.1 and 3.2 that the distributionally robust predictor $\tilde{c}^\star$ with an ellipsoidal ambiguity set for θ around the scaled sample mean (5.4) and the corresponding prescriptor $\tilde{x}^\star$ provide Pareto dominant solutions for the restricted meta-optimization problems (2.6). As the scaled sample mean fails to be a sufficient statistic for θ, however, we are unable to find Pareto dominant solutions for the original meta-optimization problems (2.4). Details are omitted for brevity.

5.3 Scalar autoregressive processes with unknown coefficient

Assume now that the observable data $\{\xi_t\}_{t=1}^T$ follows a scalar autoregressive process of the form

$$\xi_{t+1} = \theta_\star \xi_t + \varepsilon_{t+1} \quad \forall t \in \mathbb{N} \tag{5.6}$$

with state space $\Xi = \mathbb{R}$, where the autoregressive coefficient $\theta_\star \in (-1,1)$ is deterministic but unknown. Assume further that the disturbances $\{\varepsilon_t\}_{t\in\mathbb{N}}$ are normally distributed with known mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ and that the initial state ξ_1 and all disturbances are mutually independent under $\mathbb{P}_\star$. As in Section 5.2, we finally assume that the process $\{\xi_t\}_{t=1}^T$ is stationary under $\mathbb{P}_\star$. In this case $\mathbb{P}_\star$ belongs to an ambiguity set $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, where each $\theta \in \Theta = (-1,1)$ encodes a probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ under which the observations $\{\xi_t\}_{t\in\mathbb{N}}$ are jointly normally distributed with mean $\mathbb{E}_\theta[\xi_t] = \mu/(1-\theta)$ for all $t \in \mathbb{N}$ and autocovariance $\mathbb{E}_\theta[(\xi_t - \mathbb{E}_\theta[\xi_t])(\xi_s - \mathbb{E}_\theta[\xi_s])] = \sigma^2\theta^{|t-s|}/(1-\theta^2)$ for all $s, t \in \mathbb{N}$.

In the following we investigate two complementary estimators for the autoregressive coefficient θ. We first study the least squares estimator, which is defined through

$$\widehat{S}_T = \frac{\sum_{t=2}^T \xi_t \xi_{t-1}}{\sum_{t=2}^T \xi_{t-1}^2} \quad \forall T \in \mathbb{N}. \tag{5.7}$$

By construction, the state space of $\widehat{S}$ is given by $\mathbb{S} = \mathbb{R}$, and it is well known that $\widehat{S}_T$ converges $\mathbb{P}_\theta$-almost surely to θ for all $\theta \in \Theta$; see, *e.g.*, [10]. Thus, $S_\infty(\theta) = \theta$ represents a local homeomorphism, the set of asymptotic estimator realizations simplifies to $\mathbb{S}_\infty = \Theta$ and $\widehat{S}$ is a consistent model estimator in the sense of Definition 2.5. Moreover, it is also well known that $\widehat{S}$ satisfies an LDP.

Proposition 5.3 (LDP for stationary autoregressive processes with unknown coefficient (I))**.** *If $\{\xi_t\}_{t\in\mathbb{N}}$ follows a stationary autoregressive process of the form* (5.6) *with autoregressive parameter $\theta \in \Theta$, then the least squares estimator* (5.7) *satisfies an LDP with regular rate function*

$$I(s,\theta) = \begin{cases} \frac{1}{2}\log\left(\frac{1-2\theta s+\theta^2}{1-s^2}\right) & \text{if } s \in [a(\theta), b(\theta)], \\ \log\left(|\theta - 2s|\right) & \text{otherwise}, \end{cases} \tag{5.8}$$

where $a(\theta) = \frac{1}{4}(\theta - \sqrt{\theta^2+8})$ and $b(\theta) = \frac{1}{4}(\theta + \sqrt{\theta^2+8})$.

Proof. By [10, Proposition 8], the least squares estimator satisfies an LDP with rate function (5.8). To see that this rate function is regular, note first that $a(\theta) > -1$ and $b(\theta) < +1$ for all $\theta \in \Theta$. This implies that $(1-2\theta s+\theta^2)/(1-s^2) \geq 1$ whenever $s \in [a(\theta), b(\theta)]$. Note also that $|\theta - 2s| > 1$ whenever $s \notin [a(\theta), b(\theta)]$. Hence, the rate function (5.8) is indeed non-negative on $\mathbb{S} \times \mathrm{cl}\,\Theta$. In addition, one readily verifies that the two pieces of $I(s,\theta)$ described in the first and the second line of (5.8), respectively, match whenever $s = a(\theta)$ or $s = b(\theta)$, and thus $I(s,\theta)$ is continuous on $\mathbb{S} \times \mathrm{cl}\,\Theta$. This in turn implies that all sublevel sets of $I(x,\theta)$ are closed. As $\mathrm{cl}\,\Theta = [-1,1]$ is bounded and as $\log(|\theta - 2s|)$ is bounded below by the coercive function $\log(|2s|-1)$ uniformly across all $s \notin [a(\theta, b(\theta)]$ and $\theta \in \mathrm{cl}\,\Theta$, the sublevel sets of $I(x,\theta)$ are also bounded. Thus, $I(x,\theta)$ satisfies the level-compactness condition of Definition 3.3. To prove radial monotonicity, fix any $s \in \mathbb{S}_\infty = \Theta$ and any $\theta \in \mathrm{cl}\,\Theta$, and define $\theta(\lambda) = (1-\lambda)s + \lambda\theta$ for all $\lambda \in [0,1)$. If $s = \theta$, then $I(s,\theta(\lambda)) = 0$ for all $\lambda \in [0,1)$, and thus (3.2) is satisfied. If $s \neq \theta$, on the other hand, then $I(s,\theta(\lambda))$ is strictly monotonically increasing in λ, which implies that (3.2) holds as a strict inequality. Indeed, if $s \in [a(\theta), b(\theta)]$, then we have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} I(s,\theta(\lambda)) = \frac{(\theta(\lambda)-s)(\theta - s)}{1 - 2\theta(\lambda)s + \theta(\lambda)^2} > 0 \quad \forall \lambda \in (0,1).$$

Note that the denominator of the above fraction is strictly positive because $s, \theta \in (-1, 1)$ for $s \in [a(\theta), b(\theta)]$. The numerator is also strictly positive because $\theta(\lambda) - s$ and $\theta - s$ must have the same sign and because $s \neq \theta$. If $s < a(\theta)$, on the other hand, then we have $\theta - 2s > 1$, which implies that $s < 0$ because $\theta < 1$. In addition, as $a(\theta) \leq \theta$, the assumption $s < a(\theta)$ also ensures that $s < \theta$. Using $s < 0$ and $\theta - 2s > 1$, we thus find

$$\theta(\lambda) - 2s \geq \min\{s - 2s, \theta - 2s\} > \min\{1, 0\} = 0 \quad \forall \lambda \in (0, 1).$$

These observations imply that $\frac{\mathrm{d}}{\mathrm{d}\lambda} I(s, \theta(\lambda)) = (\theta - s)/(\theta(\lambda) - 2s) > 0$ for all $\lambda \in (0, 1)$. If $s > b(\theta)$, finally, then we have $2s - \theta > 1$, which implies that $s > 0$ because $\theta > -1$. In addition, as $\theta \leq b(\theta)$, the assumption $s > b(\theta)$ also ensures that $s > \theta$. Using $s > 0$ and $2s - \theta > 1$, we thus find

$$2s - \theta(\lambda) \geq \min\{2s - s, 2s - \theta\} > \min\{0, 1\} = 0 \quad \forall \lambda \in (0, 1).$$

These observations imply that $\frac{\mathrm{d}}{\mathrm{d}\lambda} I(s, \theta(\lambda)) = (s - \theta)/(2s - \theta(\lambda)) > 0$ for all $\lambda \in (0, 1)$. Thus, $I(s, \theta(\lambda))$ is indeed strictly increasing in λ whenever $s \neq \theta$. In summary, $I(s, \theta)$ satisfies the inequality (3.2), which is sufficient for radial monotonicity; see the discussion after Definition 3.3. Thus, the claim follows. □

An alternative estimator for the autoregressive coefficient θ is the Yule-Walker estimator defined through

$$\widehat{S}_T = \frac{\sum_{t=2}^{T} \xi_t \xi_{t-1}}{\sum_{t=1}^{T} \xi_t^2} \quad \forall T \in \mathbb{N}. \tag{5.9}$$

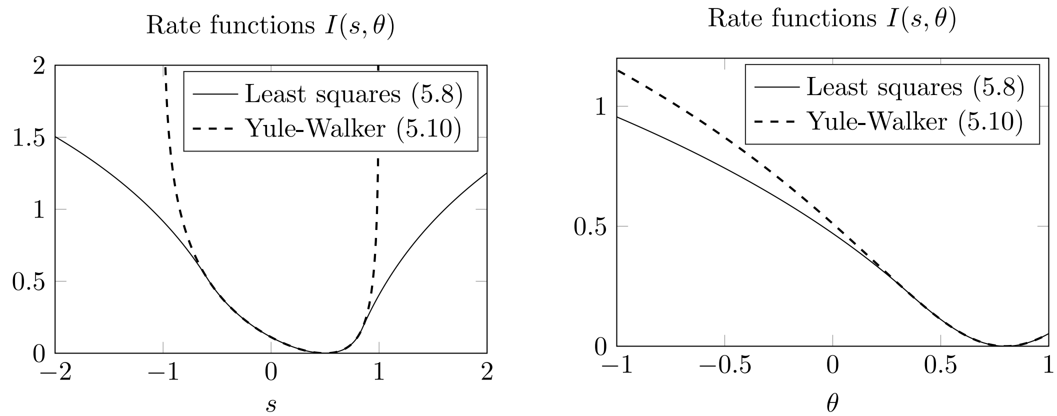
By construction, the state space of $\widehat{S}$ is given by $\mathbb{S} = \mathbb{R}$, and $\widehat{S}_T$ converges $\mathbb{P}_\theta$-almost surely to θ for all $\theta \in \Theta$ [10]. Hence, $S_\infty(\theta) = \theta$ and $\mathbb{S}_\infty = \Theta$, which means that the Yule-Walker estimator constitutes a consistent model estimator. Interestingly, it satisfies a different LDP than the least squares estimator.

Proposition 5.4 (LDP for stationary autoregressive processes with unknown coefficient (II))**.** *If $\{\xi_t\}_{t \in \mathbb{N}}$ follows a stationary autoregressive process of the form* (5.6) *with autoregressive parameter $\theta \in \Theta$, then the Yule-Walker estimator* (5.9) *satisfies an LDP with regular rate function*

$$I(s, \theta) = \begin{cases} \frac{1}{2} \log \left(\frac{1 - 2\theta s + \theta^2}{1 - s^2} \right) & \text{if } s \in (-1, 1), \\ 0 & \text{if } s = \theta = 1 \text{ or } s = \theta = -1, \\ +\infty & \text{otherwise.} \end{cases} \tag{5.10}$$

Proof. By [10, Proposition 8], the Yule-Walker estimator satisfies an LDP with rate function (5.10). By construction, this rate function is non-negative and continuous throughout $\mathbb{S} \times \mathrm{cl}\, \Theta$ except at the two points $(1, 1)$ and $(-1, -1)$, where the function is only lower semi-continuous. All sublevel sets of $I(x, \theta)$ are closed (as the function is lower semicontinuous) and bounded (as the function evaluates to ∞ outside of the bounded set $[-1, 1]^2$), and thus $I(s, \theta)$ satisfies the level-compactness condition of Definition 3.3. Radial monotonicity can be proved as in Proposition 5.3 with obvious minor modifications. Details are omitted for brevity. □

Figure 2 visualizes the rate functions (5.8) and (5.10) for fixed values of the estimator realization s and the model θ. As is also evident from their definitions, the two rate functions coincide whenever $s \in [a(\theta, b(\theta)]$. In general, however, the rate function corresponding to the Yule-Walker estimator majorizes the one corresponding to the least squares estimator. This indicates that the probability of unlikely estimator realizations decays faster when we use the Yule-Walker estimator. One can show that neither the least squares nor the Yule-Walker estimator represent a sufficient statistic for θ. Therefore, the corresponding distributionally robust predictors and prescriptors cannot be used to construct Pareto dominant solutions for the original meta-optimization problems (2.4). As both statistics satisfy an LDP with a regular rate function (as shown in Propositions 3.1 and 3.2), however, the corresponding distributionally robust predictors and prescriptors are strongly optimal in the respective *restricted* meta-optimization problems (2.6). We also emphasize that these predictor-prescriptor pairs are *not* equivalent. Indeed, for any desired decay rate $r \geq 0$ of the out-of-sample disappointment and for any fixed estimator realization $s \in \mathbb{S}$, the predictor induced by the Yule-Walker estimator is less conservative than the one induced by the least squares estimator because the rate ball of radius r around s corresponding to the Yule-Walker estimator is always contained in the rate ball of radius r around s corresponding to the least squares estimator. Intuitively, the Yule-Walker estimator thus results in a less conservative predictor with the same guarantees on the out-of-sample disappointment.



(a) $I(s,\theta)$ for fixed $\theta = 0.5$; see also [10, Figure 1].

(b) $I(s,\theta)$ for fixed $s = 0.8$.

Figure 2: Comparison of the rate functions corresponding to the least squares and Yule-Walker estimators.

5.4 I.i.d. processes with parametric distribution functions

As a last example, assume that the observations $\{\xi_t\}_{t=1}^T$ are valued in $\mathbb{R}^m$ and that they are serially independent and share the same distribution function $F_{\theta_\star}$ under $\mathbb{P}_\star$, that is, we have $\mathbb{P}_\star[\xi_t \leq z] = F_{\theta_\star}(z)$ for all $z \in \mathbb{R}^m$ and $t \in \mathbb{N}$. Here, F_θ, $\theta \in \Theta$, represents a family of distribution functions with common support $\Xi \subseteq \mathbb{R}^m$, where the parameter θ ranges over a relatively open convex subset Θ of $\mathbb{R}^d$, and $\theta_\star$ denotes the unknown true parameter. Clearly, the mean value of F_θ must be a function of θ and can thus be expressed as $S_\infty(\theta)$. Throughout this section we assume that the function S_∞ constitutes a homeomorphism from the set Θ to its image $\mathbb{S}_\infty = \{S_\infty(\theta) : \theta \in \Theta\}$. As any homeomorphism is invertible, this assumption means that the parameter θ is uniquely determined by the mean value of F_θ. We may then conclude that $\mathbb{P}_\star$ belongs to an ambiguity set $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where each $\theta \in \Theta$ encodes a probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ satisfying

$$\mathbb{P}_\theta[\xi_t \leq z_t \ \forall t = 1, \ldots, T] = \prod_{t=1}^T F_\theta(z_t) \qquad \forall z \in \mathbb{R}^{mT},\ T \in \mathbb{N}.$$

In order to estimate the mean value $S_\infty(\theta)$ (and thereby implicitly also θ) we use the sample mean

$$\widehat{S}_T = \frac{1}{T}\sum_{t=1}^T \xi_t \quad \forall T \in \mathbb{N}. \tag{5.11}$$

By our standard conventions, the state space $\mathbb{S}$ of $\widehat{S}$ is given by the closure of the convex hull of Ξ. In the following we assume that the distribution function F_θ has exponentially bounded tails for every $\theta \in \Theta$. The strong law of large numbers then implies that $\widehat{S}_T$ converges $\mathbb{P}_\theta$-almost surely to $S_\infty(\theta)$. More specifically, we henceforth focus on several popular families of distribution functions that are susceptible to analytical treatment:

(a) normal distributions on $\mathbb{R}^m$ with an unknown mean vector $\theta \in \mathbb{R}^m$ and a positive definite covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$;

(b) exponential distributions on $\mathbb{R}_+$ with an unknown rate parameter $\theta > 0$;

(c) Gamma distributions on $\mathbb{R}_+$ with an unknown scale parameter $\theta > 0$ and a shape parameter $k > 0$;

(d) Poisson distributions on $\mathbb{N} \cup \{0\}$ with an unknown rate parameter $\theta \in \mathbb{R}_{++}$;

(e) Bernoulli distributions on $\{0, 1\}$ with an unknown success probability $\theta \in (0, 1)$;

(f) geometric distributions on $\mathbb{N}$ with an unknown success probability $\theta \in (0, 1)$;

(g) binomial distributions on $\mathbb{N} \cup \{0\}$ with an unknown success probability $\theta \in (0, 1)$ and $N \in \mathbb{N}$ trials.

Clearly, each of these examples satisfies Assumption 2.1. It is also well known that each of these examples gives rise to a time-homogeneous exponential family in the sense of Assumption 4.1 and that the sample mean (5.11) is a sufficient statistic for θ. To see that the sample mean also satisfies an LDP with a regular rate function, note that for i.i.d. data the limiting log-moment generating function simplifies to

$$\Lambda(\lambda, \theta) = \lim_{T\to\infty} \frac{1}{T} \log \mathbb{E}_\theta \left[\exp(\langle T\lambda, \widehat{S}_T\rangle)\right] = \log\left(\int_{\mathbb{R}^m} e^{\lambda^\top \xi}\, \mathrm{d}F_\theta(\xi)\right). \tag{5.12}$$

As F_θ is assumed to have exponentially bounded tails, $\Lambda(\lambda, \theta)$ is finite on a neighborhood of $\lambda = 0$ for every fixed $\theta \in \Theta$. Moreover, $\Lambda(\lambda, \theta)$ is available in closed form for all families of distribution functions listed above; see Table 1. In each case one can therefore verify by inspection that the gradient $\nabla_\lambda \Lambda(\lambda, \theta)$ exists on the interior of $\operatorname{dom} \Lambda(\cdot, \theta)$ and that its norm tends to infinity when λ approaches the boundary of $\operatorname{dom} \Lambda(\cdot, \theta)$. Thus, Assumption 4.2 holds, which ensures via the Gärtner-Ellis theorem that $\widehat{S}$ satisfies an LDP. The corresponding rate function $I(s, \theta)$ coincides with the Cramér function $\Lambda^*(s, \theta)$, that is, the convex conjugate of the limiting log-moment generating function (5.12) with respect to λ. The Cramér function is again available in closed form for all examples listed above; see Table 1. In each case one can verify by inspection that $\Lambda^*(s, \theta)$ represents in fact a regular rate function. By Theorems 3.1 and 3.2, the distributionally robust predictor $\tilde{c}^\star$ constructed from the Cramér function and the corresponding prescriptor $\tilde{x}^\star$ thus provide Pareto dominant solutions for the restricted meta-optimization problems (2.6). Moreover, by Theorems 4.2 and 4.3 the data-driven predictor $\widehat{c}^\star$ and the corresponding prescriptor $\widehat{x}^\star$ induced by $\tilde{c}^\star$ and $\tilde{x}^\star$, respectively, provide Pareto dominant solutions for the original meta-optimization problems (2.4).

Remark 5.1 (I.i.d. processes as degenerate autoregressive processes)**.** *Any i.i.d. process of multivariate normal random variables can alternatively be interpreted as a degenerate vector autoregressive process of the type studied in Section 5.2 with a vanishing coefficient matrix $A = 0$. It is therefore not surprising that if $A = 0$, then the scaled sample mean* (5.4) *coincides with the ordinary sample mean* (5.11)*, and the rate function derived in Proposition 5.2 coincides with the Cramér function in Table 1(a).*

6 Concluding reflections on efficient estimators

This paper proposes a rigorous framework for identifying optimal estimators for the objective functions and the optimal solutions of data-driven decision problems. The study of the fundamental performance limitations and the efficiency properties of various estimators has of course a long and distinguished history in statistics. To close this paper, we highlight several connections between the Pareto dominance properties of data-driven predictors studied in Sections 3 and 4 and some classical efficiency concepts.

Throughout this section we fix an arbitrary $x \in X$. Any data-driven predictor $\widehat{c}_T(x)$ can be regarded as an estimator for the cost $c(x,\theta)$ of x under the probability measure $\mathbb{P}_\theta$. If one is only interested in cost prediction, then the symmetric error probability $\mathbb{P}_\theta[(\widehat{c}_T(x) - c(x,\theta))^2 > \varepsilon_T^2]$ for some prescribed error tolerance $\varepsilon_T > 0$ represents a more appropriate performance measure than the asymmetric out-of-sample disappointment introduced in Definition 2.4. There is indeed a vast literature on quantifying the statistical efficiency of estimators based on how fast this error probability decays to zero as the sample size T grows. Estimators for which this decay is in some precise sense as fast as possible are designated as efficient. There are two classical notions of efficiency that correspond to different asymptotic regimes of the error tolerance ε_T.

The Cramér-Rao inequality guarantees that the variance of $T^{\frac{1}{2}}(\widehat{c}_T(x) - c(x,\theta))$ is bounded below by the inverse Fisher information whenever $\widehat{c}_T(x)$ represents an unbiased estimator for $c(x,\theta)$ and some standard regularity conditions are met. Estimators that attain this bound asymptotically are termed *relatively Pitman efficient* [31]. For such estimators the error probability $\mathbb{P}_\theta[(\widehat{c}_T(x) - c(x,\theta))^2 > \varepsilon_T^2]$ can be guaranteed to remain uniformly small if the error tolerances decay as $\varepsilon_T = \mathcal{O}(T^{-\frac{1}{2}})$.

When focusing on constant error tolerances $\varepsilon_T = \varepsilon > 0$, on the other hand, then the error rate

$$e(\varepsilon,\theta,\widehat{c}(x)) = \limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta \left[(\widehat{c}_T(x) - c(x,\theta))^2 > \varepsilon^2 \right]$$

may be used as an appropriate yardstick for comparing estimators [5, 8]. Bahadur proved under standard regularity conditions that the error rate $e(\varepsilon,\theta,\widehat{c}(x))$ of any consistent estimator $\widehat{c}(x)$ is bounded below by a function $b(\varepsilon,\theta)$ and thus established a constant error counterpart to the Cramér-Rao bound [5]. As small error tolerances are particularly important, it is sometimes reasonable to measure the quality of an estimator by its error rate in the limit when ε tends to 0. Accordingly, an estimator is called *locally Bahadur efficient* if

$$\lim_{\varepsilon\to 0} \frac{e(\varepsilon,\theta,\widehat{c}(x))}{b(\varepsilon,\theta)} = 1 \quad \forall \theta \in \Theta.$$

Such estimators attain Bahadur's lower bound for small values of ε. Similarly, an estimator is called *globally Bahadur efficient* if $e(\varepsilon,\theta,\widehat{c}(x)) = b(\varepsilon,\theta)$ for all $\varepsilon > 0$ and $\theta \in \Theta$. As $e(\varepsilon,\theta,\widehat{c}(x))$ is never smaller than $b(\varepsilon,\theta)$, such an estimator constitutes a Pareto dominant solution of the multi-objective optimization problem

$$\underset{\widehat{c}}{\text{minimize}} \; \{e(\varepsilon,\theta,\widehat{c}(x))\}_{\varepsilon>0,\,\theta\in\Theta},$$

which is reminiscent of (2.4a). A globally efficient estimator, should it exist, enjoys an optimal error rate for constant errors of any size $\varepsilon > 0$. As most multi-objective optimization problems admit no Pareto dominant solutions, the existence of Bahadur efficient estimators can not be taken for granted. They are in fact only known to exist if the ambiguity set $\mathcal{P}$ constitutes an exponential family, and there is strong evidence suggesting that they do not exist for more general ambiguity sets [31]. These findings are in line with the strong optimality results presented in Section 4, which also require $\mathcal{P}$ to represent an exponential family.

Acknowledgements. This research was supported by the Swiss National Science Foundation under the NCCR Automation, grant agreement 51NF40_180545.

References

[1] Charalambos Aliprantis and Kim Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2007.

[2] Panos J. Antsaklis and Anthony N. Michel. *Linear Systems*. Birkhäuser, 2006.

[3] Robert Ash and Catherine A. Doléans-Dade. *Probability and Measure Theory*. Academic Press, 2000.

[4] Jean-Pierre Aubin and Hélène Frankowska. *Set-Valued Analysis*. Birkhäuser, 2009.

[5] Raghu R. Bahadur. Stochastic comparison of tests. *Annals of Mathematical Statistics*, 31(2):276–295, 1960.

[6] Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

[7] Ole Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, 2014.

[8] Debabrata Basu. The concept of asymptotic efficiency. *Sankhyā: The Indian Journal of Statistics*, 17(2):193–196, 1956.

[9] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[10] Bernard Bercu, Fabrice Gamboa, and Alain Rouault. Large deviations for quadratic forms of stationary Gaussian processes. *Stochastic Processes and their Applications*, 71(1):75–90, 1997.

[11] Claude Berge. *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*. Courier Corporation, 1997.

[12] Dimitri Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[13] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Robust sample average approximation. *Mathematical Programming*, 171(1-2):217–282, 2018.

[14] Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

[15] Patrick Billingsley. Statistical methods in Markov chains. *Annals of Mathematical Statistics*, 32:12–40, 1961.

[16] David Blackwell. Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, 18:105–110, 1947.

[17] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[18] Woodrow Bledsoe. Neighborly functions. *Proceedings of the American Mathematical Society*, 3:114–115, 1952.

[19] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[20] Thomas Cover and Joy Thomas. *Elements of Information Theory*. Wiley, 2006.

[21] Erik Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[22] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2009.

[23] Esther Derman and Shie Mannor. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

[24] Xialiang Dou and Mihai Anitescu. Distributionally robust optimization with correlated data from vector autoregressive processes. *Operations Research Letters*, 47(4):294–299, 2019.

[25] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *Mathematics of Operations Research*, 2021. Ahead of Print.

[26] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2010.

[27] Rui Gao. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, 2020.

[28] Vishal Gupta. Near-optimal Bayesian ambiguity sets for distributionally robust optimization. *Management Science*, 65(9):4242–4260, 2019.

[29] Grani Hanasusanto, Daniel Kuhn, and Wolfram Wiesemann. A comment on "computational complexity of stochastic programming problems". *Mathematical Programming*, 159(1):557–569, 2016.

[30] Leonid Vasilevich Kantorovich and Gennady S. Rubinshtein. On a space of totally additive functions. *Vestnik Leningradskogo Universiteta*, 13(7):52–59, 1958.

[31] Arnold Kester. *Some Large Deviation Results in Statistics*. Centrum voor Wiskunde en Informatica, 1985.

[32] Alan King and R. Tyrrell Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18(1):148–162, 1993.

[33] Alan King and Roger J.-B. Wets. Epi-consistency of convex stochastic programs. *Stochastics and Stochastics Reports*, 34(1-2):83–92, 1991.

[34] Bernard Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39(3):399–409, 1936.

[35] Uwe Küchler and Michael Sørensen. *Exponential Families of Stochastic Processes*. Springer, 2006.

[36] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

[37] Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

[38] Olivier Le Maître and Omar Knio. *Introduction: Uncertainty Quantification and Propagation*. Springer, 2010.

[39] Erich Leo Lehmann and George Casella. *Theory of Point Estimation*. Springer, 1998.

[40] Mengmeng Li, Tobias Sutter, and Daniel Kuhn. Distributionally robust optimization based on Markovian data. In preparation, 2021.

[41] William S. Massey. *A Basic Course in Algebraic Topology*. Springer, 1991.

[42] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[43] Whitney Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, pages 2111–2245. Elsevier, 1994.

[44] Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

[45] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[46] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, 1998.

[47] Sheldon Ross. *Introduction to Probability Models*. Elsevier, 2010.

[48] Alexander Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *Annals of Statistics*, 17(2):841–858, 1989.

[49] Alexander Shapiro. On differential stability in stochastic programming. *Mathematical Programming*, 47(1):107–116, 1990.

[50] Alexander Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30(1-4):169–186, 1991.

[51] Alexander Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18(4):829–845, 1993.

[52] Alexander Shapiro. Monte Carlo sampling methods. In *Stochastic Programming*, pages 353–425. Elsevier, 2003.

[53] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014.

[54] James Smith and Robert Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

[55] Bart Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 2021. Articles in Advance.

[56] Halbert White. *Asymptotic Theory for Econometricians.* Emerald Group Publishing, 2001.

A Proof of Proposition 3.1

In order to prove Proposition 3.1, we have to recall several notions of continuity for set-valued mappings.

Definition A.1 (Continuity of set-valued mappings [11, Chapter VI, Section 1])**.** *Consider a set-valued mapping $\Gamma : X \rightrightarrows Y$ between two topological spaces.*

(i) Γ *is called lower semi-continuous (lsc) at* x_0 *if for every open set* $V \subseteq Y$ *with* $\Gamma(x_0) \cap V \neq \emptyset$ *there exists an open neighborhood* $U \subseteq X$ *of* x_0 *such that* $\Gamma(x) \cap V \neq \emptyset$ *for all* $x \in U$*;*

(ii) Γ *is called upper semi-continuous (usc) at* x_0 *if for every open set* $V \subseteq \Theta$ *with* $\Gamma(x_0) \subseteq V$ *there exists an open neighborhood* $U \subseteq Y$ *of* x_0 *such that* $\Gamma(x) \subseteq V$ *for all* $x \in U$*;*

(iii) Γ *is called continuous at* x_0 *if it is both lsc and usc at* x_0*;*

(iv) Γ *is called lsc if it is lsc at every point* $x_0 \in X$*;*

(v) Γ *is called usc if it is compact-valued and usc at every point* $x_0 \in X$*;*

(vi) Γ *is called continuous if it is lsc and usc.*

Proof of Proposition 3.1. We first show that the distributionally robust predictor $\tilde{c}^\star(x, s)$ is continuous in (x, s) on $X \times \mathbb{S}_\infty$. To this end, define the set-valued mapping $\Gamma : \mathbb{S} \rightrightarrows \operatorname{cl}\Theta$ through $\Gamma(s) = \{\theta \in \operatorname{cl}\Theta : I(s, \theta) \leq r\}$ for every $s \in \mathbb{S}$. Note that the graph $\{(s, \theta) \in \mathbb{S} \times \operatorname{cl}\Theta : I(s, \theta) \leq r\}$ of Γ is compact because the regular rate function $I(s, \theta)$ has compact sublevel sets. Hence, Γ has a closed graph and is compact-valued, which implies via [4, Proposition 1.4.8] that Γ is usc. Recall now that $\operatorname{cl}\Theta$ is equipped with the subspace topology induced by the Euclidean topology on $\mathbb{R}^d$, and choose any $s_0 \in \mathbb{S}_\infty$ and any open set $V \subseteq \operatorname{cl}\Theta$ with $\Gamma(s_0) \cap V \neq \emptyset$. As Θ is a relatively open convex subset of $\mathbb{R}^d$ (see Assumption 2.1), it is open with respect to the subspace topology on $\operatorname{cl}\Theta$. Thus, both V and Θ are open, which implies that $V \subseteq \operatorname{int}\operatorname{cl}\Theta = \Theta$. Hence, there exists $\theta_0 \in V \subseteq \Theta$ with $I(s_0, \theta_0) \leq r$. In the following we may assume without loss of generality that $I(s_0, \theta_0) < r$. Suppose to the contrary that $I(s_0, \theta_0) = r$. Since $s_0 \in \mathbb{S}_\infty$ and $r > 0$ and since the regular rate function $I(s, \theta)$ is radially monotonic in θ, there exist $\theta_k \in \Theta$, $k \in \mathbb{N}$, such that $I(s_0, \theta_k) < r$ for all $k \in \mathbb{N}$ and $\lim_{k\to\infty} \theta_k = \theta_0$. As $\theta_0 \in V$ and V is open, there further exists $k_V \in \mathbb{N}$ such that $\theta_k \in V$ and $I(s_0, \theta_k) < r$ for all $k \geq k_V$. We may thus re-define θ_0 as θ_k for any $k \geq k_V$. Next, define $U = \{s \in \mathbb{S}_\infty : I(s, \theta_0) < r\}$, and note that U is open because $\mathbb{S}_\infty$ is open and the regular rate function $I(s, \theta)$ is continuous on $\mathbb{S} \times \Theta$. By construction, we have that $s_0 \in U$ and $\theta_0 \in \Gamma(s) \cap V$ for all $s \in U$. Thus, Γ is lsc at s_0. As $s_0 \in \mathbb{S}_\infty$ was chosen arbitrarily, Γ is indeed lsc on $\mathbb{S}_\infty$. Being lsc as well as usc, Γ represents a continuous set-valued mapping on $\mathbb{S}_\infty \subseteq \mathbb{S}$. In addition, $\Gamma(s)$ is non-empty for every $s \in \mathbb{S}_\infty$ because there exists $\theta \in \Theta$ with $s = S_\infty(\theta)$. Indeed, by the discussion after Definition 3.2, we have $I(s, \theta) = 0$ and thus $\theta \in \Gamma(s)$. As the model-based predictor $c(x, \theta)$ is continuous on $X \times \operatorname{cl}\Theta$ due to the arguments outlined after Assumption 2.2, we may finally invoke Berge's maximum theorem [11, pp. 115–116]) to conclude that the distributionally robust predictor $\tilde{c}^\star(x, s) = \max_{\theta\in\Gamma(s)} c(x, \theta)$ is continuous on $X \times \mathbb{S}_\infty$.

We may use a similar but significantly simpler reasoning to demonstrate that $\tilde{c}^\star(x,s)$ is continuous in x on $X \times \mathbb{S}$. The simplification arises because the set-valued mapping $\Gamma(s)$ is constant and thus trivially continuous in x for any fixed $s \in \mathbb{S}$. Finally, $\tilde{c}^\star(x,s)$ inherits boundedness from $c(x,\theta)$; see Assumption 2.2. $\square$

B Auxiliary probabilistic results

Throughout this section we assume that all random objects are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a fixed probability measure $\mathbb{P}$, and we denote the expectation operator with respect to $\mathbb{P}$ by $\mathbb{E}[\cdot]$.

Lemma B.1. *If the real-valued random variables z_T, $T \in \mathbb{N}$, converge in probability to a random variable z_∞ and if there exists a random variable $\bar{z}$ with $|z_T| \leq \bar{z}$ for all $T \in \mathbb{N}$ and $\mathbb{E}[\bar{z}] < \infty$, then $\lim_{T\to\infty} \mathbb{E}[z_T] = \mathbb{E}[z_\infty]$.*

Proof. If $\{z_{T(m)}\}_{m\in\mathbb{N}}$ is a subsequence of $\{z_T\}_{T\in\mathbb{N}}$, then [26, Theorem 2.3.2] ensures that there exists a further subsequence $\{z_{T(m_k)}\}_{k\in\mathbb{N}}$ that converges almost surely to z_∞. By the dominated convergence theorem, we may thus conclude that $\lim_{k\to\infty} \mathbb{E}[z_{T(m_k)}] = \mathbb{E}[z_\infty]$. As the subsequence $\{z_{T(m)}\}_{m\in\mathbb{N}}$ was chosen arbitrarily, this finally implies via [26, Theorem 2.3.3] that $\lim_{T\to\infty} \mathbb{E}[z_T] = \mathbb{E}[z_\infty]$. $\square$

Lemma B.2. *If the random variables z_T, $T \in \mathbb{N}$, take values in a compact state space $Z \subseteq \mathbb{R}^n$, then there exists a deterministic vector $z_\infty \in Z$ such that $\limsup_{T\to\infty} \mathbb{P}[\|z_T - z_\infty\| < \rho] > 0$ for all $\rho > 0$.*

If the sequence $\{z_T\}_{T\in\mathbb{N}}$ converges in probability, then z_∞ may be set to any point in the support of the limiting random variable. We emphasize, however, that Lemma B.2 remains valid even if the sequence $\{z_T\}_{T\in\mathbb{N}}$ fails to converge. One can thus think of z_∞ as a probabilistic accumulation point of $\{z_T\}_{T\in\mathbb{N}}$.

Proof of Lemma B.2. As Z is bounded, we may assume without loss of generality that $\|z\| \leq 1$ for all $z \in Z$. We then construct z_∞ as follows. First, we define $z_\infty^0 = 0$. Next, for each $k \in \mathbb{N}$, we set $\rho^k = 1/2^{k-1}$, and we recursively use the procedure described below to construct $z_\infty^k \in Z$ with $\|z_\infty^k - z_\infty^{k-1}\| \leq \rho^k$ and

$$\limsup_{T\to\infty} \mathbb{P}\left[\|z_T - z_\infty^k\| < \rho^k\right] > 0. \tag{B.1}$$

Before showing how one can construct iterates z_∞^k with these properties, we explain how they can be used to prove the the lemma. To this end, note that $\{z_\infty^k\}_{k\in\mathbb{N}}$ represents a Cauchy sequence because

$$\|z_\infty^k - z_\infty^{k'}\| \leq \sum_{\ell=k}^{k'-1} \|z_\infty^{\ell+1} - z_\infty^\ell\| \leq \sum_{\ell=k}^{k'-1} \frac{1}{2^\ell} \leq \sum_{\ell=k}^{\infty} \frac{1}{2^\ell} = \frac{1}{2^k} = \frac{\rho^k}{2} \quad \forall k < k'$$

and because ρ^k converges to 0 as k grows. As $Z \subseteq \mathbb{R}^n$ is compact, the sequence $\{z_\infty^k\}_{k\in\mathbb{N}}$ thus converges to a point $z_\infty \in Z$ that satisfies $\|z_\infty^k - z_\infty\| \leq \rho^k/2$ for all $k \in \mathbb{N}$. Next, select any $\rho > 0$ and an arbitrary $k \in \mathbb{N}$ with $\rho^k < 2\rho/3$. The triangle inequality then implies that for all $z \in Z$ with $\|z - z_\infty^k\| < \rho^k$ we have

$$\|z - z_\infty\| \leq \|z - z_\infty^k\| + \|z_\infty^k - z_\infty\| \leq 3\rho^k/2 < \rho.$$

We may thus conclude that

$$\limsup_{T\to\infty} \mathbb{P}\left[\|z_T - z_\infty\| < \rho\right] \geq \limsup_{T\to\infty} \mathbb{P}\left[\|z_T - z_\infty^k\| < \rho^k\right] > 0,$$

where the strict inequality follows from (B.1). As $\rho > 0$ was chosen arbitrarily, the claim follows.

It remains to be shown that one can always construct iterates $z_\infty^k \in Z$ with $\|z_\infty^k - z_\infty^{k-1}\| \leq \rho^k$ that satisfy (B.1). To see this, initialize the iteration counter as $k = 1$, and set $Z^k = Z$. As Z^k is bounded, there exist a finite index set $\mathcal{J}^k$ and finitely many points $\overline{z}_j^k \in Z^k$, $j \in \mathcal{J}^k$, such that the open balls $B_j^k = \{z \in Z : \|z - \overline{z}_j^k\| < \rho^k\}$, $j \in \mathcal{J}^k$, cover Z^k. Next, select $j^k \in \mathcal{J}^k$ with $\limsup_{T\to\infty} \mathbb{P}[z_T \in B_{j^k}^k] > 0$, and set $z_\infty^k = \overline{z}_{j^k}^k$. Note that j^k exists because $\mathbb{P}[z_T \in \cup_{j\in\mathcal{J}^k} B_j^k] = \mathbb{P}[z_T \in Z^k]$ for all $T \in \mathbb{N}$ and because $\limsup_{T\to\infty} \mathbb{P}[z_T \in Z^k] > 0$. Finally, define $Z^{k+1} = B_{j^k}^k$, increment the iteration counter k and repeat the above procedure. By construction, z_∞^{k+1} belongs to Z as well as to the ball of radius ρ^k around z_∞^k, and it satisfies (B.1) for every $k \in \mathbb{N}$. Hence, the sequence $\{z_\infty^k\}_{k\in\mathbb{N}}$ displays all desired properties. $\square$

Law of ξ_t	$S_\infty(\theta)$	log-MGF $\Lambda(\lambda,\theta)$	$\mathrm{dom}(\Lambda(\cdot,\theta))$	Cramér Function $\Lambda^*(s,\theta)$	$\mathrm{dom}(\Lambda^*(\cdot,\theta))$
(a) Normal	θ	$\theta^\top\lambda + \frac{1}{2}\lambda^\top\Sigma\lambda$	$\mathbb{R}^d$	$\frac{1}{2}(s-\theta)^\top\Sigma^{-1}(s-\theta)$	$\mathbb{R}^d$
(b) Exponential	$1/\theta$	$\log(\frac{\theta}{\theta-\lambda})$	$(-\infty,\theta)$	$\theta s - 1 - \log(\theta s)$	$\mathbb{R}_{++}$
(c) Gamma	$k\theta$	$-k\log(1-\theta\lambda)$	$(-\infty,1/\theta)$	$s/\theta - k + k\log(k\theta/s)$	$\mathbb{R}_{++}$
(d) Poisson	θ	$\theta(e^\lambda - 1)$	$\mathbb{R}$	$s\log(s/\theta) - s + \theta$	$\mathbb{R}_{++}$
(e) Bernoulli	θ	$\log(1-\theta+\theta e^\lambda)$	$\mathbb{R}$	$s\log(\frac{s(1-\theta)}{\theta(1-s)}) - \log(\frac{1-\theta}{1-s})$	$(0,1)$
(f) Geometric	$1/\theta$	$\lambda + \log(\frac{\theta}{1-(1-\theta)e^\lambda})$	$(-\infty,-\log(1-\theta))$	$(s-1)\log(\frac{1-s}{s(\theta-1)}) - \log(\theta s)$	$(1,\infty)$
(g) Binomial	$N\theta$	$N\log(1-\theta+\theta e^\lambda)$	$\mathbb{R}$	$s\log(\frac{s(\theta-1)}{\theta(s-N)}) - N\log(\frac{N(1-\theta)}{N-s})$	$(0,N)$

Table 1: Log-moment generating functions (log-MGFs) and their conjugates for popular families of distribution functions.