Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Pareto Dominance Principle for Data-Driven Optimization

Tobias Sutter, Bart P. G. Van Parys, Daniel Kuhn

OPERATIONS RESEARCH
Vol. 72, No. 5, September–October 2024, pp. 1976–1999
ISSN 0030-364X (print), ISSN 1526-5463 (online)
https://pubsonline.informs.org/journal/opre

Crosscutting Areas

A Pareto Dominance Principle for Data-Driven Optimization

Tobias Sutter,[a,*] Bart P. G. Van Parys,[b] Daniel Kuhn[c]

[a] Department of Computer Science, University of Konstanz, 78464 Konstanz, Germany; [b] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142; [c] Risk Analytics and Optimization Chair, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

*Corresponding author

Contact: tobias.sutter@uni-konstanz.de, https://orcid.org/0000-0003-1226-6845 (TS); vanparys@mit.edu, https://orcid.org/0000-0003-4177-4849 (BPGVP); daniel.kuhn@epfl.ch, https://orcid.org/0000-0003-2697-8886 (DK)

Received: September 14, 2021
Revised: December 2, 2022; August 25, 2023
Accepted: December 4, 2023
Published Online in Articles in Advance: January 19, 2024

Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2021.0609

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 in-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 are not independent and identically distributed. 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.

Funding: This research was supported by the Swiss National Science Foundation under the NCCR Automation [Grant Agreement 51NF40_180545].
Supplemental Material: The online appendices are available at https://doi.org/10.1287/opre.2021.0609.

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 capture 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, the long-run average expected cost of a parametric control policy, and so on. 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. 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 nontrivial 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. However, $\widehat{x}_T$ may be induced by many different data-driven predictors $\widehat{c}_T$, and 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 out-of-sample risk strictly exceeds the in-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. 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 tradeoff 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

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

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 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 multiobjective 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 $\theta \in \Theta$. The risk-aversion parameter r is chosen by the decision maker. Although it plays the role of a hyperparameter, it is directly interpretable thanks to its link to the out-of-sample disappointment. Multiobjective 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. Thus, such Pareto dominant solutions simultaneously minimize all objective functions of the meta-optimization problem. 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 $\widehat{S}_T$ and 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 problem 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 nontrivial extension of the celebrated Rao-Blackwell theorem (Rao 1945, Blackwell 1947) 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, and i.i.d. processes with parametric distribution functions.

Our results 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 are not 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 Van Parys et al. (2021). 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.

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. (Shapiro 1989, 1990, 1991, 1993; King and Wets 1991; King and Tyrrell Rockafellar 1993; Shapiro et al. 2014). In particular, the optimal value of the SAA problem is known to be strongly consistent and asymptotically normal (Shapiro et al. 2014, 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 (Shapiro 2003) or the optimizer's curse (Smith and Winkler 2006). 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 (Delage and Ye 2010). 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. Because 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 (Ben-Tal et al. 2013), for ambiguity sets containing all distributions that pass a statistical goodness-of-fit test against the observed training data (Bertsimas et al. 2018), or for ambiguity sets containing all distributions that are close to the empirical distribution with respect to some Wasserstein distance (Mohajerin Esfahani and Kuhn 2018, Kuhn et al. 2019). If these ambiguity sets are scaled 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 (Lam 2019, Duchi et al. 2021).

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 (Lam 2019, Duchi et al. 2021) and for Wasserstein ambiguity sets (Blanchet et al. 2019, Gao 2023). 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 Gupta (2019), 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 Lam (2019). 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 (Van Parys et al. 2021). 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 extend the main results of Van Parys et al. (2021) to more general (not necessarily risk-neutral) stochastic optimization problems, more general (not necessarily finitely supported) parametric distribution families and more general (not necessarily i.i.d.) data-generating stochastic processes. As a byproduct of our general theory of data-driven decision making, we discover several new DRO schemes that are statistically optimal for different structures of the data-generating stochastic process. Our theory thus provides practical guidance for choosing the best decision model in different data-driven decision situations. We also stress that Van Parys et al. (2021) assumes the predictors and prescriptors to depend on the training data only indirectly through the empirical distribution. Here, we do not impose such an implicit structure. Instead, we consider a much larger class of prescriptors that essentially depend on the training data in an arbitrary manner.

All statistical guarantees reviewed thus far rely indeed on the assumption that the training samples are i.i.d. Moreover, the literature on data-driven DRO with non-i.i.d. data are 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 (Duchi et al. 2021). 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 Dou and Anitescu (2019). Although 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 Derman and Mannor (2020). In this case, the training data set 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.

Although this paper was under review, its main results were extended along several dimensions. For example, if the training data are generated by a Markov chain, then the statistically optimal DRO models derived in Section 5.1 of this paper give rise to high-dimensional nonconvex optimization problems. An efficient Frank-Wolfe algorithm to solve these problems is developed in Li et al. (2021). In addition, a critical assumption of this paper is that the training and the test data are generated by the same stochastic process. This assumption is relaxed in Sutter et al. (2021), where the large deviation-type results of this paper for i.i.d. data are combined with the principle of minimum discriminating information to address data-driven decision problems suffering from a distribution shift. Another basic assumption of this paper is that the decision maker requires the out-of-sample disappointment to decay at a fixed exponential rate. This assumption can be relaxed using ideas from moderate deviations theory if the training samples are generated by a finite state i.i.d. process (Bennouna and Van Parys 2021). Specifically, it is shown that if the out-of-sample disappointment must decay superexponentially, then the Pareto dominant data-driven prescriptor is obtained by solving a classical robust optimization model that minimizes the worst-case risk with respect to all possible uncertainty realizations. Conversely, if the out-of-sample disappointment must decay subexponentially, then the Pareto dominant data-driven prescriptor is obtained by solving an empirical risk minimization problem with a variance penalty. Finally, we assume in this paper that the decision maker has access to noise-free training samples. This assumption is relaxed in Van Parys (2021), where a DRO model based on an entropic optimal tranport distance is shown to provide Pareto dominant data-driven prescriptors when the training samples are corrupted by noise.

The out-of-sample disappointment and the in-sample risk are by no means the only performance criteria for which the best representatives within a certain class of prescriptors are accessible. Another performance criterion of interest is the regret convergence rate. For example, in the context of data-driven linear optimixation with side information, it has recently been shown that the naïve "estimate and then optimize" approach is markedly superior to the "induced empirical risk method" with respect to this criterion (Hu et al. 2022).

The paper develops as follows. Section 2 formalizes our approach to data-driven decision-making and constructs the meta-optimization problems that will be used to find optimal data-driven predictors and prescriptors. Sections 3 and 4 establish sufficient conditions under which the restricted and original meta-optimization problems have Pareto dominant solutions, respectively, and Section 5 showcases practically relevant examples in which these conditions hold. All proofs, along with several auxiliary results, are provided in the online appendices.

1.1. Notation

A multiobjective 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 $f_\alpha(x^\star)\le f_\alpha(x)$ 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 $f_\alpha(x)\le f_\alpha(x^\star)$ 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 nonempty 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 $1_\mathcal{E}$ evaluates to one if $\mathcal{E}$ is true and to zero otherwise, and for any $A,B\in\mathbb{R}^{n\times m}$, the trace inner product is denoted by $\langle A,B\rangle=\mathrm{tr}(A^\top B)$.

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 following 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 (Shapiro et al. 2014, 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 (Ban and Rudin 2019, Bertsimas and Kallus 2020, Hu et al. 2022). If the matrix $C \in \mathbb{R}^{m \times m_C}$ filters out mC 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 parametrized 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 Objective Function (2.2d) both directly through the policy π_x and 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 Objective Functions (2.2a)–(2.2d) to be well defined. However, the previous 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 (Le Maître and Knio 2010); 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 nonnegative 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 $\mathbb{P}_\star$ (Hanasusanto et al. 2016, 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. 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 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 $\xi_t \in \Xi$, $t = 1, \ldots, T$, 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 $\widehat{\ }$ and a subscript T indicating the size of the underlying data set. 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. Here, 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. 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, whereas 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 data set) as the *out-of-sample disappointment*. 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 (Van Parys et al. 2021, example 2).

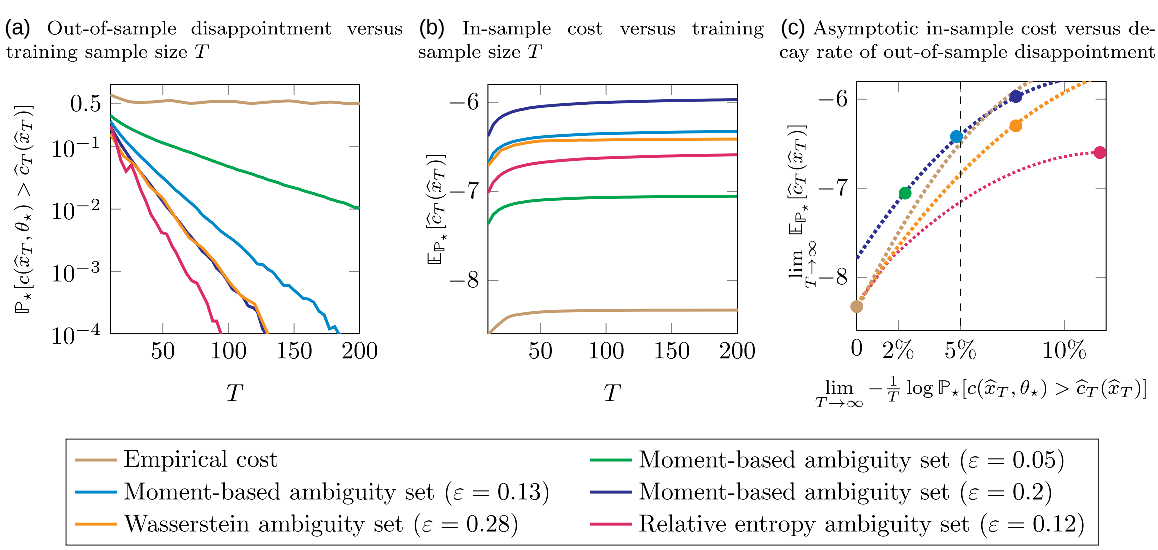
A naïve approach to force the out-of-sample disappointment to decay would be to add a constant positive penalty ε to the empirical cost estimator, thereby increasing its asymptotic in-sample cost and thus introducing a conservative bias. This reasoning suggests that the in-sample cost and the out-of-sample disappointment stand in direct competition. To provide a better intuition for the tradeoff 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 $\widehat{\Theta}_T = \{\theta \in \Delta_d : |\sum_{i\in\Xi} i^j\theta_i - \sum_{i\in\Xi} i^j(\widehat{S}_T)_i| \le \varepsilon \;\; \forall j = 1, \ldots, J\}$. 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 Delage and Ye 2010, section 3) for the first results of this kind). The recent literature has witnessed an increasing interest in Wasserstein ambiguity sets of the form $\widehat{\Theta}_T = \{\theta \in \Delta_d : \mathsf{d}_\mathsf{W}(\theta, \widehat{S}_T) \le \varepsilon\}$, where $\mathsf{d}_\mathsf{W}(\theta, \widehat{S}_T)$ denotes the first Wasserstein distance between θ and $\widehat{S}_T$ (Kantorovich and Rubinshtein 1958). This ambiguity set can be viewed as a Wasserstein ball of radius $\varepsilon \geq 0$ 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 (Mohajerin Esfahani and Kuhn 2018, section 3). Finally, we also study relative entropy ambiguity sets of the form $\widehat{\Theta}_T = \{\theta \in \Delta_d : D(\widehat{S}_T \| \theta) \leq \varepsilon\}$, 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 (Ben-Tal et al. 2013, section 3). In contrast to most of the existing literature on distributionally robust optimization, here we are *not* concerned about whether 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 and the tradeoff between the asymptotic in-sample cost and the decay rate of the out-of-sample disappointment for different estimators. Figure 1(a) 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 Figure 1, (a) and (b), correspond to the empirical cost estimator (light brown) and to distributionally robust cost estimators with a moment ambiguity set (green: $\varepsilon = 0.05$, light blue: $\varepsilon = 0.13$, dark blue: $\varepsilon = 0.2$), a Wasserstein ambiguity set (orange: $\varepsilon = 0.28$), and a relative entropy ambiguity set (magenta: $\varepsilon = 0.12$). The ε hyperparameters are chosen to ensure efficient use of the available plotting area. 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 1(c) visualize the tradeoff 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 1(c) correspond to the six estimators investigated in Figure 1, (a) and (b). As expected, the dashed lines corresponding to the distributionally robust cost estimators with a Wasserstein and a relative entropy ambiguity set intersect because both 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

Figure 1. (Color online) 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



Note. All probabilities and expectations involving random training data are evaluated empirically using 10^4 independent training sets.

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 a prescribed decay rate of the out-of-sample disappointment. For example, if we require a decay rate of at least 5%, all cost estimators on the right-hand side of the vertical dashed line in Figure 1(c) are feasible. A simple line search reveals that this includes all penalized empirical cost estimators with penalty $\varepsilon \geq 1.9$, all distributionally robust cost estimators with a moment ambiguity set of size $\varepsilon \geq 0.14$, all distributionally robust cost estimators with a Wasserstein ambiguity set of radius $\varepsilon \geq 0.23$, and all distributionally robust cost estimators with a relative entropy ambiguity set of radius $\varepsilon \geq 0.05$. However, many other estimators not considered in this experiment are also feasible. We endeavor to identify the least conservative of *all* such feasible estimators. In the remainder, we address this fundamental challenge 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 the relative interior of a convex subset of the finite-dimensional parameter space $\mathbb{R}^d$, 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 Van Parys et al. (2021) 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 $\mathbb{P}_\theta[\xi_t = i_t \; \forall t = 1, \ldots, T] = \prod_{t=1}^T \theta_{i_t} \;\; \forall i_t \in \Xi,\, t = 1, \ldots, T,\, T \in \mathbb{N}$.

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 $X \times \Theta$.*

If $c(x, \theta)$ is uniformly continuous and bounded on $X \times \Theta$, then it admits a unique uniformly continuous and bounded extension to $X \times \mathrm{cl}\Theta$ (Aliprantis and Border 2007, theorem 5.15). By slight abuse of notation, we will denote this extension by $c(x, \theta)$. 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. 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. **Nonanticipativity.** 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 nonnegative random variable $\overline{c}$ such that $\mathbb{E}_\theta[\overline{c}] < \infty$ for all $\theta \in \Theta$ and $|\widehat{c}_T(x)| \le \overline{c}\mathbb{P}_\theta$-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 1 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 nonanticipativity condition 2 implies via Ash and Doléans-Dade (2000) (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 3 is of technical nature and nonrestrictive in all examples studied in this paper. The convergence condition 4 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 are generated under $\mathbb{P}_\theta$. 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 5 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 are generated under $\mathbb{P}_\theta$. Thus, it may be a *biased* estimator for the optimal value of the stochastic optimization problem $\min_{x\in X} c(x,\theta)$. 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. **Nonanticipativity.** The process $\{\widehat{x}_T\}_{T\in\mathbb{N}}$ is adapted to the filtration $\{\mathcal{F}_T\}_{T\in\mathbb{N}}$.

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 nonanticipativity condition 1 implies via Ash and Doléans-Dade (2000) (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 2 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 because X is compact and because $\widehat{c}_T(x)$ 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 (Rockafellar and Roger 1998, 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 $\{\xi_t\}_{t\in\mathbb{N}}$. The newsvendor problem of Section 2.1 satisfies all these assumptions. We now define the empirical predictor $\widehat{c}$ through $\widehat{c}_T(x) = \frac{1}{T}\sum_{t=1}^{T}\ell(x,\xi_t)\ \ \forall T \in \mathbb{N}$. The empirical predictor simply evaluates the sample average of the loss across the observed data set and represents a data-driven predictor in the sense of Definition 2.2. Although the continuity condition 1 and the nonanticipativity assumption 2 hold by construction, the boundedness condition 3 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 4 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. Also,

$c_\infty(x,\theta)$ is continuous by Assumption 2.2. As ℓ is continuous and X is compact, the uniform law of large numbers (Newey and McFadden 1994, 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 5 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 are 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 are 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 are generated by a simple i.i.d. processes, such meta-optimization problems were already studied in Van Parys et al. (2021). 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 $\theta\in\Theta$. However, 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 $\theta\in\Theta$.

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 later aims to minimize the in-sample risk. As $\widehat{c}_T(x)$ for $x\in X$ and $\widehat{c}_T(\widehat{x}_T)$ 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 are 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 G.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 previous reasoning indicates that both the out-of-sample disappointment and 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 and 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* $\theta\in\Theta$. Because 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 multiobjective optimization problem*.

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

$$\begin{aligned}&\underset{\widehat{c}\in\widehat{\mathcal{C}}}{\text{minimize}}\ \left\{\lim_{T\to\infty}\mathbb{E}_\theta[\widehat{c}_T(x)]\right\}_{x\in X,\,\theta\in\Theta}\\ &\text{subject to}\ \limsup_{T\to\infty}\frac{1}{T}\log\mathbb{P}_\theta[c(x,\theta)>\widehat{c}_T(x)]\le -r\quad \forall x\in X,\ \theta\in\Theta.\end{aligned}\tag{2.4a}$$

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 $\mathbb{P}_\theta[c(x,\theta)>\widehat{c}_T(x)]\le e^{-rT+o(T)}$ 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

$$\begin{aligned}&\underset{(\widehat{c},\widehat{x})\in\widehat{\mathcal{X}}}{\text{minimize}}\ \left\{\lim_{T\to\infty}\mathbb{E}_\theta[\widehat{c}_T(\widehat{x}_T)]\right\}_{\theta\in\Theta}\\ &\text{subject to}\ \limsup_{T\to\infty}\frac{1}{T}\log\mathbb{P}_\theta[c(\widehat{x}_T,\theta)>\widehat{c}_T(\widehat{x}_T)]\le -r\quad \forall\theta\in\Theta.\end{aligned}\tag{2.4b}$$

As previously stated, $\lim_{T\to\infty}\mathbb{E}_\theta[\widehat{c}_T(\widehat{x}_T)]$ 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 $\mathbb{P}_\theta[c(\widehat{x}_T,\theta)>\widehat{c}_T(x)]\le e^{-rT+o(T)}$ 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 $\lim_{T\to\infty}\mathbb{P}_\theta[c(x,\theta)>\widehat{c}_T(x)]=1_{c(x,\theta)>c_\infty(x,\theta)}$ for all $x\in X$, $\theta\in\Theta$ 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.

Multiobjective optimization problems such as (2.4a) and (2.4b) typically only admit Pareto optimal solutions, that is, 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, that is, 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 Ash and Doléans-Dade (2000) (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 $\mathbb{S} = \text{cl}(\cup_{T\in\mathbb{N}} \Delta_d \cap (\mathbb{Z}^d/T)) = \text{cl}(\Delta_d \cap \mathbb{Q}^d) = \Delta_d = \text{cl}\Theta$.

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, the multifunction $\arg\min_{x\in X} \tilde{c}(x, s)$ is nonempty 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 Van Parys et al. (2021), which applies here because $\tilde{c}$ is continuous on $X \times \mathbb{S}_\infty$ and X is compact. 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 1–3 follow directly from the definitions of the statistic $\widehat{S}$ and the compressed data-driven predictor $\tilde{c}$. To check condition 4, 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 (Durrett 2010, 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 4 holds with $c_\infty(x, \theta) = \tilde{c}(x, S_\infty(\theta))$. To check condition 5, 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 (Berge 1997, pp. 115–116). Invoking the continuous mapping theorem as previously done, 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 5 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 1 follows directly from the defining properties of a statistic. To check condition 2, 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

$$\begin{aligned}\widehat{x}_T = \tilde{x}(\widehat{S}_T) &\in \arg\min_{x\in X} \tilde{c}(x, \widehat{S}_T)\\ &= \arg\min_{x\in X} \widehat{c}_T(x) \qquad \forall T \in \mathbb{N},\end{aligned}$$

where the two equalities follow from the definitions of $\widehat{x}_T$ and $\widehat{c}_T$, respectively, whereas the membership relation holds by assumption. Hence, $\widehat{x}$ is induced by the data-driven predictor $\widehat{c}$, and thus condition 2 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 $\text{cl}(\Theta)$, 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 $X \times \text{cl}(\Theta)$ because of 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 $\mathbb{S} = \text{cl}(\Theta)$, 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 $\theta \in \Theta$ because of the continuous mapping theorem (Durrett 2010, theorem 3.2.4) and Lemma G.1, the restricted meta-optimization problem can be formulated as follows.

$$\begin{aligned} &\underset{\tilde{c}\in\tilde{\mathcal{C}}}{\text{minimize}} \;\; \{\tilde{c}(x, S_\infty(\theta))\}_{x\in X,\,\theta\in\Theta} \\ &\text{subject to} \;\; \limsup_{T\to\infty} \frac{1}{T}\log \mathbb{P}_\theta[c(x,\theta) > \tilde{c}(x,\widehat{S}_T)] \le -r \\ &\qquad\qquad\qquad\qquad\qquad\qquad \forall x \in X,\, \theta\in\Theta \qquad (2.6a) \end{aligned}$$

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 G.1, we obtain the following restriction of the meta-optimization problem (2.4b).

$$\begin{aligned} &\underset{(\tilde{c},\tilde{x})\in\tilde{\mathcal{X}}}{\text{minimize}} \;\; \{\tilde{c}(\tilde{x}(S_\infty(\theta)), S_\infty(\theta))\}_{\theta\in\Theta} \\ &\text{subject to} \;\; \limsup_{T\to\infty}\frac{1}{T}\log\mathbb{P}_\theta[c(\tilde{x}(\widehat{S}_T),\theta) > \tilde{c}(\tilde{x}(\widehat{S}_T),\widehat{S}_T)] \\ &\qquad\qquad \le -r \quad \forall\theta\in\Theta \qquad (2.6b) \end{aligned}$$

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

3.1. Large Deviation Principles

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. 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 (Dembo and Zeitouni 2009, section 2.1)). A function $I : \mathbb{S}\times\text{cl}\Theta \to [0,\infty]$ is called a rate function if $I(s,\theta)$ is lower semicontinuous in s throughout $\mathbb{S}\times\text{cl}\Theta$.

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

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

$$\le \limsup_{T\to\infty}\frac{1}{T}\log\mathbb{P}_\theta[\widehat{S}_T\in\mathcal{D}]$$

$$\le -\inf_{s\in\text{cl}\mathcal{D}} I(s,\theta). \qquad (3.1b)$$

As the subspace topology on $\mathbb{S}$ induced by the Euclidean topology on $\mathbb{R}^d$ is Hausdorff, we know from Dembo and Zeitouni (2009, lemma 4.1.4) that if $\widehat{S}$ satisfies an LDP, then 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 $\mathbb{S}\times\text{cl}\Theta$. Although its values on the boundary of Θ are immaterial, extending the rate function to $\mathbb{S}\times\text{cl}\Theta$ will simplify some of the derivations in Section 3, provided the extension preserves the regularity conditions of Definition 3.3.

Before defining regular rate functions, we discuss a few immediate consequences of 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

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

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 nonnegativity and lower semicontinuity 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 zero 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 one as T grows. In general, 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 $\overline{r} = \inf_{s\in\mathrm{int}\mathcal{D}} I(s,\theta))$ represents the I distance between θ and the interior of $\mathcal{D}$, and bounded above by $e^{-\underline{r}T+o(T)}$, where $\underline{r} = \inf_{s\in\mathrm{cl}\mathcal{D}} I(s,\theta)$ represents the I distance between θ and the closure of $\mathcal{D}$.

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

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

iii. **Level-compactness.** $\{(s,\theta) \in \mathbb{S}\times\mathrm{cl}\Theta : I(s,\theta) \le r\}$ 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 $\theta \in \Theta$ (Dembo and Zeitouni 2009, section 1.2). As $\mathbb{S}_\infty$ is a subset of $\mathbb{S}$ and as $\mathbb{S}$ is closed, condition 3 of Definition 3.3 implies indeed that every regular rate function is good. Also, the radial monotonicity condition 1 may be difficult to check in practice. A more easily checkable sufficient condition for radial monotonicity is that for any $s \in \mathbb{S}_\infty$, $\theta \in \mathrm{cl}\Theta$ and $\theta_s \in \Theta$ with $S_\infty(\theta_s) = s$ (θ_s exists because $s \in \mathbb{S}_\infty$), we have

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

where the inequality is strict if $I(s,\theta) > 0$. Inequality (3.2) actually inspired the name radial monotonicity for condition 1. We will now prove that (3.2) together with condition 3 implies condition 1. To this end, fix any $s \in \mathbb{S}_\infty$, and set $A(s) = \{\theta \in \Theta : I(s,\theta) < r\}$ and $B(s) = \{\theta \in \mathrm{cl}\Theta : I(s,\theta) \le r\}$. By construction, we have $A(s) \subseteq B(s)$. As $B(s)$ is compact thanks to condition 3, this even implies that $\mathrm{cl}A(s) \subseteq B(s)$. It remains to be shown that (3.2) implies the converse inclusion $B(s) \subseteq \mathrm{cl}A(s)$. 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 $\theta \in \mathrm{cl}\Theta$, and as Θ is open and convex, the line segment principle (Bertsekas 2009, proposition 1.3.1) implies that $\theta(\lambda) \in \Theta$ for all $\lambda \in [0,1)$. By (3.2), we also have $I(s,\theta(\lambda)) \le I(s,\theta) \le r$ 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 toward one, we may finally conclude that $\theta \in \mathrm{cl}A(s)$. We have therefore shown that condition 1 holds.

Example 3.1 (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 (Dembo and Zeitouni 2009, 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 $p \ge 0$ 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 (Cover and Thomas 2006, theorem 2.6.3), it is nonnegative 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, $D(s\|\theta)$ is continuous on $\mathbb{S}\times\Theta$ and level-compact (Dembo and Zeitouni 2009, pp. 13–18). In addition, $D(s\|\theta)$ is jointly convex in s and θ because of (Cover and Thomas 2006, theorem 2.7.2). Recalling from Example 2.4 that S_∞ is the identity function, for any $s \in \mathbb{S}_\infty = \Theta$ and $\theta \in \mathrm{cl}\Theta$, we thus have $D(s\|(1-\lambda)s + \lambda\theta) \le (1-\lambda)D(s\|s) + \lambda D(s\|\theta) \le D(s\|\theta)$ $\forall \lambda \in [0,1)$, where the second inequality holds because $D(s\|s) = 0$ and $D(s\|\theta) \ge 0$. 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

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

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

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

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 $X \times \mathrm{cl}\Theta$ because of 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 $X \times \mathrm{cl}\Theta$. 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 $X \times \mathbb{S}_\infty$.*

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 $\theta \in \mathrm{cl}\Theta$ 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 $\tilde{c}^\star(x,s) = \min_{\alpha \ge \overline{\ell}(x)} \alpha - e^{-r} \prod_{i=1}^d (\alpha - \ell(x,i))^{s_i}$ with $\overline{\ell}(x) = \max_{i \in \Xi} \ell(x,i)$, which can be solved efficiently via line search methods (Van Parys et al. 2021, proposition 2).

The following theorem establishes that the distributionally robust predictor (3.3) strikes 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).

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

$$\tilde{x}^\star(s) \in \arg\min_{x \in X} \tilde{c}^\star(x,s) \quad \forall s \in \mathbb{S} \tag{3.4}$$

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, the multifunction $\arg\min_{x \in X} \tilde{c}^\star(x,s)$ is nonempty valued because X is compact and $\tilde{c}^\star(x,s)$ is continuous in x on $X \times \mathbb{S}$ (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 Van Parys et al. (2021), 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).

An interesting question arises as to whether an alternative parametrization for either the statistic $\widehat{S}$ or the model class Θ would impact the optimal data-driven predictor-prescriptor pair. Notably, an invariance principle can be demonstrated, indicating that the optimal solution remains unchanged under homeomorphic coordinate transformations. A detailed discussion of this invariance is relegated to Online Appendix A.

4. Separation of Estimation and Optimization

We are now ready to tackle a fundamental question in data-driven decision making that is of theoretical and 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 (Lehmann and Casella 1998). 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 (Koopman 1936) implies that if the observed data are 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 semicontinuous in λ because of Barndorff-Nielsen (2014) (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 $\overline{\theta} \in \Theta$, 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 $A_T(\lambda) = \Lambda_T(\lambda, \overline{\theta})$ such that $Tg(\theta) \in \mathrm{dom}(A_T)$ for all $\theta \in \Theta$ and*

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

Exponential families that obey Assumption 4.1 are called time-homogeneous because the parametrization function g is independent of T (Küchler and Sørensen 2006, 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 semicontinuity from the log-moment generating function Λ_T. Although the log-partition function was defined as the log-moment generating function corresponding to the baseline model $\overline{\theta}$, 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 Equation (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.,

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

Assumption 4.1 guarantees via the Fisher-Neyman factorization theorem (Lehmann and Casella 1998, 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 $\mathbb{P}_\theta^T$ and $\mathbb{P}_{\overline{\theta}}^T$; 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 $\mathrm{dom}\Lambda(\cdot, \theta)$ for all $\theta \in \Theta$. Finally, 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 $\mathrm{dom}\Lambda(\cdot, \theta)$ 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 $\mathrm{dom}\Lambda(\cdot, \theta)$. Assumption 4.2 imposes the stronger condition that the origin belongs to the interior of $\mathrm{dom}\Lambda(\cdot, \theta)$. Recall next that the log-moment generating functions $\Lambda_T(\lambda, \theta)$ are convex in λ for all $T \in \mathbb{N}$. By Dembo and Zeitouni (2009, lemma 2.3.9), their asymptotic counterpart $\Lambda(\lambda,\theta)$ inherits convexity in λ. Assumption 4.2 further stipulates that $\mathrm{dom}\Lambda_T(\cdot,\theta)=\mathbb{R}^d$, which implies via Barndorff-Nielsen (2014, theorem 7.2) that $\Lambda_T(\lambda,\theta)$ is analytical in λ, throughout all $\mathbb{R}^d$, 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* $\theta\in\Theta$.

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 are 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 $\overline{\theta}$, we then find $\mathrm{d}\mathbb{P}_\theta^T/\mathrm{d}\mathbb{P}_{\overline{\theta}}^T=d^T\prod_{t=1}^T\theta_{\xi_t}=d^T\prod_{j=1}^d\theta_j^{\sum_{t=1}^T 1_{\xi_t=j}}=\exp(\langle T\log\theta,\widehat{S}_T\rangle+T\log d)$, where $\log\theta$ is evaluated component-wise. In addition, the Tth log-moment generating function is given by

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

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 $\lambda\in\mathbb{R}^d$, $\Lambda(\lambda,\theta)=\lim_{T\to\infty}\frac{1}{T}\Lambda_T(T\lambda,\theta)=\log\sum_{i=1}^d\theta_i e^{\lambda_i}<\infty$ 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 (Dembo and Zeitouni 2009, 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*

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

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 semicontinuous 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 (Rockafellar 1970, 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* $S_\infty(\theta)=\nabla_\lambda\Lambda(0,\theta)$.

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

Example 4.2 (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 (Dembo and Zeitouni 2009, 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 semicontinuous 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 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 previous 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.

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 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

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

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.

Theorems 4.2 and 4.3 are reminiscent of the celebrated Rao-Blackwell theorem (Rao 1945, Blackwell 1947), 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 noninferior 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

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

that is, the row sums of $\theta_\star$ coincide with the respective column sums. These properties of $\theta_\star$ prompt us to define $\Theta = \{\theta \in \mathbb{R}^{m\times m}_{++} : \sum_{i,j\in\Xi} \theta_{ij} = 1,\ \sum_{j\in\Xi}\theta_{ij} = \sum_{j\in\Xi}\theta_{ji}\ \forall i \in \Xi\}$ as the set of all strictly positive doublet probability mass functions with balanced marginals. 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 2010, 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

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

Also, 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

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

By construction, $\widehat{S} = \{\widehat{S}_T\}_{T\in\mathbb{N}}$ constitutes a statistic with state space $\mathbb{S} = \mathrm{cl}(\cup_{T\in\mathbb{N}}\Delta_{m\times m} \cap (\mathbb{Z}^{m\times m}/T)) = \mathrm{cl}(\Delta_{m\times m} \cap \mathbb{Q}^{m\times m}) = \Delta_{m\times m}$. 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 (Ross 2010, 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, S_∞ is clearly a local homeomorphism.

We now follow the reasoning in Billingsley (1961) to show that the ambiguity set $\mathcal{P}$ represents a time-homogeneous exponential family. Specifically, we define the baseline model $\overline{\theta} \in \Theta$ through $\overline{\theta}_{ij} = 1/m^2$ for all $i, j \in \Xi$. The observations ξ_t, $t \in \mathbb{N}$, are thus serially independent and uniformly distributed under $\mathbb{P}_{\overline{\theta}}$, and the corresponding transition probability matrix satisfies $(P_{\overline{\theta}})_{ij} = 1/m$ for all $i, j \in \Xi$. In addition, the probability of observing $\xi_{[T]}$ under $\mathbb{P}_{\overline{\theta}}$ is given by $1/m^T$, and $\mathrm{d}\mathbb{P}^T_\theta / \mathrm{d}\mathbb{P}^T_{\overline{\theta}} = m^T \prod_{t=1}^T (P_\theta)_{\xi_{t-1}\xi_t} = m^T \prod_{i,j\in\Xi} (P_\theta)_{ij}^{\sum_{t=1}^T \mathbb{1}_{(\xi_{t-1},\xi_t)=(i,j)}} = m^T \prod_{i,j\in\Xi} (P_\theta)_{ij}^{T(\widehat{S}_T)_{ij}} = \exp(\langle T\log(P_\theta), \widehat{S}_T\rangle + T\log m)$, where the logarithm of the matrix P_θ is evaluated elementwise. 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 Dembo and Zeitouni (2009) (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. Hence, Assumption 4.2 holds, which ensures via the Gärtner-Ellis theorem that $\widehat{S}$ satisfies an LDP (Dembo and Zeitouni 2009, 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 (Dembo and Zeitouni 2009, 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,[1] 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." Also, Definition 5.1 specifies $D_c(s\|\theta)$ only on $\mathbb{S}\times\Theta$ and that $D_c(s\|\theta)$ is continuous on $\mathbb{S}\times\Theta$ because of our standard 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 semicontinuous extension to $\mathbb{S}\times\text{cl}\Theta$, which is obtained by setting

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

see also (Rockafellar and Roger 1998, definition 1.5). In the following, we will always mean this lower semicontinuous extension to $\mathbb{S}\times\text{cl}\Theta$ 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.*

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 (Li et al. 2021). In Online Appendix C, we show that the restricted meta-optimization problems sometimes admit Pareto dominant solutions even if the training data are generated by an autoregressive process with an uncountable state space instead of a finite-state Markov chain.

5.2. Independent Observations with Identical 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 $\mathbb{P}_\star[\xi_t \le z] = F_{\theta_\star}(z)$ for all $z\in\mathbb{R}^m$ and $t\in\mathbb{N}$. Here, F_θ, $\theta\in\Theta$, is a family of distribution functions with common support $\Xi\subseteq\mathbb{R}^m$, where the parameter θ ranges over the relative interior Θ of a convex subset 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

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

To estimate the mean value $S_\infty(\theta)$ (and thereby implicitly also θ), we use the sample mean

$$\widehat{S}_T = \frac{1}{T}\sum_{t=1}^T \xi_t \quad \forall T\in\mathbb{N}. \tag{5.2}$$

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 (5.2) is a sufficient statistic for θ. To see that the sample mean also satisfies an LDP with a regular rate function, for i.i.d. data, the limiting log-moment generating function simplifies to

$$\Lambda(\lambda,\theta) = \lim_{T\to\infty}\frac{1}{T}\log\mathbb{E}_\theta[\exp(\langle T\lambda, \widehat{S}_T\rangle)]$$
$$= \log\left(\int_{\mathbb{R}^m} e^{\lambda^\top \xi}\,\mathrm{d}F_\theta(\xi)\right). \tag{5.3}$$

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 previously (Online Appendix B). In each case, one can therefore verify by inspection that the gradient $\nabla_\lambda\Lambda(\lambda,\theta)$ exists on the interior of $\mathrm{dom}\Lambda(\cdot,\theta)$ and that its norm tends to infinity when λ approaches the boundary of $\mathrm{dom}\Lambda(\cdot,\theta)$. 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 (5.3) with respect to λ. The Cramér function is again available in closed form for all examples listed above (Table 1 in Online Appendix). 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).

6. Conclusions

This paper proposes a rigorous framework for identifying optimal estimators for the objective functions and the optimal solutions of data-driven decision problems. To conclude, we provide recommendations for practitioners and discuss potential generalizations of our results.

Our paper offers the following three-step guideline for practitioners faced with a data-driven decision problem. First, users should identify a finitely parametrized time series model consistent with the observable data. Second, they should find a statistic for the unknown parameters of the time series model that satisfies an LDP. Third, they should construct efficient data-driven predictors and prescriptors by solving DRO Problems (3.3) and (3.4), which involve an ambiguity set constructed form the rate function of the LDP. The out-of-sample disappointment of these predictors and prescriptors is guaranteed to be equal to $e^{-rT+o(T)}$, where r is the radius of the ambiguity set. Because of its direct physical interpretation, we believe that it is natural for decision makers to *choose* r in view of their risk tolerance instead of *calibrating* it algorithmically. Nevertheless, some decision makers may want to calibrate r via cross-validation with the goal to minimize the out-of-sample risk. In doing so, however, direct control over the out-of-sample disappointment is lost.

The main results of this paper rely on several assumptions, some of which could be generalized. Assumption 2.1 requires that Θ constitutes a *finitely* parametrized ambiguity set. However, we believe that the results of Section 3 extend to *infinitely* parametrized (i.e., nonparametric) ambiguity sets. For example, in Van Parys et al. (2021), our results for finite-state i.i.d. processes are extended to i.i.d. processes with a continuous state space. This generalization does not require fundamentally new ideas but requires more sophisticated topological arguments that make the proofs less accessible. Assumption 2.2 requires $c(x,\theta)$ to be uniformly continuous and bounded. It is nonrestrictive for practical purposes. We believe that it can be relaxed to requiring that $c(x,\theta)$ be lower semicontinuous at the expense of complicating the proofs of Proposition 3.1, Theorem 3.1, and Theorem 3.2. Assumption 3.1 requires the statistic $\widehat{S}$ to satisfy an LDP with a regular rate function and thus guarantees that the restricted meta-optimization problems (2.6) are solvable. This assumption seems more difficult to relax as our results critically rely on large deviations theory. Assumption 4.1 requires $\mathcal{P} = \{\mathbb{P}_\theta : \theta\in\Theta\}$ to represent an exponential family, and Assumption 4.2 captures standard technical conditions required for the Gärtner-Ellis Theorem (Theorem 4.1). Together, these assumptions imply that $\widehat{S}$ is a sufficient statistic satisfying an LDP, and thus they imply Assumption 3.1. Clearly, the statistic $\widehat{S}$ must satisfy some notion of sufficiency for Theorems 4.2 and 4.3 to hold. Nevertheless, we believe that Assumptions 4.1 and 4.2 can be relaxed and that Theorems 4.2 and 4.3 remain valid if $\widehat{S}$ is only sufficient in an *asymptotic* sense. Finally, the meta-optimization problems (2.4) and (2.6) involve two *asymptotic* performance criteria, that is, the *asymptotic* in-sample risk and the *asymptotic* decay rate of the out-of-sample disappointment. Although the asymptotic nature of these performance criteria is undesirable from a modeling perspective, the meta-optimization problems corresponding to a fixed sample size T may no longer admit Pareto dominant solutions. However, if the statistic $\widehat{S}_T$ enjoys a finite sample guarantee, then the distributionally robust predictors and prescriptors (3.3) and (3.4) may still be *approximately* Pareto dominant.

Endnote

[1] 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.

References

Aliprantis C, Border K (2007) *Infinite Dimensional Analysis: A Hitchhiker's Guide* (Springer, Berlin).

Ash R, Doléans-Dade CA (2000) *Probability and Measure Theory* (Academic Press, New York).

Ban G-Y, Rudin C (2019) The big data newsvendor: Practical insights from machine learning. *Oper. Res.* 67(1):90–108.

Barndorff-Nielsen O (2014) *Information and Exponential Families in Statistical Theory* (John Wiley & Sons, Hoboken, NJ).

Ben-Tal A, den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Sci.* 59(2):341–357.

Bennouna A, Van Parys BPG (2021) Learning and decision-making with data: Optimal formulations and phase transitions. Preprint, submitted September 14, https://arxiv.org/abs/2109.06911.

Berge C (1997) *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity* (Courier Corporation, North Chelmsford, MA).

Bertsekas D (2009) *Convex Optimization Theory* (Athena Scientific, Nashua, NH).

Bertsimas D, Kallus N (2020) From predictive to prescriptive analytics. *Management Sci.* 66(3):1025–1044.

Bertsimas D, Gupta V, Kallus N (2018) Robust sample average approximation. *Math. Programming* 171(1–2):217–282.

Billingsley P (1961) Statistical methods in Markov chains. *Ann. Math. Statist.* 32:12–40.

Blackwell D (1947) Conditional expectation and unbiased sequential estimation. *Ann. Math. Statist.* 18:105–110.

Blanchet J, Kang Y, Murthy K (2019) Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probabilities* 56(3):830–857.

Cover T, Thomas J (2006) *Elements of Information Theory* (Wiley, New York).

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

Dembo A, Zeitouni O (2009) *Large Deviations Techniques and Applications* (Springer, Berlin).

Derman E, Mannor S (2020) Distributional robustness and regularization in reinforcement learning. Preprint, submitted March 5, https://arxiv.org/abs/2003.02894.

Dou X, Anitescu M (2019) Distributionally robust optimization with correlated data from vector autoregressive processes. *Oper. Res. Lett.* 47(4):294–299.

Duchi J, Glynn P, Namkoong H (2021) Statistics of robust optimization: A generalized empirical likelihood approach. *Math. Oper. Res.* 46(3):946–969.

Durrett R (2010) *Probability: Theory and Examples* (Cambridge University Press, Cambridge, UK).

Gao R (2023) Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Oper. Res.* 71(6):2291–2306.

Gupta V (2019) Near-optimal Bayesian ambiguity sets for distributionally robust optimization. *Management Sci.* 65(9):4242–4260.

Hanasusanto G, Kuhn D, Wiesemann W (2016) A comment on "computational complexity of stochastic programming problems". *Math. Programming* 159(1):557–569.

Hu Y, Kallus N, Mao X (2022) Fast rates for contextual linear optimization. *Management Sci.* 68(6):4236–4245.

Kantorovich LV, Rubinshtein GS (1958) On a space of totally additive functions. *Vestnik Leningradskogo Universiteta* 13(7):52–59.

King A, Tyrrell Rockafellar R (1993) Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res.* 18(1):148–162.

King A, Wets RJ-B (1991) Epi-consistency of convex stochastic programs. *Stochastic Stochastic Rep.* 34(1–2):83–92.

Koopman B (1936) On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.* 39(3):399–409.

Küchler U, Sørensen M (2006) *Exponential Families of Stochastic Processes* (Springer, Berlin).

Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research & Management Science in the Age of Analytics* (INFORMS, Catonsville, MD), 130–166.

Lam H (2019) Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Oper. Res.* 67(4):1090–1105.

Le Maître O, Knio O (2010) *Introduction: Uncertainty Quantification and Propagation* (Springer, Berlin).

Lehmann EL, Casella G (1998) *Theory of Point Estimation* (Springer, Berlin).

Li M, Sutter T, Kuhn D (2021) Distributionally robust optimization with Markovian data. Meila M, Zhang T, eds. *Proc. 38th Internat. Conf. Machine Learn.*, (PMLR, New York), 6493–6503.

Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.* 171(1–2): 115–166.

Newey W, McFadden D (1994) Large sample estimation and hypothesis testing. *Handbook of Econometrics* (Elsevier, New York), 2111–2245.

Rao R (1945) Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* 37:81–91.

Rockafellar RT (1970) *Convex Analysis* (Princeton University Press, Princeton, NJ).

Rockafellar RT, Roger J-B (1998) *Wets Variational Analysis* (Springer, Berlin).

Ross S (2010) *Introduction to Probability Models* (Elsevier, New York).

Shapiro A (1989) Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.* 17(2):841–858.

Shapiro A (1990) On differential stability in stochastic programming. *Math. Programming* 47(1):107–116.

Shapiro A (1991) Asymptotic analysis of stochastic programs. *Ann. Oper. Res.* 30(1–4):169–186.

Shapiro A (1993) Asymptotic behavior of optimal solutions in stochastic programming. *Math. Oper. Res.* 18(4):829–845.

Shapiro A (2003) Monte Carlo sampling methods. *Stochastic Programming* (Elsevier, New York), 353–425.

Shapiro A, Dentcheva D, Ruszczyński A (2014) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM, Philadelphia).

Smith J, Winkler R (2006) The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Sci.* 52(3):311–322.

Sutter T, Krause A, Kuhn D (2021) Robust generalization despite distribution shift via minimum discriminating information. Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. *Advances in Neural Information Processing Systems*, vol. 34 (Curran Associates, Inc., Red Hook, NY), 29754–29767.

Van Parys B (2021) Efficient data-driven optimization with noisy data. Preprint, submitted February 8, https://arxiv.org/abs/2102.04363.

Van Parys B, Esfahani PM, Kuhn D (2021) From data to decisions: Distributionally robust optimization is optimal. *Management Sci.* 67(6):3387–3402.

Tobias Sutter received a BSc and MSc degree in mechanical engineering in 2010 and 2012 from ETH Zürich, and a PhD in electrical engineering at the Automatic Control Laboratory, ETH Zürich in 2017. He currently is an assistant professor at the Computer Science

Department in Konstanz, Germany. Prior to joining University of Konstanz, he held a research and lecturer appointment with EPFL at the chair of risk analytics and optimization and at the Institute of Machine Learning at ETH Zürich.

Bart P. G. Van Parys is an assistant professor in the operations research and statistics group at the MIT Sloan School of Management. He obtained his doctoral degree in control engineering at ETH Zurich. His current research is located on the interface between robust optimization and machine learning. In particular, he develops novel mathematical methodologies and algorithms with which can turn data into better informed decisions. Although most of his research is methodological, he does enjoy applications related to problems in renewable energy.

Daniel Kuhn holds the chair of risk analytics and optimization at EPFL. Before joining EPFL, he was a faculty member at Imperial College London (2007–2013) and a postdoctoral researcher at Stanford University (2005–2006). He received a PhD in economics from the University of St. Gallen in 2004 and an MSc in theoretical physics from ETH Zürich in 1999. His research interests revolve around robust optimization and stochastic programming.