Optimization with Data: Large Deviation Limits

Bart P.G. Van Parys Peyman Mohajerin Esfahani Daniel Kuhn

December 1, 2016

1 Introduction

Notation: The natural logarithm of $x \in \mathrm{R}_+$ is denoted by $\log(x)$, where we use the conventions $0\log(\frac{0}{q}) = 0$ for any $q > 0$ and $p\log(\frac{p}{0}) = \infty$ for any $p > 0$. The set of positive (semi)define symmetric matrices in $\mathrm{R}^{d\times d}$ is represented by S^d_{++} (S^d_+). For any logical statement $\mathcal{E}$, the indicator function $\mathbb{1}_{\mathcal{E}}$ evaluates to 1 if $\mathcal{E}$ is true and to 0 otherwise.

2 Stochastic programming

Stochastic programming is a powerful modeling paradigm for taking informed decisions in an uncertain environment. A generic single-stage stochastic program can be represented as

$$\underset{x\in X}{\text{minimize}} \int_{\Xi} \gamma(x,\xi)\,\mathrm{d}F^\star(\xi). \tag{2.1}$$

Here, the goal is to select a decision $x \in X$ that minimizes the expected value of a cost function $\gamma(x,\xi) \in \mathrm{R}$, which depends both on x and an exogenous random vector $\xi \in \Xi$ governed by a probability distribution $F^\star$. Below we will assume that X is closed and $\gamma(x,\xi)$ is continuous jointly in both arguments. A wide spectrum of decision problems can be cast as instances of (2.1). Shapiro *et al.* [13] point out, for example, that (2.1) can be viewed as the first stage of a two-stage stochastic program, where the cost function $\gamma(x,\xi)$ embodies the optimal value of a subordinate second-stage problem. Alternatively, problem (2.1) may also be interpreted as a generic learning problem in the spirit of Vapnik's [16] statistical learning theory.

In the following, we distinguish the *prediction problem*, which merely aims to predict the expected cost associated with a fixed decision x, and the *prescription problem*, which seeks to identify a decision $x^\star$ that minimizes the expected cost across all $x \in X$.

Any attempt to solve the prescription problem seems futile unless there is a procedure for solving the corresponding prediction problem. However, the prediction problem requires the evaluation of a potentially high-dimensional integral, which is already hard even if $\gamma(x,\xi)$ represents the non-negative part of an affine function, see Hanasusanto *et al.* [**?**]. The generic prediction problem is closely related to what Le Maître and Knio [9] call an uncertainty quantification problem and is therefore of prime interest in its own right. Throughout the rest of the paper, we thus analyze prediction and prescription problems on equal footing.

In the what follows we formalize the notion of a data-driven solution to the prescription and prediction problems, respectively. Furthermore, we introduce the basic assumptions as well as the notation used throughout the remainder of this paper.

2.1 Data-driven predictors and prescriptors

Unfortunately, the distribution $F^\star$ of ξ is hardly ever observable but must be estimated from time series data, that is, a finite (possibly small) number of samples that all follow the same marginal distribution $F^\star$.

Thus, we lack essential information to evaluate the expected cost of any fixed decision and—*a fortiori*—to solve the stochastic program (2.1). The standard approach to overcome this deficiency is to approximate $F^\star$ with a parametric or non-parametric estimate $\hat{F}$ inferred from the samples and to minimize the expected cost under $\hat{F}$ instead of the true expected cost under $F^\star$. However, if we calibrate a stochastic program to a training dataset and evaluate its optimal decision on a test dataset, then the resulting test performance is often disappointing—even if the two datasets are sampled independently from $F^\star$. This phenomenon has been observed in many different contexts. It is particularly pronounced in finance, where Michaud [**?**] refers to it as the 'error maximization effect' of portfolio optimization, and in statistics or machine learning, where it is known as 'overfitting'. In decision analysis, Smith and Winkler [14] refer to it as the 'optimizer's curse'. Thus, when working with data instead of exact probability distributions, one should safeguard against solutions that display promising in-sample performance but lead to out-of-sample disappointment.

In the following we assume that the distribution $F^\star$ must be estimated from a finite sequence of consecutive samples of a stationary stochastic process $\{\xi_t\}_{t\in\mathbb{N}}$ with state space Ξ. We assume that this stochastic process is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P}^\star)$ whose probability measure $\mathbb{P}^\star$ is unknown but belongs to a known parametric ambiguity set in the sense of the following definition.

Assumption 2.1 (Parametric ambiguity set)**.** *The unknown true probability measure $\mathbb{P}^\star$ is known to belong to a parametric ambiguity set $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where Θ is a convex closed subset of $\mathbb{R}^d$, while $\mathbb{P}_\theta$ represents a probability measure on $(\Omega, \mathcal{F})$ for every $\theta \in \Theta$.*

As each parameter $\theta \in \Theta$ encodes a different probabilistic model $\mathbb{P}_\theta$, by slight abuse of terminology, we will henceforth refer to θ as a *model* and to Θ as the *model class*. Assumption 2.1 implies that there exists $\theta^\star \in \Theta$ with $\mathbb{P}_{\theta^\star} = \mathbb{P}^\star$. If the process $\{\xi_t\}_{t\in\mathbb{N}}$ is known to be stationary under $\mathbb{P}^\star$ in the sense that each ξ_t follows the same unknown marginal distribution $F^\star$, it makes sense to assume that the parametric ambiguity set $\{\mathbb{P}_\theta : \theta \in \Theta\}$ contains only probability measures under which the data generating process is indeed stationary. We formalize this requirement in the following assumption.

Assumption 2.2 (Stationarity)**.** *The stochastic process $\{\xi_t\}_{t\in\mathbb{N}}$ is stationary under any model, that is, for any given $\theta \in \Theta$ there exits a distribution function F_θ with $\mathbb{P}_\theta(\xi_t \leq z) = F_\theta(z)$ for all $z \in \mathbb{R}^{\dim \Xi}$ and $t \in \mathbb{N}$.*

We emphasize that stationarity does not entail serial independence. However, it ensures that a sample path of the stochastic process $\{\xi_t\}_{t\in\mathbb{N}}$ carries *some* information about the marginal distribution of the samples. Thus, stationarity can be viewed as a minimal requirement for the meaningfulness and feasibility of any approach to data-driven prediction or prescription. We now present three model classes that satisfy Assumption 2.2.

Example 2.1 (Model classes)**.** *The following model classes serve as running examples throughout the text.*

(i) **Continuous state i.i.d. processes with unknown means:** *Assume that $\Xi = \mathbb{R}^d$, the samples are serially independent under $\mathbb{P}^\star$, and $\mathbb{P}^\star(\xi_t \leq z) = F_0(z - \theta^\star)$ for all $z \in \Xi$ and $t \in \mathbb{N}$, were F_0 is a known distribution function with zero mean, and $\theta^\star$ denotes the unknown mean vector of ξ_t. Thus, $\mathbb{P}^\star$ belongs to an ambiguity set $\{\mathbb{P}_\theta : \theta \in \Theta_{\rm loc}\}$, where $\Theta_{\rm loc} = \mathbb{R}^d$ represents the set of all possible mean vectors, and each $\theta \in \Theta_{\rm loc}$ encodes a probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ satisfying*

$$\mathbb{P}_\theta\left(\xi_t \leq z_t \ \forall t = 1, \ldots, T\right) = \prod_{t=1}^{T} F_0(z_t - \theta) \qquad \forall z \in \Xi^T,\ T \in \mathbb{N}.$$

(ii) **Finite state i.i.d. processes:** *Assume that $\Xi = \{1, \ldots, d\}$, the samples ξ_t are serially independent under $\mathbb{P}^\star$, and $\mathbb{P}^\star(\xi_t = i) = \theta_i^\star$ for all $i \in \Xi$ and $t \in \mathbb{N}$, where the vector $\theta^\star$ denotes the unknown probability mass function of ξ_t. Thus, $\mathbb{P}^\star$ belongs to an ambiguity set $\{\mathbb{P}_\theta : \theta \in \Theta_{\rm iid}\}$, where $\Theta_{\rm iid} = \{\theta \in \mathbb{R}^d_+ : \sum_{i=1}^d \theta_i = 1\}$ represents the simplex of all possible probability mass functions, and each $\theta \in \Theta_{\rm iid}$ encodes a probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ satisfying*

$$\mathbb{P}_\theta\left(\xi_t = i_t \ \forall t = 1, \ldots, T\right) = \prod_{t=1}^{T} \theta_{i_t} \qquad \forall i \in \Xi^T,\ T \in \mathbb{N}.$$

(iii) **Finite state stationary Markov chains:** *Assume that $\Xi = \{1, \ldots, d\}$, the samples ξ_t follow a stationary Markov chain under $\mathbb{P}^\star$, and $\mathbb{P}^\star(\xi_t = i,\, \xi_{t+1} = j) = \theta^\star_{ij}$ for all $i, j \in \Xi$ and $t \in \mathbb{N}$, where the matrix $\theta^\star$ denotes the unknown probability mass function of the doublet (ξ_t, ξ_{t+1}). This implies that*

$$\sum_{j\in\Xi} \theta^\star_{ij} \;=\; \sum_{j\in\Xi} \mathbb{P}^\star(\xi_t = i,\, \xi_{t+1} = j) \;=\; \mathbb{P}^\star(\xi_t = i) \;=\; \sum_{j\in\Xi} \mathbb{P}^\star(\xi_{t-1} = j,\, \xi_t = i) \;=\; \sum_{j\in\Xi} \theta^\star_{ji},$$

where t is any integer larger than 1. Hence, the row sums of $\theta^\star$ coincide with the corresponding column sums. In the following, we denote by

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

the set of all doublet probability mass functions with balanced marginals. Every $\theta \in \Theta_{\rm mc}$ gives rise to a vector $\pi_\theta \in \mathbb{R}_+^{1\times d}$ of stationary probabilities and a transition probability matrix $P_\theta \in \mathbb{R}_+^{d\times d}$ defined through $(\pi_\theta)_i = \sum_{j\in\Xi} \theta_{ij}$ and $(P_\theta)_{ij} = \theta_{ij}/(\pi_\theta)_i$, respectively.[1] *By construction, P_θ is a stochastic matrix with each row summing to 1, and π_θ is a non-negative row vector summing to 1 with $\pi_\theta P_\theta = \pi_\theta$; see Ross [12, Chapter 4] for more details on Markov chains. We conclude that $\mathbb{P}^\star$ belongs to the ambiguity set $\{\mathbb{P}_\theta : \theta \in \Theta_{\rm mc}\}$, where each $\theta \in \Theta_{\rm mc}$ encodes a probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$ with*

$$\mathbb{P}_\theta\left(\xi_t = i_t\; \forall t = 1, \ldots, T\right) = (\pi_\theta)_{i_1} \prod_{t=1}^{T-1} (P_\theta)_{i_t i_{t+1}} \qquad \forall i \in \Xi^T,\; T \in \mathbb{N}.$$

Note that

We emphasize that the ambiguity set $\mathcal{P}_\Theta$ is meant to capture all a priori information on $\mathbb{P}^\star$ that is available before observing any statistical data. Therefore, we assume that $\mathcal{P}_\Theta$ is known to contain $\mathbb{P}^\star$ with certainty (and not just with high confidence). The ambiguity set $\mathcal{P}_\Theta$ thus only determines the structure of the data-generating process but not the values of its parameters. As the true probability measure $\mathbb{P}^\star$ is known to reside within $\{\mathbb{P}_\theta : \theta \in \Theta\}$, the true probability distribution $F^\star$ belongs to the set $\{F_\theta : \theta \in \Theta\}$.

Next, we introduce parametric predictors and prescriptors corresponding to the stochastic program (2.1), where the true unknown distribution $F^\star$ is replaced with F_θ.

Definition 2.1 (Parametric predictors and prescriptors)**.** *For any fixed model $\theta \in \Theta$, we define the predictor $c(x, \theta) = \int_\Xi \gamma(x, \xi)\, \mathrm{d}F_\theta(\xi)$ as the expected cost of a given decision $x \in X$ and the 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 expected cost 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 continuity assumption.

Assumption 2.3 (Continuity of the predictor)**.** *The predictor c is jointly continuous in both its arguments.*

Assumption 2.3 is satisfied for all continuous state i.i.d. processes with unknown means studied in Example 2.1(i) provided that the continuous cost function $\gamma(x, \xi)$ is bounded in $\xi \in \Xi$ for every fixed $x \in X$. Moreover, Assumption 2.3 is automatically satisfied for all finite state i.i.d. processes and all finite state irreducible Markov chains discussed in Examples 2.1(i) and 2.1(ii), respectively, regardless of the cost function.

Note that neither the prediction nor the prescription problem can be solved for the true distribution $F^\star = F_{\theta^\star}$ as both the predictor $c(x, \theta^\star)$ and the prescriptor $x^\star(\theta^\star)$ depend explicitly on the unknown parameter $\theta^\star$. If one has access to a sample path $\{\xi_t\}_{t=1}^T$ of length T drawn at random under the probability measure $\mathbb{P}^\star = \mathbb{P}_{\theta^\star}$, however, one may construct an empirical estimator $\hat{\theta}_T$ for $\theta^\star$. The precise definition of $\hat{\theta}_T$ will depend on the underlying model class, but we generally require that there exists a measurable function $f_T : \Xi^T \to \Theta$ such that $\hat{\theta}_T(\omega) = f_T(\xi_1(\omega), \ldots, \xi_T(\omega))$ for all $\omega \in \Omega$.

[1] If $(\pi_\theta)_i = 0$, without loss of generality, we may set $(P_\theta)_{ij} = 1$ if $j = i$ and $(P_\theta)_{ij} = 0$ otherwise.

Example 2.2 (Empirical estimators)**.** *For the model classes described in Example 2.1 we will use the following empirical estimators.*

(i) **Continuous state i.i.d. processes with unknown means:** *If the ξ_t are mutually independent and follow the same distribution with known shape but unknown mean, we define $\hat{\theta}_T$ as the sample mean*

$$\hat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \xi_t.$$

(ii) **Finite state i.i.d. processes:** *If the ξ_t are mutually independent and follow the same discrete distribution on $\Xi = \{1, \ldots, d\}$, we define $\hat{\theta}_T$ as the vector of empirical state frequencies, that is,*

$$(\hat{\theta}_T)_i = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\xi_t = i} \quad \forall i \in \Xi.$$

(iii) **Finite state irreducible Markov chains:** *If the ξ_t follow an irreducible Markov chain on $\Xi = \{1, \ldots, d\}$, we define $\hat{\theta}_T$ as the matrix of empirical transition frequencies, that is,*

$$(\hat{\theta}_T)_{ij} = \frac{1}{T}\left(\sum_{t=1}^{T-1} \mathbb{1}_{\xi_t = i} \cdot \mathbb{1}_{\xi_{t+1} = j} + \mathbb{1}_{\xi_T = i} \cdot \mathbb{1}_{\xi_1 = j}\right) \quad \forall i, j \in \Xi.$$

The "ghost" transition from ξ_T to ξ_1 in the above definition ensures that $\hat{\theta}_T \in \Theta_{\mathrm{mc}}$.

As we lack essential information to solve the true prediction and prescription problems, we will henceforth approximate the unknown predictor $c(x, \theta^\star)$ as well as the unknown prescriptor $x^\star(\theta^\star)$ by suitable functions of the empirical estimator $\hat{\theta}_T$.

Definition 2.2 (Data-driven predictors and prescriptors)**.** *A Carathéodory integrand $\hat{c} : X \times \Theta \to \mathrm{R}$ is called a data-driven predictor if $\hat{c}(x, \hat{\theta}_T)$ is used as an approximation for $c(x, \theta^\star)$. A Borel measurable function $\hat{x} : \Theta \to X$ is called a data-driven prescriptor if there exists a data-driven predictor $\hat{c}$ for which*[2]

$$\hat{x}(\theta) \in \arg\min_{x \in X} \hat{c}(x, \theta) \quad \forall \theta \in \Theta,$$

and $\hat{x}(\hat{\theta}_T)$ is used as an approximation for $x^\star(\theta^\star)$.

Example 2.3 (Naïve data-driven predictor)**.** *The parametric predictor c introduced in Definition 2.1 constitutes a trivial data-driven predictor, that is, $c(x, \hat{\theta}_T)$ can be used as a naïve approximation for $c(x, \theta^\star)$. Note that c is indeed a Carathéodory integrand as it is continuous by virtue of Assumption 2.3. Both in the case of the finite state i.i.d. processes and the finite state stationary Markov chains described in Example 2.1, the naïve predictor reduces to the well-known sample average estimator, that is, we have*

$$c(x, \hat{\theta}_T) = \frac{1}{T}\sum_{t=1}^{T} \gamma(x, \xi_t).$$

The estimates $\hat{c}(x, \hat{\theta}_T)$ and $\hat{x}(\hat{\theta}_T)$ inherit the randomness from the empirical estimator $\hat{\theta}_T$, which is constructed from the (random) samples $\xi_1, \ldots, \xi_T$. Note that the prediction and prescription problems are naturally interpreted as instances of statistical estimation problems. Indeed, data-driven prediction aims to estimate the expected cost $c(x, \theta^\star)$ from data. Standard statistical estimation theory would typically endeavor to find a data-driven predictor $\hat{c}$ that (approximately) minimizes the mean squared error

$$\int_\Omega \left\| c(x, \theta^\star) - \hat{c}(x, \hat{\theta}_T(\omega)) \right\|_2^2 \mathbb{P}_{\theta^\star}(\mathrm{d}\omega).$$

[2]By Corollary 14.6 and Theorem 14.37 in [**?**], every data-driven predictor $\hat{c}$ induces at least one data-driven predictor if the arg min mapping is non-empty for every $\theta \in \Theta$.

The mean squared error penalizes the mismatch between the actual cost $c(x,\theta)$ and its estimator $\hat{c}(x,\hat{\theta}_T)$. Events in which we are left disappointed ($c(x,\theta) > \hat{c}(x,\hat{\theta}_T)$) are not treated differently from positive surprises ($c(x,\theta) < \hat{c}(x,\hat{\theta}_T)$). In a decision-making context where the goal is to minimize costs, however, disappointments (underestimated costs) are more harmful than positive surprises (overestimated costs). While statisticians strive for accuracy by minimizing a symmetric estimation error, decision makers endeavor to limit the one-sided prediction disappointment.

Definition 2.3 (Out-of-sample disappointment)**.** *For any data-driven predictor $\hat{c}$ the probability*

$$\mathbb{P}_\theta \left(c(x,\theta) > \hat{c}(x,\hat{\theta}_T) \right) \tag{2.2a}$$

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

$$\mathbb{P}_\theta \left(c(\hat{x}(\hat{\theta}_T),\theta) > \hat{c}(\hat{x}(\hat{\theta}_T),\hat{\theta}_T) \right) \tag{2.2b}$$

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

The out-of-sample prediction disappointment quantifies the probability (with respect to $\mathbb{P}_\theta$ for some $\theta \in \Theta$) that the expected cost $c(x,\theta)$ of a fixed decision x exceeds the predicted cost $\hat{c}(x,\hat{\theta}_T)$. Thus, the out-of-sample prediction disappointment is independent of the actual realization of the empirical estimator $\hat{\theta}_T$ but depends on the hypothesized model θ. A similar statement holds for the out-of-sample prescription disappointment.

The main objective of this paper is to construct attractive data-driven predictors and prescriptors, which are optimal in a sense to be made precise below. We first developa notion of optimality for data-driven predictors and extended it later to data-driven prescriptors in a straightforward way. As indicated above, a crucial requirement for any data-driven predictor is that it must limit the the out-of-sample disappointment. This informal requirement can be operationalized either in an asymptotic sense or in a finite sample sense.

(i) **Asymptotic guarantee:** As T grows, the out-of-sample prediction disappointment (2.2a) decays exponentially at a rate at least equal to $r \geq 0$ up to first order in the exponent, that is,

$$\limsup_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta \left(c(x,\theta) > \hat{c}(x,\hat{\theta}_T) \right) \leq -r \quad \forall x \in X,\ \theta \in \Theta. \tag{2.3}$$

(ii) **Finite sample guarantee:** The out-of-sample prediction disappointment (2.2a) is bounded above by a *known* function $g(T)$ that decays exponentially at rate $r \geq 0$ to first order in the exponent, that is,

$$\mathbb{P}_\theta \left(c(x,\theta) > \hat{c}(x,\hat{\theta}_T) \right) \leq g(T) \quad \forall x \in X,\ \theta \in \Theta,\ T \in \mathbb{N} \quad \text{and} \quad \limsup_{T\to\infty} \frac{1}{T} \log(g(T)) = -r. \tag{2.4}$$

The inequalities (2.3) and (2.4) are imposed uniformly across all models $\theta \in \Theta$. This ensures that they are satisfied under the true model $\theta^\star$, which is only known to reside within Θ. By requiring the inequalities to hold for all $x \in X$, we further ensure that the out-of-sample prediction disappointment is uniformly small irrespective of the chosen decision. Note that the finite sample guarantee (2.4) is sufficient but not necessary for the asymptotic guarantee (2.3). In fact, knowing $g(T)$ enables us to determine the sample complexity

$$\min \left\{ T \in \mathbb{N} : g(T') \leq \beta \ \forall T' \geq T \right\},$$

that is, the minimum number of samples needed to certify that the out-of-sample prediction disappointment does not exceed a prescribed significance level $\beta \in [0,1]$.

At first sight the requirements (2.3) and (2.4) may seem unduly restrictive, and the existence of data-driven predictors with exponentially decaying out-of-sample disappointment may be questioned. Indeed, the naïve data-driven predictor $c(x,\hat{\theta}_T)$ of Example 2.3 typically violates the conditions (2.3) and (2.4).

Example 2.4 (Large out-of-sample disappointment)**.** *Let $\{\xi_t\}_{t\in\mathbb{N}}$ be a finite state i.i.d. process of the type described in Example 2.1 (ii), and set the cost function to $\gamma(x,\xi)=\xi$. Assume further that the data process is governed by a probability measure $\mathbb{P}_\theta$ induced by some $\theta\in\Theta_{\text{iid}}$. In this setting, the naïve data-driven predictor approximates the expected cost $c(x,\theta)=\sum_{i\in\Xi} i\theta_i$ by its sample mean $c(x,\hat{\theta}_T)=\frac{1}{T}\sum_{t=1}^T \xi_t$; see Example 2.2 (ii). As the sample size T tends to infinity, the central limit theorem implies that*

$$\sqrt{T}\left(c(x,\hat{\theta}_T)-c(x,\theta)\right)$$

converges in law to a normal distribution with mean 0 and variance $\sigma^2=\sum_{i\in\Xi} i^2\theta_i-(\sum_{i\in\Xi} i\theta_i)^2$. Thus,

$$\lim_{T\to\infty}\mathbb{P}_\theta\left(c(x,\theta)>\hat{c}(x,\hat{\theta}_T)\right)=\lim_{T\to\infty}\mathbb{P}_\theta\left(\sqrt{T}\left(\hat{c}(x,\hat{\theta}_T)-c(x,\theta)\right)/\sigma<0\right)=\frac{1}{2},$$

which means that the out-of-sample prediction disappointment remains large for all sample sizes.

Example 2.4 suggests that the out-of-sample disappointment of a predictor $\hat{c}$ cannot be expected to decay at an exponential rate unless $\hat{c}$ is *conservative*, that is, unless $\hat{c}(x,\theta)>c(x,\theta)$ for all $x\in X$ and $\theta\in\Theta$. If the predictor is conservative and if the empirical estimator $\hat{\theta}_T$ obeys a strong law of large numbers under $\mathbb{P}_\theta$ (meaning that $\hat{\theta}_T$ converges $\mathbb{P}_\theta$-almost surely to θ), then—maybe surprisingly—an exponential decay of the prediction disappointment is to be expected under rather generic conditions. In fact, asymptotic guarantees of the type (2.3) hold whenever the empirical estimator $\hat{\theta}_T$ satisfies a *weak* large deviation principle, while finite sample guarantees of the type (2.4) hold when $\hat{\theta}_T$ satisfies a *strong* large deviation principle. As will be shown below, all empirical estimators of Example 2.2 satisfy suitable large deviation principles.

For ease of exposition, we henceforth denote by $\mathcal{C}$ the set of all data-driven predictors, that is, all normal integrands that map $X\times\Theta$ to the reals. Moreover, we introduce a partial order $\preceq_{\mathcal{C}}$ on $\mathcal{C}$ defined through

$$\hat{c}_1\preceq_{\mathcal{C}}\hat{c}_2 \quad\Longleftrightarrow\quad \hat{c}_1(x,\theta)\leq\hat{c}_2(x,\theta)\quad\forall x\in X,\,\theta\in\Theta$$

for any $\hat{c}_1,\hat{c}_2\in\mathcal{C}$. Thus, $\hat{c}_1\preceq_{\mathcal{C}}\hat{c}_2$ means that $\hat{c}_1$ is (weakly) less conservative than $\hat{c}_2$. The problem of finding the least conservative predictor among all data-driven predictors whose out-of-sample disappointment decays at rate at least $r\geq 0$ can thus be formalized as the following *vector optimization problem*.

$$\begin{array}{ll}\underset{\hat{c}\in\mathcal{C}}{\text{minimize}}\,{\scriptstyle\preceq_{\mathcal{C}}} & \hat{c}\\ \text{subject to} & \limsup\limits_{T\to\infty}\dfrac{1}{T}\log\mathbb{P}_\theta\left(c(x,\theta)>\hat{c}(x,\hat{\theta}_T)\right)\leq -r\quad\forall x\in X,\,\theta\in\Theta\end{array}\tag{2.5}$$

We highlight that the minimization in (2.5) is understood with respect to the partial order $\preceq_{\mathcal{C}}$. Thus, the relation $\hat{c}_1\preceq_{\mathcal{C}}\hat{c}_2$ between two feasible decision means that $\hat{c}_1$ is weakly preferred to $\hat{c}_2$. However, not all pairs of feasible decisions are comparable, that is, it is possible that both $\hat{c}_1\not\preceq_{\mathcal{C}}\hat{c}_2$ and $\hat{c}_2\not\preceq_{\mathcal{C}}\hat{c}_1$. A predictor $\hat{c}^\star$ is a *strongly* optimal solution for (2.5) if it is feasible and weakly preferred to every other feasible solution (*i.e.*, every $\hat{c}\neq\hat{c}^\star$ feasible in (2.5) satisfies $\hat{c}^\star\preceq_{\mathcal{C}}\hat{c}$). Similarly, $\hat{c}^\star$ is a *weakly* optimal solution for (2.5) if it is feasible and if every other solution preferred to $\hat{c}^\star$ is infeasible (*i.e.*, every $\hat{c}\neq\hat{c}^\star$ with $\hat{c}\preceq_{\mathcal{C}}\hat{c}^\star$ is infeasible in (2.5)). Generic vector optimization problems typically only admit weak solutions. In the next section we will show, however, that (2.5) admits an explicit strong solution in closed form.

We henceforth denote by $\mathcal{X}$ the set of all data-driven predictor-prescriptor-pairs $(\hat{c},\hat{x})$, where $\hat{c}\in\mathcal{C}$, and $\hat{x}$ is a prescriptor induced by $\hat{c}$ as per Definition (2.2). The problem of finding the least conservative predictor-prescriptor-pair whose out-of-sample prescription disappointment decays at rate at least $r\geq 0$ can thus be formalized as the following vector optimization problem akin to (2.5).

$$\begin{array}{ll}\underset{(\hat{c},\hat{x})\in\mathcal{X}}{\text{minimize}}\,{\scriptstyle\preceq_{\mathcal{C}}} & \hat{c}\\ \text{subject to} & \limsup\limits_{T\to\infty}\dfrac{1}{T}\log\mathbb{P}_\theta\left(c(\hat{x}(\hat{\theta}_T),\theta)>\hat{c}(\hat{x}(\hat{\theta}_T),\hat{\theta}_T)\right)\leq -r\quad\forall\theta\in\Theta\end{array}\tag{2.6}$$

We will demonstrate below that (2.6) also admits an explicit strong solution in closed form.

3 Large deviation principles

Large deviations theory aims to quantify the exact exponential rate at which the probabilities of unlikely estimator realizations decay as the sample size T tends to infinity. In the following, we denote by $\mathcal{S}_\theta$ the smallest convex closed subset of Θ with the property that $\mathbb{P}_\theta(\hat{\theta}_T \in \mathcal{S}_\theta) = 1$ for all $T \in \mathbb{N}$. Thus, $\mathcal{S}_\theta$ covers $\mathbb{P}_\theta$-almost all possible estimator realizations.

Example 3.1 (Possible estimator realizations)**.** *For the empirical estimators described in Example 2.2, which correspond to the respective model classes in Example 2.1, the sets $\mathcal{S}_\theta$, $\theta \in \Theta$, can be constructed explicitly.*

(i) **Continuous state i.i.d. processes with unknown means:** *Denote by $\mathcal{S}_0$ as the smallest convex closed subset of R^d that has probability 1 under the distribution F_0. Then, we have*

$$\mathcal{S}_\theta = \{\theta' \in \Theta_{\mathrm{loc}} : \theta' - \theta \in \mathcal{S}_0\} \qquad \forall \theta \in \Theta_{\mathrm{loc}}.$$

(ii) **Finite state i.i.d. processes:** *Denote by $\mathrm{supp}(\theta)$ the set of all $i \in \Xi$ with $\theta_i > 0$. Then, we have*

$$\mathcal{S}_\theta = \{\theta' \in \Theta_{\mathrm{iid}} : \theta'_i = 0 \ \forall i \notin \mathrm{supp}(\theta)\} \qquad \forall \theta \in \Theta_{\mathrm{iid}}.$$

(iii) **Finite state stationary Markov chains:** *Denote by $\mathrm{supp}(\theta)$ the set of all $(i,j) \in \Xi^2$ with $\theta_{ij} > 0$. Then, we have*

$$\mathcal{S}_\theta = \left\{\theta' \in \Theta_{\mathrm{mc}} : \theta'_{ij} = 0 \ \forall (i,j) \notin \mathrm{supp}(\theta)\right\} \qquad \forall \theta \in \Theta_{\mathrm{mc}}.$$

By construction, all estimator realizations outside of $\mathcal{S}_\theta$ have probability 0 under $\mathbb{P}_\theta$. Large deviations theory provides bounds on the exponential rate at which the probabilities of possible but unlikely estimator realizations inside of $\mathcal{S}_\theta$ decay under $\mathbb{P}_\theta$. These bounds are expressed through a problem-specific *rate function*.

Definition 3.1 (Rate function)**.** *A lower semi-continuous mapping $I : \Theta \times \Theta \to [0, \infty]$ is termed a rate function. Moreover, we call a rate function I regular if the following conditions hold:*

(i) **Convexity in θ':** *For all $\theta_1, \theta_2, \theta' \in \Theta$ with we have*

$$I(\theta', (1-\lambda)\theta_1 + \lambda\theta_2) < (1-\lambda) I(\theta', \theta_1) + \lambda I(\theta', \theta_2) \quad \lambda \in [0,1].$$

(ii) **Radial monotonicity in θ:** *For all $\theta, \theta' \in \Theta$ with $I(\theta', \theta) > 0$ we have*

$$I(\theta', (1-\lambda)\theta' + \lambda\theta) < I(\theta', \theta) \quad \lambda \in [0,1).$$

(iii) **Continuity in (θ, θ'):** *$I(\theta', \theta)$ is continuous*[3] *on $\mathcal{S} = \{(\theta, \theta') \in \Theta \times \Theta : \theta' \in \mathcal{S}_\theta\}$.*

We are now ready to define the fundamental notion of a large deviation principle.

Definition 3.2 (Weak large deviation principle)**.** *A sequence of empirical estimators $\{\hat{\theta}_T\}_{T \in \mathbb{N}}$ is said to satisfy a weak large deviation principle (WLDP) under model $\theta \in \Theta$ if there exists a rate function I with*

$$-\inf_{\theta' \in \mathrm{int}\,(\mathcal{D} \cap \mathcal{S}_\theta)} I(\theta', \theta) \leq \liminf_{T \to \infty} \frac{1}{T} \log \mathbb{P}_\theta\left(\hat{\theta}_T \in \mathcal{D}\right) \tag{3.1a}$$

$$\leq \limsup_{T \to \infty} \frac{1}{T} \log \mathbb{P}_\theta\left(\hat{\theta}_T \in \mathcal{D}\right) \leq -\inf_{\theta' \in \mathrm{cl}\,(\mathcal{D} \cap \mathcal{S}_\theta)} I(\theta', \theta) \tag{3.1b}$$

for all $\mathcal{D} \in \mathcal{B}(\Theta)$. Similarly, $\{\hat{\theta}_T\}_{T \in \mathbb{N}}$ is said to satisfy a WLDP under the model class Θ if there exists a rate function I such that (3.1) holds for each $\theta \in \Theta$.

[3]Continuity of extended real-valued functions is understood in the sense of [**?**]. For example, $\log(x)$ is a continuous function on the non-negative reals under our standing convention that $\log(0) = -\infty$.

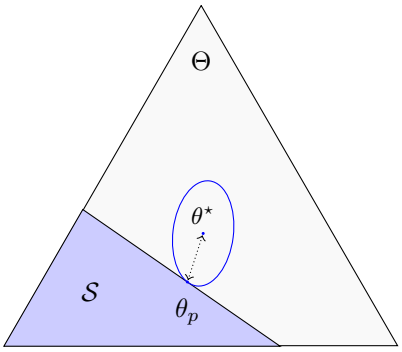


Figure 1: Visualization of a large deviation principle. The gray triangle represents the set $\mathcal{S}_\theta$ covering all realizations of $\hat{\theta}_T$ under $\mathbb{P}_\theta$. The probability $\mathbb{P}_\theta(\hat{\theta}_T \in \mathcal{D})$ that the empirical estimator $\hat{\theta}_T$ falls within the blue triangle $\mathcal{D}$ diminishes at an exponential rate bracketed by $\inf_{\theta' \in \operatorname{int}(\mathcal{D}\cap\mathcal{S}_\theta)} I(\theta',\theta)$ and $\inf_{\theta' \in \operatorname{cl}(\mathcal{D}\cap\mathcal{S}_\theta} I(\theta',\theta)$. If $\mathcal{D}$ is I-continuous, then this exponential decay rate coincides with the distance $\inf_{\theta' \in \mathcal{D}\cap\mathcal{S}_\theta} I(\theta',\theta)$ of θ from $\mathcal{D}$. The blue ellipse visualizes a contour of the rate function $I(\cdot,\theta)$ governing this decay rate.

We will see below that many practically relevant empirical estimators satisfy a large deviation principle. Before studying such examples, we discuss a few immediate consequences of the inequalies (3.1). First, the estimator $\hat{\theta}_T$ resides within the convex set $\mathcal{S}_\theta$ almost surely with respect to $\mathbb{P}_\theta$. As any rate function is non-negative, the upper bound in (3.1b) thus implies that $\inf_{\theta' \in \Theta\cap\mathcal{S}_\theta} I(\theta',\theta) = 0$ for any $\theta \in \Theta$. Moreover, if $\hat{\theta}_T$ converges in probability to θ with respect to $\mathbb{P}_\theta$ (which is the case for all empirical estimators considered in this paper), then (3.1b) implies

$$0 = \lim_{T\to\infty} \frac{1}{T} \log \mathbb{P}_\theta\left(\|\hat{\theta}_T - \theta\| \leq \frac{1}{n}\right) \leq -\inf_{\theta' \in \mathcal{S}_\theta} \left\{ I(\theta',\theta) \;:\; \|\theta' - \theta\| \leq \frac{1}{n} \right\} \quad \forall n \in \mathbb{N}.$$

Thus, there exists a sequence $\{\theta'_n\}_{n\in\mathbb{N}}$ in $\mathcal{S}_\theta$ that converges to θ and satisfies $\liminf_{n\in\mathbb{N}} I(\theta'_n,\theta) \leq 0$, which implies via the lower semi-continuity and non-negativity of the rate function that $I(\theta,\theta) = 0$ for every $\theta \in \Theta$.

A WLDP provides asymptotic bounds on the probability that the estimator $\hat{\theta}_T$ belongs to some measurable set $\mathcal{D}$. If the true parameter θ resides within the closure of $\mathcal{D}$, then $\inf_{\theta' \in \operatorname{cl}(\mathcal{D}\cap\mathcal{S}_\theta)} I(\theta',\theta) = I(\theta,\theta) = 0$, which leads to the trivial upper bound $\mathbb{P}_\theta(\hat{\theta}_T \in \mathcal{D}) \leq 1$. On the other hand, if $\mathcal{D}$ has empty interior (*e.g.*, if $\mathcal{D} = \{\theta\}$ is a singleton containing only the true parameter), then $\inf_{\theta' \in \operatorname{int}(\mathcal{D}\cap\mathcal{S}_\theta)} I(\theta',\theta) = \infty$, which leads to the trivial lower bound $\mathbb{P}_\theta(\hat{\theta}_T \in \mathcal{D}) \geq 0$. Non-trivial bounds are obtained if $\mathcal{D}$ has non-empty interior and does not contain θ. In these cases the rate function bounds the exponential rate at which the probability of the 'rare' event $\{\hat{\theta}_T \in \mathcal{D}\}$ decays as T tends to infinity. For some sets $\mathcal{D}$ this rate of decay is precisely determined by the rate function. Specifically, a set $\mathcal{D} \subseteq \Theta$ is called I-continuous under model θ if

$$\inf_{\theta' \in \operatorname{int}(\mathcal{D}\cap\mathcal{S}_\theta)} I(\theta',\theta) = \inf_{\theta' \in \mathcal{D}\cap\mathcal{S}_\theta} I(\theta',\theta) = \inf_{\theta' \in \operatorname{cl}(\mathcal{D}\cap\mathcal{S}_\theta)} I(\theta',\theta).$$

For example, if I is a regular rate function and thus continuous in θ' on $\mathcal{S}_\theta$, any set $\mathcal{D} \subseteq \mathcal{S}_\theta$ with $\mathcal{D} \subseteq \operatorname{cl}\operatorname{int}\mathcal{D}$ is I-continuous under θ. The large deviation principle (3.1) implies that for large T the probability of an I-continuous set $\mathcal{D}$ decays at rate $\inf_{\theta' \in \mathcal{D}} I(\theta',\theta)$ to first order in the exponent, that is, we have

$$\mathbb{P}_\theta(\hat{\theta}_T \in \mathcal{D}) = e^{-T \inf_{\theta' \in \mathcal{D}\cap\mathcal{S}_\theta} I(\theta',\theta) + o(T)}.$$

If we interpret $I(\theta',\theta)$ as the distance between θ' and θ, then the decay rate of $\mathbb{P}_\theta(\hat{\theta}_T \in \mathcal{D})$ coincides with the distance of the true parameter θ from the rare event set $\mathcal{D}$; see Figure 1.

Weak large deviation principles only provide *asymptotic* bounds on the decay rates of rare events. However, many estimators also satisfy a *strong* large deviation principle, which offers *finite sample guarantees*.

Definition 3.3 (Strong large deviation principle)**.** *A sequence of empirical estimators $\{\hat{\theta}_T\}_{T\in\mathbb{N}}$ is said to satisfy a strong large deviation principle (SLDP) under model $\theta \in \Theta$ if there exists a rate function I with*

$$-\inf_{\theta'\in\operatorname{int}(\mathcal{D}\cap\mathcal{S}_\theta)} I(\theta',\theta) - \dim\Theta\cdot\frac{\log(T+1)}{T} \;\leq\; \frac{1}{T}\log\mathbb{P}_\theta\left(\hat{\theta}_T\in\mathcal{D}\right) \tag{3.2a}$$

$$\leq\; -\inf_{\theta'\in\operatorname{cl}(\mathcal{D}\cap\mathcal{S}_\theta)} I(\theta',\theta) + \dim\Theta\cdot\frac{\log(T+1)}{T} \tag{3.2b}$$

for all $\mathcal{D}\in\mathcal{B}(\Theta)$. Similarly, $\{\hat{\theta}_T\}_{T\in\mathbb{N}}$ is said to satisfy a SLDP under the model class Θ if there exists a rate function I such that (3.2) *holds for each $\theta\in\Theta$.*

Clearly, any sequence of empirical estimators obeying a *strong* large deviation principle also satisfies a *weak* large deviation principle with the same rate function. However, the converse implication may be false.

In the remainder of this section we will prove that the empirical estimators of all model classes introduced in Examples 2.1 and 2.2 satisfy an SLDP with a rate function that is regular in the sense of Definition 3.1.

3.1 Continuous state i.i.d. processes with unknown means

Throughout this section we assume that the ξ_t are mutually independent and follow the same distribution with known shape but unknown mean as described in Example 2.1(i). Before establishing that the sample mean estimators of Example 2.2(i) satisfy an SLDP, we recall the definition of the Legendre transform.

Definition 3.4 (Legendre transform)**.** *The Legendre transform of a convex function $\Lambda_0:\mathrm{R}^d\to(-\infty,\infty]$ is the convex lower semi-continuous function $\Lambda_0^\star:\mathrm{R}^d\to(-\infty,\infty]$ defined through $\Lambda_0^\star(z)=\sup_{x\in\mathrm{R}^d}\lambda^\top z-\Lambda_0(\lambda)$.*

The classical strong law of large numbers implies that the sample means $\hat{\theta}_T$ converge $\mathbb{P}_\theta$-almost surely to θ as T grows. The SLDP portrayed in the following theorem essentially quantifies the speed of convergence.

Theorem 3.1 (SLDP for i.i.d. processes with unknown mean)**.** *Let $\{\xi_t\}_{t\in\mathbb{N}}$ be an i.i.d. process where $\xi_t-\theta$ follows a known distribution F_0 on $\Xi=\mathrm{R}^d$ with $\int_\Xi \xi\,\mathrm{d}F_0(\xi)=0$, and θ is an unknown location parameter. Then, the sample means $\hat{\theta}_T=\frac{1}{T}\sum_{t=1}^T\xi_t$ satisfy an SLDP with rate function $I(\theta',\theta)=\Lambda_0^\star(\theta'-\theta)$, where*

$$\Lambda_0(\lambda)=\log\left(\int_\Xi e^{\lambda^\top\xi}\,\mathrm{d}F_0(\xi)\right)$$

represents the cumulant-generating function of F_0.

Remark 3.1. *The cumulant-generating function $\Lambda_0(\lambda)$ is defined as the logarithm of the moment-generating function of F_0. One can show that it is convex, which implies that the moment-generating function is log-convex; see e.g. Lemma 2.2.5 in Dembo and Zeitouni [4]. The Legendre transform $\Lambda_0^\star(z)$ of $\Lambda_0(\lambda)$ is usually referred to as the Cramér function. Table 1 lists several popular distributions with zero mean, their cumulant-generating functions as well as the corresponding Cramér functions along with their respective domains. Note that $\Lambda_0^\star(z)$ and $\Lambda_0(\lambda)$ are interpreted as ∞ outside of their domains.*

Proof of Theorem 3.1. TODO □

Proposition 3.1 (Properties of the Cramér function)**.** *The Cramér function $\Lambda_0^\star(z)$ is non-negative and satisfies $\Lambda_0^\star(0)=0$. Moreover, it is convex and radially monotonic in the sense that*

$$\Lambda_0^\star((1-\lambda)z)<\Lambda_0^\star(z)\quad\forall\lambda\in[0,1)\quad\textit{and}\quad z\in\mathrm{R}^d\quad\textit{with}\quad\Lambda_0^\star(z)>0.$$

Proof. Jensen's inequality implies that the cumulant-generating function is non-negative, that is,

$$\Lambda_0(\lambda)\geq\log\left(e^{\int_\Xi\lambda^\top\xi\,\mathrm{d}F_0(\xi)}\right)=0,$$

F_0	$\Lambda(\lambda)$	$\mathrm{dom}(\Lambda_0)$	$\Lambda^\star(z)$	$\mathrm{dom}(\Lambda_0^\star)$
(a) Normal	$\frac{1}{2}\lambda^\top\Sigma\lambda$	R^d	$\frac{1}{2}z^\top\Sigma^{-1}z$	R^d
(b) Exponential	$-\lambda\mu - \log(1-\lambda\mu)$	$(-\infty, \frac{1}{\mu})$	$\frac{z}{\mu} - \log(1+\frac{z}{\mu})$	$(-\mu, \infty)$
(c) Poisson	$\mu(e^\lambda - 1 - \lambda)$	R	$(z+\mu)\log(1+\frac{z}{\mu}) - z$	$[-\mu, \infty)$
(d) Bernoulli	$\log(\cosh(\lambda))$	R	$z\log\sqrt{\frac{1+z}{1-z}} + \log\sqrt{1-z^2}$	$[-1, 1]$

Table 1: The cumulant-generating functions and their Legendre transforms for several common distributions: (a) a normal distribution with mean 0 and covariance matrix $\Sigma \in \mathbb{S}_+^d$; (b) an exponential distribution with rate parameter $\frac{1}{\mu}$, $\mu > 0$, shifted to set its mean to 0; (c) a Poisson distribution with rate parameter μ, $\mu > 0$, shifted to set its mean to 0; (d) the uniform Bernoulli distribution on $\{-1, 1\}$.

where the equality holds because F_0 has mean zero. Moreover, it is easy to verify that $\Lambda_0(0) = 0$ since F_0 assigns unit probability to Ξ. The Cramér function inherits non-negativity from Λ_0 because

$$\Lambda_0^\star(z) = \sup_{\lambda \in \mathbb{R}^d} \lambda^\top z - \Lambda_0(\lambda) \geq \Lambda_0(0) = 0,$$

where the inequality holds due to the feasibility of $\lambda = 0$. Moreover, as $\Lambda_0(0) = 0$, it is easy to verify that the Cramér function attains its minimum at 0, that is, $\Lambda_0^\star(0) = 0$. Note that $\Lambda_0^\star(z)$ constitutes a pointwise supremum of affine functions and is therefore convex in z. Moreover, for any $z \in \mathrm{R}^d$ with $\Lambda_0^\star(z) > 0$ we have

$$\Lambda_0^\star(\lambda 0 + (1-\lambda)z) \leq \lambda\Lambda_0^\star(0) + (1-\lambda)\Lambda_0^\star(z) < \Lambda_0^\star(z) \quad \forall \lambda \in (0, 1],$$

where the weak inequality follows from convexity, while the strict inequality holds because $\Lambda_0^\star(0) = 0$. □

Proposition 3.1 implies that $I(\theta', \theta) = \Lambda_0^\star(\theta' - \theta)$ is a regular rate function in the sense of Definition 3.1 whenever $\Lambda_0^\star(z)$ is continuous on $\mathcal{S}_0$. Note that, by construction, $\mathcal{S}_0$ coincides with the closure of $\mathrm{dom}(\Lambda_0^\star)$. Thus, all Cramér functions listed in Table 1 are continuous on $\mathcal{S}_0$, which implies that they give rise to regular rate functions. We emphasize that the Cramér function corresponding to the exponential distribution with rate $\frac{1}{\mu}$ in Table 1 is indeed continuous at the boundary of $\mathcal{S}_0$ because $\lim_{z \downarrow -\mu} \Lambda_0^\star(z) = \infty = \Lambda_0^\star(-\mu)$. We also remark that the Cramér functions of discrete distributions such as the Poisson or Bernoulli distribution jump on the boundary of $\mathcal{S}_0$, that is, they are discontinuous on R^d but continuous when restricted to $\mathcal{S}_0$. Even though one can prove that $\Lambda_0^\star(z)$ is always infinite on the complement of $\mathcal{S}_0$ and both finite and—owing to convexity—continuous on $\mathrm{int}(\mathcal{S}_0)$, it can generically be discontinuous on $\mathcal{S}_0$ in dimensions $d > 1$.

Example 3.2 (Discontinuous Cramér function)**.** *Set F_0 to the equally weighted mixture of the uniform distribution on the closed unit ball in R^2 and the Dirac distribution at $(1, 0)^\top$. In this case, one can show that*

$$\mathrm{dom}(\Lambda_0^\star) = \{z \in \mathrm{R}^2 : \|z\|_2 < 1\} \cup \{(1, 0)^\top\}.$$

Thus, $\Lambda_0^\star(z)$ is finite only at one single boundary point of $\mathcal{S}_0 = \mathrm{cl}(\mathrm{dom}(\Lambda_0^\star))$. This implies that the Cramér function is discontinuous on the boundary of $\mathcal{S}_0$ and—a fortiori—on $\mathcal{S}_0$ itself.

3.2 Finite state i.i.d. processes

Throughout this section we assume that $\{\xi_t\}_{t \in \mathbb{N}}$ follows an i.i.d. process with a finite state space as described in Example 2.1(ii). In this case, the vectors of empirical state frequencies of Example 2.2(ii) can be shown to satisfy an SLDP where the rate function is given by the relative entropy.

Definition 3.5 (Relative entropy)**.** *For any two probability mass functions $\theta, \theta' \in \Theta_{\mathrm{iid}}$, the relative entropy of θ with respect to θ' is defined as*

$$D(\theta' \| \theta) = \sum_{i \in \Xi} \theta_i' \log\left(\frac{\theta_i'}{\theta_i}\right).$$

The relative entropy is also known as information for discrimination, cross-entropy, information gain or Kullback-Leibler divergence. Note that Kullback and Leibler [7] originally used the term "divergence" to refer to the Jeffreys divergence, which is defined as the symmetrized relative entropy $(D(\theta'\|\theta) + D(\theta\|\theta'))/2$.

The strong law of large numbers implies again that the vectors $\hat{\theta}_T$ of empirical frequencies converge $\mathbb{P}_\theta$-almost surely to θ as T grows, and the SLDP laid out in the following theorem quantifies the speed of convergence.

Theorem 3.2 (SLDP for i.i.d. processes)**.** *Let $\{\xi_t\}_{t\in\mathbb{N}}$ be an i.i.d. process on $\Xi = \{1,\ldots,d\}$, and let θ be the unknown probability mass function of the ξ_t. Then, the vectors $\hat{\theta}_T$ of empirical state frequencies, which are defined as in Example 2.2(i), satisfy an SLDP with rate function $I(\theta',\theta) = D(\theta'\|\theta)$.*

Proof. TODO. □

Proposition 3.2 (Properties of the relative entropy)**.** *The relative entropy $D(\theta'\|\theta)$ is a regular rate function in the sense of Definition 3.1.*

Proof. The relative entropy is non-negative on its entire domain due to [3, Theorem 2.6.3], and it is immediate to verify that it vanishes for $\theta = \theta'$. Next, $D(\theta'\|\theta)$ is jointly convex in θ and θ' by [3, Theorem 2.7.2] and, *a fortiori*, convex in θ. Moreover, for any $\theta,\theta' \in \Theta_{\text{iid}}$ with $D(\theta'\|\theta) > 0$ we have

$$D(\theta'\|\lambda\theta' + (1-\lambda)\theta) \leq \lambda D(\theta'\|\theta') + (1-\lambda)D(\theta'\|\theta) < D(\theta'\|\theta) \quad \forall \lambda \in (0,1],$$

where the weak inequality follows from convexity, while the strict inequality holds because $D(\theta'\|\theta') = 0$ and $D(\theta'\|\theta) > 0$. Thus, $D(\theta'\|\theta)$ is radially monotonic in θ. By construction, finally, the relative entropy constitutes a continuous extended real-valued function on Θ_{iid}^2. This observation completes the proof. □

3.3 Finite state stationary Markov chains

Throughout this section we assume that $\{\xi_t\}_{t\in\mathbb{N}}$ follows a stationary Markov chain with a finite state space as described in Example 2.1(iii). In this case, the matrices of empirical transition frequencies of Example 2.2(iii) can be shown to satisfy an SLDP where the rate function is given by the conditional relative entropy.

Definition 3.6 (Conditional relative entropy)**.** *For two balanced doublet probability mass functions $\theta,\theta' \in \Theta_{\text{mc}}$ with corresponding transition probability matrices $P_\theta, P_{\theta'} \in \mathbb{R}_+^{d\times d}$ and invariant distributions $\pi_\theta, \pi_{\theta'} \in \mathbb{R}_+^{1\times d}$, respectively, the conditional relative entropy is defined as*

$$D_c(\theta'\|\theta) = \sum_{i\in\Xi} (\pi_{\theta'})_i \, D((P_{\theta'})_{i\cdot}\|(P_\theta)_{i\cdot}) = \sum_{i,j\in\Xi} \theta'_{ij} \left(\log\left(\frac{\theta'_{ij}}{\sum_{k\in\Xi}\theta'_{ik}} \right) - \log\left(\frac{\theta_{ij}}{\sum_{k\in\Xi}\theta_{ik}} \right) \right).$$

The ergodic theorem for irreducible Markov chains ensures that the matrix $\hat{\theta}_T$ of empirical transition frequencies converges $\mathbb{P}_\theta$-almost surely to the true doublet probability mass function θ as T grows; see Ross [12, Theorem 4.1].

Hence, given enough data the ambiguity concerning the distribution underlying the Markov data eventually disappears as well. The exact exponential rate of convergence from $\hat{\mathbb{Q}}_n$ to its limit $\mathbb{Q}^\star$ is will be shown to be governed by the conditional relative entropy.

Theorem 3.3 (SLDP for stationary Markov chains)**.** *Let $\{\xi_t\}_{t\in\mathbb{N}}$ be an i.i.d. process on $\Xi = \{1,\ldots,d\}$, and let θ be the unknown probability mass function of the ξ_t. Then, the matrices $\hat{\theta}_T$ of empirical transition frequencies, which are defined as in Example 2.2(ii), satisfy an SLDP with rate function $I(\theta',\theta) = D(\theta'\|\theta)$.*

Proof of Theorem 3.3. TODO. □

Proposition 3.3 (Properties of the conditional relative entropy)**.** *The conditional relative entropy $D_c(\theta'\|\theta)$ is a regular rate function in the sense of Definition 3.1.*

Proof. By definition, the conditional relative entropy inherits non-negativity from the relative entropy. To show convexity in θ', we note that the perspective function $x\log(x/y)$, defined for $(x,y) \in \mathbb{R}^2_+$, is convex and continuous on its domain. As convexity and continuity are preserved under composition with a linear function, the mapping $\theta'_{ij}\log(\theta'_{ij}/\sum_{k\in\Xi}\theta'_{ik})$, defined for $\theta' \in \Theta_{\rm mc}$, is also convex and continuous on its domain for every $i,j \in \Xi$. Thus, $D_c(\theta'\|\theta)$ is convex and continuous in θ' as a sum of finitely many convex and continuous functions. To prove radial monotonicity in θ, we first use Jensen's inequality to argue that

$$\sum_{j\in\Xi} \frac{w_j}{\sum_{k\in\Xi} w_k} \left(\frac{v_j}{w_j}\right)^2 \geq \left(\sum_{j\in\Xi} \frac{w_j}{\sum_{k\in\Xi} w_k} \frac{v_j}{w_j}\right)^2 \quad \Longleftrightarrow \quad \sum_{j\in\Xi} \frac{v_j^2}{w_j} \geq \frac{\left(\sum_{j\in\Xi} v_j\right)^2}{\sum_{k\in\Xi} w_k} \tag{3.3}$$

for any two vectors $v, w \in \mathbb{R}^d_+$. Note that the above inequalities are strict unless v and w are parallel. Next, fix $\theta, \theta' \in \Theta_{\rm mc}$ and define $\theta(\lambda) = (1-\lambda)\theta' + \lambda\theta$ for any $\lambda \in (0,1)$. Basic algebra then implies that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda} D_c(\theta'\|\theta(\lambda)) &= \sum_{i,j\in\Xi} \theta'_{ij} \left(\frac{\sum_{k\in\Xi}\theta_{ik} - \theta'_{ik}}{\sum_{k\in\Xi}(1-\lambda)\theta'_{ik} + \lambda\theta_{ik}} - \frac{\theta_{ij} - \theta'_{ij}}{(1-\lambda)\theta'_{ij} + \lambda\theta_{ij}} \right) \\
&= \frac{1}{\lambda} \sum_{i,j\in\Xi} \theta'_{ij} \left(\frac{\sum_{k\in\Xi}\theta_{ik}(\lambda) - \theta'_{ik}}{\sum_{k\in\Xi}\theta_{ik}(\lambda)} - \frac{\theta_{ij}(\lambda) - \theta'_{ij}}{\theta_{ij}(\lambda)} \right) \\
&= \frac{1}{\lambda} \sum_{i\in\Xi} \left(\sum_{j\in\Xi} \frac{\left(\theta'_{ij}\right)^2}{\theta_{ij}(\lambda)} - \frac{\left(\sum_{j\in\Xi}\theta'_{ij}\right)^2}{\sum_{j\in\Xi}\theta_{ij}(\lambda)} \right) \geq 0,
\end{aligned}$$

where the inequality follows from (3.3). Note that this inequality reduces to an equality iff each row of θ is parallel to the corresponding row of θ', that is, iff the transition probability matrices P_θ and $P_{\theta'}$ induced by θ and θ', respectively, are identical. In that case, $D_c(\theta'\|\theta(\lambda)) = 0$ for all $\lambda \in [0,1]$. In all other cases, $D_c(\theta'\|\theta(\lambda))$ is strictly monotonically increasing in λ. Thus, $D_c(\theta'\|\theta)$ is radially monotonic in θ. □

We emphasize that $D_c(\theta'\|\theta)$ is *not* convex in θ.

Determining for the probability of the event in which $\hat{\theta}_n$ takes value in a given set $\mathcal{S}$ reduces from computing a very high dimensional integral to merely solving a (convex) optimization problem. We also remark that the relation (**??**) for I-continuous ambiguity sets provides us with an extremely powerful tool to understand the smallest number of samples that are necessary to ensure that the probability of the event $\hat{\theta}_n \in \mathcal{S}$ is smaller than a given probability β. It does indeed follow from the large deviation Theorem **??** that the number of samples n necessary to guarantee that $\mathbb{Q}^\star_n(\hat{\theta}_n \in \mathcal{S}) \leq \beta$ should scale with - sample complexity

$$\frac{1}{r} \cdot \log\left(\frac{1}{\beta}\right) \tag{3.4}$$

where $r = \inf_{\theta\in\mathcal{S}} I(\theta, \theta^\star)$ denotes the I-distance of the set $\mathcal{S}$ to the parameter $\theta^\star$.

The strong large deviation principle presented in Theorem **??** will allow us to make complete abstraction of how the data was generated! Instead, the rate function is what distinguishes the data processes from one another. A Markov data process will differ from its independent identically distributed (i.i.d.) counterpart only in that exponential convergence is governed by the relative conditional entropy instead of the conditional entropy. This powerful large deviation property will be used in Section 4 to construct data-driven predictions and prescriptions which can be analyzed for all data classes in a disciplined and straightforward fashion.

3.4 A coin tossing experiment

To showcase the power of large deviations theory, we investigate an illustrative coin tossing experiment involving a fictitious Markovian coin.

In the first experiment, we are confronted with a sequence of n i.i.d. random variables $\hat{\xi}_i \in \Xi = \{\text{tail}, \text{head}\}$ resulting from coin tossing. The coin underlying the i.i.d. process is for the sake of the first experiment assumed fair $\mathbb{Q}^\star(\text{tail}) = 0.5$. It will be of interest here to understand how well the stationary marginal distribution $\mathbb{Q}^\star$ can be determined using historical data with n data samples. In what follows we will try to characterize how close the empirical estimator $\hat{\mathbb{Q}}_n(\text{tail})$ is to the actual probability $\mathbb{Q}^\star(\text{tail})$ of tossing tails. In particular, we will try to bound the probability $\beta(\epsilon, n)$ of the estimator $\hat{\mathbb{Q}}_n(\text{tail})$ underestimating the probability $\mathbb{Q}^\star(\text{tail})$ by more than some amount $\epsilon > 0$.

The probability of estimator failure can in case of this very simplistic coin tossing example be exactly calculated as

$$\begin{aligned}\beta(\epsilon, T) &= \mathbb{P}(\mathbb{Q}^\star(\text{tail}) \geq \hat{\mathbb{Q}}_T(\text{tail}) + \epsilon), \\ &= \sum_{i=0}^{k} \binom{T}{i} \cdot \mathbb{Q}^\star(\text{tail})^i \cdot \mathbb{Q}^\star(\text{head})^{T-i} \quad \text{with} \quad k = \lfloor T(\mathbb{Q}^\star(\text{tail}) - \epsilon) \rfloor.\end{aligned}$$

This exact expression for the probability of estimator failure is given in Figure 2 for a fair coin and $\epsilon = 0.15$ as a function of the sample size n. As pointed out before, the probability of out-of-sample disappointment diminishes at an exponential rate to zero. However, it is easily seen that expression (**??**) is (i) not monotone in in the number of samples n, (ii) hard to evaluate numerically, and (iii) does not admit a straightforward generalization to practically relevant problems in a higher dimensional setting. We will now illustrate that large deviation theory can be used to analyze exactly with which exponential rate the probability of estimator failure $\beta(\epsilon, n)$ diminishes as the number of data points n increases.

The event of estimator failure corresponds to the empirical estimator $\hat{\mathbb{Q}}_n$ realizing in the convex set

$$\mathcal{S} = \left\{ \mathbb{Q} \in \mathcal{P}_\Xi \ : \ \mathbb{Q}^\star(\text{tail}) \geq \mathbb{Q}(\text{tail}) + \epsilon \right\}.$$

According to Theorem **??** the probability of estimator failure must now diminish at an exponential rate determined by the I-distance between the distribution $\mathbb{Q}^\star$ and the ambiguity set $\mathcal{S}$. The probability of estimator failure $\beta(\epsilon, n)$ can thus be bounded in terms of the relative entropy, i.e.

$$\frac{1}{n} \log \beta(\epsilon, n) \leq 2\frac{\log(n+1)}{n} - \inf_{\mathbb{Q} \in \text{cl}\, \mathcal{C}} D(\mathbb{Q} \| \mathbb{Q}^\star), \quad \forall n \in \mathrm{N}. \tag{3.5}$$

The main advantage of the approximate expression (3.5) over its exact counterpart (**??**) is that (i) it is monotone in n, and (ii) only requires a convex optimization problem over distributions which does lend itself to a straightforward generalization to more practically relevant problems. The divergence between the ambiguity set $\mathcal{S}$ and the underlying stationary marginal distribution $\mathbb{Q}^\star$ can in this particular case even be explicitly characterized as

$$\inf_{\mathbb{Q} \in \mathcal{S}} D(\mathbb{Q} \| \mathbb{Q}^\star) = 1 + (\mathbb{Q}^\star(\text{tail}) - \epsilon) \log(\mathbb{Q}^\star(\text{tail}) - \epsilon) + (\mathbb{Q}^\star(\text{head}) + \epsilon) \log(\mathbb{Q}^\star(\text{head}) + \epsilon).$$

In Figure 2, the difference between the exact expression (**??**) and its large deviation approximation in equation (3.5) is visually illustrated for a fair coin and $\epsilon = 0.15$ as a function of the sample size n. By visual inspection, expression (3.5) offers an exact expression for the dominating rate the convergence with which $\hat{\mathbb{Q}}_n$ tends to $\mathbb{Q}^\star$.

For the second experiment, we consider two families of irreducible Markov chains and illustrate that it can be very hard to distinguish between the two even though their stationary distributions are dramatically different. This seemingly harmless observation will be shown to have far reaching consequences in the last section of this paper. Let us define two families of irreducible Markov coins parametrized in $\epsilon \in (0, 1)$ explicitly given by the following transition matrices

$$\mathbb{Q}_\epsilon(\cdot|\cdot) = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix} \quad \text{and} \quad \mathbb{Q}^\star_\epsilon(\cdot|\cdot) = \begin{pmatrix} \epsilon & 1-\epsilon \\ \epsilon & 1-\epsilon \end{pmatrix},$$

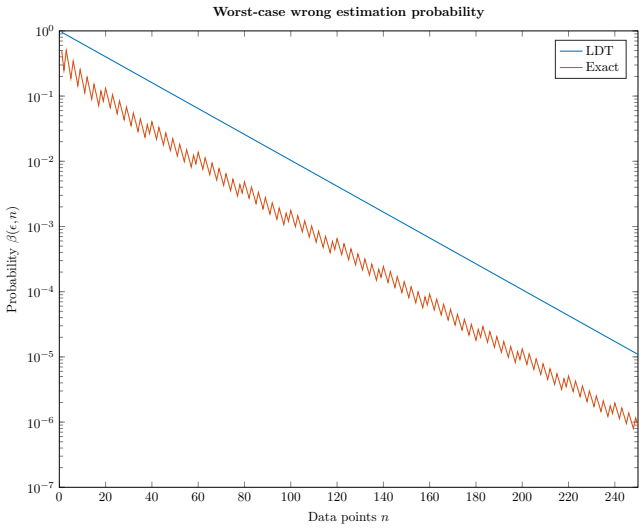
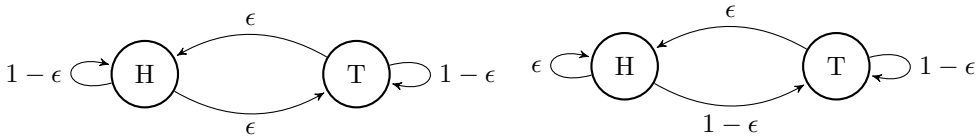


Figure 2: One can visually check here that the large deviation bound (3.5) provides an exact expression of the exponential decay of the probability $\beta(\epsilon, n)$ of estimator failure.



(a) Markov coin $\mathbb{Q}_\epsilon(\cdot|\cdot)$ (b) Markov coin $\mathbb{Q}^\star_\epsilon(\cdot|\cdot)$

Figure 3: Visualization of the Markov coins $\mathbb{Q}_\epsilon(\cdot|\cdot)$ and $\mathbb{Q}^\star_\epsilon(\cdot|\cdot)$. Although the Markov coins are dramatically different for small ϵ, it is nevertheless very hard to distinguish between both coins based solely on data.

with unique stationary distributions $\mathbb{Q}_\epsilon = [1/2, 1/2]$ and $\mathbb{Q}^\star_\epsilon = [\epsilon, 1-\epsilon]$, respectively. It is worth pointing out that the Markov coin defined through the transition matrices $\mathbb{Q}^\star_\epsilon(\cdot|\cdot)$ results in an i.i.d. process. Both families of Markov coins are depicted graphically in Figure 3.

The conditional relative entropy between both families of Markov chains can be explicitly characterized as

$$D_c(\mathbb{Q}_\epsilon(\cdot,\cdot)\|\mathbb{Q}^\star_\epsilon(\cdot,\cdot)) = \epsilon(1-2\epsilon)\log\left(\frac{1-\epsilon}{\epsilon}\right),$$

and is shown in Figure 4. Curiously, it can be remarked that $\lim_{\epsilon\to 0} D_c(\mathbb{Q}_\epsilon(\cdot,\cdot)\|\mathbb{Q}^\star_\epsilon(\cdot,\cdot)) = 0$. Theorem **??** thus implies that although the Markov coins are dramatically different for small ϵ, it is nevertheless very hard to distinguish between both coins based solely on data samples. This remarkable phenomenon will be discussed in greater detail at the end of Section 6.

4 Distributionally robust predictions and prescriptions

In this section we return at last to the problem of data-driven prescription and prediction put forward in the introduction. Here, the reader will come to appreciate the true power of large deviation theory in providing a clarifying and unifying perspective to prediction and prescription with data.

Recall that we want to construct data-driven solutions $\hat{c}_n$ which are both asymptotically consistent **??** and which furthermore enjoy finite sample guarantees **??**.

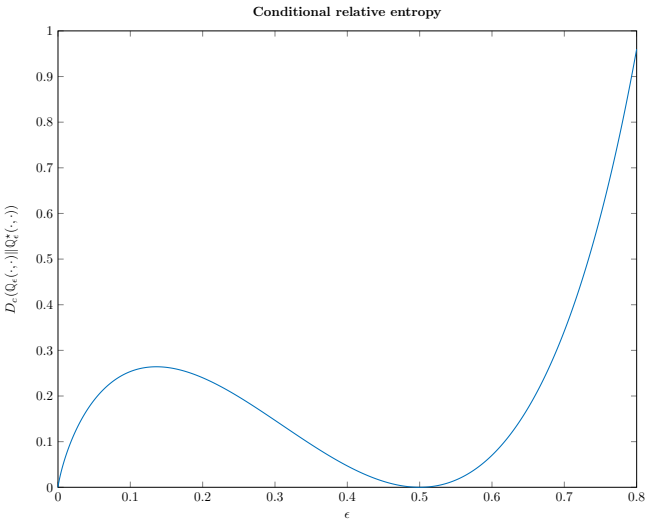


Figure 4: The relative entropy between the families of Markov coins defined through the transition matrices $\mathbb{Q}_\epsilon(\cdot|\cdot)$ and $\mathbb{Q}^\star_\epsilon(\cdot|\cdot)$ depicted in Figure 3. Observe that for $\epsilon = 0.5$ both coins are equal and thus $D_c(\mathbb{Q}_{0.5}(\cdot,\cdot)\|\mathbb{Q}^\star_{0.5}(\cdot,\cdot)) = 0$ as expected. However, also for ϵ tending to zero the conditional relative entropy between both Markov coins tends to zero although the Markov coins are dramatically different.

The sample average approximation is obtained by setting the data-driven predictor to

$$\hat{c}_0(x, \hat{\theta}_t) = c(x, \hat{\theta}_t) \tag{4.1}$$

has received a lot of attention. Despite its simplicity, the sample average approximation $\hat{c}_0$ results in data-driven predictions and prescriptions which are asymptotically consistent under rather mild technical conditions. The law of large numbers guarantees that the empirical estimator $\hat{\theta}_t$ converges to its limit θ. Under mild conditions this weak convergence guarantees the consistency of the stochastic average approximation as well. Unfortunately, the sample average approximation comes without any finite sample guarantees. For that indeed we need to have that $\hat{\theta}_t$ converges to its limit θ with a certain rate.

Assumption 4.1 (Strong large deviation principle)**.** *The empirical estimators $\{\hat{\theta}_T\}_{T\in\mathbb{N}}$ satisfy an SLDP of the form* (3.2) *under the model class Θ with a regular rate function I.*

The large deviation Assumption 4.1 will make it possible to construct data-driven solutions which are not only asymptotically consistent, but satisfy the finite sample requirement **??** as well. In order to do so, we consider data-driven solutions as defined by the optimization problem

$$\hat{c}_r(x, \theta) = \sup_{\theta' \in \Theta} \left\{ c(x, \theta') : I(\theta, \theta') \leq r \right\}. \tag{4.2}$$

recall that $I(\theta, \theta') = \infty$ whenever $\theta' \notin \mathcal{S}_\theta$

$$\hat{x}_r(\theta) \in \arg\min_{x \in X} \hat{c}_r(x, \theta) \tag{4.3}$$

Careful observation of our data-driven solutions reveals that it can be considered as the robust counterpart of the sample average approximation given in equation (4.1) with respect to the set

$$\mathcal{R}_r(\theta) = \left\{ \theta' \in \Theta \ : \ I(\theta, \theta') \leq r \right\}.$$

The sample average approximation is recovered for the trivial rate ($r = 0$) for which it follows that $\mathcal{R}_0(\theta) = \{\theta\}$. The previous set should be put in sharp contrast to the closely related set

$$\mathcal{I}_r(\theta') = \left\{ \theta \in \Theta \ : \ I(\theta, \theta') \leq r \right\}$$

as the rate function I is not symmetric in general. Curiously enough though, we will have necessity for the latter set as well.

For the independent data class $\mathcal{Q}_n^{\text{iid}}$, our data-driven solutions reduce to distributionally robust optimization problems over the relative entropy ball as explicitly pointed out already in (4.14). In that sense, they are of the same type of the data-driven solutions as discussed in for instance by Bertsimas [2] and Esfahani and Kuhn [5]. The power of the large deviation approach taken here is that we can carry out the analysis of the data-driven solution (4.2) without needing to specify the data class $\mathcal{Q}_n$ but instead work with its corresponding rate function directly. In this way, whether we work with i.i.d. or Markov processes will pose no difference to the analysis presented here.

The current section will discuss the merits of the data-driven solution $\hat{c}_{n,r}$ as a predictor in Section 4.1 and as a prescriptor in Section 4.2. The rate $r > 0$ will come to determine the exponential rate with which the out-of-sample disappointment of our data-driven predictions and prescriptions $\hat{c}_{n,r}$ drop to zero with an increasing number of samples n. Furthermore, we will be able to argue that our data-driven solutions are in some sense optimal and can not be improved upon. Although distributionally robust optimization has experienced a lot of attention, see Zymler et al. [17] and Van Parys [15], we are the first to discover its optimality in a data-driven context.

4.1 Data-driven predictions

We now analyze the performance of our data-driven approach (4.2) in providing data-driven predictions using the large deviation perspective discussed in the previous section. The following theorem indicates that the parameter r in the data-driven predictions $\hat{c}_{n,r}$ determines the rate with which their out-of-sample prediction disappointment drops to zero in function of the number of data samples t.

Theorem 4.1 (Data-driven predictors)**.** *Assume that the model class Θ satisfies a strong large deviation principle with good rate function I. Then, the out-of-sample disappointment of the data-driven predictor $\hat{c}_r$ defined in* (4.2) *enjoys the following finite sample guarantee under any model $\theta \in \Theta$ and for any $x \in X$.*

$$\mathbb{P}_\theta \left(c(x,\theta) > \hat{c}_r(x, \hat{\theta}_T) \right) \leq (T+1)^{\dim \Theta} \cdot e^{-rT} \quad \forall T \in \mathbb{N} \tag{4.4}$$

Proof. We have $c(x,\theta) > \hat{c}_r(x,\hat{\theta}_T)$ if and only if the estimator $\hat{\theta}_T$ falls within the disappointment set

$$\mathcal{D}(x,\theta) = \{\theta' \in \mathcal{S}_\theta : c(x,\theta) > \hat{c}_r(x,\theta')\}.$$

Note that by the definition of $\hat{c}_r$, we have

$$I(\theta',\theta) \leq r \quad \Longrightarrow \quad \hat{c}_r(x,\theta') = \sup_{\theta'' \in \Theta} \{c(x,\theta'') : I(\theta',\theta'') \leq r\} \geq c(x,\theta).$$

By contraposition, the above implication is equivalent to

$$c(x,\theta) > \hat{c}_r(x,\theta') \quad \Longrightarrow \quad I(\theta',\theta) > r.$$

Therefore, $\mathcal{D}(x,\theta)$ is a subset of

$$\mathcal{D}(\theta) = \{\theta' \in \Theta : I(\theta',\theta) > r\}.$$

irrespective of $x \in X$. For any fixed $\theta \in \Theta$ we thus have

$$\begin{aligned} \frac{1}{T} \log \mathbb{P}_\theta \left(\hat{\theta}_T \in \mathcal{D}(x,\theta) \right) &\leq \frac{1}{T} \log \mathbb{P}_\theta \left(\hat{\theta}_T \in \mathcal{D}(\theta) \right) \\ &\leq \dim \Theta \cdot \frac{\log(T+1)}{T} - \inf_{\theta' \in \operatorname{cl} \mathcal{D}(\theta)} I(\theta',\theta) \\ &\leq \dim \Theta \cdot \frac{\log(T+1)}{T} - r, \end{aligned} \tag{4.5}$$

where the first inequality follows from the inclusion $\mathcal{D}(x,\theta) \subseteq \mathcal{D}(\theta)$, and the second inequality holds because the model class Θ obeys a strong large deviation principle with good rate function I. The last inequality holds because I is continuous in its first argument, which implies that $\operatorname{cl} \mathcal{D}(\theta) \subseteq \{\theta' \in \Theta : I(\theta',\theta) \geq r\}$ and

$$\inf_{\theta' \in \operatorname{cl} \mathcal{D}(\theta)} I(\theta',\theta) \;\geq\; \inf_{\theta' \in \Theta} \{I(\theta',\theta) : I(\theta',\theta) \geq r\} \;\geq\; r.$$

Multiplying (4.5) by t and exponentiating both sides of the resulting inequality yields the postulated finite sample guarantee (4.4). □

The rate parameter r encoding the predictor $\hat{c}_r$ captures the inherent trade-off between out-of-sample disappointment and accuracy which must be made by any data-driven attempt. With an increasing rate function r, our predictor indeed become more reliable when measured its out-of-state prediction disappointment. However, increasing the rate function r translates at the same time into more conservative predictions as they become more robust. The stochastic average approximation ($r = 0$) takes no interest in out-of-sample disappointment and consequently is less conservative then predictors who do ($r > 0$). We intend to show now that our data-driven predictors $\hat{c}_{n,r}$ are all optimal in the sense that they make this inherent trade-off between both interest optimally.

The previous theorem guarantees that the out-of-sample disappointment of our data-driven predictor $\hat{c}_{n,r}$ decreases with the number of samples n exponentially at rate r to zero. The following theorem indicates that no estimator exists with a better out-of-sample disappointment rate than our optimal predictor $\hat{c}_{n,r}$ amongst a fairly large class of predictors.

If an arbitrary data-driven predictor $\hat{c}$ predicts a lower expected cost than $\hat{c}_r$ only in one state of the world $\omega \in \Omega$ and only for one sample size t, then $\hat{c}$ must suffer from a higher out-of-sample disappointment than $\hat{c}_r$ to first order in the exponent.

Suppose that $\hat{c}$ is less conservative than $\hat{c}_r$ in at least one scenario

Theorem 4.2 (Pareto efficiency of the predictor $\hat{c}_r$)**.** *Assume that the model class Θ satisfies a weak large deviation principle with good rate function I. Fix some decision $x \in X$ and decay rate $r > 0$, and consider a data-driven predictor $\hat{c}$ that is more optimistic than $\hat{c}_r$ for some outcome $\omega_0 \in \Omega$ and sample size $t_0 \in \mathbb{N}$, that is, $\hat{c}(x, \hat{\theta}_{t_0}(\omega_0)) < \hat{c}_r(x, \hat{\theta}_{t_0}(\omega_0))$. Then, there exists $\theta^\star \in \Theta$ and $r^\star \in (0, r)$ with*

$$-r^\star \leq \liminf_{t\to\infty} \frac{1}{t} \log \mathbb{P}_{\theta^\star} \left(c(x,\theta^\star) > \hat{c}(x, \hat{\theta}_t) \right).$$

Proof. Fix $x \in X$, $r > 0$, $\omega_0 \in \Omega$ and $t_0 \in \mathbb{N}$ as in the theorem statement. Throughout the proof we use θ' as a notational shorthand for $\hat{\theta}_{t_0}(\omega_0)$. Similarly, we use ϵ to denote the prediction difference $\hat{c}_r(x,\theta') - \hat{c}(x,\theta') > 0$. Note that both $\theta' \in \Theta$ and $\epsilon > 0$ are deterministic quantities because ω_0 and t_0 are fixed. By the definition of $\hat{c}_r$ as a worst-case expectation, there exists an $\frac{\epsilon}{2}$-suboptimal model $\theta \in \Theta$ in (4.2) with $I(\theta',\theta) \leq r$ and

$$\hat{c}_r(x,\theta') \leq c(x,\theta) + \frac{\epsilon}{2}. \tag{4.6}$$

Consider now a model $\theta(\lambda) = (1-\lambda)\theta' + \lambda\theta$, $\lambda \in [0,1]$, on the line segment between θ' and θ. As r is strictly positive by assumption, the radial monotonicity of the good rate function I implies that $I(\theta', \theta(\lambda)) < r$ for every $\lambda \in (0,1)$. Moreover, as the expected cost $c(x,\theta_\lambda)$ changes continuously in λ, there exists $\lambda^\star \in (0,1)$ such that $\theta^\star = \theta(\lambda^\star)$ and $r^\star = I(\theta',\theta^\star)$ satisfy $0 < r^\star < r$ and

$$c(x,\theta) < c(x,\theta^\star) + \frac{\epsilon}{2}. \tag{4.7}$$

In summary, we thus have

$$\hat{c}(x,\theta') = \hat{c}_r(x,\theta') - \epsilon \leq c(x,\theta) - \frac{\epsilon}{2} < c(x,\theta^\star) \leq \hat{c}_r(x,\theta'), \tag{4.8}$$

where the first and second inequalities follow from (4.6) and (4.7), respectively, while the third inequality follows from the definition of $\hat{c}_r$ and the fact that $I(\theta',\theta^\star) \leq r$.

In the remainder of the proof we will argue that the prediction disappointment $\mathbb{P}_{\theta^\star}(c(x,\theta^\star) > \hat{c}(x,\hat{\theta}_t))$ under model $\theta^\star$ decays at a rate of at most $r^\star$ as the sample size t tends to infinity. In analogy to the proof of Theorem 4.1, we define the set of disappointing estimator realizations as

$$\mathcal{D}(x,\theta^\star) = \{\theta \in \Theta \ : \ c(x,\theta^\star) > \hat{c}(x,\theta)\}\,.$$

This set contains θ' due to the inequalities (4.8) and is open in the subspace topology on Θ because $\hat{c}(x,\theta)$ is continuous in θ. Thus, we find

$$\inf_{\theta \in \operatorname{int} \mathcal{D}(x,\theta^\star)} I(\theta,\theta^\star) = \inf_{\theta \in \mathcal{D}(x,\theta^\star)} I(\theta,\theta^\star) \leq I(\theta',\theta^\star) = r^\star,$$

where the inequality holds because $\theta' \in \mathcal{D}(x,\theta^\star)$, and the last inequality follows from the definition of $r^\star$. As model $\theta^\star$ satisfies a weak large deviation principle with rate function I, we finally conclude that

$$-r^\star \leq - \inf_{\theta \in \operatorname{int} \mathcal{D}(x,\theta^\star)} I(\theta,\theta^\star) \leq \liminf_{t\to\infty} \frac{1}{t} \log \mathbb{P}_{\theta^\star}\left(\hat{\theta}_t \in \mathcal{D}(x,\theta^\star)\right),$$

and thus the claim follows. □

The previous theorem tells us that among a fairly large class of predictors, our distributionally robust solutions (4.4) can not be improved upon. We remark that the class of predictors which depend on the data directly through the empirical estimator indeed includes all known data-driven predictors available in the literature. In the dependent case, this functional restriction translates to predictors which are not sensitive to the order in which the data samples are presented. Among all those, our data-driven predictors $\hat{c}_{n,r}$ are established in Theorem 4.2 as optimal. That is, any attempt to make them less conservative invariable means elevating the out-of-sample prediction disappointment and vice versa. In other words, the predictors $\hat{c}_{n,r}$ are Pareto optimal with respect to out-of-sample prediction disappointment and accuracy of their provided predictions.

4.2 Data-driven prescriptions

We now argue that the data-driven solutions (4.2) provide trustworthy prescriptions as well. In fact, the following theorem provides a verbatim guarantee when compared to its counterpart Theorem 4.1 for the prediction problem.

Theorem 4.3 (Data-driven prescriptors)**.** *Assume that the model class Θ satisfies a strong large deviation principle with good rate function I. Then, the out-of-sample disappointment of the data-driven prescriptor $\hat{x}_r$ defined in* (4.3) *enjoys the following finite sample guarantee under any model $\theta \in \Theta$.*

$$\mathbb{P}_\theta\left(c(\hat{x}_r(\hat{\theta}_t),\theta) > \hat{c}_r(\hat{x}_r(\hat{\theta}_t),\hat{\theta}_t)\right) \leq (t+1)^{\dim \Theta} \cdot e^{-rt} \quad \forall t \in \mathbb{N} \tag{4.9}$$

Proof. Define $\mathcal{D}(\theta)$ and $\mathcal{D}(x,\theta)$ as in the proof of Theorem 4.1 and recall that $\mathcal{D}(x,\theta) \subseteq \mathcal{D}(\theta)$ for every decision $x \in X$ and $\theta \in \Theta$. Thus, for every fixed outcome $\omega \in \Omega$ we have

$$\begin{array}{lll} c(\hat{x}_r(\hat{\theta}_t(\omega)),\theta) > \hat{c}_r(\hat{x}_r(\hat{\theta}_t(\omega)),\hat{\theta}_t) & \implies & \exists x \in X \text{ with } c(x,\theta) > \hat{c}_r(x,\hat{\theta}_t(\omega)) \\ & \implies & \hat{\theta}_t(\omega) \in \cup_{x\in X} \mathcal{D}(x,\theta) \\ & \implies & \hat{\theta}_t(\omega) \in \mathcal{D}(\theta), \end{array}$$

which in turn implies

$$\frac{1}{t}\log \mathbb{P}_\theta\left(c(\hat{x}_r(\hat{\theta}_t),\theta) > \hat{c}_r(\hat{x}_r(\hat{\theta}_t),\hat{\theta}_t)\right) \leq \frac{1}{t}\log \mathbb{P}_\theta\left(\hat{\theta}_t \in \mathcal{D}(\theta)\right) \leq \dim\Theta \cdot \frac{\log(t+1)}{t} - r$$

for every model $\theta \in \Theta$ and sample size $t \in \mathbb{N}$. Note that the second inequality in the above expression has already been established in the proof of Theorem 4.1. Thus, the claim follows. □

We remark that our data-driven solution $c_{n,r}$ enjoys the same finite sample guarantee for both the prediction and prescription problem. Both the finite sample guarantees (4.4) and (4.9) are pre-decision and sample independent. They guarantee that the data-driven solution $c_{n,r}$ delivers trustworthy predictions and prescriptions before the data is revealed. The out-of-sample disappointment is a property characterizing the quality of a predictor or prescription. In subsequent section, we investigate what can be said concerning the quality of the data-driven solution after the data is revealed.

4.3 Outline and contributions

Any data-driven approach must overcome the ambiguity concerning the distribution underlying the data by observing sufficiently many historical data samples. If we are to have asymptotically consistent predictions or prescriptions, then we must be able to tell in the limit which of the distributions in the data class $\mathcal{Q}_n$ did in fact generate our particular data. Whether this is at all possible is related to data class $\mathcal{Q}_n$ itself. If $\mathcal{Q}_n$ is a singleton then of course our work would be done. However if the data class $\mathcal{Q}_n$ is simply too large no data-driven approach can ever succeed. The most commonly made restriction is that the process underlying the data is i.i.d.. This well studied independent data class is given as the set of all n-fold distributions

$$\mathcal{P}_{[T]} = \left\{ \mathbb{P}_{[T]} = \mathbb{P}^T \ : \ \mathbb{P} \text{ is any distribution on } \Xi \right\}. \tag{4.10}$$

We will go here well beyond the independent data by introducing several other types of data classes which from a first glance are quite distinct. Using a large deviation perspective however, we shall indicate that a unified approach to each of the distinct data classes is within reach. A major contribution of this paper is hence the introduction of a novel large deviation perspective which we believe is both fundamental and intuitive to any data-driven approach. We shall highlight the structure of the paper by anticipating the results obtained for the independent data class $\mathcal{Q}_n^{\text{iid}}$ and remark to what extent they generalize to the other settings as well.

Any asymptotically consistent approach in a data-driven setting (implicitly) hinges on the fact whether one can determine the unknown marginal distribution $\mathbb{Q}^\star$ given sufficiently many historical observations. In Section 5 we introduce several classes of data processes for which this is the case. Independent data and more general Markov data in which each data sample can depend on its predecessor, are arguably the most noteworthy discussed in this paper. We do assume in these settings that the event space of values which the uncertainty can take is finite. Although the assumption is not always strictly necessary, it makes the exposition of the paper that much more digestible.

Assumption 4.2 (Finite state space)**.** *The state space Ξ has finite cardinality.*

Notice that the distribution underlying independent data could be discovered if we had access to a distribution estimator $\hat{\mathbb{Q}}_n$ which converges to the unknown distribution $\mathbb{Q}^\star$. Stated more precisely, whether we can construct an estimator such that

$$\hat{\mathbb{Q}}_n \to \mathbb{Q}^\star, \tag{4.11}$$

whatever was the unknown marginal $\mathbb{Q}^\star$ with the number of samples n tending to infinity. All data classes discussed in Section 5 will admit such an estimator. A variation on the weak law of large numbers for $\hat{\mathbb{Q}}_n := \frac{1}{n}\sum_{i=1}^n \delta_{\hat{\xi}_i}$ with independent data usually does the trick. We end Section 5 with remarking that naively replacing the unknown marginal distribution with its estimate yields the popular stochastic average approximation

$$\hat{c}_{n,0}(x,\xi) := \int_\Xi \ell(x,\xi)\, \hat{\mathbb{Q}}_n(\mathrm{d}\xi) = c(x, \hat{\mathbb{Q}}_n) \quad \text{and} \quad \hat{x}_{n,0} := \arg\min \ c(x, \hat{\mathbb{Q}}_n). \tag{4.12}$$

The consistency of the empirical distribution $\hat{\mathbb{Q}}_n$ in estimating $\mathbb{Q}^\star$ ensures under mild technical conditions that the stochastic average approximation $\hat{c}_{n,0}$ provides asymptotically consistent **??** predictions and prescriptions as well. The situation changes profoundly though when the finite sample guarantees **??** enter the picture. When asymptotically consistency is all that is envisioned, a data process in which the convergence (4.11) takes place is all what is needed. When finite sample guarantees are needed as well, then not only must $\hat{\mathbb{Q}}_n$

converge to $\mathbb{Q}^\star$, it must do so at a certain speed. It is well known that for independent data the convergence (4.11) takes place at exponential speed. For reasons beyond the scope of this outline, the same can be said for a much larger class of data processes as well. An excellent intuitive explanation of why exponential convergence occurs with independent data is offered by Cover [3, Chapter 11]. For the more general case, the interested reader is referred to the classic book [4] by Dembo and Zeitouni.

Large deviation theory is concerned with characterizing the exact exponential rate with which the empirical estimator $\hat{\mathbb{Q}}_n$ converges to its limit $\mathbb{Q}^\star$ as a function of the number of samples n in a certain data class $\mathcal{Q}_n$. Notice that this rate is dependent both on data class $\mathcal{Q}_n$ and estimator $\hat{\mathbb{Q}}_n$. In case of independent data, large deviation theory loosely establishes that the empirical distribution converges at exponential[4] speed

$$\mathbb{Q}_n^\star(\hat{\mathbb{Q}}_n \in \mathcal{S}) \approx e^{-n \cdot \inf_{\mathbb{Q}\in\mathcal{S}} D(\mathbb{Q},\mathbb{Q}^\star)}, \quad \forall \mathbb{Q}_n^\star \in \mathcal{Q}_n^{\text{iid}} \tag{4.13}$$

where $\mathcal{S}$ is any set of distributions on Ξ with rate function D the Kullback-Leibler divergence [3]. We say that convergence takes place with rate $\inf_{\mathbb{Q}\in\mathcal{S}} D(\mathbb{Q},\mathbb{Q}^\star)$. The rate function thus entirely determines the exponential convergence behavior of $\hat{\mathbb{Q}}_n \to \mathbb{Q}^\star$. Different data processes $\mathcal{Q}_n$ have different rate functions. For instance in Section 3 we indicate that the rate of convergence in case of Markov processes will determined by the conditional relative entropy instead. Our novel large deviation perspective is powerful in that is makes abstraction of the particular data process $\mathcal{Q}_n$ underlying the data and merely makes use of its corresponding rate function. The large deviation perspective will enable us to relax the commonly made independence assumption on the data without much additional effort compared to the i.i.d. case.

We will show in Section 4 that distributionally robust optimization can be used to construct data-driven predictors and prescriptors which satisfy both finite sample guarantees **??** and are asymptotically consistent **??**. For independent data our novel data-driven solution specializes to

$$\begin{array}{rl} \hat{c}_{n,r}(x,\xi) := \sup\limits_{\mathbb{Q}} & \displaystyle\int_\Xi \ell(x,\xi)\,\mathbb{Q}(\mathrm{d}\xi), \\ \text{s.t.} & D(\hat{\mathbb{Q}}_n,\mathbb{Q}) \leq r, \end{array} \tag{4.14}$$

which is directly based on the rate function of the data class $\mathcal{Q}_n^{\text{iid}}$. It is worth pointing out that our data-driven results are recognized as the distributionally robust counterpart to the well known stochastic average approximation solution (4.12) with respect to the distributional ambiguity set

$$\mathcal{R}(\hat{\mathbb{Q}}_n) := \left\{ \mathbb{Q} \;:\; D(\hat{\mathbb{Q}}_n,\mathbb{Q}) \leq r \right\}.$$

The results here follow from a simple yet powerful observation concerning previously defined ambiguity set. The set $\mathcal{R}(\hat{\mathbb{Q}}_n)$ is the smallest set around the empirical distribution $\hat{\mathbb{Q}}_n$ with respect to set inclusion for which the probability $\mathbb{Q}_n^\star(\mathbb{Q}^\star \notin \mathcal{S})$ diminishes at least with rate r. For the sake of argument lets pretend for the moment that (4.13) holds with equality for all n. This is not the case exactly but will prove insightful in the discussion here. By letting r scale as $\frac{1}{n}\log(\frac{1}{\beta})$ our data-driven prediction (4.14) and its corresponding prescription will both enjoy a finite sample disappointment probability less than β. From the fact that the set to which we are robust is the smallest conceivable, we shall in Section 4 be able to establish a notion in which the predictor (4.14) is optimal.

We are not the first to consider distributionally robust optimization for the construction of data-driven solutions as both Bertsimas et al. [2] and Esfahani and Kuhn [5] employed a similar strategy. We do go beyond what is currently known in two essential ways. Using large deviation theory, we show that indeed (i) our constructed predictor (4.14) is optimal and that (ii) the results anticipated here admit an intuitive extension beyond independent data equally well. We want to point out that the Kullback-Leibler divergence has been used in the context of independent data already in for instance Hu and Hong [6] and Lam [8]. However, the distributional robust counterpart of the stochastic average approximation with respect to the set

$$\mathcal{I}(\hat{\mathbb{Q}}_n) := \left\{ \mathbb{Q} \;:\; D(\mathbb{Q},\hat{\mathbb{Q}}_n) \leq r \right\}$$

[4]Here the sign $\approx$ means up to polynomial terms or to be more precise $\frac{1}{n}\log \mathbb{Q}_n^\star(\hat{\mathbb{Q}}_n \in \mathcal{S}) \to \inf_{\mathbb{Q}\in\mathcal{S}} D(\mathbb{Q},\mathbb{Q}^\star)$.

was considered instead. However, as the rate function is not symmetric those results do not enjoy any similar advantageous properties as the ones discussed before.

One can remark here that the out-of-sample disappointment β for the predictions and prescriptions defined in Definition **??** and 2.2 is an a priori quantity independent of the observed data. It is a good indicator of trustworthy of the data-driven solution as a process from data to prediction or prescription *before* the data is revealed. However, in a practical context one also might be interested in assessing the a posteriori accuracy of a data-driven solution *after* the data has been disclosed. Indeed, consider for instance the problem of prediction with data. The achieved a posteriori accuracy is determined partly by the a priori quality of the predictor and partly, no doubt, by the "luck of the draw". Although the proportions in which these two factors affect the estimate may very widely, it is the presence of both that characterizes the quality of the solution. To put it in another way – for one may be troubled by the time element the argument seems to involve – the out-of-sample disappointment is an attribute of the data-driven predictor, the rule by which we predict whatever may be the data; whereas the accuracy of the solution is an attribute of the prediction, the evaluation of the predictor specific to an observed sample.

The accuracy of a prediction after the data is observed for our optimal predictor (4.14) we will come to denote as the value of data (VoD). In full agreement with the results presented by Lindsay and Li [10], we will show that the observed Fisher information gives the best assessment of the accuracy of the estimate after the data is observed. Using the fact that the results here do not hinge on the data class $\mathcal{Q}_n^{\text{iid}}$, we also present *en route* a counterpart to the Fisher information in the Markov setting as well.

5 In the beginning there was data

In this section we will make the case that if we are to entertain any hope to solve the classical prescription problem (2.1) or prediction problem (**??**) in a data-driven fashion, then the data class $\mathcal{Q}_n$ must be restricted beyond what is suggested by Assumption 2.2. We will do so by enforcing various notions of inter-sample independence. We discuss three data classes introducing distinct independence conditions on the data.

We discuss in every such class what form the weak law of large numbers (4.11) takes. Each data class discussed admits an empirical estimator which discovers the unknown true distribution underlying the data process given a sufficient amount of data samples. Although the different data types discussed here look quite dissimilar, we will show in subsequent Section 3 that in fact they are governed by the same underlying principle. Alluding to this profound result, we already introduce for each data class a function which will characterize how fast introduced empirical estimators unveils the unknown true distributions exactly.

The stationarity Assumption 2.2 forces the data to be identically distributed but is in general not sufficient to guarantee any hopes of constructing data-driven solutions. This last observation can be made by considering the following counterexample. Consider a degenerate data process in which all data samples are equal across time, i.e. $\hat{\xi}_i = \hat{\xi}_j$ for all i and j, but nevertheless distributed identically as $\mathbb{Q}^\star$. As it is evident that a sample path in this degenerate case does not contain much information concerning its unknown underlying stationary distribution $\mathbb{Q}^\star$, it is clear that further conditions on the data process $\hat{\xi}$ and its distribution $\mathbb{Q}_n^\star$ are necessary. In other words, the ambiguity set $\mathcal{Q}_n$ is at the moment too big to admit data-driven predictors or prescriptors. Some sort of independence notion is called for. In what follows we introduce different three types of data classes for which we will show in Section 4 that they admit data-driven solutions.

Independent data with unknown location parameter

We consider here data which originates from independent identically distributed process in which the marginal distribution $\mathbb{Q}^\star$ is known up to a certain location parameter $\mu^\star$. In this simple data class which will serve to illustrate the power of large deviation theory later in the paper, we do not even have need for the finite event space Assumption 4.2. All we know is thus that the marginal distribution $\mathbb{Q}^\star$ is a shifted version of a known centered distribution $\mathbb{Q}_0 := \mathbb{Q}^\star - \mu^\star$. For instance, we might assume the data to be a standard normal with unknown mean μ. The sample path distribution $\mathbb{Q}_n^\star$ for this i.i.d. process satisfies the following

simple expression

$$\mathbb{Q}_n^\star(\hat{\xi}_1 = \xi_1, \ldots, \hat{\xi}_n = \xi_n) = \prod_{i=1}^n \mathbb{Q}^\star(\xi_i), \quad \forall n \in \mathrm{N} \tag{5.1}$$

Recall that as only data is given, neither the sample path distribution $\mathbb{Q}_n^\star$ nor its marginal distribution $\mathbb{Q}^\star$ is known. As discussed in the introduction, the data class described here can be described uniquely through its induced ambiguity set $\mathcal{Q}_n^{\mathrm{loc}}$. From the simple expression (5.1) it follows immediately that the ambiguity set can be parametrized in the location as

$$\mathcal{Q}_n^{\mathrm{loc}} := \{\mathbb{Q}_n := (\mathbb{Q}_0 + \mu)^n \ : \ \mu \in \Xi\}\,. \tag{5.2}$$

From the previous characterization of the ambiguity we face in the data it is quite clear that figuring out which distribution $\mathbb{Q}_n^\star$ in $\mathcal{Q}_n^{\mathrm{loc}}$ does underly the data boils down to working out its corresponding unknown location parameter $\mu^\star$. To that end we define its empirical estimator as the average of the observed samples

$$\hat{\mu}_n := \tfrac{1}{n} \textstyle\sum_{i=1}^n \hat{\xi}_i.$$

The classical (weak) law of large numbers ensures that this empirical estimator converges (weakly) to its limit $\mu^\star$ for an increasing number of data samples n. In Section 3 we shall discuss exactly how fast the convergence of $\hat{\mu}_n$ to its limit $\mu^\star$ takes place. Anticipating those results here, we shall have a necessity for the cumulant generating function of the distribution $\mathbb{Q}_0$.

The cumulant generating function associate to any distribution $\mathbb{Q}_0$ on the event set Ξ a convex function $\Lambda(\lambda) := \log \int \exp \lambda^\top \xi \, \mathbb{Q}_0(\mathrm{d}\xi)$. The exponential rate of convergence from $\hat{\mu}_n$ to its limit $\mu^\star$ shall be determined by the Fenchel-Legendre transformation of the moment generating function Λ as we will show in Section 3. The moment generating function and its Fenchel-Legendre transformation for a few common distributions are given in Table 1.

Definition 5.1 (The Fenchel-Legendre transformation)**.** *The Fenchel-Legendre transformation of a function* Λ *is given as the convex function* $\Lambda^\star : \Xi \to \mathrm{R},\ x \mapsto \sup_{\lambda \in \Xi} \lambda^\top x - \Lambda(\lambda)$.

Independent data with unknown marginal distribution

By far the most common notion of independence is that of the i.i.d. process as already introduced in the introduction. Most literature discussing data-driven approaches give the i.i.d. assumption a central role in their work. Following Assumption 4.2, we will have need here for the assumption that the event set Ξ has finite cardinality.

In this context, all distinct random variables $\hat{\xi}_i$ making up the stochastic process are considered to be mutually independent. Under previous assumption, the sample path distribution $\mathbb{Q}_n^\star$ for an i.i.d. process satisfies again the simple expression 5.1. As discussed in the introduction, the data class described here can be described uniquely through its induced ambiguity set $\mathcal{Q}_n^{\mathrm{iid}}$. From the simple expression (5.1) it follows immediately that the ambiguity set can be parametrized in the marginal distribution as

$$\mathcal{Q}_n^{\mathrm{iid}} := \{\mathbb{Q}_n := (\mathbb{Q})^n \ : \ \mathbb{Q} \in \mathcal{P}_\Xi\}\,. \tag{5.3}$$

Again, from previous characterization of the ambiguity we face in the data it is quite clear that figuring out which distribution $\mathbb{Q}_n^\star$ in $\mathcal{Q}_n^{\mathrm{iid}}$ does underly the data boils down to working out its corresponding unknown marginal distribution $\mathbb{Q}^\star$. The empirical distribution of the first n realizations of the data process is defined as

$$\hat{\mathbb{Q}}_n := \tfrac{1}{n} \textstyle\sum_{i=1}^n \delta_{\hat{\xi}_i}.$$

Convergence of the empirical distribution to the unknown marginal distribution is ensured through the law of large numbers as found in for instance Baum and Katz [1]. We have indeed $\hat{\mathbb{Q}}_n \to \mathbb{Q}^\star$ as stated in equation (4.11). That is, the empirical distribution converges in distribution to the underlying stationary distribution $\mathbb{Q}^\star$ given sufficient data samples. Hence, given an unlimited amount of data the ambiguity concerning the distribution underlying the data eventually disappears.

As discussed in the introduction, the exponential rate of convergence from $\hat{\mathbb{Q}}_n$ to its limit $\mathbb{Q}^\star$ shall be determined by the Kullback-Leibler divergence. Another name for the same function is the relative entropy.

Markov data with unknown doublet distribution

A weaker independence notion is that the data process ξ is first-order Markov. In a first-order Markov chain each data sample is allowed to depend on the value of its direct predecessor. Once again, we assume that the event set Ξ is finite a voiced by Assumption 4.2. It should be remarked that a higher-order Markov chain can be transformed to a first-order Markov chain through an appropriate lifting of the state. From this perspective, limiting attention to first-order Markov chains is hence not a real limitation.

The data process $\hat{\xi}$ is formally said to be a first-order Markov chain with transition matrix $\mathbb{Q}^\star(\cdot|\cdot)$ if

$$\mathbb{Q}_n^\star(\hat{\xi}_1 = \xi_1, \ldots, \hat{\xi}_n = \xi_n) = \mathbb{Q}^\star(\sigma|\xi_1) \prod_{i=1}^{n-1} \mathbb{Q}^\star(\xi_i|\xi_{i+1}), \quad \forall n \in \mathrm{N}. \tag{5.4}$$

The marginal distributions of a Markov chain data process are closely related to its stationary distributions which we define now.

Definition 5.2 (Stationary distribution)**.** *A distribution $\mathbb{Q} \in \mathcal{P}_\Xi$ is denoted as a stationary distribution of a Markov chain with transition matrix $\mathbb{Q}(\cdot|\cdot)$ if it holds that $\sum_{i\in\Xi} \mathbb{Q}(i)\mathbb{Q}(i|j) = \mathbb{Q}(j)$ for all $j \in \Xi$.*

From the definition of a first-order Markov process it is quite clear that a necessary condition for having a unique stationary distribution is that the Markov chain is irreducible. That is, all states Ξ can be reached regardless of the initial state σ.

Definition 5.3 (Irreducible Markov chain)**.** *A Markov chain with transition matrix $\mathbb{Q}(\cdot|\cdot)$ is irreducible if there exists for all $(i,j) \in \Xi \times \Xi$ a $k \in \mathrm{N}$ such that the k step transition matrix $\mathbb{Q}^k(i|j) > 0$. There is thus a non-zero probability to go from any state i to any other state j. Slightly abusing notation, we will write $\mathbb{Q}(\cdot|\cdot) > 0$ when dealing with an irreducible transition matrix.*

In what follows, we shall therefore assume that the Markov chain producing the data is irreducible. By the Perron-Frobenius Theorem found for instance in Dembo and Zeitouni [4, Theorem 3.1.1], an irreducible finite state Markov chain has an unique stationary distribution independent of the initial condition σ. An irreducible Markov process makes that the produced data is identically distributed in the limit for large n. It should be remarked however that the stationarity Assumption 2.2 is violated as $\hat{\xi}_n \sim \mathbb{Q}^\star$ only holds in the limit for n tending to infinity. Nevertheless, the irreducibility assumption will make that none of the results stated in this paper depend on the initial state σ. Consequently, we have no need to discuss the initial state σ any further.

We will denote with $\mathcal{Q}_n^{\mathrm{im}}$ the set of data distributions satisfying the first-order Markov property (5.4) for an irreducible Markov chain. The set of unknown underlying distributions $\mathcal{Q}_n^{\mathrm{im}}$ corresponding to irreducible first-order Markov chains can be defined as

$$\mathcal{Q}_n^{\mathrm{im}} := \left\{ \mathbb{Q}_n := \mathbb{Q}(\sigma|\cdot)(\mathbb{Q}(\cdot|\cdot))^{n-1} \ : \ \mathbb{Q}(\cdot,\cdot) \in \mathcal{M}_\Xi \right\} \tag{5.5}$$

where the relationship between the doublet distribution $\mathbb{Q}(\cdot,\cdot)$ and the corresponding transition kernel $\mathbb{Q}(\cdot|\cdot)$ is discussed in the notation section. Again, from previous characterization of the ambiguity we face in the data it is quite clear that figuring out which distribution does underly the data boils down to working out its corresponding unknown doublet distribution $\mathbb{Q}(\cdot,\cdot)$. Define the empirical doublet distribution as

$$\hat{\mathbb{Q}}_n(\cdot,\cdot) := \tfrac{1}{n}(\textstyle\sum_{k=1}^{n-1} \delta_{[\hat{\xi}_k, \hat{\xi}_{k+1}]} + \delta_{[\hat{\xi}_n, \hat{\xi}_1]}).$$

Notice that the ghost last transition in the definition of the empirical doublet distribution $\hat{\mathbb{Q}}_n(\cdot,\cdot)$ makes the empirical doublet distribution consistent with the empirical distribution $\hat{\mathbb{Q}}_n$ defined earlier. That is, we have the relationship $\hat{\mathbb{Q}}_n(j) = \sum_{i\in\Xi} \hat{\mathbb{Q}}_n(i,j) = \sum_{i\in\Xi} \hat{\mathbb{Q}}_n(j,i)$.

6 The value of data

References

[1] L.E. Baum and M. Katz. Convergence rates in the law of large numbers. *Transactions of the American Mathematical Society*, 120(1):108–123, 1965.

[2] D. Bertsimas, V. Gupta, and N. Kallus. Robust SAA. *arXiv preprint arXiv:1408.4445*, 2014.

[3] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

[4] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 38. Springer Science & Business Media, 2009.

[5] P.M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

[6] Z. Hu and L.J. Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available on optimization online*, 2012.

[7] S. Kullback and R.A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86, 1951.

[8] H. Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 2013.

[9] O.P. Le Maître and O.M. Knio. *Introduction: Uncertainty Quantification and Propagation*. Springer, 2010.

[10] B.G. Lindsay and B. Li. On second-order optimality of the observed Fisher information. *The Annals of Statistics*, 25(5):2172–2199, 1997.

[11] A. Prekopa. On probabilistic constrained programming. In *Proceedings of the Princeton Symposium on Mathematical Programming*, pages 113–138. Princeton University Press, 1970.

[12] M.R. Ross. *Introduction to Probability Models*. Elsevier, 10 edition, 2010.

[13] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2 edition, 2014.

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

[15] B.P.G. Van Parys. *Distributionally robust control and optimization*. PhD thesis, ETH Zürich, August 2015. Accepted on the recommendation of Prof. M. Morari, Prof. D. Bertsimas and Prof. E. Delage.

[16] V.N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[17] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013.