# Robust Generalization despite Distribution Shift via Minimum Discriminating Information

**Tobias Sutter**
University of Konstanz, Germany
tobias.sutter@uni-konstanz.de

**Andreas Krause**
ETH Zurich, Switzerland
krausea@ethz.ch

**Daniel Kuhn**
EPFL, Switzerland
daniel.kuhn@epfl.ch

## Abstract

Training models that perform well under distribution shifts is a central challenge in machine learning. In this paper, we introduce a modeling framework where, in addition to training data, we have partial structural knowledge of the shifted test distribution. We employ the principle of *minimum discriminating information* to embed the available prior knowledge, and use *distributionally robust optimization* to account for uncertainty due to the limited samples. By leveraging large deviation results, we obtain explicit generalization bounds with respect to the unknown shifted distribution. Lastly, we demonstrate the versatility of our framework by demonstrating it on two rather distinct applications: (1) training classifiers on systematically biased data and (2) off-policy evaluation in Markov Decision Processes.

## 1 Introduction

Developing machine learning-based systems for real world applications is challenging, particularly because the conditions under which the system was trained are rarely the same as when using the system. Unfortunately, a standard assumption in most machine learning methods is that test and training distribution are the *same* [78, 59, 12]. This assumption, however, rarely holds in practice, and the performance of many models suffers in light of this issue, often called *dataset shift* [52] or equivalently *distribution shift*. Consider building a model for diagnosing a specific heart disease, and suppose that most participants of the study are middle to high-aged men. Further suppose these participants have a higher risk for the specific disease, and as such do not reflect the general population with respect to age and gender. Consequently, the training data suffers from the so-called *sample selection bias* inducing a *covariate shift* [62, 52]. Many other reasons lead to distribution shifts, such as non-stationary environments [67], imbalanced data [52], domain shifts [3], label shifts [83] or observed contextual information [8, 9]. A specific type of distribution shift takes center stage in off-policy evaluation (OPE) problems. Here, one is concerned with the task of estimating the resulting cost of an *evaluation policy* for a sequential decision making problem based on historical data obtained from a different policy known as *behavioral policy* [73]. This problem is of critical importance in various applications of reinforcement learning—particularly, when it is impossible or unethical to evaluate the resulting cost of an evaluation policy by running it on the underlying system. Solving a learning problem facing an arbitrary and unknown distribution shift based on training data in general is hopeless. Oftentimes, fortunately, partial knowledge about the distribution shift is available. In the medical example above, we might have prior information how the demographic attributes in our sample differ from the general population. Given a training distribution and partial knowledge about the shifted test distribution, one might ask what is the “most natural" distribution

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

shift mapping the training distribution into a test distribution consistent with the available structural information. Here, we address this question, interpreting "most natural" as maximizing the underlying Shannon entropy. This concept has attracted significant interest in the past in its general form, called *principle of minimum discriminating information* dating back to Kullback [37], which can be seen as a generalization of Jaynes' *maximum entropy principle* [31]. While these principles are widely used in tasks ranging from economics [27] to systems biology [63] and regularized Markov decision processes [48, 25, 2], they have not been investigated to model general distribution shifts as we consider in this paper.

Irrespective of the underlying distribution shift, the training distribution of any learning problem is rarely known, and one typically just has access to finitely many training samples. It is well-known that models can display a poor out-of-sample performance if training data is sparse. These overfitting effects are commonly avoided via regularization [12]. A regularization technique that has become popular in machine learning during the last decade and provably avoids overfitting is *distributionally robust optimization (DRO)* [36].

**Contributions.** We highlight the following main contributions of this paper:

- We introduce a *new modelling framework* for distribution shifts via the *principle of minimum discriminating information*, which encodes prior structural information on the resulting test distribution.
- Using our framework and the available training samples, we provide *generalization bounds* via a DRO program and prove that the introduced DRO model is *optimal* in a precise statistical sense.
- We show that the optimization problems characterizing the distribution shift and the DRO program can be *efficiently solved* by exploiting convex duality and recent accelerated first order methods.
- We demonstrate the *versatility* of the proposed *Minimum Discriminating based DRO* (MDI-DRO) method on two distinct problem classes: Training classifiers on systematically biased data and the OPE for Markov decision processes. In both problems MDI-DRO outperforms existing approaches.

## 2 Related work

For supervised learning problems, there is a rich literature in the context of covariate shift adaptation [62, 68]. A common approach is to address this distribution shift via importance sampling, more precisely by weighting the training loss with the ratio of the test and training densities and then minimize the so-called importance weighted risk (IWERM), see [62, 82, 68, 69]. While this importance weighted empirical risk is an unbiased estimator of the test risk, the method has two major limitations: It tends to produce an estimator with high variance, making the resulting test risk large. Further, the ratio of the training and test densities must be estimated which in general is difficult as the test distribution is unknown. There are modifications of IWERM reducing the resulting variance [15, 13, 65], for example by exponentially flattening the importance ratios [62]. For the estimation of the importance weights several methods have been presented, see for example [81]. These methods, however crucially rely on having data from both training and test distribution. Liu and Ziebart [40] and Chen et al. [14] propose a minimax approach for regression problems under covariate shift. Similar to our approach taken in this paper, they consider a DRO framework, which however, optimizes over so-called moment-based ambiguity sets. Distribution shifts play a key role in causal inference. In particular, the connection between causal predictors and distributional robustness under shifts arising from interventions has been widely studied [58, 42, 56, 66]. Oftentimes, a causal graph is used to represent knowledge about the underlying distribution shift induced by an intervention [49, 50]. Distribution shifts have been addressed in a variety of different settings [35], we refer the reader to the comprehensive textbook [52] and references therein.

There is a vast literature on OPE methods which we will not attempt to summarize. In a nutshell, OPE methods can be grouped into three classes: a first class of approaches that aims to fit a model from the available data and uses this model then to estimate the performance of the given evaluation policy [41, 1, 38]. A second class of methods are based on invoking the idea of importance sampling to model the underlying distribution shift from behavioral to evaluation policy [51, 29, 74]. The third, more recent, class of methods combines the first two classes [24, 32, 76, 33].

Key reasons for the popularity of DRO in machine learning are the ability of DRO models to regularize learning problems [36, 60, 61] and the fact that the underlying optimization problems can often be exactly reformulated as finite convex programs solvable in polynomial time [4, 10]. Such reformulations hold for a variety of ambiguity sets such as: regions defined by moments [20, 26, 80, 11],

$\phi$-divergences [5, 44, 39], Wasserstein ambiguity sets [43, 36], or maximum mean discrepancy ambiguity sets [64, 34]. DRO naturally seems a convenient tool when analyzing "small" distribution shifts as it seeks models that perform well "sufficiently close" to the training sample. However, modelling a general distribution shift via DRO seems difficult, and recent interest has focused on special cases such as adversarial example shifts [23] or label shifts [83]. To the best of our knowledge, the idea of combining DRO with the principle of minimum discriminating information is new.

## 3 Problem statement and motivating examples

We study learning problems of the form

$$\min_{\theta\in\Theta} R(\theta, \mathbb{P}^\star), \tag{1}$$

where $R(\theta, \mathbb{P}^\star) = \mathbb{E}_{\mathbb{P}^\star}[L(\theta, \xi)]$ denotes the risk of an uncertain real-valued loss function $L(\theta, \xi)$ that depends on a parameter $\theta \in \Theta \subset \mathbb{R}^n$ to be estimated as well as a random vector $\xi \in \Xi \subset \mathbb{R}^m$ governed by the probability distribution $\mathbb{P}^\star$. In order to avoid technicalities, we assume from now on that $\Theta$ and $\Xi$ are compact and $L$ is continuous. In statistical learning, it is usually assumed that $\mathbb{P}^\star$ is unknown but that we have access to independent samples from $\mathbb{P}^\star$. This paper departs from this standard scenario by assuming that there is a distribution shift. We first state our formal assumption about the shift and provide concrete examples below. Specifically, we assume to have access to samples from a distribution $\mathbb{P} \neq \mathbb{P}^\star$ and that $\mathbb{P}^\star$ is only known to belong to the distribution family

$$\Pi = \{\mathbb{Q} \in \mathcal{P}(\Xi) \ : \ \mathbb{E}_{\mathbb{Q}}[\psi(\xi)] \in E\} \tag{2}$$

encoded by a measurable feature map $\psi : \Xi \to \mathbb{R}^d$ and a compact convex set $E \subset \mathbb{R}^d$. In view of the principle of minimum discriminating information, we identify $\mathbb{P}^\star$ with the I-projection of $\mathbb{P}$ onto $\Pi$.

**Definition 3.1** (Information projection)**.** *The I-projection of $\mathbb{P} \in \mathcal{P}(\Xi)$ onto $\Pi$ is defined as*

$$\mathbb{P}^f = f(\mathbb{P}) = \arg\min_{\mathbb{Q}\in\Pi} \mathsf{D}(\mathbb{Q}\|\mathbb{P}), \tag{3}$$

*where $\mathsf{D}(\mathbb{Q}\|\mathbb{P})$ denotes the relative entropy of $\mathbb{Q}$ with respect to $\mathbb{P}$.*

One can show that the I-projection exists whenever $\Pi$ is closed with respect to the topology induced by the total variation distance [17, Theorem 2.1]. As $E$ is closed, this is the case whenever $\psi$ is bounded. Note that $f(\mathbb{P}) = \mathbb{P}$ if $\mathbb{P} \in \Pi$. In the remainder, we assume that $\mathbb{P} \notin \Pi$ and that $\mathbb{P}$ is only indirectly observable through independent training samples $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ drawn from $\mathbb{P}$.

**Example 3.2** (Logistic regression)**.** *Assume that $\xi = (x, y)$, where $x \in \mathbb{R}^{m-1}$ is a feature vector of patient data (e.g., a patient's age, sex, chest pain type, blood pressure, etc.), and $y \in \{-1, 1\}$ a label indicating the occurrence of a heart disease. Logistic regression models the conditional distribution of $y$ given $x$ by a logistic function* $\text{Prob}(y|x) = [1 + \exp(-y \cdot \theta^\top x)]^{-1}$ *parametrized by $\theta \in \mathbb{R}^{m-1}$. The maximum likelihood estimator for $\theta$ is found by minimizing the empirical average of the logistic loss function $L(\theta, \xi) = \log(1 + \exp(-y \cdot \theta^\top x))$ on the training samples. If the samples pertain to a patient cohort, where elderly males are overrepresented with respect to the general population, then they are drawn from a training distribution $\mathbb{P}$ that differs from the test distribution $\mathbb{P}^\star$. Even if sampling from $\mathbb{P}^\star$ is impossible, we may know that the expected age of a random individual in the population falls between 40 and 45 years. This information can be modeled as $\mathbb{E}_{\mathbb{P}^\star}[\psi(\xi)] \in E$, where $E = [\ell, u]$, $\ell = 40$, $u = 45$ and $\psi(\xi)$ projects $\xi$ to its 'age'-component. Other available prior information can be encoded similarly. Inspired by the principle of minimum discriminating information, we then minimize the expected log-loss under the I-projection $\mathbb{P}^f$ of the data-generating distribution $\mathbb{P}$ onto the set $\Pi$ defined in* (2)*.*

**Example 3.3** (Production planning)**.** *Assume that $\theta \in \mathbb{R}$ and $\xi \in \mathbb{R}$ denote the production quantity and the demand of a perishable good, respectively, and that the loss function $L(\theta, \xi)$ represents the sum of the production cost and a penalty for unsatisfied demand. To find the optimal production quantity, one could minimize the average loss in view of training samples drawn from the historical demand distribution $\mathbb{P}$. However, a disruptive event such as the beginning of a recession might signal that demand will decline by at least $\eta$%. The future demand distribution $\mathbb{P}^\star$ thus differs from $\mathbb{P}$ and belongs to a set $\Pi$ of the form* (2) *defined through $\psi(\xi) = \xi$ and $E = [0, (1 - \eta)\mu]$, where $\mu$ denotes the historical average demand. By the principle of minimum discriminating information it then makes again sense to minimize the expected loss under the I-projection $\mathbb{P}^f$ of $\mathbb{P}$ onto $\Pi$.*

Loosely speaking, the principle of minimum discriminating information identifies the I-projection $\mathbb{P}^f$ of $\mathbb{P}$ as the least prejudiced and thus most natural model for $\mathbb{P}^\star$ in view of the information that $\mathbb{P}^\star \in \Pi$. The principle of minimum discriminating information is formally justified by the conditional limit theorem [18], which we paraphrase below using our notation.

**Proposition 3.4** (Conditional limit theorem)**.** *If the interior of the compact convex set $E$ overlaps with the support of the pushforward measure $\mathbb{P} \circ \psi^{-1}$, the I-projection $\mathbb{P}^f = f(\mathbb{P})$ exists and the moment-generating function $\mathbb{E}_{\mathbb{P}^f}[e^{tL(\theta,\xi)}]$ is finite for all $t$ in a neighborhood of $0$, then we have*

$$\lim_{N\to\infty} \mathbb{E}_{\mathbb{P}^N}[L(\theta,\xi_1)|\tfrac{1}{N}\textstyle\sum_{i=1}^N \psi(\xi_i) \in E] = \mathbb{E}_{\mathbb{P}^f}[L(\theta,\xi)] \quad \forall\theta \in \Theta.$$

In the context of Examples 3.2 and 3.3, the conditional limit theorem provides an intuitive justification for modeling distribution shifts via I-projections. More generally, the following proposition suggests that *any* distribution shift can be explained as an I-projection onto a suitably chosen set $\Pi$.

**Proposition 3.5** (Every distribution is an I-projection)**.** *If $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\Xi)$ are such that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$ and if $\Pi$ is a set of the form* (2) *defined through $\psi(\xi) = \log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\xi)$ and $E = \{\mathsf{D}(\mathbb{Q}\|\mathbb{P})\}$, then $\mathbb{Q} = f(\mathbb{P})$.*

The modelling of arbitrary distribution shifts via the I-projection according to Proposition 3.5 has an interesting application in the off-policy evaluation problem for Markov decision processes (MDPs).

**Example 3.6** (Off-policy evaluation)**.** *Consider an MDP $(\mathcal{S}, \mathcal{A}, Q, c, s_0)$ with finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$, respectively, transition kernel $Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, cost-per-stage function $c : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and initial state $s_0$. A stationary Markov policy $\pi$ is a stochastic kernel that maps states to probability distributions over $\mathcal{A}$. We use $\pi(a|s)$ to denote the probability of selecting action $a$ in state $s$ under policy $\pi$. The long-run average cost generated by $\pi$ can be expressed as*

$$V_\pi = \lim_{T\to\infty} \tfrac{1}{T} \textstyle\sum_{t=0}^{T-1} \mathbb{E}_{s_0}^\pi[c(s_t, a_t)].$$

*Each policy induces an occupation measure $\mu_\pi$ on $\mathcal{S}\times\mathcal{A}$ defined through the state-action frequencies*

$$\mu_\pi(x, a) = \lim_{T\to\infty} \tfrac{1}{T} \textstyle\sum_{t=0}^{T-1} \mathbb{P}_{s_0}^\pi[(s_t, a_t) = (s, a)] \quad \forall s \in \mathcal{S},\ a \in \mathcal{A},$$

*see [28, Chapter 6]. One can additionally show that $\mu_\pi$ belongs to the polytope*

$$\mathcal{M} = \big\{\mu \in \Delta_{\mathcal{S}\times\mathcal{A}} : \textstyle\sum_{a'\in\mathcal{A}} \mu(s', a') - \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q(s'|s, a)\mu(s, a) = 0\ \forall s' \in \mathcal{S}\big\},$$

*where $\Delta_{\mathcal{S}\times\mathcal{A}}$ represents the simplex of all probability mass functions over $\mathcal{S}\times\mathcal{A}$. Conversely, each occupation measure $\mu \in \mathcal{M}$ induces a policy $\pi_\mu$ defined through $\pi_\mu(a|s) = \mu(s, a)/\sum_{a'\in\mathcal{A}} \mu(s, a')$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Assuming that all parameters of the MDP except for the cost $c$ are known, the off-policy evaluation problem asks for an estimate of the long-run average cost $V_{\pi_\mathsf{e}}$ of an evaluation policy $\pi_\mathsf{e}$ based on a trajectory of states, actions and costs generated by a behavioral policy $\pi_\mathsf{b}$. This task can be interpreted as a degenerate learning problem without a parameter $\theta$ to optimize if we define $\xi = c(s, a)$ and set $L(\theta, \xi) = \xi$. Here, a distribution shift emerges because we must evaluate the expectation of $\xi$ under $\mathbb{Q} = \mu_\mathsf{e} \circ c^{-1}$ given training samples from $\mathbb{P} = \mu_\mathsf{b} \circ c^{-1}$, where $\mu_\mathsf{b}$ and $\mu_\mathsf{e}$ represent the occupation measures corresponding to $\pi_\mathsf{b}$ and $\pi_\mathsf{e}$, respectively. Note that $\mathbb{P}$ and $\mathbb{Q}$ are unknown because $c$ is unknown. Moreover, as the policy $\pi_\mathsf{e}$ generates different state-action trajectories than $\pi_\mathsf{b}$, the costs generated under $\pi_\mathsf{e}$ cannot be inferred from the costs generated under $\pi_\mathsf{b}$ even though $\pi_\mathsf{b}$ and $\pi_\mathsf{e}$ are known. Note also that $\mathbb{Q}$ coincides with the I-projection $\mathbb{P}^f$ of $\mathbb{P}$ onto the set $\Pi$ defined in Proposition 3.5. The corresponding feature map $\psi$ as well as the set $E$ can be computed without knowledge of $c$ provided that $c$ is invertible. Indeed, in this case we have*

$$\psi(\xi_i) = \log \tfrac{\mathrm{d}\mu_\mathsf{e}\circ c^{-1}}{\mathrm{d}\mu_\mathsf{b}\circ c^{-1}}(\xi_i) = \log \tfrac{\mu_\mathsf{e}(s_i,a_i)}{\mu_\mathsf{b}(s_i,a_i)} \quad \textit{and} \quad E = \big\{\mathsf{D}(\mu_\mathsf{e} \circ c^{-1} \| \mu_\mathsf{b} \circ c^{-1})\big\}$$

*for any $s_i \in \mathcal{S}$, $a_i \in \mathcal{A}$ and $\xi_i = c(s_i, a_i)$. Note that as $\mathcal{S}$ and $\mathcal{A}$ are finite, $c$ is generically invertible, that is, $c$ can always be rendered invertible by an arbitrarily small perturbation. In summary, we may conclude that the off-policy evaluation problem reduces to an instance of* (1).

Given $N$ training samples $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$, we henceforth use $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_i}$ and $\widehat{\mathbb{P}}_N^f$ to denote the empirical distribution on and its I-projection onto $\Pi$, respectively. As the true data-generating distribution $\mathbb{P}$ and its I-projection $\mathbb{P}^f$ are unknown, it makes sense to replace them by their empirical counterparts. However, the resulting empirical risk minimization problem is susceptible to overfitting if the number of training samples is small relative to the feature dimension. In order to combat overfitting, we propose to solve the DRO problem

$$J_N^\star = \min_{\theta\in\Theta} \ R^\star(\theta, \widehat{\mathbb{P}}_N^f), \tag{4}$$

which minimizes the worst-case risk over all distributions close to $\widehat{\mathbb{P}}_N^f$. Here, $R^\star$ is defined through

$$R^\star(\theta, \mathbb{P}') = \sup_{\mathbb{Q}\in\Pi} \{R(\theta, \mathbb{Q}) : \mathsf{D}(\mathbb{P}' \| \mathbb{Q}) \le r\} \tag{5}$$

and thus evaluates the worst-case risk of a given parameter $\theta \in \Theta$ in view of all distributions $\mathbb{Q}$ that have a relative entropy distance of at most $r$ from a given nominal distribution $\mathbb{P}' \in \Pi$. In the remainder we use $J_N^\star$ and $\theta_N^\star$ to denote the minimum and a minimizer of problem (4), respectively.

**Main results.** The main theoretical results of this paper can be summarized as follows.

(i) *Out-of-sample guarantee.* We show that the optimal value of the DRO problem (4) provides an upper confidence bound on the risk of its optimal solution $\theta_N^\star$. Specifically, we prove that

$$\mathbb{P}^N \left(R(\theta_N^\star, \mathbb{P}^f) > J_N^\star\right) \le e^{-rN+o(N)}, \tag{6}$$

where $\mathbb{P}^f = f(\mathbb{P})$ is the I-projection of $\mathbb{P}$. If $\Xi$ is finite, then (6) can be strengthened to a finite sample bound that holds for every $N$ if the right hand side is replaced with $e^{-rN}(N+1)^{|\Xi|}$.

(ii) *Statistical efficiency.* In a sense to be made precise below, the DRO problem (4) provides the least conservative approximation for (1) whose solution satisfies the out-of-sample guarantee (6).

(iii) *Computational tractability.* We prove that the I-projection $\widehat{\mathbb{P}}_N^f$ can be computed via a regularized fast gradient method whenever one can efficiently project onto $E$. Given $\widehat{\mathbb{P}}_N^f$, we then show that $\theta_N^\star$ can be found by solving a tractable convex program whenever $\Theta$ is a convex and conic representable set, while $L(\theta, \xi)$ is a convex and conic representable function of $\theta$ for any fixed $\xi$.

## 4 Statistical guarantees

Throughout this section, we equip $\mathcal{P}(\Xi)$ with the topology of weak convergence. As $L(\theta, \xi)$ is continuous on $\Theta \times \Xi$ and $\Xi$ is compact, this implies that the risk $R(\theta, \mathbb{Q})$ is continuous on $\Theta \times \mathcal{P}(\Xi)$. The DRO problem (4) is constructed from the I-projection of the empirical distribution, which, in turn, is constructed from the given training samples. Thus, $\theta_N^\star$ constitutes a data-driven decision. Other data-driven decisions can be obtained by solving surrogate optimization problems of the form

$$\widehat{J}_N = \min_{\theta\in\Theta} \ \widehat{R}(\theta, \widehat{\mathbb{P}}_N^f), \tag{7}$$

where $\widehat{R} : \Theta \times \Pi \to \mathbb{R}$ is a continuous function that uses the empirical I-projection $\widehat{\mathbb{P}}_N^f$ to predict the true risk $R(\theta, \mathbb{P}^f)$ of $\theta$ under the true I-projection $\mathbb{P}^f$. From now on we thus refer to $\widehat{R}$ as a predictor, and we use $\widehat{J}_N$ and $\widehat{\theta}_N$ to denote the minimum and a minimizer of problem (7), respectively. We call a predictor $\widehat{R}$ *admissible* if $\widehat{J}_N$ provides an upper confidence bound on the risk of $\widehat{\theta}_N$ in the sense that

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^N \left(R(\widehat{\theta}_N, \mathbb{P}^f) > \widehat{J}_N\right) \le -r \quad \forall \mathbb{P} \in \mathcal{P}(\Xi) \tag{8}$$

for some prescribed $r > 0$. The inequality (8) requires the true risk of the minimizer $\widehat{\theta}_N$ to exceed the optimal value $\widehat{J}_N$ of the surrogate optimization problem (7) with a probability that decays exponentially at rate $r$ as the number $N$ of training samples tends to infinity. The following theorem asserts that the DRO predictor $R^\star$ defined in (5), which evaluates the worst-case risk of any given $\theta$ across a relative entropy ball of radius $r$, almost satisfies (8) and is thus essentially admissible.

**Theorem 4.1** (Out-of-sample guarantee)**.** *If $r > 0$, $0 \in \mathrm{int}(E)$ and for every $z \in \mathbb{R}^d$ there exists an uncertainty realization $\xi \in \Xi$ such that $z^\top \psi(\xi) > 0$, then the DRO predictor $R^\star$ defined in* (5) *is continuous on $\Theta \times \Pi$. In addition, $\widehat{R} = R^\star + \varepsilon$ is an admissible data-driven predictor for every $\varepsilon > 0$.*

Theorem 4.1 implies that, for any fixed $\varepsilon > 0$, the DRO predictor $R^\star$ provides an upper confidence bound $J_N^\star + \varepsilon$ on the true risk $R(\theta_N^\star, \mathbb{P}^f)$ of the data-driven decision $\theta_N^\star$ that becomes increasingly reliable as $N$ grows. Of course, the reliability of *any* upper confidence bound trivially improves if it is increased. Finding *some* upper confidence bound is thus easy. The next theorem shows that the DRO predictor actually provides the *best possible* (asymptotically smallest) upper confidence bound.

**Theorem 4.2** (Statistical efficiency)**.** *Assume that all conditions of Theorem 4.1 hold. If $J_N^\star$ and $R^\star$ are defined as in* (4) *and* (5)*, while $\widehat{J}_N$ is defined as in* (7) *for any admissible data-driven predictor $\widehat{R}$, then we have $\lim_{N\to\infty} J_N^\star \leq \lim_{N\to\infty} \widehat{J}_N$ $\mathbb{P}^\infty$-almost surely irrespective of $\mathbb{P} \in \mathcal{P}(\Xi)$.*

One readily verifies that the limits in Theorem 4.2 exist. Indeed, if $\widehat{R}$ is an arbitrary data-driven predictor, then the optimal value $\widehat{J}_N$ of the corresponding surrogate optimization problem converges $\mathbb{P}$-almost surely to $\min_{\theta\in\Theta} \widehat{R}(\theta, \mathbb{P}^f)$ as $N$ tends infinity provided that the training samples are drawn independently from $\mathbb{P}$. This is a direct consequence of the following three observations. First, the optimal value function $\min_{\theta\in\Theta} \widehat{R}(\theta, \mathbb{P}^f)$ is continuous in $\mathbb{P}^f \in \Pi$ thanks to Berge's maximum theorem [7, pp. 115–116], which applies because $\widehat{R}$ is continuous and $\Theta$ is compact. Second, the I-projection $\mathbb{P}^f = f(\mathbb{P})$ is continuous in $\mathbb{P} \in \mathcal{P}(\Xi)$ thanks to [70, Theorem 9.17], which applies because the relative entropy is strictly convex in its first argument [21, Lemma 6.2.12]. Third, the strong law of large numbers implies that the empirical distribution $\widehat{\mathbb{P}}_N$ converges weakly to the data-generating distribution $\mathbb{P}$ as the sample size $N$ grows. Therefore, we have

$$\lim_{N\to\infty} \widehat{J}_N = \lim_{N\to\infty} \min_{\theta\in\Theta} \widehat{R}\left(\theta, f(\widehat{\mathbb{P}}_N)\right) = \min_{\theta\in\Theta} \widehat{R}\left(\theta, f\left(\lim_{N\to\infty} \widehat{\mathbb{P}}_N\right)\right) = \min_{\theta\in\Theta} \widehat{R}(\theta, \mathbb{P}^f) \quad \mathbb{P}\text{-a.s.}$$

In summary, Theorems 4.1 and 4.2 assert that the DRO predictor $R^\star$ is (essentially) admissible and that it is the least conservative of all admissible data-driven predictors, respectively. Put differently, the DRO predictor makes the most efficient use of the available data among all data-driven predictors that offer the same out-of-sample guarantee (8). In the special case when $\Xi$ is finite, the asymptotic out-of-sample guarantee (8) can be strengthened to a finite sample guarantee that holds for every $N \in \mathbb{N}$.

**Corollary 4.3** (Finite sample guarantee)**.** *If $R^\star$ is defined as in* (5)*, then*

$$\frac{1}{N} \log \mathbb{P}^N \left(R^\star(\theta_N^\star, \mathbb{P}^f) > J_N^\star\right) \leq \frac{\log(N+1)}{N}|\Xi| - r \quad \forall N \in \mathbb{N}. \tag{9}$$

We now temporarily use $R_r^\star$ to denote the DRO predictor defined in (5), which makes its dependence on $r$ explicit. Note that if $r > 0$ is kept constant, then $R_r^\star(\theta, \widehat{\mathbb{P}}_N^f)$ is neither an unbiased nor a consistent estimator for $R(\theta, \mathbb{P}^f)$. Consistency can be enforced, however, by shrinking $r$ as $N$ grows.

**Theorem 4.4** (Asymptotic consistency)**.** *Let the assumptions of Proposition 3.4 hold and $\{r_N\}_{N\in\mathbb{N}}$ be a sequence of non-negative reals with $\lim_{N\to\infty} r_N = 0$. If the loss function $L(\theta, \xi)$ is Lipschitz continuous in $\xi$ with Lipschitz constant $\Lambda > 0$ uniformly across all $\theta \in \Theta$, then we have*

$$\lim_{N\to\infty} R_{r_N}^\star(\theta, \widehat{\mathbb{P}}_N^f) = R(\theta, \mathbb{P}^f) \quad \mathbb{P}^\infty\text{-a.s. } \forall \theta \in \Theta, \tag{10a}$$

$$\lim_{N\to\infty} \min_{\theta\in\Theta} R_{r_N}^\star(\theta, \widehat{\mathbb{P}}_N^f) = \min_{\theta\in\Theta} R(\theta, \mathbb{P}^f) \quad \mathbb{P}^\infty\text{-a.s.} \tag{10b}$$

**Remark 4.5** (Choice of radius)**.** *Theorem 4.2 shows that the ambiguity set used in our paper displays a strong Pareto-optimality property, i.e., it leads to the least conservative predictor, uniformly across all estimator realizations, for which the out-of-sample disappointment probability is guaranteed to decay exponentially at rate $r$. Therefore, the radius $r$ has a direct operational interpretation that captures the risk tolerance of the decision maker—it is chosen subjectively. Since the statistical guarantees of Theorem 4.1 are asymptotic, selecting the radius $r$ when we only have access to finitely many samples is challenging, and in practice $r$ is usually selected via cross validation.*

We now exemplify our DRO approach and its statistical guarantees in the context of the off-policy evaluation problem introduced in Section 3.

**Example 4.6** (Off-policy evaluation)**.** *Consider again the OPE problem introduced in Example 3.6. We now aim to construct an estimator for the performance of the evaluation policy $V_{\pi_\mathsf{e}} = \mathbb{E}_{f(\mathbb{P})}[\xi]$ based on the available behavioral policy and its empirical cost. As described in Example 3.6, we choose $\Pi$ such that $\mu_\mathsf{e} \circ c^{-1} = f(\mathbb{P})$, where $\mathbb{P} = \mu_\mathsf{b} \circ c^{-1} \in \mathcal{P}(\Xi)$. Given the behavioral data $(\widehat{s}_i, \widehat{a}_i) \sim \mu_\mathsf{b}$ for $i = 1, \ldots N$, we then construct the empirical distribution $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{c(\widehat{s}_i, \widehat{a}_i)}$. Our statistical results require the samples $(\widehat{s}_i, \widehat{a}_i)$ to be i.i.d., which can be enforced approximately by discarding a sufficient number of intermediate samples, for example. We emphasize, however, that the proposed large deviation framework readily generalizes to situations in which there is a single trajectory of correlated data [39, 72]. Details are omitted for brevity. The value function $V_{\pi_\mathsf{e}}$*

*under the evaluation policy can now be approximated by $J_N^\star = R^\star(\widehat{\mathbb{P}}_N^f)$, where $R^\star$ denotes the DRO predictor* (5). *As $\Xi$ is finite, Corollary 4.3 provides the generalization bound*

$$\mathbb{P}^N \left(V_{\pi_\mathsf{e}} \leq J_N^\star\right) \geq 1 - (N+1)^{|\mathcal{S}|+|\mathcal{A}|} e^{-rN} \quad \forall \mathbb{P} \in \mathcal{P}(\Xi), \tag{11}$$

*which holds for all $N \in \mathbb{N}$.*

# 5 Efficient computation

We now outline an efficient procedure to solve the DRO problem (4). This procedure consists of two steps. First, we propose an algorithm to compute the I-projection $\widehat{\mathbb{P}}_N^f = f(\widehat{\mathbb{P}}_N)$ of the empirical distribution $\widehat{\mathbb{P}}_N$ corresponding to the training samples $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$. Given $\widehat{\mathbb{P}}_N^f$, we then show how to compute the worst-case risk $R^\star(\theta, \widehat{\mathbb{P}}_N^f)$ and a corresponding optimizer $\theta_N^\star$ over the search space $\Theta$.

**Computation of the I-projection.** Computing the I-projection of the empirical distribution $\widehat{\mathbb{P}}_N$ is a non-trivial task because it requires solving the infinite-dimensional optimization problem (3). Generally, one would expect that the difficulty of evaluating $f(\widehat{\mathbb{P}}_N)$ depends on the structure of the set $\Pi$, which is encoded by $\psi$ and $E$; see (2). Thanks to the discrete nature of the empirical distribution $\widehat{\mathbb{P}}_N$, however, we can leverage recent advances in convex optimization together with an algorithm proposed in [71] to show that $f(\widehat{\mathbb{P}}_N)$ can be evaluated efficiently for a large class of sets $\Pi$.

In the following we let $\eta = (\eta_1, \eta_2)$ be a smoothing parameter with $\eta_1, \eta_2 > 0$, and we let $L_\eta > 0$ be a learning rate that may depend on $\eta$. In addition, we denote by $z \in \mathbb{R}^d$ the vector of dual variables of the constraint $\mathbb{E}_\mathbb{Q}[\psi(\xi)] \in E$ in problem (3), and we define $G_\eta : \mathbb{R}^d \to \mathbb{R}^d$ with

$$G_\eta(z) = -\pi_E(\eta_1^{-1} z) - \eta_2 z + \frac{\sum_{i=1}^N \psi(\widehat{\xi}_i) \exp\left(-\sum_{j=1}^d z_j \, \psi_j(\widehat{\xi}_i)\right)}{\sum_{i=1}^N \exp\left(-\sum_{j=1}^d z_j \, \psi_j(\widehat{\xi}_i)\right)} \tag{12}$$

as a smoothed gradient of the dual objective, where $\pi_E$ denotes the projection operator onto $E$ defined through $\pi_E(z) = \arg\min_{x \in E} \|x - z\|_2^2$. The corresponding smoothed dual of the I-projection problem (3) can then be solved with the fast gradient method described in Algorithm 1. The complexity of evaluating $G_\eta$, and thus the per-iteration complexity of Algorithm 1, is determined by the projection operator onto $E$. For simple sets (e.g., 2-norm balls or hybercubes) the solution is available in closed form, and for many other sets (e.g., simplices or 1-norm balls) it can be computed cheaply, see [53, Section 5.4] for a comprehensive survey.

**Algorithm 1:** Fast gradient method for smooth & strongly convex optimization [47]

Choose $w_0 = z_0 \in \mathbb{R}^d$ and $\eta \in \mathbb{R}^2_{++}$

**For** $k \in \mathbb{N}$    **Step 1:** Set $z_{k+1} = w_k + \frac{1}{L_\eta} G_\eta(w_k)$

**Step 2:** Compute $w_{k+1} = z_{k+1} + \frac{\sqrt{L_\eta} - \sqrt{\eta_2}}{\sqrt{L_\eta} + \sqrt{\eta_2}}(z_{k+1} - z_k)$

Any output $z_k$ of Algorithm 1 after $k$ iterations can be used to construct a candidate solution

$$\widehat{\mathbb{Q}}_k = \frac{\sum_{j=1}^N \exp\left(-\sum_{i=1}^d (z_k)_i \psi_i(\widehat{\xi}_j)\right) \delta_{\widehat{\xi}_j}}{\sum_{j=1}^N \exp\left(-\sum_{i=1}^d (z_k)_i \, \psi_i(\widehat{\xi}_j)\right)} \tag{13}$$

for problem (3) that approximates the I-projection $\widehat{\mathbb{P}}_N^f$. The convergence guarantees for Algorithm 1 and, in particular, the approximation quality of (13) with respect to $\widehat{\mathbb{P}}_N^f$ detailed in Theorem 5.2 below require that problem (3) admits a Slater point $\mathbb{P}^\circ$ in the sense of the following assumption.

**Assumption 5.1** (Slater point)**.** *Problem* (3) *admits a Slater point $\mathbb{P}^\circ \in \Pi$ that satisfies*

$$\delta = \min_{y \notin E} \|\mathbb{E}_{\mathbb{P}^\circ}[\psi(\xi)] - y\|_2 > 0.$$

Finding a Slater point $\mathbb{P}^\circ$ may be difficult in general. However, $\mathbb{P}^\circ$ can be constructed systematically if $\psi$ is a polynomial [71, Remark 8], for example. Given $\mathbb{P}^\circ$ and a tolerance $\varepsilon > 0$, we then define

$$
\begin{aligned}
&C = \mathsf{D}(\mathbb{P}^\circ \| \widehat{\mathbb{P}}_N), \qquad D = \tfrac{1}{2}\max_{y\in E}\|y\|_2, \qquad \eta_1 = \tfrac{\varepsilon}{4D}, \qquad \eta_2 = \tfrac{\varepsilon\delta^2}{2C^2},\\
&\alpha = \max_{\xi\in\Xi}\|\psi(\xi)\|_\infty, \quad L_\eta = 1/\eta_1 + \eta_2 + (\max_{\xi\in\Xi}\|\psi(\xi)\|_\infty)^2,\\
&M_1(\varepsilon) = 2\left(\sqrt{\tfrac{8DC^2}{\varepsilon^2\delta^2} + \tfrac{2\alpha^2C^2}{\varepsilon\delta^2} + 1}\right)\log\left(\tfrac{10(\varepsilon+2C)}{\varepsilon}\right), \qquad (14)\\
&M_2(\varepsilon) = 2\left(\sqrt{\tfrac{8DC^2}{\varepsilon^2\delta^2} + \tfrac{2\alpha^2C^2}{\varepsilon\delta^2} + 1}\right)\log\left(\tfrac{C}{\varepsilon\delta(2-\sqrt{3})}\sqrt{4\left(\tfrac{4D}{\varepsilon} + \alpha^2 + \tfrac{\varepsilon\delta^2}{2C^2}\right)\left(C + \tfrac{\varepsilon}{2}\right)}\right).
\end{aligned}
$$

**Theorem 5.2** (Almost linear convergence rate)**.** *If Assumption 5.1 holds and $\varepsilon > 0$, then the candidate solution* (13) *obtained after $k = \lceil \max\{M_1(\varepsilon), M_2(\varepsilon)\}\rceil$ iterations of Algorithm 1 satisfies*

$$\textit{Optimality:} \qquad |\mathsf{D}(\widehat{\mathbb{Q}}_k\|\widehat{\mathbb{P}}_N) - \mathsf{D}(\widehat{\mathbb{P}}_N^f\|\widehat{\mathbb{P}}_N)| \le 2(1+2\sqrt{3})\varepsilon, \qquad (15a)$$

$$\textit{Feasibility:} \qquad \mathsf{d}\Big(\mathbb{E}_{\widehat{\mathbb{Q}}_k}[\psi(\xi)], E\Big) \le \tfrac{2\varepsilon\delta}{C}, \qquad (15b)$$

*where we use the definitions* (14)*, and the function $\mathsf{d}(\cdot, E)$ denotes the Euclidean distance to the set $E$ defined through $\mathsf{d}(x, E) = \min_{y\in E}\|x - y\|_2$.*

Theorem 5.2 implies that Algorithm 1 needs at most $O(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$ iterations to find an $O(\varepsilon)$-suboptimal and $O(\varepsilon)$-feasible solution for the I-projection problem (3). These results are derived via convex programming and duality by using the double smoothing techniques introduced in [22] and [71].

**Computation of the DRO predictor.** Equipped with Algorithm 1 to efficiently approximate $\widehat{\mathbb{P}}_N^f$ via $\widehat{\mathbb{Q}}_k$, the DRO predictor $R^\star(\theta, \widehat{\mathbb{P}}_N^f)$ defined in (4) can be approximated by $R^\star(\theta, \widehat{\mathbb{Q}}_k)$ because the function $R^\star$ is continuous. We now show that the worst-case risk evaluation problem (5) admits a dual representation, which generalizes [77, Proposition 5].

**Proposition 5.3** (Dual representation of $R^\star$)**.** *If $r > 0$, then the DRO predictor $R^\star$ satisfies*

$$R^\star(\theta, \mathbb{P}') = \left\{\begin{array}{cl} \inf\limits_{\alpha\in\mathbb{R}, z\in\mathbb{R}^d} & \alpha + \sigma_E(z) - e^{-r}\exp\left(\mathbb{E}_{\mathbb{P}'}[\log(\alpha - L(\theta,\xi) + z^\top\psi(\xi))]\right) \\ \text{s.t.} & \alpha \ge \max_{\xi\in\Xi} L(\theta,\xi) - z^\top\psi(\xi) \end{array}\right. \qquad (16)$$

*for ever $\theta\in\Theta$ and $\mathbb{P}'\in\Pi$, where $\sigma_E(z) = \max_{x\in E} x^\top z$ denotes the support function of $E$.*

Proposition 5.3 implies that if $L(\theta, \xi)$ is convex in $\theta$ for every $\xi$, then the DRO predictor (5) coincides with the optimal value of a finite-dimensional convex program. Note that the objective function of (16) can be evaluated cheaply whenever the support function of $E$ is easy to compute and $\mathbb{P}'$ has finite support (e.g., if $\mathbb{P}'$ is set to an output $\widehat{\mathbb{Q}}_k$ of Algorithm 1). In addition, the robust constraint in (16) can be expressed in terms of explicit convex constraints if $L$, $\Xi$ and $\psi$ satisfy certain regularity conditions. A trivial condition is that $\Xi$ is finite. More general conditions are described in [6].

## 6 Experimental results

We now assess the empirical performance of the MDI-DRO method in our two running examples.[1]

**Synthetic dataset — covariate shift adaptation.** The first two experiments revolve around the logistic regression problem with a distribution shift described in Example 3.2. Specifically, we consider a synthetic dataset where the test data is affected by a covariate shift, which constitutes a special case of a distribution shift. Detailed information about the data generation process is provided in Appendix 7.4. Our numerical experiments reveal that the proposed MDI-DRO method significantly outperforms the naive ERM method in the sense that its out-of-sample risk has both a lower mean as well as a lower variance; see Figures 1a and 1b. We also compare MDI-DRO against the IWERM method, which accounts for the distribution shift by assigning importance weights $p^\star(\cdot)/p(\cdot)$ to the training samples, where $p^\star(\cdot)$ and $p(\cdot)$ denote the densities of the test distribution $\mathbb{P}^\star$ and training distribution $\mathbb{P}$, respectively. These importance weights are assumed to be known in IWERM. In

[1] All simulations were implemented in MATLAB and run on a 4GHz CPU with 16Gb RAM. The Matlab code for reproducing the plots is available from `https://github.com/tobsutter/PMDI_DRO`.

contrast, MDI-DRO does *not* require any knowledge of the test distribution other than its membership in $\Pi$. Nevertheless, MDI-DRO displays a similar out-of-sample performance as IWERM even though it has less information about $\mathbb{P}^\star$, and it achieves a lower variance than IWERM; see Figures 1c-1d. Figure 1e shows how the reliability of the upper confidence bound $J_N^\star$ and the out-of-sample risk $R(\theta_N^\star, \mathbb{P}^\star)$ change with the regularization parameter $r$. Additional results are reported in Figure 4 in the appendix. These results confirm that small regularization parameters $r$ lead to small out-of-sample risk and that increasing $r$ improves the reliability of the upper confidence bound $J_N^\star$.

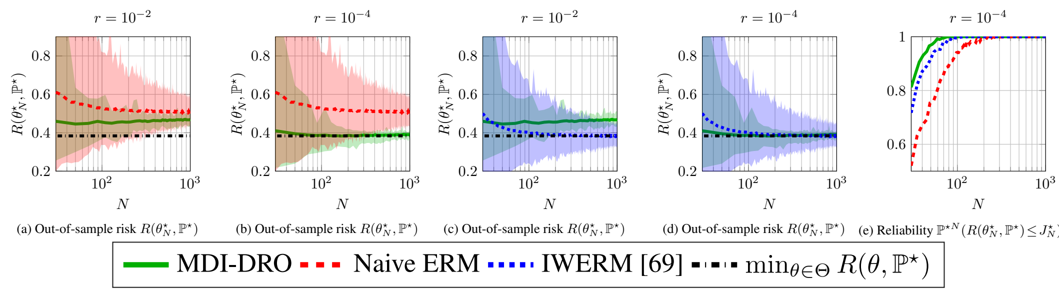


Figure 1: Results for a synthetic dataset with $m = 6$. Shaded areas and lines represent ranges and mean values across 1000 independent experiments, respectively.

**Real data — classification under sample bias.** The second experiment addresses the heart disease classification task of Example 3.2 based on a real dataset[2] consisting of $N^\star$ i.i.d. samples from an unknown test distribution $\mathbb{P}^\star$. To assess the effects of a distribution shift, we construct a biased training dataset $\{(\widehat{x}_1, \widehat{y}_1), \ldots, (\widehat{x}_N, \widehat{y}_N)\}$, $N < N^\star$, in which male patients older than 60 years are substantially over-represented. Specifically, the $N$ training samples are drawn randomly from the set of the 20% oldest male patients. Thus, the training data follows a distribution $\mathbb{P} \neq \mathbb{P}^\star$. Even though the test distribution $\mathbb{P}^\star$ is unknown, we assume to know the empirical mean $m = \frac{1}{N^\star} \sum_{i=1}^{N^\star} (\widehat{x}_i, \widehat{y}_i)$ of the entire dataset to within an absolute error $\Delta m > 0$. The test distribution thus belongs to the set $\Pi$ defined in (2) with $E = [m - \Delta m \mathbf{1}, m + \Delta m \mathbf{1}]$ and with $\psi(x, y) = (x, y)$. We compare the proposed MDI-DRO method for classification against the naive ERM method that ignores the sample bias. In addition, we use a logistic regression model trained on the entire dataset as an (unachievable) ideal benchmark. Figure 2a shows the out-of-sample cost, Figure 2b the upper confidence bound $J_N^\star$ and Figure 2c the misclassification rates of the different methods as the radius $r$ of the ambiguity set is swept. Perhaps surprisingly, for some values of $r$ the classification performance of MDI-DRO is comparable to that of the logistic regression method trained on the entire dataset.

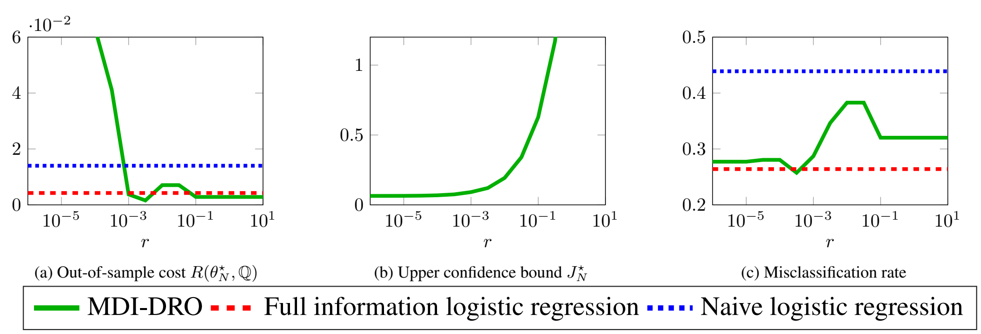


Figure 2: Heart disease classification example with $m = 6$, $N = 20$, $N^\star = 303$ and $\Delta m = 10^{-3}$.

**OPE for MDPs — inventory control.** We now consider the OPE problem of Examples 3.6 and 4.6. A popular estimator for the cost $V_{\pi_\mathsf{e}}$ of the evaluation policy is the inverse propensity score (IPS) [57]

$$\widehat{J}_N^{\mathrm{IPS}} = \frac{1}{N} \sum_{i=1}^{N} c(\widehat{s}_i, \widehat{a}_i) \frac{\mu_\mathsf{e}(\widehat{s}_i, \widehat{a}_i)}{\mu_\mathsf{b}(\widehat{s}_i, \widehat{a}_i)}.$$

Hoeffding's inequality then gives rise to the simple concentration bound

$$\mathbb{P}^N \left( V_{\pi_\mathsf{e}} \leq \widehat{J}_N^{\mathrm{IPS}} + \varepsilon \right) \geq 1 - e^{\frac{-2N\varepsilon^2}{b^2}} \quad \forall \varepsilon > 0,\ \forall N \in \mathbb{N}, \tag{17}$$

[2] `https://www.kaggle.com/ronitf/heart-disease-uci`

where $b = \max_{s\in\mathcal{S},a\in\mathcal{A}} c(s,a)\mu_{\mathsf{e}}(s,a)/\mu_{\mathsf{b}}(s,a)$. As $b$ is typically a large constant, the finite sample bound (11) for $J_N^\star$ is often more informative than (17). In addition, the variance of $\widehat{J}_N^{\mathrm{IPS}}$ grows exponentially with the sample size $N$ [15, 13, 65]. As a simple remedy, one can cap the importance weights beyond some threshold $\beta > 0$ and construct the modified IPS estimator as

$$\widehat{J}_N^{\mathrm{IPS}_\beta} = \frac{1}{N}\sum_{i=1}^N c(\widehat{s}_i,\widehat{a}_i)\min\left\{\beta, \frac{\mu_{\mathsf{e}}(\widehat{s}_i,\widehat{a}_i)}{\mu_{\mathsf{b}}(\widehat{s}_i,\widehat{a}_i)}\right\}.$$

Decreasing $\beta$ reduces the variance of $\widehat{J}_N^{\mathrm{IPS}_\beta}$ but increases its bias. An alternative estimator for $V_{\pi_{\mathsf{e}}}$ is the doubly robust (DR) estimator $\widehat{J}_N^{\mathrm{DR}}$, which uses a control variate to reduce the variance of the IPS estimator. The DR estimator was first developed for contextual bandits [24] and then generalized to MDPs [32, 75]. We evaluate the performance of the proposed MDI-DRO estimator on a classical inventory control problem. A detailed problem description is relegated to Appendix 7.4. We sample both the evaluation policy $\pi_{\mathsf{e}}$ and the behavioral policy $\pi_{\mathsf{b}}$ from the uniform distribution on the space of stationary policies. The decision maker then has access to the evaluation policy $\pi_{\mathsf{e}}$ and to a sequence of i.i.d. state action pairs $\{\widehat{s}_i,\widehat{a}_i\}_{i=1}^N$ sampled from $\mu_{\mathsf{b}}$ as well as the observed empirical costs $\{\widehat{c}_i\}_{i=1}^N$, where $\widehat{c}_i = c(\widehat{s}_i,\widehat{a}_i)$. Figure 3 compares the proposed MDI-DRO estimator against the original and modified IPS estimators, the DR estimator and the ground truth expected cost of the evaluation policy. Figures 3a and 3b show that for small radii $r$, the MDI-DRO estimator outperforms the IPS estimators both in terms of accuracy and precision. Figure 3c displays the disappointment probabilities $\mathbb{P}^N(V_{\pi_{\mathsf{e}}} > \widehat{J}_N)$ analyzed in Theorem 4.1, where $\widehat{J}_N$ denotes any of the tested estimators.

**Acknowledgments.** We thank Mengmeng Li for helpful discussions. This research was supported by the Swiss National Science Foundation under the NCCR Automation, grant agreement 51NF40_180545.

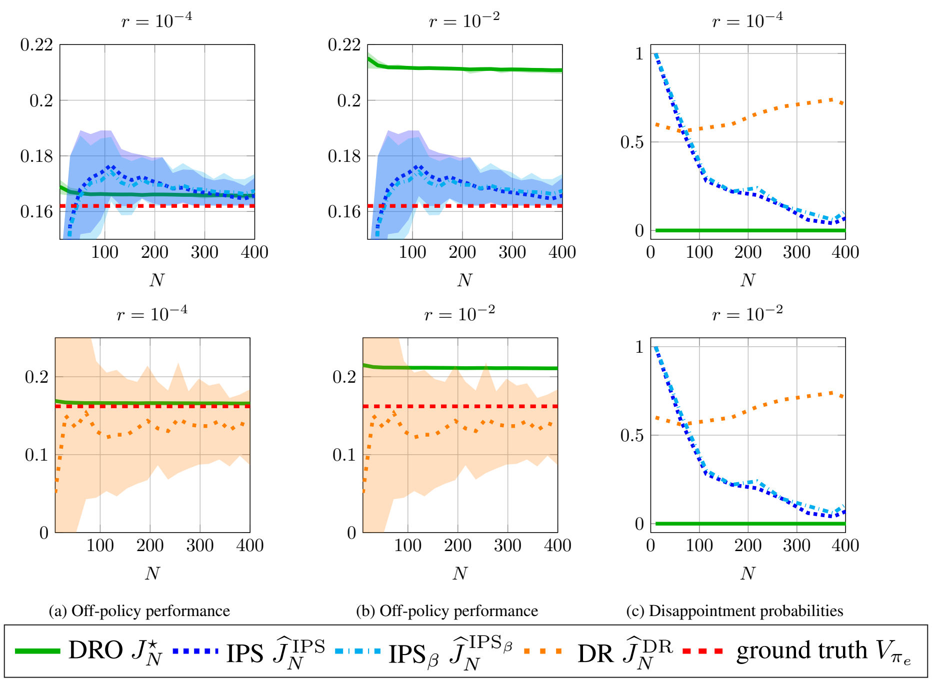


Figure 3: Shaded areas and lines represent 90% confidence intervals and mean values across 1000 independent experiments, respectively.

# References

[1] A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. In **Learning Theory**, pages 574–588. Springer, 2006.

[2] B. Belousov and J. Peters. Entropic regularization of Markov decision processes. **Entropy**, 21 (7), 2019.

[3] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In **Advances in Neural Information Processing Systems**, 2007.

[4] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. **Robust Optimization**. Princeton University Press, 2009.

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

[6] A. Ben-Tal, D. den Hertog, and J.-P. Vial. Deriving robust counterparts of nonlinear uncertain inequalities. **Mathematical Programming**, 149(1):265–299, 2015.

[7] C. Berge. **Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces, and Convexity**. Courier Corporation, 1997.

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

[9] D. Bertsimas and B. V. Parys. Bootstrap robust prescriptive analytics. **arXiv preprint arXiv:1711.09974, Accepted in Mathematical Programming**, 2017.

[10] D. Bertsimas and M. Sim. The price of robustness. **Operations Research**, 52(1):35–53, 2004.

[11] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. **Mathematical Programming**, 167(2):235–292, 2018.

[12] C. M. Bishop. **Pattern Recognition and Machine Learning**. Springer, 2006.

[13] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. **Journal of Machine Learning Research**, 14(65):3207–3260, 2013.

[14] X. Chen, M. Monfort, A. Liu, and B. D. Ziebart. Robust covariate shift regression. In **International Conference on Artificial Intelligence and Statistics**, 2016.

[15] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In **Advances in Neural Information Processing Systems**, 2010.

[16] T. Cover and J. Thomas. **Elements of Information Theory**. Wiley, 2006.

[17] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. **Annals of Probability**, 3(1):146–158, 02 1975.

[18] I. Csiszár. Sanov property, generalized $I$-projection and a conditional limit theorem. **The Annals of Probability**, 12(3):768–793, 1984.

[19] I. Csiszár and J. Korner. **Information Theory: Coding Theorems for Discrete Memoryless Systems**. Academic Press, 1982.

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

[21] A. Dembo and O. Zeitouni. **Large Deviations Techniques and Applications**. Springer, 2009.

[22] O. Devolder, F. Glineur, and Y. Nesterov. Double smoothing technique for large-scale linearly constrained convex optimization. **SIAM Journal on Optimization**, 22(2):702–727, 2012.

[23] J. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. **arXiv preprint, arXiv.1810.08750, Accepted in Annals of Statistics**, 2020.

[24] M. Dudik, D. Erhan, J. Langford, and L. Li. Doubly Robust Policy Evaluation and Optimization. **Statistical Science**, 29(4):485–511, 2014.

[25] M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized Markov decision processes. In **International Conference on Machine Learning**, 2019.

[26] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. **Operations Research**, 58(4):902–917, 2010.

[27] A. Golan. Information and Entropy Econometrics: Review and Synthesis. **Foundations and Trends in Econometrics**, 2(1-2):1–145, 2008.

[28] O. Hernández-Lerma and J. Lasserre. **Discrete-Time Markov Control Processes: Basic Optimality Criteria**. Springer, 1996.

[29] K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. **Econometrica**, 71(4):1161–1189, 2003.

[30] J. Horowitz and R. L. Karandikar. Mean rates of convergence of empirical measures in the Wasserstein metric. **Journal of Computational and Applied Mathematics**, 55(3):261–273, 1994.

[31] E. T. Jaynes. Information theory and statistical mechanics. **Physical Review**, 108:171–190, 1957.

[32] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In **International Conference on Machine Learning**, 2016.

[33] N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. **Journal of Machine Learning Research**, 21(167):1–63, 2020.

[34] J. Kirschner, I. Bogunovic, S. Jegelka, and A. Krause. Distributionally robust Bayesian optimization. In **Artificial Intelligence and Statistics**, 2020.

[35] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li. Stable prediction across unknown environments. In **International Conference on Knowledge Discovery & Data Mining**, 2018.

[36] D. Kuhn, P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. **INFORMS TutORials in Operations Research**, pages 130–166, 2019.

[37] S. Kullback. **Information Theory and Statistics**. Wiley, 1959.

[38] M. G. Lagoudakis and R. Parr. Least-squares policy iteration. **Journal on Machine Learning Research**, 4:1107–1149, 2003.

[39] M. Li, T. Sutter, and D. Kuhn. Distributionally robust optimization with Markovian data. In **International Conference on Machine Learning**, 2021.

[40] A. Liu and B. Ziebart. Robust classification under sample selection bias. In **Advances in Neural Information Processing Systems**, 2014.

[41] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis. Bias and variance approximation in value function estimates. **Management Science**, 53(2):308–322, 2007.

[42] N. Meinshausen. Causality from a distributional robustness point of view. In **IEEE Data Science Workshop (DSW)**, 2018.

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

[44] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In **Advances in Neural Information Processing Systems**, 2016.

[45] A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. **SIAM Journal on Optimization**, 19(4):1757–1780, 2008.

[46] Y. Nesterov. Smooth minimization of non-smooth functions. **Mathematical Programming**, 103(1):127–152, 2005.

[47] Y. Nesterov. **Introductory Lectures on Convex Optimization: A Basic Course**. Springer, 2014.

[48] G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized Markov decision processes. **arXiv preprint arXiv:1705.07798**, 2017.

[49] J. Pearl and E. Bareinboim. Transportability of causal and statistical relations: A formal approach. In **IEEE International Conference on Data Mining Workshops**, 2011.

[50] J. Peters, D. Janzing, and B. Schölkopf. **Elements of Causal Inference: Foundations and Learning Algorithms**. MIT Press, 2017.

[51] D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In **International Conference on Machine Learning**, 2000.

[52] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. **Dataset Shift in Machine Learning**. MIT Press, 2009.

[53] S. Richter. **Computational Complexity Certification of Gradient Methods for Real-Time Model Predictive Control**. PhD thesis, ETH Zurich, 2012.

[54] R. T. Rockafellar. **Convex Analysis**. Princeton University Press, 1997.

[55] R. T. Rockafellar and R. J.-B. Wets. **Variational Analysis**. Springer, 1998.

[56] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters. Invariant models for causal transfer learning. **Journal of Machine Learning Research**, 19(36):1–34, 2018.

[57] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. **Biometrika**, 70(1):41–55, 1983.

[58] D. Rothenhäusler, P. Bühlmann, N. Meinshausen, and J. Peters. Anchor regression: Heterogeneous data meet causality. **Journal of the Royal Statistical Society: Series B**, 83(2):215–246, 2021.

[59] B. Schölkopf and A. J. Smola. **Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond**. MIT Press, 2001.

[60] S. Shafieezadeh Abadeh, P. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In **Advances in Neural Information Processing Systems**, 2015.

[61] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via mass transportation. **Journal of Machine Learning Research**, 20(103):1–68, 2019.

[62] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. **Journal of Statistical Planning and Inference**, 90(2):227–244, 2000.

[63] P. Smadbeck and Y. N. Kaznessis. On a theory of stability for nonlinear stochastic chemical reaction networks. **The Journal of Chemical Physics**, 142(18):184101, 2015.

[64] M. Staib and S. Jegelka. Distributionally robust optimization and generalization in kernel methods. In **Advances in Neural Information Processing Systems**, 2019.

[65] A. L. Strehl, J. Langford, L. Li, and S. M. Kakade. Learning from logged implicit exploration data. In **Advances in Neural Information Processing Systems**, pages 2217–2225, 2010.

[66] A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In **International Conference on Artificial Intelligence and Statistics**, 2019.

[67] M. Sugiyama and M. Kawanabe. **Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation**. MIT Press, 2012.

[68] M. Sugiyama and K.-R. Müller. Input-dependent estimation of generalization error under covariate shift. **Statistics & Decisions**, 23:249–279, 2005.

[69] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. **Journal of Machine Learning Research**, 8(35):985–1005, 2007.

[70] R. K. Sundaram. **A First Course in Optimization Theory**. Cambridge University Press, 1996.

[71] T. Sutter, D. Sutter, P. Mohajerin Esfahani, and J. Lygeros. Generalized maximum entropy estimation. **Journal of Machine Learning Research**, 20(138):1–29, 2019.

[72] T. Sutter, B. P. G. V. Parys, and D. Kuhn. A general framework for optimal data-driven optimization. **arXiv preprint, 2010.06606**, 2020.

[73] R. S. Sutton and A. G. Barto. **Reinforcement Learning: An Introduction**. MIT Press, 2018.

[74] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In **Advances in Neural Information Processing Systems**, 2015.

[75] Z. Tang, Y. Feng, L. Li, D. Zhou, and Q. Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In **International Conference on Learning Representations**, 2020.

[76] P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In **International Conference on Machine Learning**, 2016.

[77] B. Van Parys, P. Mohajerin Esfahani, and D. Kuhn. From data to decisions: Distributionally robust optimization is optimal. **Management Science**, 67(6):3387–3402, 2021.

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

[79] C. Villani. **Optimal Transport: Old and New**. Springer, 2008.

[80] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. **Operations Research**, 62(6):1358–1376, 2014.

[81] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. In **Advances in Neural Information Processing Systems**, 2011.

[82] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In **International Conference on Machine Learning**, 2004.

[83] J. Zhang, A. Menon, A. Veit, S. Bhojanapalli, S. Kumar, and S. Sra. Coping with label shift via distributionally robust optimisation. **arXiv preprint, arXiv.2010.12230**, 2020.

## 7 Appendix

The appendix details all proofs and provides some auxiliary results grouped by section.

### 7.1 Proofs of Section 3

*Proof of Proposition 3.4.* Denote by $\mathbb{P}^N_{\xi_1|\Pi}$ the probability distribution of $\xi_1$ with respect to $\mathbb{P}^N$ conditional on the event $\widehat{\mathbb{P}}_N \in \Pi$. By [18, Theorem 4], we then have

$$\lim_{N\to\infty} \mathsf{D}(\mathbb{P}^N_{\xi_1|\Pi} \| \mathbb{P}^f) = 0,$$

i.e., the conditional distribution $\mathbb{P}^N_{\xi_1|\Pi}$ converges in information to $\mathbb{P}^f$. As the moment-generating function $\mathbb{E}_{\mathbb{P}^f}[e^{tL(\theta,\xi)}]$ is finite for all $t$ in a neighborhood of 0, [17, Lemma 3.1] ensures that

$$\lim_{N\to\infty} \mathbb{E}_{\mathbb{P}^N}[L(\theta,\xi_1)|\widehat{\mathbb{P}}_N \in \Pi] = \lim_{N\to\infty} \mathbb{E}_{\mathbb{P}^N_{\xi_1|\Pi}}[L(\theta,\xi_1)] = \mathbb{E}_{\mathbb{P}^\star}[L(\theta,\xi_1)].$$

Thus, the claim follows. ☐

*Proof of Proposition 3.5.* Proposition 3.5 can be seen as a generalization of [16, Exercise 12.6]. To simplify notation, we define $\alpha = \mathsf{D}(\mathbb{Q}\|\mathbb{P})$. Then, we have

$$\min_{\bar{\mathbb{Q}}\in\Pi} \mathsf{D}(\bar{\mathbb{Q}}\|\mathbb{P}) = \min_{\bar{\mathbb{Q}}\in\mathcal{P}(\Xi)} \sup_{\lambda\in\mathbb{R}} \mathsf{D}(\bar{\mathbb{Q}}\|\mathbb{P}) - \lambda \left( \int_\Xi \log\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right) \mathrm{d}\bar{\mathbb{Q}} - \alpha \right) \tag{18a}$$

$$= \max_{\lambda\in\mathbb{R}} \min_{\bar{\mathbb{Q}}\in\mathcal{P}(\Xi)} \mathsf{D}(\bar{\mathbb{Q}}\|\mathbb{P}) - \lambda \int_\Xi \log\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right) \mathrm{d}\bar{\mathbb{Q}} + \lambda\alpha \tag{18b}$$

$$= \max_{\lambda\in\mathbb{R}} -\log \int_\Xi \left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right)^\lambda \mathrm{d}\mathbb{P} + \lambda\alpha = \alpha, \tag{18c}$$

where (18a) holds by the definition of the set $\Pi$, and (18b) follows from Sion's minimax theorem. The latter applies because the relative entropy $\mathsf{D}(\bar{\mathbb{Q}}\|\mathbb{P})$ is convex in $\bar{\mathbb{Q}}$ and the distribution family $\mathcal{P}(\Xi)$ is convex and weakly compact thanks to the compactness of $\Xi$. Finally, (18c) holds because of [71, Lemma 2], which implies that the inner minimization problem in (18b) is uniquely solved by the probability distribution $\bar{\mathbb{Q}}^\star_\lambda \in \mathcal{P}(\Xi)$ defined through

$$\bar{\mathbb{Q}}^\star_\lambda(B) = \frac{\int_B e^{\lambda \log\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right)} \mathrm{d}\mathbb{P}}{\int_\Xi e^{\lambda \log\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right)} \mathrm{d}\mathbb{P}} = \frac{\int_B \left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right)^\lambda \mathrm{d}\mathbb{P}}{\int_\Xi \left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right)^\lambda \mathrm{d}\mathbb{P}} \quad \forall B \in \mathcal{B}(\Xi).$$

By inspecting the first-order optimality condition of the convex maximization problem in (18c) and remembering that $\alpha = \mathsf{D}(\mathbb{Q}\|\mathbb{P})$, one can then show that (18c) is solved by $\lambda^\star = 1$. The Nash equilibrium of the zero-sum game in (18b) is therefore given by $\lambda^\star$ and its unique best response $\bar{\mathbb{Q}}^\star_{\lambda^\star} = \mathbb{Q}$, and the solution $f(\mathbb{P})$ of the I-projection problem in (18a) coincides with $\mathbb{Q}$. ☐

### 7.2 Proofs of Section 4

*Proof of Theorem 4.1.* The continuity of $R^\star$ on $\Theta \times \Pi$ is established in Corollary 7.1 below.

In order to prove that the DRO predictor $R^\star$ is also admissible, we first prove that the following inequality holds for any fixed $\theta \in \Theta$ and $\mathbb{P} \in \mathcal{P}(\Xi)$.

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^N \left( R(\theta, \mathbb{P}^f) > R^\star(\theta, \widehat{\mathbb{P}}^f_N) \right) \le -r \tag{19}$$

For the sake of concise notation, we then define the disappointment set

$$\mathcal{D}(\theta, \mathbb{P}) = \{\mathbb{P}' \in \mathcal{P}(\Xi) : R(\theta, f(\mathbb{P})) > R^\star(\theta, f(\mathbb{P}'))\}$$

containing all realizations $\mathbb{P}'$ of the empirical distribution $\widehat{\mathbb{P}}_N$, for which the true risk $R(\theta, f(\mathbb{P}))$ under the I-projection of the unknown true distribution exceeds the risk $R^\star(\theta, f(\mathbb{P}'))$ predicted by the distributionally robust predictor under the I-projection of the empirical distribution. Hence, $\bar{\mathcal{D}}(\theta, \mathbb{P})$

contains all realizations of $\widehat{\mathbb{P}}_N$ under which the distributionally robust predictor is too optimistic and thus leads to disappointment. Similarly, we define the weak disappointment set

$$\bar{\mathcal{D}}(\theta, \mathbb{P}) = \{\mathbb{P}' \in \mathcal{P}(\Xi) : R(\theta, f(\mathbb{P})) \geq R^\star(\theta, f(\mathbb{P}'))\},$$

which simply replaces the strict inequality in the definiton of $\bar{\mathcal{D}}(\theta, \mathbb{P})$ with a weak inequality. Recall now that $R^\star$ is continuous. In addition, note that $f$ is continuous thanks to [70, Theorem 9.17], which follows from the strict convexity of the relative entropy in its first argument [21, Lemma 6.2.12]. Therefore the set $\bar{\mathcal{D}}(\theta, \mathbb{P})$ is closed, and $\mathrm{cl}\,\mathcal{D}(\theta, \mathbb{P}) \subset \bar{\mathcal{D}}(\theta, \mathbb{P})$. The left hand side of (19) thus satisfies

$$\begin{aligned}
\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^N \left( R(\theta, f(\mathbb{P})) > R^\star(\theta, f(\widehat{\mathbb{P}}_N)) \right) &= \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}^N \left( \widehat{\mathbb{P}}_N \in \mathcal{D}(\theta, \mathbb{P}) \right) \\
&\leq - \inf_{\mathbb{P}' \in \mathrm{cl}\,\mathcal{D}(\theta, \mathbb{P})} \mathsf{D}(\mathbb{P}' \| \mathbb{P}) \\
&\leq - \inf_{\mathbb{P}' \in \bar{\mathcal{D}}(\theta, \mathbb{P})} \mathsf{D}(\mathbb{P}' \| \mathbb{P}) \\
&\leq -r,
\end{aligned}$$

where the first inequality follows from Sanov's Theorem, which asserts that $\widehat{\mathbb{P}}_N$ satisfies a large deviation principle with the relative entropy as the rate function [21, Theorem 6.2.10]. The second inquality exploits the inclusion $\mathrm{cl}\,\mathcal{D}(\theta, \mathbb{P}) \subset \bar{\mathcal{D}}(\theta, \mathbb{P})$, and the last inequality holds because

$$\mathbb{P}' \in \bar{\mathcal{D}}(\theta, \mathbb{P}) \quad \Longrightarrow \quad \mathsf{D}(f(\mathbb{P}') \| f(\mathbb{P})) \geq r \quad \Longrightarrow \quad \mathsf{D}(\mathbb{P}' \| \mathbb{P}) \geq r,$$

where the first implication has been established in te proof of [77, Theorem 10], and the second implication follows from the data-processing inequality [19, Lemma 3.11]. This proves (19).

In the last step of the proof, we fix an arbitrary $\varepsilon > 0$ and show that

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P} \left( R(\theta^\star_N, \mathbb{P}^f) > R^\star(\theta^\star_N, \widehat{\mathbb{P}}^f_N) + \varepsilon \right) \leq -r$$

for any $\mathbb{P} \in \mathcal{P}(\Xi)$, where $\theta^\star_N$ is defined as usual as a minimizer of (4). The proof of this generalized statement widely parallels that of [77, Theorem 11] and exploits the data processing inequality in a similar manner as in the proof of (19). Details are omitted for brevity. □

*Proof of Theorem 4.2.* The proof is inspired by [77, Theorems 7 & 11]. We first show that any continuous admissible data-driven predictor $\widehat{R}$ satisfies the inequality

$$\lim_{N\to\infty} R^\star(\theta, f(\widehat{\mathbb{P}}_N)) \leq \lim_{N\to\infty} \widehat{R}(\theta, f(\widehat{\mathbb{P}}_N)) \quad \mathbb{P}^\infty\text{-a.s.} \tag{20}$$

for all $\theta \in \Theta$ and $\mathbb{P} \in \mathcal{P}(\Xi)$. As the empirical distribution $\widehat{\mathbb{P}}_N$ converges weakly to $\mathbb{P}$ and as $R^\star$, $\widehat{R}$ and $f$ represent continuous mappings, the inequality (20) is equivalent to

$$R^\star(\theta, f(\mathbb{P})) \leq \widehat{R}(\theta, f(\mathbb{P}))$$

for all $\theta \in \Theta$ and $\mathbb{P} \in \mathcal{P}(\Xi)$. Suppose now for the sake of contradiction there exists a continuous admissible predictor $\widehat{R}$, a parameter $\theta_0 \in \Theta$ and an asymptotic estimator realization $\mathbb{P}'_0 \in \mathcal{P}(\Xi)$ with

$$\widehat{R}(\theta_0, f(\mathbb{P}'_0)) < R^\star(\theta_0, f(\mathbb{P}'_0)).$$

In fact, as $\widehat{R}$, $R^\star$ and $f$ are continuous functions, this strict inequality holds on a neighborhood of $\mathbb{P}'_0$. Next, define $\varepsilon = R^\star(\theta_0, f(\mathbb{P}'_0)) - \widehat{R}(\theta_0, f(\mathbb{P}'_0)) > 0$ and denote by $\bar{\mathbb{P}} \in \Pi$ an optimizer of the worst-case risk evaluation problem (5) for $\mathbb{P}' = f(\mathbb{P}'_0)$, which satisfies $R^\star(\theta_0, f(\mathbb{P}'_0)) = R(\theta_0, \bar{\mathbb{P}})$ and $\mathsf{D}(f(\mathbb{P}'_0) \| \bar{\mathbb{P}}) \leq r$. By using a continuity argument as in the proof of [77, Theorem 10] and by exploiting the convexity of $\Pi$, one can then show that there exists a model $\mathbb{P}_0 \in \Pi$ with

$$R(\theta_0, \bar{\mathbb{P}}) < R(\theta, \mathbb{P}_0) + \varepsilon \quad \text{and} \quad \mathsf{D}(f(\mathbb{P}'_0) \| \mathbb{P}_0) = r_0 < r. \tag{21}$$

All of this implies that

$$\widehat{R}(\theta_0, f(\mathbb{P}'_0)) = R^\star(\theta_0, f(\mathbb{P}'_0)) - \varepsilon = R(\theta_0, \bar{\mathbb{P}}) - \varepsilon < R(\theta_0, \mathbb{P}_0) = R(\theta_0, f(\mathbb{P}_0)), \tag{22}$$

where the three equalities follow from the definition of $\varepsilon$, the construction of $\bar{\mathbb{P}}$ and the observation that $f$ reduces to the identity mapping when restricted to $\Pi$. The inequality holds due to the first relation in (21). In analogy to the proof of Theorem 4.1, we now introduce the disappointment set for the data-driven predictor $\widehat{R}$ under the data-generating distribution $\mathbb{P}_0$, that is,

$$\mathcal{D}(\theta_0, \mathbb{P}_0) = \left\{ \mathbb{P}' \in \mathcal{P}(\Xi) : R(\theta_0, f(\mathbb{P}_0)) > \widehat{R}(\theta_0, f(\mathbb{P}')) \right\}.$$

The relation (22) readily implies that $\mathbb{P}'_0 \in \mathcal{D}(\theta_0, \mathbb{P}_0)$. As the I-projection is idempotent (that is, $f \circ f = f$), one can further verify that $f(\mathbb{P}'_0) \in \mathcal{D}(\theta_0, \mathbb{P}_0)$. Denoting the empirical distribution of $N$ training samples drawn independently from $\mathbb{P}_0$ by $\widehat{\mathbb{P}}_{0,N}$, we thus find

$$\begin{aligned}
\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}_0^N \left( R(\theta_0, f(\mathbb{P}_0)) > \widehat{R}(\theta_0, f(\widehat{\mathbb{P}}_{0,N})) \right) &= \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}_0^N \left( \widehat{\mathbb{P}}_{0,N} \in \mathcal{D}(\theta_0, \mathbb{P}_0) \right) \\
&\geq - \inf_{\mathbb{P}' \in \operatorname{int} \mathcal{D}(\theta_0, \mathbb{P}_0)} \mathsf{D}(\mathbb{P}' \| \mathbb{P}_0) \\
&= - \inf_{\mathbb{P}' \in \mathcal{D}(\theta_0, \mathbb{P}_0)} \mathsf{D}(\mathbb{P}' \| \mathbb{P}_0) \\
&\geq -\mathsf{D}(f(\mathbb{P}'_0) \| \mathbb{P}_0) \\
&= -r_0 > -r,
\end{aligned}$$

where the first inequality follows from Sanov's Theorem, which ensures that $\widehat{\mathbb{P}}_N$ satisfies a large deviation principle with the relative entropy as the rate function. The second equality holds because $\mathcal{D}(\theta_0, \mathbb{P}'_0)$ is open thanks to the continuity of $\widehat{R}$ and $f$, and the second inequality exploits our earlier insight that $f(\mathbb{P}'_0) \in \mathcal{D}(\theta_0, \mathbb{P}'_0)$. The last inequality, finally, follows from the second relation in (21). The above reasoning shows that $\widehat{R}$ fails to be admissible, and hence a data-driven predictor $\widehat{R}$ with the advertised properties cannot exist. Thus, $R^\star$ indeed satisfies the efficiency property (20).

To show that $\lim_{N\to\infty} J^\star_N \leq \lim_{N\to\infty} \widehat{J}_N$ $\mathbb{P}^\infty$-almost surely for all $\mathbb{P} \in \mathcal{P}(\Xi)$, we use (20) and adapt the proof of [77, Theorem 11] with obvious modifications. Details are omitted for brevity. □

*Proof of Corollary 4.3.* Recalling that Sanov's Theorem for finite state spaces offers finite sample bounds [16, Theorem 11.4.1], the claim can be established by repeating the proof of Theorem 4.1. □

*Proof of Theorem 4.4.* It suffices to prove (10b) because (10a) can be seen as a special case of (10b) when $\Theta = \{\theta\}$. In the remainder we denote by $\mathsf{d}_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q})$ the total variation distance and by $\mathsf{d}_{\mathsf{W}_p}(\mathbb{P}, \mathbb{Q})$ the $p$-th Wasserstein distance ($p \in \mathbb{N}$) between two probability distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\Xi)$. To make its dependence on the radius $r$ explicit, throughout this proof we temporarily use $R^\star_r$ to denote the DRO predictor (5). As usual, we use $\theta^\star_N \in \Theta$ to denote a minimizer of the DRO problem (4) with $\mathbb{P}' = \widehat{\mathbb{P}}^f_N$. In addition, we use $\widehat{\mathbb{Q}}^\star_{N,\theta} \in \Pi$ to denote a maximizer of the worst-case risk evaluation problem (5) with $\mathbb{P}' = \widehat{\mathbb{P}}^f_N$. By definition, this maximizer must satisfy the relations

$$R^\star_{r_N}(\theta, \widehat{\mathbb{P}}^f_N) = R(\theta, \widehat{\mathbb{Q}}^\star_{N,\theta}) \quad \text{and} \quad \mathsf{D}(\widehat{\mathbb{P}}^f_N \| \widehat{\mathbb{Q}}^\star_{N,\theta}) \leqslant r_N$$

for all $\theta \in \Theta$ and $N \in \mathbb{N}$. Pinsker's inequality then implies that

$$\sup_{\theta\in\Theta} \mathsf{d}_{\mathsf{TV}} \left( \widehat{\mathbb{P}}^f_N, \widehat{\mathbb{Q}}^\star_{N,\theta} \right) \leq \sup_{\theta\in\Theta} \sqrt{\frac{1}{2} \mathsf{D}(\widehat{\mathbb{P}}^f_N \| \widehat{\mathbb{Q}}^\star_{N,\theta})} \leq \sqrt{\frac{r_N}{2}} \quad \forall N \in \mathbb{N}. \tag{23}$$

Thus, we find

$$\begin{aligned}
&\sup_{\theta\in\Theta} \left\{ \left| R^\star_{r_N}(\theta, \widehat{\mathbb{P}}^f_N) - R(\theta, \mathbb{P}^f) \right| \right\} \\
&\quad = \sup_{\theta\in\Theta} \left\{ \left| \mathbb{E}_{\widehat{\mathbb{Q}}^\star_{N,\theta}}[L(\theta, \xi)] - \mathbb{E}_{\mathbb{P}^f}[L(\theta, \xi)] \right| \right\} \\
&\quad \leq \sup_{\theta\in\Theta} \left\{ \left| \mathbb{E}_{\widehat{\mathbb{Q}}^\star_{N,\theta}}[L(\theta, \xi)] - \mathbb{E}_{\widehat{\mathbb{P}}^f_N}[L(\theta, \xi)] \right| + \left| \mathbb{E}_{\widehat{\mathbb{P}}^f_N}[L(\theta, \xi)] - \mathbb{E}_{\mathbb{P}^f}[L(\theta, \xi)] \right| \right\} \\
&\quad \leq \sup_{\theta\in\Theta} \left\{ \left| \mathbb{E}_{\widehat{\mathbb{Q}}^\star_{N,\theta}}[L(\theta, \xi)] - \mathbb{E}_{\widehat{\mathbb{P}}^f_N}[L(\theta, \xi)] \right| \right\} + \sup_{\theta\in\Theta} \left\{ \left| \mathbb{E}_{\widehat{\mathbb{P}}^f_N}[L(\theta, \xi)] - \mathbb{E}_{\mathbb{P}^f}[L(\theta, \xi)] \right| \right\} \\
&\quad \leq \Lambda \sup_{\theta\in\Theta} \mathsf{d}_{\mathsf{W}_1} \left( \widehat{\mathbb{Q}}^\star_{N,\theta}, \widehat{\mathbb{P}}^f_N \right) + \Lambda \mathsf{d}_{\mathsf{W}_1} \left( \widehat{\mathbb{P}}^f_N, \mathbb{P}^f \right) \\
&\quad \leq \Lambda C \sup_{\theta\in\Theta} \mathsf{d}_{\mathsf{TV}} \left( \widehat{\mathbb{Q}}^\star_{N,\theta}, \widehat{\mathbb{P}}^f_N \right) + \Lambda \mathsf{d}_{\mathsf{W}_2} \left( \widehat{\mathbb{P}}^f_N, \mathbb{P}^f \right),
\end{aligned}$$

where the first three inequalities follow from the triangle inequality, the subadditivity of the supremum operator and the Kantorovich-Rubinstein theorem [79, Theorem 5.10], respectively. The last inequality holds because $\Xi$ is compact, which implies that the first Wasserstein distance can be bounded above by the total variation distance scaled with a positive constant $C$ [79, Theorem 6.15] and because $\mathsf{d}_{\mathsf{W}_1}(\cdot,\cdot) \leq \mathsf{d}_{\mathsf{W}_2}(\cdot,\cdot)$ thanks to Jensen's inequality. By (23), the first term in the above expression decays deterministically to zero as $N$ grows. The second term converges $\mathbb{P}^\infty$-almost surely to zero as $N$ increases because the empirical distribution converges $\mathbb{P}^\infty$-almost surely to the data-generating distribution in the second Wasserstein distance [30]. In summary, we thus have

$$\lim_{N\to\infty} \sup_{\theta\in\Theta} \left| R^\star_{r_N}(\theta, \widehat{\mathbb{P}}^f_N) - R(\theta, \mathbb{P}^f) \right| = 0 \quad \mathbb{P}^\infty\text{-a.s.} \tag{24}$$

Put differently, for $\mathbb{P}^\infty$-almost every trajectory of training samples, the functions $R^\star_{r_N}(\cdot, \widehat{\mathbb{P}}^f_N)$ converge uniformly to $R(\cdot, \mathbb{P}^f)$. The claim then follows from [55, Proposition 7.15 and Theorem 7.31]. □

### 7.3 Proofs and auxiliary results for Section 5

*Proof of Theorem 5.2.* The key enabling mechanism to prove (15a) and (15b) is the so-called double smoothing method for linearly constrained convex programs [22]. Our proof parallels that of [71, Theorem 5] and is provided here to keep the paper self contained. Throughout the proof, we denote by $\mathcal{M}(\Xi)$ the vector space of all finite signed Borel measures on $\Xi$, and we equip $\mathcal{M}(\Xi)$ with the total variation norm $\|\cdot\|_{\mathsf{TV}}$. Choosing the total variation norm has the benefit that the function $g : \mathcal{P}(\Xi) \to \mathbb{R}_+$ defined through $g(\mathbb{Q}) = \mathsf{D}(\mathbb{Q}\|\widehat{\mathbb{P}}_N)$ is strongly convex with convexity parameter 1. Indeed, Pinsker's inequality implies that $d(\mathbb{Q}) \geq \frac{1}{2}\|\mathbb{Q} - \widehat{\mathbb{P}}_N\|^2_{\mathsf{TV}}$ for all $\mathbb{Q} \in \mathcal{P}(\Xi)$. To prove (15a) and (15b), we consider the primal and dual optimization problems

$$J^\star_{\mathsf{P}} = \min_{\mathbb{Q}\in\mathcal{P}(\Xi)} \left\{ \mathsf{D}(\mathbb{Q}\|\widehat{\mathbb{P}}_N) + \sup_{z\in\mathbb{R}^d} \left\{ \mathbb{E}_{\mathbb{Q}}[\psi(\xi)]^\top z - \sigma_E(z) \right\} \right\} \tag{25a}$$

$$J^\star_{\mathsf{D}} = \sup_{z\in\mathbb{R}^d} \left\{ -\sigma_E(z) + \min_{\mathbb{Q}\in\mathcal{P}(\Xi)} \left\{ \mathsf{D}(\mathbb{Q}\|\widehat{\mathbb{P}}_N) + \mathbb{E}_{\mathbb{Q}}[\psi(\xi)]^\top z \right\} \right\}, \tag{25b}$$

where $\sigma_E : \mathbb{R}^d \to \mathbb{R}$ defined through $\sigma_E(z) = \max_{x\in E} z^\top x$ denotes the support function of $E$. As the convex conjugate of the support function $\sigma_E$ is the indicator function $\delta_E : \mathbb{R}^d \to [0,\infty]$ defined through $\delta_E(x) = 0$ if $x \in E$ and $\delta_E(x) = \infty$ if $x \notin E$, the optimal value of the maximization problem over $z$ in (25a) equals $\delta_E(\mathbb{E}_{\mathbb{Q}}[\psi(\xi)])$. Hence, the unique minimizer of (25a) coincides with the I-projection of the empirical distribution onto the set $\Pi$. We also remark that $\sigma_E$ is continuous because $E$ is non-empty and compact [54, Corollary 13.2.2]. Assumption 5.1 then ensures via [71, Lemma 3] that there is no duality gap, i.e, $J^\star_{\mathsf{P}} = J^\star_{\mathsf{D}}$. Next, we introduce the shorthand

$$F(z) = -\sigma_E(z) + \min_{\mathbb{Q}\in\mathcal{P}(\Xi)} \left\{ \mathsf{D}(\mathbb{Q}\|\widehat{\mathbb{P}}_N) + \mathbb{E}_{\mathbb{Q}}[\psi(\xi)]^\top z \right\}$$

for the dual objective function. While the primal problem (25a) is an infinite-dimensional optimization problem, the dual problem (25b) can be solved via first-order methods provided that the gradient of the dual objective function $F$ can be evaluated at low cost. Unfortunately, this function fails to be smooth. Consequently, an optimal first-order method would require $O(1/\varepsilon^2)$ iterations, where $\varepsilon$ denotes the desired additive accuracy [47, Section 3.2]. However, the computation can be accelerated by smoothing the dual objective function as in [22, 46] and by exploiting structural properties. To this end, we introduce a smoothed version $F_\eta$ of the dual objective function defined through

$$F_\eta(z) = -\max_{x\in E} \left\{ x^\top z - \frac{\eta_1}{2}\|x\|_2^2 \right\} + \min_{\mathbb{Q}\in\mathcal{P}(\Xi)} \left\{ \mathsf{D}(\mathbb{Q}\|\widehat{\mathbb{P}}_N) + \mathbb{E}_{\mathbb{Q}}[\psi(\xi)]^\top z \right\} - \frac{\eta_2}{2}\|z\|_2^2,$$

where $\eta = (\eta_1, \eta_2) \in \mathbb{R}^2_{++}$ is a smoothing parameter. One readily verifies that $x^\star_z = \pi_E(\eta_1^{-1} z)$ solves the optimization problem in the first term. The optimization problem in the second term minimizes the sum of a relative entropy function and a linear function. Therefore, it is reminiscent of an entropy maximization problem, and one can show that it is solved by the Gibbs distribution

$$\mathbb{Q}^\star_z = \frac{\sum_{j=1}^N \exp\left(-z^\top \psi(\widehat{\xi}_j)\right) \delta_{\widehat{\xi}_j}}{\sum_{j=1}^N \exp\left(-z^\top \psi(\widehat{\xi}_j)\right)},$$

see [71, Lemma 2]. By construction, the smoothed dual objective function $F_\eta$ is $\eta_2$-strongly concave and differentiable. Its gradient can be expressed in terms of the parametric optimizers $x_z^\star$ and $\mathbb{Q}_z^\star$ as

$$\nabla F_\eta(z) = -x_z^\star + \mathbb{E}_{\mathbb{Q}_z^\star}[\psi(\xi)] - \eta_2 z = G_\eta(z),$$

where $G_\eta$ is defined in (12); see also [46, Theorem 1]. In addition, as shown in [46, Theorem 1], the gradient function $G_\eta$ is Lipschitz continuous with a Lipschitz constant $L_\eta$ that satisfies

$$\begin{aligned} L_\eta &= 1/\eta_1 + \eta_2 + \left( \sup_{\lambda \in \mathbb{R}^d, \mathbb{Q} \in \mathcal{M}(\Xi)} \left\{ \lambda^\top \mathbb{E}_{\mathbb{Q}}[\psi(\xi)] \;:\; \|\lambda\|_2 = 1, \|\mathbb{Q}\|_{\mathsf{TV}} = 1 \right\} \right)^2 \\ &\le 1/\eta_1 + \eta_2 + \left( \sup_{\lambda \in \mathbb{R}^d, \mathbb{Q} \in \mathcal{M}(\Xi)} \{ \|\lambda\|_2 \|\mathbb{E}_{\mathbb{Q}}[\psi(\xi)]\|_2 \;:\; \|\lambda\|_2 = 1, \|\mathbb{Q}\|_{\mathsf{TV}} = 1 \} \right)^2 \\ &= 1/\eta_1 + \eta_2 + (\max_{\xi \in \Xi} \|\psi(\xi)\|_\infty)^2 < \infty. \end{aligned}$$

Therefore, the smoothed dual optimization problem

$$\sup_{z \in \mathbb{R}^d} F_\eta(z) \tag{26}$$

has a smooth and strongly concave objective function, implying that it can be solved highly efficiently via fast gradient methods. When solving (26) by Algorithm 1, we can use its outputs $z_k$ to construct candidate solutions $\widehat{\mathbb{Q}}_k$ for the primal (non-regularized) problem (25a) as described in (13). These candidate solutions satisfy the optimality and feasibility guarantees (15a) and (15b), which can be derived by using the techniques developed in [22]. A detailed derivation using our notation is also provided in [71, Appendix A]. We highlight that (15a) and (15b) critically rely on Assumption 5.1, which implies via [45, Lemma 1] that the norm of the unique maximizer of the regularized dual problem (26) is bounded above by $C/\delta$, where $C$ and $\delta$ are defined as in (14). □

*Proof of Proposition 5.3.* By the definition of the DRO predictor $R^\star$ in (5), we have

$$\begin{aligned} R^\star(\theta, \mathbb{P}') &= \sup_{\mathbb{Q} \in \mathcal{P}(\Xi)} \{ \mathbb{E}_{\mathbb{Q}}[L(\theta, \xi)] \;:\; \mathsf{D}(\mathbb{P}' \| \mathbb{Q}) \le r, \; \mathbb{E}_{\mathbb{Q}}[\psi(\xi)] \in E \} \\ &= \sup_{\mathbb{Q} \in \mathcal{P}(\Xi)} \left\{ \mathbb{E}_{\mathbb{Q}}[L(\theta, \xi)] - \sup_{z \in \mathbb{R}^d} \{ z^\top \mathbb{E}_{\mathbb{Q}}[\psi(\xi)] - \sigma_E(z) \} \;:\; \mathsf{D}(\mathbb{P}' \| \mathbb{Q}) \le r \right\} \\ &= \inf_{z \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathcal{P}(\Xi)} \left\{ \mathbb{E}_{\mathbb{Q}}[L(\theta, \xi) - z^\top \psi(\xi)] + \sigma_E(z) \;:\; \mathsf{D}(\mathbb{P}' \| \mathbb{Q}) \le r \right\} \\ &= \inf_{z \in \mathbb{R}^d} \left\{ \begin{array}{ll} \inf\limits_{\alpha \in \mathbb{R}} & \alpha + \sigma_E(z) - e^{-r} \exp\left( \mathbb{E}_{\mathbb{P}'}[\log(\alpha - L(\theta, \xi) + z^\top \psi(\xi))] \right) \\ \text{s.t.} & \alpha \ge \max_{\xi \in \Xi} L(\theta, \xi) - z^\top \psi(\xi) \end{array} \right. \end{aligned}$$

where the second equality holds because the convex conjugate of the support function $\sigma_E$ is the indicator function $\delta_E : \mathbb{R}^d \to [0, \infty]$ defined through $\delta_E(x) = 0$ if $x \in E$ and $\delta_E(x) = \infty$ if $x \notin E$, and the third equality follows from Sion's minimax theorem, which applies because the relative entropy $\mathsf{D}(\mathbb{P}' \| \mathbb{Q})$ is convex in $\mathbb{Q}$, while the distribution family $\mathcal{P}(\Xi)$ is convex and weakly compact. Finally, the fourth equality follows from [77, Proposition 5], which applies because $r > 0$ and because the modified loss function $L(\theta, \xi) - z^\top \psi(\xi)$ is continuous in $\xi$ for any fixed $\theta$ and $z$. The last expression is manifestly equivalent to (16), and thus the claim follows. □

The following corollary of Proposition 5.3 establishes that the DRO predictor $R^\star$ is continuous. This result is relevant for Theorem 4.1.

**Corollary 7.1** (Continuity of $R^\star$)**.** *If $r > 0$, $0 \in \mathrm{int}(E)$ and for every $z \in \mathbb{R}^d$ there exists $\xi \in \Xi$ such that $z^\top \psi(\xi) > 0$, then the DRO predictor $R^\star$ is continuous on $\Theta \times \Pi$.*

*Proof.* Since $r > 0$, we may use Proposition 5.3 to express the DRO predictor as

$$R^\star(\theta, \mathbb{P}') = \inf_{z \in \mathbb{R}^d} \varphi_E(\theta, z, \mathbb{P}') \tag{27}$$

for all $\theta \in \Theta$ and $\mathbb{P}' \in \Pi$, where the parametric objective function $\varphi_E$ is defined through

$$\varphi_E(\theta, z, \mathbb{P}') = \inf_{\alpha \geq \underline{\alpha}(\theta,z)} \alpha + \sigma_E(z) - e^{-r} \exp\left(\mathbb{E}_{\mathbb{P}'}[\log(\alpha - L(\theta,\xi) + z^\top \psi(\xi))]\right)$$

with $\underline{\alpha}(\theta, z) = \max_{\xi\in\Xi} L(\theta, \xi) - z^\top \psi(\xi)$. Note that the support function $\sigma_E$ is continuous because $E$ is compact. Applying [77, Proposition 6] to the modified loss function $L(\theta, \xi) - z^\top \psi(\xi)$ thus implies that $\varphi_E$ is continuous on $\Theta \times \mathbb{R}^d \times \Pi$. To bound $\varphi_E$ from below by a coercive function, we define

$$\kappa = \min_{\|z\|_2=1} \min_{\mathbb{Q}\in\Pi} \sigma_E(z) - e^{-r} z^\top \mathbb{E}_{\mathbb{Q}}[\psi(\xi)],$$

which is a finite constant. Indeed, $\sigma_E$ is continuous because $E$ is compact, and $\mathbb{E}_{\mathbb{P}'}[\psi(\xi)]$ is weakly continuous in $\mathbb{P}'$ because $\psi$ is a continuous and bounded function on the compact set $\Xi$. In addition, the unit sphere in $\mathbb{R}^d$ is compact, and the set $\Pi$ is weakly compact. Therefore, both minima in the definition of $\kappa$ are attained at some $z^\star \in \mathbb{R}^d$ with $\|z^\star\|_2 = 1$ and some $\mathbb{Q}^\star \in \Pi$, respectively. As $0 \in \mathrm{int}(E)$ and $z^\star \neq 0$, we have $\sigma_E(z^\star) > 0$. In addition, as $\mathbb{Q}^\star \in \Pi$, we have $\mathbb{E}_{\mathbb{Q}^\star}[\psi(\xi)] \in E$, which implies that $(z^\star)^\top \mathbb{E}_{\mathbb{Q}^\star}[\psi(\xi)] \leq \sigma_E(z^\star)$. Again as $r > 0$, this reasoning ensures that

$$\kappa = \sigma_E(z^\star) - e^{-r}(z^\star)^\top \mathbb{E}_{\mathbb{Q}^\star}[\psi(\xi)] > 0.$$

Similarly, we introduce the finite constant

$$\underline{L} = \min_{\theta\in\Theta} \min_{z\in\mathbb{R}^d} \min_{\xi\in\Xi} \; (1 - e^{-r})\underline{\alpha}(\theta, z) + e^{-r} L(\theta, \xi).$$

To see that $\underline{L}$ is bounded below, note that the definition of $\underline{\alpha}$ and the subadditivity of the minimum operator lead to the estimate

$$\begin{aligned} \underline{L} &\geq (1 - e^{-r}) \min_{\theta\in\Theta} \min_{\xi\in\Xi} \; L(\theta, \xi) + (1 - e^{-r}) \min_{z\in\mathbb{R}^d} \max_{\xi\in\Xi} \; (-z)^\top \psi(\xi) + e^{-r} \min_{\theta\in\Theta} \min_{\xi\in\Xi} \; L(\theta, \xi) \\ &= \min_{\theta\in\Theta} \min_{\xi\in\Xi} \; L(\theta, \xi) + (1 - e^{-r}) \min_{z\in\mathbb{R}^d} \max_{\xi\in\Xi} \; (-z)^\top \psi(\xi). \end{aligned}$$

The first term in the resulting lower bound is finite because $L$ is continuous, while $\Theta$ and $\Xi$ are compact. The second term is also finite because the convex function $\max_{\xi\in\Xi} \; (-z)^\top \psi(\xi)$ is continuous in $z$ thanks to the continuity of $\psi$ and the compactness of $\Xi$. In addition, $\max_{\xi\in\Xi} \; (-z)^\top \psi(\xi)$ is also coercive in $z$ because of the assumption that for every $z \in \mathbb{R}^d$ there exists $\xi \in \Xi$ with $z^\top \psi(\xi) > 0$.

The above preparatory arguments imply that

$$\begin{aligned} \varphi_E(\theta, z, \mathbb{P}') &\geq \inf_{\alpha \geq \underline{\alpha}(\theta,z)} (1 - e^{-r})\alpha + \sigma_E(z) + e^{-r}\mathbb{E}_{\mathbb{P}'}[L(\theta, \xi)] - e^{-r} z^\top \mathbb{E}_{\mathbb{P}'}[\psi(\xi)] \\ &= (1 - e^{-r})\underline{\alpha}(\theta, z) + e^{-r}\mathbb{E}_{\mathbb{P}'}[L(\theta, \xi)] + \left(\sigma_E\left(\frac{z}{\|z\|_2}\right) - e^{-r}\left(\frac{z}{\|z\|_2}\right)^\top \mathbb{E}_{\mathbb{P}'}[\psi(\xi)]\right)\|z\|_2 \\ &\geq \underline{L} + \kappa\|z\|_2, \end{aligned}$$

where the first inequality exploits Jensen's inequality, the equality holds thanks to the positive homogeneity of the support function $\sigma_E$ and the trivial observation that $e^{-r} < 1$, and the second inequality follows from the definitions of $\underline{L}$ and $\kappa$ and the assumption that $\mathbb{P}' \in \Pi$. We thus have

$$\varphi_E(\theta, z, \mathbb{P}') \geq \underline{L} + \kappa\|z\|_2 \quad \forall \theta \in \Theta, \; \forall z \in \mathbb{R}^d, \; \forall \mathbb{P}' \in \Pi. \tag{28a}$$

Next, define

$$\overline{L} = \max_{\theta\in\Theta} \max_{\xi\in\Xi} L(\theta, \xi),$$

and note that

$$\inf_{z\in\mathbb{R}^d} \varphi_E(\theta, z, \mathbb{P}') = R^\star(\theta, \mathbb{P}^\star) \leq \overline{L} \quad \forall \theta \in \Theta, \; \forall \mathbb{P}' \in \Pi. \tag{28b}$$

Taken together, the estimates (28a) and (28b) imply that

$$R^\star(\theta, \mathbb{P}') = \inf_{z\in\mathbb{R}^d} \left\{\varphi_E(\theta, z, \mathbb{P}') : \|z\|_2 \leq \frac{\overline{L} - \underline{L}}{\kappa}\right\},$$

which in turn implies via Berge's maximum theorem [7, pp. 115–116] and the continuity of the objective function $\varphi_E$ on $\Theta \times \mathbb{R}^d \times \Pi$ that the DRO predictor $R^\star$ is indeed continuous on $\Theta \times \Pi$. □

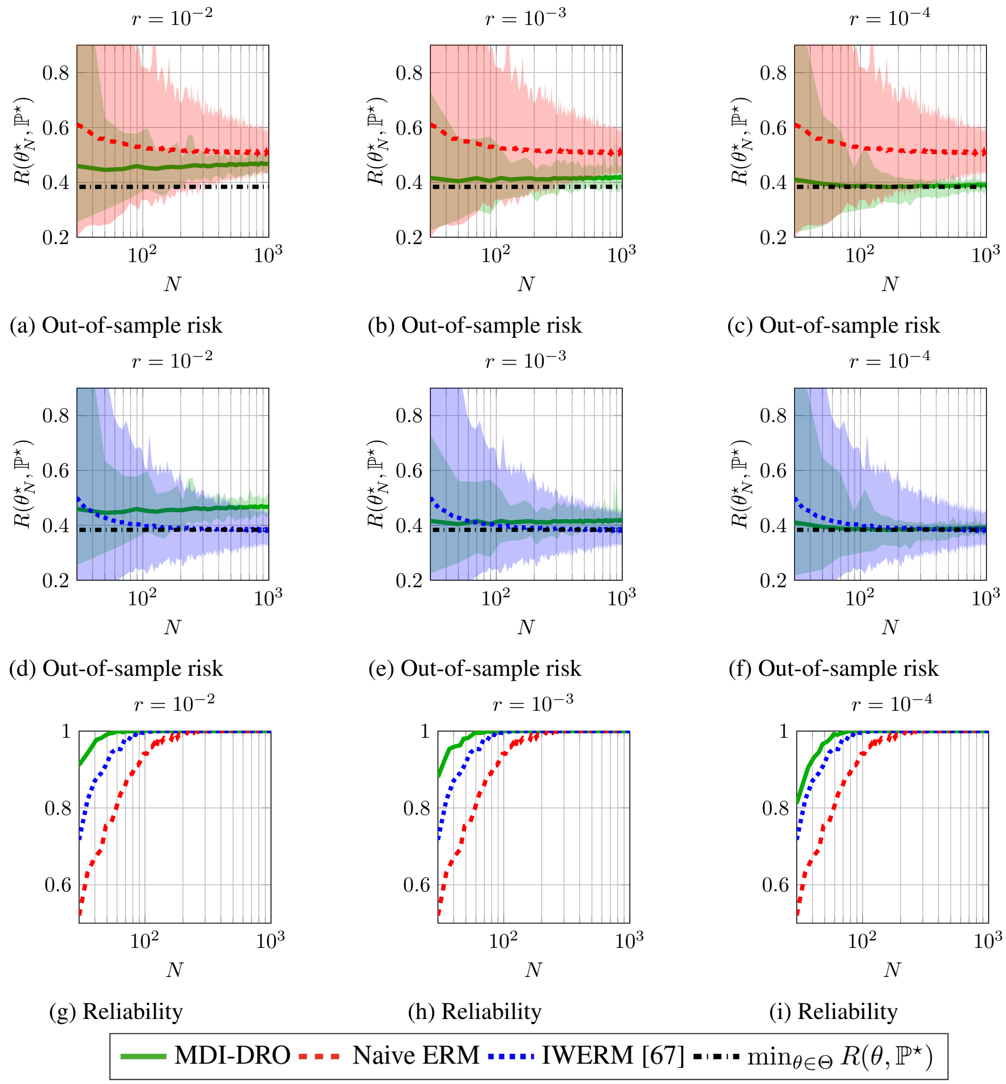


Figure 4: Additional results for the synthetic dataset with $m = 6$ (see also Figure 1). Shaded areas and lines represent ranges and mean values across 1000 independent experiments, respectively.

### 7.4 Auxiliary results for Section 6

**Classification under covariate shift.** We construct a synthetic training data consisting of feature vectors $\widehat{x}_i$ and corresponding labels $\widehat{y}_i$. Under the training distribution $\mathbb{P}$, the feature vectors are uniformly distributed on $[0,1]^{m-1}$, where $m \geq 2$, and the labels are set to $\widehat{y}_i = 1$ if $\frac{1}{m-1}\sum_{j=1}^{m-1}(\widehat{x}_i)_j > \frac{1}{2}$ and $\widehat{y}_i = -1$ otherwise. By construction, we thus have $\mathbb{E}_{\mathbb{P}}[(x,y)] = (0,0)$. The test distribution $\mathbb{P}^\star$ differs from $\mathbb{P}$. Specifically the probability density function of the features under $\mathbb{P}^\star$ is set to

$$p^\star(x) = \tfrac{2}{m-1}\textstyle\sum_{j=1}^{m-1} x_j \quad \forall x \in [0,1]^{m-1},$$

while the conditional distribution of the labels given the features is the same under $\mathbb{P}$ and $\mathbb{P}^\star$. A direct calculation then reveals that $\mathbb{E}_{\mathbb{P}^\star}[x_j] = \frac{m-2}{2(m-1)} + \frac{2}{3(m-1)} = \mu^\star > 0$ for all $j = 1, \ldots, m-1$. Similarly, one can show that $\mathbb{E}_{\mathbb{Q}}[y] > 0$. In the numerical experiments we assume that both $\mathbb{P}$ and $\mathbb{P}^\star$ are unknown. However, we assume to have access to $N$ i.i.d. samples from $\mathbb{P}$, and we assume that $\mathbb{P}^\star$ is known to satisfy $\mathbb{E}_{\mathbb{P}^\star}[\psi(\xi)] \in E$, where $\psi(x,y) = (x,y)$ and $E = [(\mu^\star - \varepsilon)\cdot 1, (\mu^\star - \varepsilon)\cdot 1]$ for some $\varepsilon > 0$ that is sufficiently small to ensure that $0 \notin E$. This implies that $\mathbb{P} \notin \Pi$.

**Inventory control model.** Consider an inventory that stores a homogeneous good, and let the state variable $s_i$ represent the stock level at the beginning of period $i$. The control action $a_i$ reflects the order quantity in period $i$, and we assume that any orders are delivered immediately at the beginning of the respective periods. The disturbance $\zeta_i$ represents an uncertain demand revealed in period $i$. We assume that the demands are i.i.d. across periods and follow a geometric distribution on $\mathbb{N} \cup \{0\}$ with success probability $\lambda \in (0, 1)$. The inventory capacity is denoted by $\gamma \in \mathbb{N}$, and any orders that cannot be stored are lost. Similarly, we assume that any demand that cannot be satisfied is also lost. The system equation describing the dynamics of the stock level is thus given by

$$s_{i+1} = \max\{0, \min\{\gamma, s_i + a_i\} - \zeta_i\} \quad \forall i = 0, 1, 2, \ldots,$$

see also [28]. Our aim is to estimate the long-run average cost generated by a prescribed ordering policy, assuming that the (uncertain) cost incurred in period $i \in \mathbb{N}$ can be expressed as

$$r(s_i, a_i, \zeta_i) = pa_i + h(s_i + a_i) - v \min\{s_i + a_i, \zeta_i\}.$$

The three terms in the above expression capture the order cost, the inventory holding cost and the profit from sales, where $p > 0$ and $h > 0$ denote the costs for ordering or storing one unit of the good, while $v > 0$ denotes the unit sales price. The expected per period cost thus amounts to

$$c(s_i, a_i) = pa_i + h(s_i + a_i) - v\tfrac{(1-\lambda)}{\lambda}\left(1 - (1-\lambda)^{(a_i + s_i)}\right).$$

The simulation results shown in Figure 3 are based on an instance of the inventory control model with state space $\mathcal{S} = \{1, 2, \ldots, 5\}$, action space $\mathcal{A} = \{1, 2, \ldots, 4\}$, and parameters $\lambda = 0.2$, $\gamma = 5$, $p = 0.6$, $h = 0.3$ and $v = 1$. The threshold for computing the modified IPS estimator is set to $\beta = 4$. It is easy to verify that, under this model parameterization, the cost function $c(s_i, a_i)$ is invertible in the sense that $s_i$ and $a_i$ are uniquely determined by $c(s_i, a_i)$; see also Example 3.6.