Robust Generalization despite Distribution Shift via Minimum Discriminating Information

Anonymous Author(s)
Affiliation
Address
email

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* [69, 51, 11]. This assumption, however, rarely holds in practice, and the performance of many models suffers in light of this issue, often called *distribution shift* [47]. 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* [54, 47]. Many other reasons lead to distribution shifts, such as non-stationary environments [58], imbalanced data [47], domain shifts [3], label shifts [73] or observed contextual information [7, 8]. 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 *behavioural policy* [64]. 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 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 [34], which can be seen

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

as a generalization of Jaynes' *maximum entropy principle* [30]. While these principles are widely used in tasks ranging from economics [26] to systems biology [55] and regularized Markov decision processes [45, 24, 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 [11]. A regularization technique that has become popular in machine learning during the last decade and provably avoids overfitting is *distributionally robust optimization (DRO)* [33].

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*, that 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 [54, 59]. 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 [54, 72, 59, 60]. 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 [14, 12, 57], for example by exponentially flattening the importance ratios [54]. For the estimation of the importance weights several methods have been presented, see for example [71]. These methods, however crucially rely on having data from both training and test distribution. For a treatment of other distribution shifts, we refer the reader to [47] 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 [38, 1, 35]. A second class of methods are based on invoking the idea of importance sampling to model the underlying distribution shift from behavioural to evaluation policy [46, 28, 66]. The third, more recent, class of methods combines the first two classes [23, 31, 67, 7].

Key reasons for the popularity of DRO in machine learning are the ability of DRO models to regularize learning problems [33, 52, 53] and the fact that the underlying optimization problems can often be exactly reformulated as finite convex programs solvable in polynomial time [4, 9]. Such reformulations hold for a variety of ambiguity sets such as: regions defined by moments [19, 25, 70, 10], ϕ-divergences [5, 40, 37], Wasserstein ambiguity sets [39, 33], or maximum mean discrepancy ambiguity sets [56, 32]. 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 [22] or label shifts [73]. To the best of our knowledge, the proposed idea of combining DRO with the principle of minimum discriminating information has not been considered yet.

3 Problem statement and motivating examples

We study learning problems of the form

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

where $R(\theta, \mathbb{P}^f) = \mathbb{E}_{\mathbb{P}^f}[L(\theta, \xi)]$ denotes the risk of an uncertain real-valued loss function $L(\theta, \xi)$ that depends on the 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 $\mathbb{P}^f$. In statistical learning, it is usually assumed that $\mathbb{P}^f$ is unknown but that we have access to independent samples from $\mathbb{P}^f$. 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 $\mathbb{P} \neq \mathbb{P}^f$ and that $\mathbb{P}^f$ 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 $\mathbb{P}^f$ with the I-projection of $\mathbb{P}$ onto Π.

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

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

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

In the following, we equip Π with the topology induced by the total variation distance. In this case, one can show that the I-projection exists whenever Π is closed [16, Theorem 2.1]. 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* $\mathrm{Prob}(y|x) = [1 + \exp(-y \cdot \theta^\top x)]^{-1}$ *parametrized by $\theta \in \mathbb{R}^{m-1}$. The maximum likelihood estimator for θ 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 w.r.t. the general population, then they are drawn from a training distribution $\mathbb{P}$ that differs from the test distribution $\mathbb{Q}$. Even if sampling from $\mathbb{Q}$ 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 $\mathbb{E}_{\mathbb{Q}}[\psi(\xi)] \in E$, where $E = [\ell, u]$, $\ell = 40$, $u = 45$ and $\psi(\xi)$ projects ξ to its 'age'-component. Other available prior information can be encoded similarly. Via the principle of minimum discriminating information, we then minimize the expected log-loss under the I-projection of the data-generating distribution $\mathbb{P}$ onto the set Π 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 $\mathbb{Q}$ thus differs from $\mathbb{P}$ and belongs to a set Π of the form (2) defined through $\psi(\xi) = \xi$ and $E = [0, (1 - \eta)\mu]$, where μ 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 of $\mathbb{P}$ onto Π.*

Loosely speaking, the principle of minimum discriminating information identifies the I-projection $f(\mathbb{P})$ as the least prejudiced and thus most natural model for $\mathbb{P}^f$ in view of the information that $\mathbb{P}^f \in \Pi$. The principle of minimum discriminating information is formally justified by the conditional limit theorem [17], 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*

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

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 Π.

Proposition 3.5 (Every distribution is an I-projection)**.** *If $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\Xi)$ such that $\mathbb{Q} \ll \mathbb{P}$ and if Π 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 π 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 π. The long-run average cost generated by π 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 μ_π on $\mathcal{S} \times \mathcal{A}$ defined through the state-action frequencies

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

see [27, Chapter 6]. One can additionally show that μ_π 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 π_μ 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_{π_e} of an evaluation policy π_e based on a trajectory of states, actions and costs generated by a behavioral policy π_b. This task can be interpreted as a degenerate learning problem without a parameter θ 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 ξ under $\mathbb{Q} = \mu_\mathsf{e} \circ c^{-1}$ given training samples from $\mathbb{P} = \mu_\mathsf{b} \circ c^{-1}$, where μ_b and μ_e represent the occupation measures corresponding to π_b and π_e, respectively. Note that $\mathbb{P}$ and $\mathbb{Q}$ are unknown because c is unknown. Moreover, as the policy π_e generates different state-action trajectories than π_b, the costs generated under π_e cannot be inferred from the costs generated under π_b even though π_b and π_e are known. Note also that $\mathbb{Q}$ coincides with the I-projection $\mathbb{P}^f$ of $\mathbb{P}$ onto the set Π defined in Proposition 3.5. The corresponding feature map ψ as well as the set E can be computed without knowledge of c provided that c is invertible. Indeed, in this case we have

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

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).

From now on we 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 of the training samples and its I-projection onto Π, 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

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

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

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

where $\mathbb{P}^f = f(\mathbb{P})$ is the I-projection of $\mathbb{P}$. If Ξ 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 Θ is a convex and conic representable set, while $L(\theta, \xi)$ is a convex and conic representable function of θ for any fixed ξ.

4 Statistical guarantees

Throughout this section, we assume that the sets Θ and Ξ are compact and that the risk $R : \Theta \times \Pi \to \mathbb{R}$ is a continuous function to avoid technical discussions of little practical relevance. 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 θ 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

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

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 θ across a relative entropy ball of radius r, almost satisfies (8) and is thus essentially admissible.

Theorem 4.1 (Out-of-sample guarantee)**.** *If $R^\star$ is defined as in (5) and $\varepsilon > 0$, then $\widehat{R} = R^\star + \varepsilon$ is a continuous function and represents an admissible data-driven predictor.*

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)**.** *If $R^\star$ is defined as in (5) and $\widehat{R}$ is any admissible data-driven predictor, then we have $\lim_{N\to\infty} J_N^\star \leq \lim_{N\to\infty} \widehat{J}_N$ $\mathbb{P}$-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 [6, pp. 115–116], which applies because $\widehat{R}$ is continuous and Θ is compact. Second, the I-projection $\mathbb{P}^f = f(\mathbb{P})$ is continuous in $\mathbb{P} \in \mathcal{P}(\Xi)$ thanks to [61, Theorem 9.17], which applies because the relative entropy is strictly convex in its first argument [20, 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 Ξ 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*

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

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)**.** *If the assumptions of Proposition 3.4 hold and $\{r_N\}_{N\in\mathbb{N}}$ is a sequence of non-negative numbers with $\lim_{N\to\infty} r_N = 0$, then the DRO predictor satisfies*

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

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

We now continue with the off-policy evaluation example introduced in Section 3 and show the corresponding DRO approach and its statistical guarantees.

Example 4.5 (Off-policy evaluation)**.** *For the OPE problem introduced in Example 3.6, we aim to construct an estimator for the performance of the evaluation policy $V_{\pi_e}(c) = \mathbb{E}_{f_\Pi(\mathbb{P})}[\xi]$ based on the available behavioural policy and its incurred cost. As described in Example 3.6, we choose Π such that $\mu_e \circ c^{-1} = f(\mathbb{P})$, where $\mathbb{P} = \mu_b \circ c^{-1} \in \mathcal{P}(\Xi)$. Given the behavioural data $(\widehat{s}_t, \widehat{a}_t) \sim \mu_b$ for $t = 0, \ldots N-1$, consider the empirical counterpart of $\mathbb{P}$ as $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{t=0}^{N-1} \delta_{c(\widehat{s}_t, \widehat{a}_t)}$. While we assume in this paper that the samples $(\widehat{s}_t, \widehat{a}_t)$ are i.i.d., the underlying large deviation framework used in principle allows for a generalization to a single trajectory of correlated data [63, 37]. The proposed approximation of the value function under the evaluation policy V_{π_e} is provided by $J_N^\star = R^\star(\widehat{\mathbb{P}}_N^f)$, where $R^\star$ is the DRO predictor (5), the admissibility guarantees provided by Theorem 4.1 using the fact that Ξ is finite provide the generalization bound*

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

that holds for all $N \in \mathbb{N}$.

5 Efficient computation

Motivated by the generalization and optimality guarantees provided by Theorems 4.1 and 4.2, we now discuss how to efficiently compute the corresponding optimal parameter $\theta_N^\star = \arg\min_{\theta\in\Theta} R^\star(\theta, \widehat{\mathbb{P}}_N^f)$. This computation can be split into two steps: First, we aim to efficiently compute the estimator $\widehat{\mathbb{P}}_N^f$ given the training data $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ and the corresponding empirical probability measure $\widehat{\mathbb{P}}_N$. This boils down to evaluating the I-projection $\widehat{\mathbb{P}}_N^f = f(\widehat{\mathbb{P}}_N)$ for a given set Π of the form (2). Given $\widehat{\mathbb{P}}_N^f$, we then show how to compute $R^\star(\theta, \widehat{\mathbb{P}}_N^f)$ and the corresponding optimizer $\theta_N^\star$.

Computation of I-projection. Computing the I-projection of an empirical probability measure $\widehat{\mathbb{P}}_N$ built from the available data is a non-trivial task as it requires solving an infinite-dimensional optimization problem (3). Generally, one would expect that the difficulty of computing $f(\cdot)$ also depends on the structure of the set Π expressed via ψ and E, see (2). Following recent work [62], we show that for a large class of sets Π, by exploiting the fact that $\widehat{\mathbb{P}}_N$ is finitely supported and by using recent advances in convex optimization, $f(\cdot)$ can be computed in an efficient way.

Let $\eta = (\eta_1, \eta_2)$ be a smoothing parameter with $\eta_1, \eta_2 > 0$, and let $L_\eta > 0$ be a learning rate that may depend on η. In addition, define a function $G_\eta : \mathbb{R}^d \to \mathbb{R}^d$ through

$$G_\eta(z) = -\pi_E(\eta_1^{-1} z) - \eta_2 z + \frac{\sum_{j=1}^N \psi(\xi_j) \exp\left(-\sum_{i=1}^d z_i\, \psi_i(\xi_j)\right)}{\sum_{j=1}^N \exp\left(-\sum_{i=1}^d z_i\, \psi_i(\xi_j)\right)}, \tag{12}$$

where π_E is the projection operator onto the set E defined as $\pi_E(z) = \arg\min_{x\in E} \|x - z\|_2^2$. Given the function G_η, the I-projection can be computed via Algorithm 1, basically a fast gradient method.

The complexity of evaluating the function G_η, as required by Algorithm 1, is determined by the projection operator onto E; for simple sets (e.g., 2-norm balls, hybercubes) the solution is analytically available, while for more general cases (e.g., simplex, 1-norm balls) it can be computed at relatively low computational effort, see [48, Section 5.4] for a comprehensive survey.

Algorithm 1: Optimal scheme for smooth & strongly convex optimization [43]

Choose $w_0 = y_0 \in \mathbb{R}^d$ and $\eta \in \mathbb{R}^2_{++}$

For $k \geq 0$ **do**

Step 1: Set $y_{k+1} = w_k + \frac{1}{L_\eta} G_\eta(w_k)$

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

The guarantees of Algorithm 1 require the following assumption on the underlying data-generating distribution and on the set Π.

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 $\mathbb{P}^\circ$ such that Assumption 5.1 holds, in general may be difficult. A constructive approach to find such an interior point, when ψ represents a polynomial is described in [62, Remark 8]. Given Assumption 5.1, for $\varepsilon > 0$ define

$$\begin{aligned}
&C = \mathsf{D}(\mathbb{P}^\circ \| \widehat{\mathbb{P}}_N), \qquad D = \tfrac{1}{2} \max_{y \in E} \|y\|_2, \qquad \eta_1 = \tfrac{\varepsilon}{4D}, \qquad \eta_2 = \tfrac{\varepsilon \delta^2}{2C^2}, \\
&\alpha = \sup_{\lambda \in \mathbb{R}^d, \mathbb{P} \in \mathcal{P}(\Xi)} \left\{ \lambda^\top \textstyle\int_\Xi \psi(\xi) \mathrm{d}\mathbb{P}(\xi) \ : \ \|\lambda\|_2 = 1 \right\}, \quad L_\eta = 1/\eta_1 + \eta_2 + \left(\sum_{i=1}^d (2D)^i \right)^2, \\
&M_1(\varepsilon) = 2 \left(\sqrt{\tfrac{8DC^2}{\varepsilon^2 \delta^2} + \tfrac{2\alpha^2 C^2}{\varepsilon \delta^2} + 1} \right) \log \left(\tfrac{10(\varepsilon + 2C)}{\varepsilon} \right), \qquad (13) \\
&M_2(\varepsilon) = 2 \left(\sqrt{\tfrac{8DC^2}{\varepsilon^2 \delta^2} + \tfrac{2\alpha^2 C^2}{\varepsilon \delta^2} + 1} \right) \log \left(\tfrac{C}{\varepsilon \delta (2 - \sqrt{3})} \sqrt{4 \left(\tfrac{4D}{\varepsilon} + \alpha^2 + \tfrac{\varepsilon \delta^2}{2C^2} \right) \left(C + \tfrac{\varepsilon}{2} \right)} \right).
\end{aligned}$$

Due to the compactness of Ξ, when ψ is a continuous function the parameter α is finite. Indeed let $K \in \mathbb{R}$ be such that $(\psi(\xi))_i \leq K$ for all $\xi \in \Xi$ and $i = 1, \ldots, d$, then $\alpha \leq \sqrt{d} K$.

Theorem 5.2 (Almost linear convergence rate)**.** *Given Assumption 5.1 and the definitions (13), let* $\varepsilon > 0$ *and* $M(\varepsilon) = \lceil \max\{M_1(\varepsilon), M_2(\varepsilon)\} \rceil$*. Then,* $k = M(\varepsilon)$ *iterations of Algorithm 1 provide*

$$\widehat{z}_{k,\eta} = y_k \quad \textit{and} \quad \widehat{\mu}_{k,\eta}(B) = \frac{\sum_{j=1}^N \mathbf{1}_{\xi_j \in B} \exp\left(- \sum_{i=1}^d (\widehat{z}_{k,\eta})_i \psi_i(\xi_j) \right)}{\sum_{j=1}^N \exp\left(- \sum_{i=1}^d (\widehat{z}_{k,\eta})_i \psi_i(\xi_j) \right)} \quad \forall B \subset \Xi \textit{ measurable}, \qquad (14)$$

which satisfy

$$\begin{aligned}
&\varepsilon\textit{-optimality:} && |\mathsf{D}(\widehat{\mu}_{k,\eta} \| \widehat{\mathbb{P}}_N) - \mathsf{D}(\widehat{\mathbb{P}}^f_N \| \widehat{\mathbb{P}}_N)| \leq 2(1 + 2\sqrt{3})\varepsilon, && (15a) \\
&\varepsilon\textit{-feasibility:} && \mathsf{d}\left(\textstyle\int_\Xi \psi(\xi) \mathrm{d}\widehat{\mu}_{k,\eta}(\xi), E \right) \leq \tfrac{2\varepsilon\delta}{C}, && (15b) \\
&\varepsilon\textit{-optimizer:} && \|\widehat{\mu}_{k,\eta} - \widehat{\mathbb{P}}^f_N\|^2_{\mathsf{TV}} \leq 2(1 + 2\sqrt{3})\varepsilon, && (15c)
\end{aligned}$$

where $\mathsf{d}(\cdot, E)$ *denotes the distance to the set* E*, i.e.,* $\mathsf{d}(x, E) = \min_{y \in E} \|x - y\|_2$.

Theorem 5.2 directly implies that we need at most $O(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon})$ iterations of Algorithm 1 to achieve an ε-approximation to $\widehat{\mathbb{P}}^f_N$ that is also ε-feasible with respect to Π. While Assertions (15a) and (15b) are closely related to [62], Assertion (15c) to the best of our knowledge is new and actually a crucial property for numerically computing $R^\star(\theta, \widehat{\mathbb{P}}^f_N)$.

Computation of DRO predictor. Equipped with Algorithm 1 to efficiently approximate $\widehat{\mathbb{P}}^f_N$ via $\widehat{\mu}_{k,\eta}$, the DRO predictor $R^\star(\theta, \widehat{\mathbb{P}}^f_N)$, defined in (4) can be computed/approximated by $R^\star(\theta, \widehat{\mu}_{k,\eta})$ since the function $R^\star$ is continuous. The optimization problem $R^\star(\theta, \widehat{\mu}_{k,\eta})$ admits a dual representation which follows as a special case from [68, Proposition 5].

Proposition 5.3 (DRO duality). *If $r > 0$ and $\bar{L}(\theta) = \sup_{\xi \in \Xi} L(\theta, \xi)$ is the worst-case loss function, then the DRO predictor (5) evaluated at the approximate I-projection $\widehat{\mu}_{k,\eta}$ given by (14) in Theorem 5.2 admits a dual formulation*

$$R^\star(\theta, \widehat{\mu}_{k,\eta}) = \min_{\alpha \geq \bar{L}(\theta)} \alpha - e^{-r} \prod_{j=1}^{N} (\alpha - L(\theta, \xi_j))^{\gamma_j} \tag{16}$$

where $\gamma_j = \exp(-\sum_{i=1}^{d} (\widehat{z}_{k,\eta})_i \psi_i(\xi_j)) (\sum_{j=1}^{N} \exp(-\sum_{i=1}^{d} (\widehat{z}_{k,\eta})_i \psi_i(\xi_j)))^{-1}$.

For a fixed $\theta \in \Theta$, Proposition 5.3 shows that the data-driven predictor $R^\star(\theta, \widehat{\mu}_{k,\eta})$ is equivalent to a one-dimensional convex problem and as such can be computed via bisection or other line search methods. Since the measure $\widehat{\mu}_{k,\eta}$ is finitely supported, we can express the cost function $R^\star(\theta, \widehat{\mu}_{k,\eta})$ as a second-order cone program involving $O(N)$ constraints and auxiliary variables, see [44, Section 6.2.3.5]. Therefore, in the case where $L(\theta, \xi)$ is a convex and conic representable function of θ for fixed ξ and Θ is convex and conic representable, the optimization problem $\min_{\theta \in \Theta} R^\star(\theta, \widehat{\mu}_{k,\eta})$ can be expressed as a tractable convex optimization problem.

6 Experimental results

We focus on two of our running examples and show how the proposed MDI-DRO method performs empirically.[1] We first consider two experiments on training a classifier on systematically biased data in the setting introduced in Example 3.2.

Synthetic dataset — covariate shift adaptation. We consider a synthetic dataset involving a covariate shift, where the details are provided in Appendix 7.4. In the numerical experiments, we observe that the proposed MDI-DRO method significantly outperforms ERM both in terms of expected out-of-sample risk as well as in terms of the corresponding smaller variance, see Figures 1a, 1b. We then compare MDI-DRO to the IWERM method, which accounts for the underlying distribution shift via the importance weights $p_{\mathsf{te}}(\cdot)/p_{\mathsf{tr}}(\cdot)$ that we assume to know. In contrast, MDI-DRO does *not* require any knowledge from the test distribution other than the function ψ and the set E. Nevertheless, MDI-DRO shows similar out-of-sample performance than IWERM despite the lack of information, and even achieves lower variance than IWERM, see Figures 1c, 1d. Figure 1e assesses the reliability of the upper confidence bound $J_N^\star$ and the out-of-sample risk $R(\theta_N^\star, \mathbb{P}_{\mathsf{te}})$ with respect to the regularization parameter r. In the appendix (see Figure 4) we provide additional figures showing the known tradeoff stating that a small regularization parameter r leads to small out-of-sample risk, while the reliability of the upper confidence bound $J_N^\star$ grows with r.

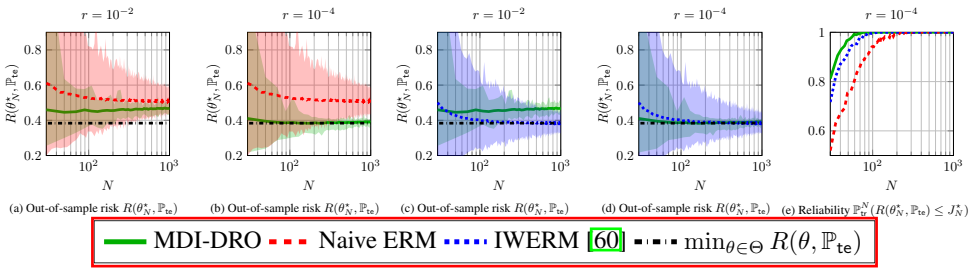


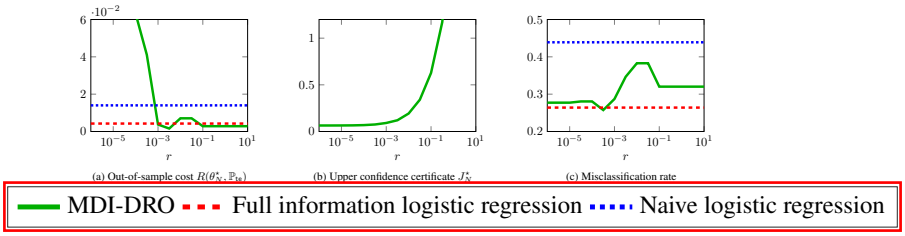
Figure 1: Synthetic dataset example for $m = 6$, $\varepsilon = 0.01$. The colored tubes represent the 100% confidence intervals of 1000 independent experiments and the lines the corresponding means.

Real-world data — classification under sample bias. We consider the heart disease classification task (cf. Example 3.2) based on a real-world dataset[2] consisting of i.i.d. samples from some unknown distribution $\mathbb{P}_{\mathsf{te}}$. To simulate the data shift, we consider training based on a *biased* subset (training data) of this data $\{(\widehat{x}_1, \widehat{y}_1), \ldots, (\widehat{x}_N, \widehat{y}_N)\}, N < N_{\mathsf{te}}$, where male patients older than 60 years are

[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/pmdidro/PMDI_DRO`

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

substantially over-represented. That is, we assume that the training data are distributed according to $\mathbb{P}_{\mathsf{tr}}$, which is different from the test distribution. While the test distribution $\mathbb{P}_{\mathsf{te}}$ is unknown, we assume that we have access to the empirical mean of the entire dataset $m = \frac{1}{N_{\mathsf{te}}} \sum_{i=1}^{N_{\mathsf{te}}} (\widehat{x}_i, \widehat{y}_i) \in \mathbb{R}^m$. To use our proposed modelling framework via the set Π in (2), we define $E = [m - \varepsilon 1, m + \varepsilon 1]$ for some $\varepsilon > 0$ and the function $\psi(x, y) = (x, y)$. We compare the proposed MDI-DRO method for classification with a "naive" logistic regression not accounting for the sample bias. In addition, we use as benchmark a logistic regression model on the entire dataset. Figure 2a displays the out-of-sample cost, Figure 2b shows the upper confidence certificate $J_N^\star$ and Figure 2c compares the misclassification rates of the different methods. Perhaps surprisingly, for a careful selection of the radius r the proposed method shows comparable classification performance to an *in-sample* logistic regression method based on the full knowledge of the entire dataset.



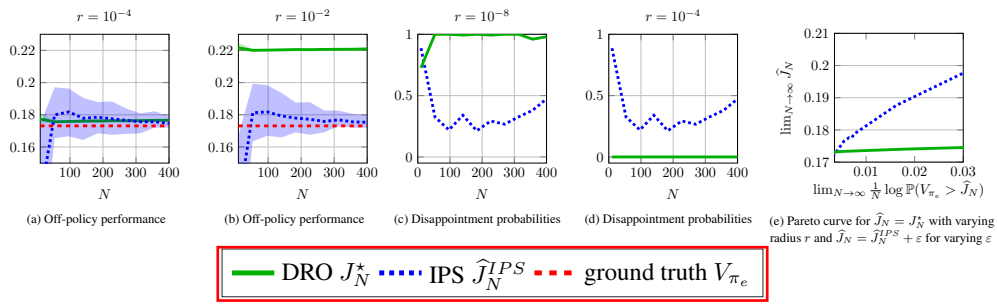
(a) Out-of-sample cost $R(\theta_N^\star, \mathbb{P}_{\mathsf{te}})$ (b) Upper confidence certificate $J_N^\star$ (c) Misclassification rate

Figure 2: Heart disease classification example for $m = 6$, $N = 20$, $N_{\mathsf{te}} = 303$, $\varepsilon = 10^{-3}$.

OPE for MDPs — inventory control example. We consider the OPE setting introduced in Examples 3.6 and 4.5. A common estimator for V_{π_e} is the inverse propensity estimator [50]

$$\widehat{J}_N^{IPS} = \frac{1}{N} \sum_{t=1}^{N} c(\widehat{s}_t, \widehat{a}_t) \frac{\mu_{\mathsf{e}}(\widehat{s}_t, \widehat{a}_t)}{\mu_{\mathsf{b}}(\widehat{s}_t, \widehat{a}_t)}, \qquad \mathbb{P}\left(V_{\pi_e} \leq \widehat{J}_N^{IPS} + \varepsilon\right) \geq 1 - e^{\frac{-2N\varepsilon^2}{b^2}}, \tag{17}$$

where the concentration bound is an application of Hoeffding's inequality and holds for any $\varepsilon > 0$ and $N \in \mathbb{N}$ where $b = \max_{s \in \mathcal{S}, a \in \mathcal{A}} c(s, a) \mu_{\mathsf{e}}(s, a) / \mu_{\mathsf{b}}(s, a)$ is typically large and as such the finite sample bounds of $J_N^\star$ provided by (11) are often more informative than (17). There are various approaches to address the unboundedness of the variance of $\widehat{J}_N^{IPS}$ [14, 12, 57] and the simplest option [65] is to cap the importance weights, which however then introduces a bias. We evaluate the performance of our proposed off-policy evaluation method on a classical inventory control problem (see Appendix 7.4 for a detailed description). We choose an evaluation policy π_{e} and a behaviour policy π_{b} at random. The decision maker then has access to the evaluation policy π_{e} and i.i.d. state action pairs $\{\widehat{s}_t, \widehat{a}_t\}_{t=1}^N$ samples according to μ_{b} as well as the observed empirical costs $\{\widehat{c}_t\}_{t=1}^N$. Figure 3 shows the results. The proposed MDI-DRO method is compared against the inverse propensity approach and the ground truth in terms of off-policy evaluation performance, see Figures 3a, 3b. For a small radius, MDI-DRO outperforms the IPS in terms of mean as well as variance. Figures 3c and 3d displays the disappointment probabilities $\mathbb{P}(V_{\pi_{\mathsf{e}}} > J_N^\star)$ and $\mathbb{P}(V_{\pi_{\mathsf{e}}} > \widehat{J}_N^{IPS})$, and confirms our theoretical result from Theorem 4.1 stating that for a larger radius r the disappointment probability decays faster. Figure 3e visualizes the statistical efficiency described in Theorem 4.2.



(a) Off-policy performance (b) Off-policy performance (c) Disappointment probabilities (d) Disappointment probabilities (e) Pareto curve for $\widehat{J}_N = J_N^\star$ with varying radius r and $\widehat{J}_N = \widehat{J}_N^{IPS} + \varepsilon$ for varying ε

Figure 3: The colored tubes are 90% confidence intervals. The numerical parameters used were $\lambda = 0.2, v = 1, p = 0.6, v = 0.3, \gamma = 5, \mathcal{S} = \{1, 2, \ldots, 6\}, \mathcal{A} = \{1, 2, \ldots, 4\}$.

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**, volume 19, 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] C. Berge. **Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity**. Courier Corporation, 1997.

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

[8] D. Bertsimas and B. V. Parys. Bootstrap robust prescriptive analytics. **arXiv preprint arXiv:1711.09974**, 2017.

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

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

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

[12] 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.

[13] G. C. Calafiore. Ambiguous risk measures and optimal robust portfolios. **SIAM Journal on Optimization**, 18(3):853–877, 2007.

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

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

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

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

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

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

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

[21] 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.

[22] J. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. **arXiv preprint, arXiv.1810.08750**, 2020.

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

[24] M. Geist, B. Scherrer, and O. Pietquin. A theory of regularized markov decision processes. In **Proceedings of the 36th International Conference on Machine Learning, ICML**, volume 97 of **Proceedings of Machine Learning Research**, pages 2160–2169. PMLR, 2019.

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

[26] A. Golan. Information and entropy econometrics: Review and synthesis. **Foundations and Trends in Econometrics**, 2(1-2):1–145, 2008.

[27] O. Hernández-Lerma and J. Lasserre. **Discrete-Time Markov Control Processes: Basic Optimality Criteria**. Applications of Mathematics Series. Springer, 1996.

[28] 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.

[29] 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.

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

[31] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In **Proceedings of The 33rd International Conference on Machine Learning**, volume 48 of **Proceedings of Machine Learning Research**, pages 652–661. PMLR, 2016.

[32] J. Kirschner, I. Bogunovic, S. Jegelka, and A. Krause. Distributionally robust bayesian optimization. In **Artificial Intelligence and Statistics**, pages 2174–2184, 2020.

[33] 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**, 2019.

[34] S. Kullback. **Information Theory and Statistics**. Wiley publication in mathematical statistics. Wiley, 1959.

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

[36] H. Lam. Robust sensitivity analysis for stochastic systems. **Mathematics of Operations Research**, 41(4):1248–1275, 2016.

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

[38] 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.

[39] 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.

[40] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In **Advances in Neural Information Processing Systems**, volume 29, pages 2208–2216, 2016.

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

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

[43] Y. Nesterov. **Introductory Lectures on Convex Optimization: A Basic Course**. Springer, 1 edition, 2014.

[44] Y. Nesterov and A. Nemirovskii. **Interior-Point Polynomial Algorithms in Convex Programming**, volume 13 of **Studies in Applied and Numerical Mathematics**. SIAM, 1994.

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

[46] D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In **Proceedings of the Seventeenth International Conference on Machine Learning**, ICML '00, pages 759–766, 2000.

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

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

[49] R. T. Rockafellar. **Convex analysis**. Princeton Landmarks in Mathematics and Physics Series. Princeton University Press, 1997.

[50] 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.

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

[52] S. Shafieezadeh Abadeh, P. M. Mohajerin Esfahani, and D. Kuhn. Distributionally robust logistic regression. In **Advances in Neural Information Processing Systems**, volume 28, pages 1576–1584, 2015.

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

[54] 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.

[55] 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.

[56] M. Staib and S. Jegelka. Distributionally robust optimization and generalization in kernel methods. In **Advances in Neural Information Processing Systems**, volume 32, pages 9134–9144, 2019.

[57] A. L. Strehl, J. Langford, L. Li, and S. M. Kakade. Learning from logged implicit exploration data. In **NIPS**, pages 2217–2225, 2010.

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

[59] M. Sugiyama and K.-R. Müller. Input-dependent estimation of generalization error under covariate shift. **Statistics & Decisions**, 23:249–279, 01 2005.

[60] 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.

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

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

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

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

[65] A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In **Proceedings of the 32nd International Conference on Machine Learning**, volume 37 of **Proceedings of Machine Learning Research**, pages 814–823. PMLR, 2015.

[66] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In **Advances in Neural Information Processing Systems**, volume 28, 2015.

[67] P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In **Proceedings of The 33rd International Conference on Machine Learning**, volume 48 of **Proceedings of Machine Learning Research**, pages 2139–2148. PMLR, 2016.

[68] B. Van Parys, P. Mohajerin Esfahani, and D. Kuhn. From data to decisions: Distributionally robust optimization is optimal. **Management Science**, 2021. Articles in Advance.

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

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

[71] 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**, volume 24, pages 594–602, 2011.

[72] B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In **Proceedings of the Twenty-First International Conference on Machine Learning**, page 114, 2004.

[73] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We address the limitations and potentially restrictive assumptions transparently throughout the paper. For example, consider the i.i.d. assumption in the off-policy evaluation scheme (Examples 3.6 and 4.5). Another example is the derivation of a Slater point, see the discussion around Assumption 5.1
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See for example Theorem's 4.4, 5.2 or Proposition 3.4.
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are provided in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see the details provided in Section 6 and in the appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] This is all explained either in the text or in the code provided.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We provide confidence intervals for our experimental results.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 6.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]