Proceedings of Machine Learning Research vol 178:1–26, 2022 35th Annual Conference on Learning Theory

Minimax Regret Optimization for Robust Machine Learning under Distribution Shift

Alekh Agarwal ALEKHAGARWAL@GOOGLE.COM
Google Research

Tong Zhang TOZHANG@GOOGLE.COM
Google Research and HKUST

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

In this paper, we consider learning scenarios where the learned model is evaluated under an unknown test distribution which potentially differs from the training distribution (i.e. distribution shift). The learner has access to a family of weight functions such that the test distribution is a reweighting of the training distribution under one of these functions, a setting typically studied under the name of Distributionally Robust Optimization (DRO). We consider the problem of deriving regret bounds in the classical learning theory setting, and require that the resulting regret bounds hold uniformly for all potential test distributions. We show that the DRO formulation does not guarantee uniformly small regret under distribution shift. We instead propose an alternative method called Minimax Regret Optimization (MRO), and show that under suitable conditions, this method achieves *uniformly low regret* across all test distributions. We also adapt our technique to have strong guarantees when the test distributions are heterogeneous in their similarity to the training data. Given the widespead optimization of worst case risks in current approaches to robust machine learning, we believe that MRO can be an attractive framework to address a broad range of distribution shift scenarios.

Keywords: Distribution shift, covariate shift, distributionally robust learning

1. Introduction

Learning good models for scenarios where the evaluation happens under a test distribution that differs from the training distribution has become an increasingly important problem in machine learning and statistics. It is particularly motivated by the widespread practical deployment of machine learning models, and the observation that model performance deteriorates when the test distribution differs from the training distribution. Some concrete scenarios where these issues are particularly salient range from class imbalance between training and test data (Galar et al., 2011; Japkowicz, 2000) to algorithmic fairness (Dwork et al., 2012; Barocas and Selbst, 2016). Consequently, several recent formalisms have been proposed to develop algorithms for such settings, such as adversarial training (Goodfellow et al., 2014; Madry et al., 2017; Wang et al., 2019), Invariant Risk Minimization (Arjovsky et al., 2019; Ahuja et al., 2020a; Rosenfeld et al., 2020) and Distributionally Robust Optimization (Duchi and Namkoong, 2021; Namkoong and Duchi, 2016; Staib and Jegelka, 2019; Kuhn et al., 2019). While the specifics differ across these formulations, at a high-level they all minimize the worst-case risk of the learned model across some family of test distributions, and spend considerable effort on finding natural classes of test distributions and designing efficient algorithms to optimize the worst-case objectives.

In this paper, we consider this problem from a learning theory perspective, by asking if it is possible to obtain regret bounds that hold uniformly across all test distributions. Recall that in supervised learning, we are given a function class $\mathcal{F}$ and a loss function $\ell(z, f(z))$, where z denotes a sample. For a test distribution P, we are interested in learning a prediction function $f \in \mathcal{F}$ using the training data, so that the risk of f under P: $R_P(f) = \mathbb{E}_{z \sim P} \ell(z, f(z))$, is small. In classical learning theory, an estimator $\hat{f}$ is learned from the training data, and we are interested in bounding the regret (or excess risk) of f as follows

$$\text{Regret}_P(\hat{f}) = R_P(\hat{f}) - \inf_{f \in \mathcal{F}} R_P(f) = R_P(\hat{f}) - R_P(f_P),$$

where f_P minimizes $R_P(f)$ over $\mathcal{F}$. If the training data $(z_1, \ldots, z_n)$ are drawn from the same distribution as the test distribution P, then standard results imply that the empirical risk minimizaiton (ERM) method, which finds an $\hat{f}$ with a small training error, achieves small regret. A natural question we would like to ask in this paper is whether there exists an estimator that allows us to achieve small regret with distribution shift? More precisely, suppose we are given a family of distributions $\mathcal{P}$ with $P_0 \in \mathcal{P}$. Given access to n samples $(z_1, \ldots, z_n)$ from P_0, we would like to find an estimator $\hat{f}$ from the training data so that the uniform regret

$$\sup_{P \in \mathcal{P}} \text{Regret}_P(\hat{f}) = \sup_{P \in \mathcal{P}} \left[R_P(\hat{f}) - \inf_{f \in \mathcal{F}} R_P(f) \right] \tag{1}$$

is small. Here the model class $\mathcal{F}$ can be statistically misspecified in that it may not contain the optimal model $f^\star$ that attains the Bayes risk in a pointwise manner.

We note that this objective is different from the objective of Distributionally Robust Optimization (henceforth DRO), which directly minimizes the risk across all test distributions as follows:

$$f_{\text{DRO}} = \operatorname*{arginf}_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}} R_P(f). \tag{2}$$

We observe that the objective (1) is less sensitive than DRO to heterogeneity in the amount of noise in test distributions. Moreover, when the model is well-specified, our criterion directly measures the closeness of the estimator $\hat{f}$ and the optimal model $f^\star$ for each distribution $P \in \mathcal{P}$, which is often desirable in applications. We call the criterion to minimize regret uniformly across test distributaions Minimax Regret Optimization (MRO), and its population formulation seeks to minimize the worst-case regret (1):

$$f_{\text{MRO}} = \operatorname*{arginf}_{f \in \mathcal{F}} \sup_{P \in \mathcal{P}} \text{Regret}_P(f). \tag{3}$$

Compared to DRO, MRO evaluates the regret of a candidate model f on each distribution $P \in \mathcal{P}$ as opposed to the raw risk. As we show in the following sections, regret is comparable across distribution in a more robust manner than the risk, since the subtraction of the minimum risk for each distribution takes away the variance due to noise. We note that a similar approach to remove the residual variance to deal with heterogenous noise arises in the offline Reinforcement Learning setting (Antos et al., 2008), which we comment more on in the discussion of related work. Similar to ERM, which replaces the regret minimization problem over a single test distribution by its empirical minimization counterpart, we consider an empirical minimization counterpart of MRO, and analyze its generalization performance.

Our contributions. With this context, our paper makes the following contributions.

1. The objective (3), which gives an alternative to DRO, alleviates its shortcomings related to noise overfitting, and provides a robust learning formulation for problems with distribution shift.
2. We show the empirical couterpart to (3) can be analyzed using classical tools from learning theory. When the loss function is bounded and Lipschitz continuous, we show that the regret of our learned model $\hat{f}_{\text{MRO}}$ on each distribution P can be bounded by that of f_{MRO} along with a deviation term that goes down as $\mathcal{O}(1/\sqrt{n})$. Concretely, we show that if $\sup_{z\,:\,dP_0(z)>0} dP(z)/dP_0(z) \le B$ for all $P \in \mathcal{P}$, then with high probability,
$$\sup_{P\in\mathcal{P}} \text{Regret}_P(\widehat{f}_{\text{MRO}}) \le \inf_{f\in\mathcal{F}} \sup_{P\in\mathcal{P}} \text{Regret}_P(f) + \mathcal{O}\left(\frac{B}{\sqrt{n}}\right),$$
where we omit the dependence on the complexities of $\mathcal{F}$ and $\mathcal{P}$ as well as the failure probability, and use a simplified assumption on dP/dP_0 than in our formal result of Theorem 2, for ease of presentation. The bound stipulates that $\widehat{f}_{\text{MRO}}$ has a bounded regret on each distribution in $\mathcal{P}$, so long as there is at least one function in our class with a uniformly small regret. For squared loss, if the class $\mathcal{F}$ is convex, we show that our rates improve to $\mathcal{O}(B/n)$ (Theorem 4). The result follows naturally as the empirical regret concentrates to population regret at a fast rate in this setting. We also show that the worst-case risk of $\widehat{f}_{\text{DRO}}$ cannot approach the worst case risk of f_{DRO} at a rate faster than $1/\sqrt{n}$ (Proposition 6), which demonstrates a theoretical benefit of our selection criterion.
3. We present SMRO, an adaptation of our basic technique which further rescales the regret of each distribution differently. We show that SMRO retains the same worst-case guarantees as MRO, but markedly improves upon it when the distributions $P \in \mathcal{P}$ satisfy an *alignment* condition, which includes all well-specified settings (Theorems 7 and 8). The rescaling parameters are fully estimated from data.
4. Algorithmically, we show that our method can be implemented using access to a (weighted) ERM oracle for the function class $\mathcal{F}$, using the standard machinery of solving minimax problems as a two player game involving a no-regret learner and a best response player. The computational complexity scales linearly in the size of $\mathcal{P}$ and as n^2 in the number of samples.

We conclude this section with a discussion of related work.

1.1. Related Work

Distributional mismatch between training and test data has been studied in many settings (see e.g. (Quinonero-Candela et al., 2009)) and under several notions of mismatch. A common setting is that of covariate shift, where only the distribution of labels can differ between training and test (Shimodaira, 2000; Huang et al., 2006; Bickel et al., 2007). Others study changes in the proportions of the label or other discrete attributes between training and test (Dwork et al., 2012; Xu et al., 2020). More general shifts are considered in domain adaptation (Mansour et al., 2009; Ben-David et al., 2010; Patel et al., 2015) and transfer learning (Pan and Yang, 2009; Tan et al., 2018).

A recent line of work on Invariant Risk Minimization (IRM) (Arjovsky et al., 2019) considers particularly broad generalization to unseen domains from which no examples are available at the training time, motivated by preventing the learning of spurious features in ERM based methods. The idea is to find an estimator which is invariant across all potential test distributions. Although most empirical works on IRM use gradient-penalty based formulations for algorithmic considerations, one interpretation of IRM leads to a minimax formulation similar to the MRO objective (3):

$$f_{\text{IRM}} = \operatorname*{arginf}_{f \in \mathcal{F}} R_{P_0}(f) \quad \text{s.t.} \sup_{P \in \mathcal{P}} R_P(f) - R_P(f_P) \leq \epsilon, \tag{4}$$

where the family $\mathcal{P}$ consists of all individual domains in the training data and P_0 is the entire training data pooled across domains. However, it is not obvious how to meaningfully analyze the distributional robustness of IRM method (4) in the classical learning theory setting, and some subsequent works (Ahuja et al., 2020b; Kamath et al., 2021) formalize IRM as minimizing a DRO style worst-case risk objective, instead of regret as in the original paper. In contrast, our MRO formulation is motivated by extending regret-driven reasoning to settings with distribution shift.

Distributionally robust optimization, which forms a starting point for this work, has a long history in the optimization literature (see e.g. (Ben-Tal et al., 2009; Shapiro, 2017)) and has gained recent prominence in the machine learning literature (Namkoong and Duchi, 2016; Duchi and Namkoong, 2021; Duchi et al., 2021; Staib and Jegelka, 2019; Kuhn et al., 2019; Zhu et al., 2020). DRO has been applied with mixed success in language modeling (Oren et al., 2019), correcting class imbalance (Xu et al., 2020) and group fairness (Hashimoto et al., 2018). For most of the theoretical works, the emphasis is on efficient optimization of the worst-case risk objective. While Duchi and Namkoong (2021) provide sharp upper and lower bounds on the risk of the DRO estimator, they do not examine regret, which is the central focus of our study.

We note that the raw risk used in DRO is sensitive to heterogeneous noise, in that larger noise leads to larger risk. Such sensitivity can be undesirable for some applications. Challenges in learning across scenarios with heterogeneous noise levels has been previously studied in supervised learning (Crammer et al., 2005), and also arise routinely in reinforcement learning (RL). In the setting of offline RL, a class of techniques is designed to minimize the Bellman error criterion (Bertsekas and Tsitsiklis, 1996; Munos and Szepesvári, 2008), which has known challenges in direct unbiased estimation (see e.g. Example 11.4 in Sutton and Barto (1998)). To counter this, Antos et al. (2008) suggest removing the residual variance, which is akin to considering regret instead of risk in our objective, and yields similar minimax objectives as this work.

On the algorithmic side, we build on the use of regret minimization strategies for solving two player zero-sum games, pioneered by Freund and Schapire (1996) to allow for general distribution families $\mathcal{P}$ as opposed to relying on closed-form maximizations for specific families, as performed in many prior works (Namkoong and Duchi, 2016; Staib and Jegelka, 2019; Kuhn et al., 2019). Similar approaches have been studied in the presence of adversarial corruption in Feige et al. (2015) and in the context of algorithmic fairness for regression in Agarwal et al. (2019).

2. Setting

We consider a class of learning problems where the learner is given a dataset sampled from a distribution P_0 of the form $z_1, \ldots, z_n$ with $z_i \in \mathcal{Z}$. We also have a class $\mathcal{P}$ of distributions such that we want to do well relative to this entire class of distributions, even though we only have access to samples from P_0. In Minimax Regret Optimization, we formalize the property of doing uniformly well for all $P \in \mathcal{P}$ through the objective (3). As mentioned in the introduction, this is closely related to the DRO objective (2). The following result shows that small risk does not lead to small regret, which means DRO does not solve the MRO objective (3). The construction also shows that under heterogenous noise, DRO tends to focus on a distribution $P \in \mathcal{P}$ with large noise level, which can be undesirable for many applications.

Proposition 1 (DRO is sensitive to noise heterogeneity) *Let* $\mathcal{P} = \{\mathrm{Ber}(\mu, 1) \ : \ \mu \in [1/2, 1]\}$*,* $\mathcal{F} = [0, 1]$ *and* $\ell(f, z) = (f - z)^2$*. Then* $\sup_{P \in \mathcal{P}} \mathrm{Regret}_P(f_{DRO}) = 1/4 = 4 \sup_{P \in \mathcal{P}} \mathrm{Regret}_P(f_{MRO})$*.*

In words, as the noise level varies from a constant noise at $\mu = 1/2$ to a vanishing noise for $\mu \to 1$, we observe that f_{DRO} ends up being governed solely by the high-noise distributions, and incurs a large worst case regret. f_{MRO} balances regret across the distribution family better, attaining a constant factor smaller regret.

Proof For this choice of $\mathcal{P}$, let $R_\mu(f)$ be the risk corresponding to Ber(μ). Then we see that

$$R_\mu(f) = f^2 - 2\mu f + \mu \quad \text{and} \quad \mathrm{Regret}_\mu(f) = (f - \mu)^2.$$

As a result, we have

$$\begin{aligned} \max_{\mu \in [1/2,1]} R_\mu(f_{\mathrm{DRO}}) &= \min_{f \in [0,1]} \max_{\mu \in [1/2,1]} f^2 - 2\mu f + \mu \\ &= \min\left(\min_{f \in [0,1/2]} f^2 + (1 - 2f), \min_{f \in [1/2,1]} f^2 + \frac{1}{2}(1 - 2f) \right) = \frac{1}{4}, \end{aligned}$$

where the solution $f_{\mathrm{DRO}} = 1/2$ is the minimizer of $R_\mu(f)$ for $\mu = 1/2$, when the distribution has the largest amount of noise. Furthermore, we have

$$\max_{\mu \in [1/2,1]} \mathrm{Regret}_\mu(f_{\mathrm{DRO}}) = \max_{\mu \in [1/2,1]} \left(\frac{1}{2} - 1 \right)^2 = \frac{1}{4}.$$

In contrast, for MRO, we see that

$$\max_{\mu \in [1/2,1]} \mathrm{Regret}_\mu(f_{\mathrm{MRO}}) = \min_{f \in [0,1]} \max_{\mu \in [1/2,1]} (f - \mu)^2 = \frac{1}{16},$$

where $f_{\mathrm{MRO}} = \frac{3}{4} = \frac{1+1/2}{2}$ focuses on the mid-point of the mean parameter family. ■

Note that while the worst-case regret of MRO is always better than that of DRO, this does not imply that the function f_{MRO} is always preferable under each distribution $P \in \mathcal{P}$ to f_{DRO}. For instance, even in the example above, it can be argued that the DRO solution is preferable for the distribution corresponding to $\mu = 0.5$, or more generally in the range $\mu \in [0.5, 5/8]$, where it attains a smaller regret than f_{MRO}. Next we give another such example where f_{DRO} achieves an arguably better trade-off in balanced performance across distributions. This situation can happen in real applications when functions achieving minimum risks vary greatly across distributions, potentially due to overfitting. The example shows that both MRO and DRO may have advantages or disadvantages under different conditions.

Example 1 (Heterogeneous regrets across domains) *Suppose the target class $\mathcal{P}$ consists of two distributions over $z \in [0, 1]$ and our function class $\mathcal{F}$ consists of 3 functions $\{f_1, f_2, f_3\}$. Let us assume that the risks of these functions under the two distributions are given by:*

$$R_1(f_1) = 0,\ \ R_1(f_2) = 0.5,\ \ R_1(f_3) = 0.5 + \epsilon,\ \ R_2(f_1) = 1, R_2(f_2) = 0.9,\ \ \textit{and}\ \ R_2(f_3) = 0.4.$$

Then f_1 is easily disregarded under both MRO and DRO, since it has poor performance under P_2. For both f_2 and f_3, their regrets on P_1 are larger than P_2, so MRO selects the better function for P_1, that is, $f_{MRO} = f_2$. DRO, on the other hand, prefers f_3 as it has a smaller worst-case risk, which is arguably the right choice in this scenario.

More generally, one can always find a distribution $P \in \mathcal{P}$ under which the regret of f_{DRO} is smaller than that of f_{MRO} and vice versa, but the ordering in the worst-case is clear by definition. In the context of these observations, we next discuss how we might estimate f_{MRO} from samples, for which it is useful to rewrite the objective (3) in an equivalent weight-based formulation we discuss next.

A weight-based reformulation. Having shown the potential benefits of our population objective, we now consider how to optimize it using samples from P_0. Notice that the objective (3) does not depend on P_0 explicitly, which makes it unclear how we should approach the problem given our dataset. We address this issue by adopting an equivalent reweighting based formulation as typically done in DRO. Concretely, let us assume that P is absolutely continuous with respect to P_0 for all $P \in \mathcal{P}$, so that there exists a weighting function $w : \mathcal{Z} \to \mathbb{R}_+$ such that $dP(z) = w(z)dP_0(z)$, with $\mathbb{E}_{P_0}[w] = 1$. We can equivalently rewrite the objective (3) as

$$f_{\text{MRO}} := \operatorname*{arginf}_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \left\{ \text{Regret}_w(f) := R_w(f) - \inf_{f'\in\mathcal{F}} R_w(f') = \ R_w(f) - R_w(f_w) \right\}, \tag{5}$$

where $\mathcal{W} = \{w \ : \ w(z) = dP(z)/dP_0(z) \ \text{ for } \ P \in \mathcal{P}\}$ and $R_w(f) = \mathbb{E}_{z\sim P_0}[w(z)\ell(z, f(z))]$. It is also straightforward to define an empirical counterpart for this objective. Given n samples $z_1, \ldots, z_n \sim P_0$, we can define the (weighted) empirical risk and its corresponding minimizer as

$$\widehat{R}_w(f) = \frac{1}{n}\sum_{i=1}^{n} w(z_i)\ell(z_i, f(z_i)) \quad \text{and} \quad \widehat{f}_w = \operatorname*{arginf}_{f\in\mathcal{F}} \widehat{R}_w(f).$$

With these notations, a natural empirical counterpart of the population objective (5) can be written as

$$\widehat{f}_{\text{MRO}} := \operatorname*{arginf}_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \left[\widehat{R}_w(f) - \inf_{f'\in\mathcal{F}} \widehat{R}_w(f')\right] = \operatorname*{arginf}_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \left[\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)\right]. \tag{6}$$

We now give a couple of concrete examples to instantiate this general formulation.

Example 2 (Covariate shift in supervised learning) *Suppose the samples consist of tuples $z = (x, y)$ where $x \in \mathbb{R}^d$ are features and $y \in \mathbb{R}$ is a prediction target. Suppose that the class $\mathcal{W}$ contains functions $\mathcal{W} = \{w(z) = w_\theta(x) \ : \ \theta \in \Theta\}$, where Θ is some parameter class. That is, we allow only the marginal distribution of x to change while the conditional $P_0(y|x)$ is identical across test distributions (Shimodaira, 2000; Huang et al., 2006). In our formulation, it is* *easy to incorporate the covariate shift setting with a general class $\mathcal{W}$ under the restriction that $\sup_x w_\theta(x) \leq B$. It is possible to add additional regularity on the class Θ, such as taking the unit ball in an RKHS (Huang et al., 2006; Staib and Jegelka, 2019). We note that most prior works on DRO imposed conditions on the joint perturbations of both (x, y), which leads to closed form solutions. However, directly instantiating $\mathcal{W}$ in DRO with a bounded f-divergence or MMD perturbation for only the marginal $P_0(x)$ (instead of the joint distribution of (x, y)) does not yield a nice closed form solution. This results in additional computational and statistical challenges as discussed in Duchi et al. (2019).*

We now demonstrate the versatility of our framework with another example from offline reinforcement learning.

Example 3 (Offline contextual bandit learning) *In offline contextual bandit (CB) learning, the data consists of $z = (x, a, r)$ with $x \in \mathbb{R}^d$ denoting a context, $a \in [K]$ being an action out of K possible choices and $r \in [0, 1]$, a reward. There is a fixed and unknown joint distribution P_0 over $(x, r(1), \ldots, r(K))$. The object of interest is a decision policy π which maps a context x to a distribution over actions, and we seek a policy which maximizes the expected reward under its action choices: $\pi^\star = argmax_{\pi \in \Pi} \mathbb{E}\mathbb{E}[r(a) \mid a \sim \pi(\cdot|x), x]$, where Π is some given policy class. In the offline setting, there is some fixed policy μ which is used to choose actions during the data collection process. Then the training data can be described as $z = (x, a, r)$ where $x \sim P_0$, $a \sim \mu(\cdot|x)$ and $r \sim P_0(\cdot|x, a)$, and this differs from the action distributions that other policies in Π induce. Existing literature on this problem typically creates an estimator η for $\mathbb{E}[r|x, a]$ and estimates the expected reward of π as $\sum_{i=1}^n \sum_a \pi(a|x_i)\eta(r|a, x_i)$, and the quality of the resulting estimate critically relies on the accuracy of η. A particularly popular choice is the class of* doubly robust estimators *(Cassel et al., 1976; Dudík et al., 2014) which solve a (weighted) regression problem to estimate $\eta \approx argmin_f \sum_i \frac{\pi(a_i|x_i)}{\mu(a_i|x_i)}(f(x_i, a_i) - r_i)^2$. Since we require the reward model to predict well on the actions selected according to different policies $\pi \in \Pi$, the ideal regression objective for any policy π reweights the data as per that policy's actions, and doing this weighting results in a significant performance gain (Su et al., 2020; Farajtabar et al., 2018). Typically this weighting is simplified or skipped during policy learning, however, as one needs to simultaneously reweight for all policies $\pi \in \Pi$, and MRO provides an approach to do exactly this by choosing $\mathcal{W} = \{w(x, a, r) = \pi(a \mid x)/\mu(a \mid x) \;:\; \pi \in \Pi\}$ in the optimization for estimating η.*

Boundedness and complexity assumptions We now make standard technical assumptions on the function class $\mathcal{F}$, the loss function ℓ and the weight class $\mathcal{W}$. To focus attention on the essence of the results, we omit relatively complex results such as chaining or local Rademacher complexity, and use an ℓ_∞- covering for uniform convergence over $\mathcal{F}$, where we use $N(\epsilon, \mathcal{F})$ to denote the ℓ_∞-covering number of $\mathcal{F}$ with an accuracy ϵ. For parametric classes, such an analysis is optimal up to a log factors. We also make a boundedness and Lipschitz continuity assumption on the loss function.

Assumption 1 (Bounded and Lipschitz losses) *The loss function $\ell(z, v)$ is bounded in $[0, 1]$ for all $z \in \mathcal{Z}$ and $v \in \{f(z) \;:\; f \in \mathcal{F}, z \in \mathcal{Z}\}$ and is L-Lipschitz continuous with respect to v.*

We also assume that the weight class satisfies a boundedness condition.

Assumption 2 (Bounded importance weights) *All weights $w \in \mathcal{W}$ satisfy $w(z) \leq B_w \leq B$ for all $z \in \mathcal{Z}$ and some constant $B \geq 1$.*

To avoid the complexity of introducing an additional covering number, throughout this paper, we assume that the weight class $\mathcal{W}$ is of finite cardinality, which is reasonable for many domain adaptation applications. We adopt the notations $d_{\mathcal{F}}(\delta) = 1 + \log \frac{N(1/(nLB),\mathcal{F})}{\delta}$ and $d_{\mathcal{F},\mathcal{W}}(\delta) = d_{\mathcal{F}}(\delta) + \ln(|\mathcal{W}|/\delta)$ to jointly capture the complexities of $\mathcal{F}$ and $\mathcal{W}$, given a failure probability δ.

3. Regret Bounds for Bounded, Lipschitz Losses

We begin with the most general case with fairly standard assumptions on the loss function, and the function and weight classes and show the following result on the regret of $\widehat{f}_{\text{MRO}}$ for any $w \in \mathcal{W}$.

Theorem 2 *Under assumptions 1 and 2, suppose further that $\mathbb{E}_{P_0}[w^2] \leq \sigma_w^2$ for any $w \in \mathcal{W}$. Then with probability at least $1 - \delta$, we have $\forall w \in \mathcal{W}$:*

$$\text{Regret}_w(\widehat{f}_{MRO}) \leq \inf_{f \in \mathcal{F}} \sup_{w' \in \mathcal{W}} \text{Regret}_{w'}(f) + \sup_{w' \in \mathcal{W}} \underbrace{\mathcal{O}\left(\sqrt{\frac{\sigma_{w'}^2 d_{\mathcal{F},\mathcal{W}}(\delta)}{n}} + \frac{B_{w'} d_{\mathcal{F},\mathcal{W}}(\delta)}{n}\right)}_{\epsilon_{w'}}.$$

We prove the theorem in Appendix A. The bound of Theorem 2 highlights the main difference of our objective compared with DRO style approaches. The bound states that we our solution $\widehat{f}_{\text{MRO}}$ has a small regret *for each* $w \in \mathcal{W}$, as long as at least one such function exists in the function class, up to the usual finite sample deviation terms. Thus, $\widehat{f}_{\text{MRO}}$ attains the uniform regret guarantee we asked for. To better interpret our result, we state a corollary in a setting where the distributions admit a common approximate minimizer, before making some additional remarks.

Corollary 3 *Under conditions of Theorem 2, suppose further that $\exists f^\star \in \mathcal{F}$ such that $R_w(f^\star) \leq R_w(f_w) + \tilde{\epsilon}_w$ for all $w \in \mathcal{W}$. Then with probability at least $1 - \delta$, we have*

$$\forall w \in \mathcal{W} \;:\; R_w(\widehat{f}_{MRO}) \leq R_w(f^\star) + \underbrace{\sup_{w' \in \mathcal{W}} \tilde{\epsilon}_{w'} + \sup_{w' \in \mathcal{W}} \epsilon_{w'}}_{\epsilon_{\mathcal{W}}}.$$

Comparison with DRO approaches. While Proposition 1 already illustrates the improved robustness of MRO to differing noise levels across target distributions, it is further instructive to compare the guarantees for the two approaches. Concretely, in our setting, it can be shown that with probability at least $1 - \delta$, we have for all $w \in \mathcal{W}$:

$$R_w(\widehat{f}_{\text{DRO}}) \leq \inf_{f \in \mathcal{F}} \sup_{w' \in \mathcal{W}} R_{w'}(f) + \sup_{w' \in \mathcal{W}} \epsilon_{w'}. \tag{7}$$

Compared with Theorem 2, this bound can be inferior when the different distributions identified by $w \in \mathcal{W}$ have vastly different noise levels. For instance, in the setting of Corollary 3, the bound (7) yields $R_w(f_{\text{DRO}}) \leq \sup_{w \in \mathcal{W}} R_w(f^\star) + \epsilon_{\mathcal{W}}$, which has additional dependence on the worst case risk of $f^\star$ across all the $w \in \mathcal{W}$. Similarly, if the data distribution is Gaussian with an identical mean but different variances across the w, then we see that the two bounds reduce to:

$$R_w(\widehat{f}_{\text{MRO}}) \leq \text{Var}_w + \epsilon_{\mathcal{W}} \quad \text{and} \quad R_w(\widehat{f}_{\text{DRO}}) \leq \sup_{w' \in \mathcal{W}} \text{Var}_{w'} + \epsilon_{\mathcal{W}},$$

where Var_w is the variance of the Gaussian corresponding to importance weights w. Overall, these examples serve to illustrate that the DRO objective is reasonable when the different distributions have similar noise levels, but is not robust to heterogeneity in noise levels.

Dependence on the complexity of $\mathcal{W}$. For discrete $\mathcal{W}$, our results incur a union bound penalty of $\ln|\mathcal{W}|$. We suspect that this dependency is unavoidable in the general case. If we define $\mathcal{W}$ using a bound constraint such as $w(z) \le B$, or an f-divergence based constraint as in the DRO literature, then the worst-case w has a simple solution in the DRO setting, with no explicit dependence on the complexity of $\mathcal{W}$. These results are easily adapted to MRO as well by observing $\sup_{w\in\mathcal{W}} \text{Regret}_w(f) = \sup_{w\in\mathcal{W}} \sup_{f'\in\mathcal{F}} R_w(f) - R_w(f')$. For instance, if $\mathcal{W}_B = \{w : 0 \le w(z) \le B, \mathbb{E}_{P_0}[w] = 1\}$ is the class of all bounded importance weights, then the results of Shapiro (2017) adapted to MRO imply that

$$\sup_{w\in\mathcal{W}_B} \text{Regret}(f) = \sup_{f'\in\mathcal{F}} \inf_{\eta\in\mathbb{R}} \left\{\eta + B\mathbb{E}_{P_0}\left[(\ell(f) - \ell(f') - \eta)_+\right]\right\}. \tag{8}$$

Similar arguments hold for the family of Cressie-Read divergences studied in Duchi and Namkoong (2021). However, these closed form maximizations rely on doing joint perturbations over (x, y) in the supervised learning setting and do not apply to covariate shifts (Duchi et al., 2019).

Remarks on standard learning theory extensions. We focus on the most classical learning theory setting in this paper to clearly communicate the key ideas. All our results are completely composable with more refined techniques, however. For instance, the logarithmic factors in our results can be improved by replacing the naïve union bound with a chaining argument to obtain a dependence on Dudley's entropy integral instead. We also focus on the parametric or VC regime where the log-covering number of $\mathcal{F}$ behaves as $d\log\frac{1}{\epsilon}$, for some function class parameter d. In non-parametric scenarios, this dependence scales as $\mathcal{O}(1/\epsilon^p)$ for some exponent p depending on the smoothness and other structural properties of $\mathcal{F}$. In such settings, the discretization at a level $\mathcal{O}(1/n)$ used in our proofs is clearly too aggressive, and we need to instead optimize over the discretization accuracy ϵ depending on the exponent p. Note that obtaining optimal results in these settings requires further use of localization arguments. We do not pursue this extension in the current work.

Proof [Sketch of the proof of Theorem 2] For a fixed f, we can bound $|\widehat{R}_w(f) - R_w(f)|$ using Bernstein's inequality. This requires bounding the range and variance of the random variable $w(z)\ell(z, f(z))$. Since the losses are bounded in $[0, 1]$, the two quantities are bounded by σ_w^2 and B_w respectively. Assumption 1 allows us to further get a uniform bound over $f \in \mathcal{F}$ with an additional dependence on $d_{\mathcal{F}}(\delta)$. Standard arguments now yield closeness of $\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)$ to $R_w(f) - R_w(f_w)$ simultaneously for all $f \in \mathcal{F}$ and $w \in \mathcal{W}$, and utilizing the definition of $\widehat{f}_{\text{MRO}}$ as the minimizer of $\sup_{w\in\mathcal{W}} \widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)$ completes the proof. ■

4. Fast Rates for Squared Loss and Convex Classes

A limitation of Theorem 2 is that it does not leverage any structure of the loss function beyond Assumption 1, and as a result obtains a $1/\sqrt{n}$ dependence on the number of samples. For the case of a fixed distribution, it is well known that self-bounding properties of the loss variance can be used to obtain faster $1/n$ bounds on the regret of the ERM solution, and here we show that this improvement extends to our setting. For ease of exposition, we specialize to the squared loss setting in this section. That is, our samples z take the form (x, y) with $x \in \mathcal{X} \subseteq \mathbb{R}^d$, $y \in [-1, 1]$, our functions f map from $\mathcal{X}$ to $[-1, 1]$ and and $\ell(z, f(z)) = (f(x) - y)^2$. We make an additional convexity assumption on the function class $\mathcal{F}$.

Assumption 3 *The class $\mathcal{F}$ is convex: $\forall\, f, f' \in \mathcal{F}$, $\alpha f + (1-\alpha) f' \in \mathcal{F}$ for all $\alpha \in [0,1]$.*

Note that convexity of $\mathcal{F}$ is quite different from convexity of f in its parameters, and can always be satisfied by taking the convex hull of a base class. The assumption can be avoided by replacing our ERM based solution with aggregation approaches (see e.g. Tsybakov, 2003; Dalalyan and Salmon, 2012). However, we focus on the convex case here to illustrate the key ideas.

Theorem 4 *Under assumptions 1, 2 and 3, with probability at least $1-\delta$, we have $\forall w \in \mathcal{W}$:*

$$\begin{aligned}\mathrm{Regret}_w(\widehat{f}_{MRO}) \leq& \mathrm{Regret}_\star + \mathcal{O}\Big(\sqrt{\mathrm{Regret}_\star \cdot B d_{\mathcal{F},\mathcal{W}}(\delta)/n} + B d_{\mathcal{F},\mathcal{W}}(\delta)/n\Big)\\ =& \mathcal{O}\Big(\mathrm{Regret}_\star + \frac{B d_{\mathcal{F},\mathcal{W}}(\delta)}{n}\Big), \quad \textit{where } \mathrm{Regret}_\star = \inf_{f\in\mathcal{F}} \sup_{w'\in\mathcal{W}} \mathrm{Regret}_{w'}(f).\end{aligned}$$

We prove the theorem in Appendix B. Theorem 4 crucially leverages the following property of convex classes $\mathcal{F}$ and squared loss.

Lemma 5 *For a convex class $\mathcal{F}$ and squared loss $\ell(y, f(x)) = (y - f(x))^2$, we have for any distribution P and $f \in \mathcal{F}$: $\mathbb{E}_P[(f(x) - f_P(x))^2] \leq \mathrm{Regret}_P(f)$.*

Note that a similar result holds for the squared loss and any class $\mathcal{F}$, if we replace f_P with the unconstrained minimizer of $R_P(f)$ over all measurable functions, but using this property in our analysis would result in an additional approximation error term in the bound of Theorem 4, which is undesirable when considering regret within the function class.

Proof [Sketch of the proof of Theorem 4] The only difference with the proof of Theorem 2 is in the handling of the variance of $A = w(z)(\ell(z, f(z)) - \ell(z, f_w(z)))$, which we use as our random variable of interest in this case. Since $w(z) \leq B_w$ almost surely, we get

$$\mathbb{E}_{P_0}[A^2] \stackrel{(a)}{\leq} 16 B_w \mathbb{E}_w[(f(x) - f_w(x))^2] \leq 16 B_w \mathrm{Regret}_w(f),$$

where the inequality (a) follows from the boundedness of $f \in \mathcal{F}$ and y and the final inequality uses Lemma 5. We can now follow the usual recipe for fast rates with a self-bounding property of the variance. Rest of the arguments mirror the proof of Theorem 2. ■

Comparison with DRO. It is natural to ask if it is possible to obtain fast rate for DRO under the conditions of Theorem 4. Of course it is not possible to show a regret bound similar to Theorem 4 for DRO, since it does not optimize the worst-case regret as Proposition 1 illustrates. Nevertheless, we investigate if DRO can achieve fast rate for the raw risk. Unfortunately, the following result shows that even under the assumptions of Theorem 4, the worst-case risk criterion for DRO is still subject to the slower $1/\sqrt{n}$ rate. The reason is that regret has a markedly smaller variance than the risk, and the latter usually deviates at a $1/\sqrt{n}$ rate even for squared loss. The following lower bound is given for the standard DRO formulation. Other variants such as self-normalization will not fix the problem.

Proposition 6 *Under the assumptions of Theorem 4, there is a family of distributions $\mathcal{W}$ with $|\mathcal{W}| = 2$, satisfying $B_w \leq B = 2$, and a function class $\mathcal{F}$ with $d_{\mathcal{F}} = \mathcal{O}(\ln n)$, such that with probability at least $2/45$, we have*

$$\sup_{w\in\mathcal{W}} R_w(\widehat{f}_{DRO}) - \sup_{w\in\mathcal{W}} R_w(f_{DRO}) = \Omega\left(\sqrt{\frac{1}{n}}\right).$$

The proof of this lower bound is given in Appendix B.2.

5. Adapting to Misspecification through Non-uniform Scaling across Distributions

Our result so far guarantees uniform regret, with learning complexity that depends on the worst case complexity over all distributions w. For example, if the complexity of the distribution family is characterized by B_w as in our analysis, then the bound depends on $\sup_w B_w$. In this section, we show that it is possible to improve such dependence using an adaptive scaling of the objective function. To better illustrate the consequences of this approach, let us assume that we can rewrite our guarantees in the form that for each $w \in \mathcal{W}$ and $f \in \mathcal{F}$, we have with probability at least $1 - \delta$,

$$\text{Regret}_w(f) \leq c(\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)) + c_w\epsilon, \tag{9}$$

where the scaling function c_w is a quantity which depends on the underlying distribution P, and is known to the algorithm. For instance, under conditions of Theorem 2, we get this bound with $c = 1$, $c_w = \sigma_w + B_w/\sqrt{n}$ and $\epsilon = d_{\mathcal{F},\mathcal{W}}(\delta)/\sqrt{n}$ (assuming $d_{\mathcal{F},\mathcal{W}}(\delta) \geq 1$). Under conditions of Theorem 4, we can use $c = 3$, $c_w = B_w$ and $\epsilon = d_{\mathcal{F},\mathcal{W}}/n$, by taking the second $O(B/n)$ bound of Theorem 4. While this does worsen our dependence on $d_{\mathcal{F},\mathcal{W}}(\delta)$ in the first case, this is preferable to assuming that $d_{\mathcal{F},\mathcal{W}}(\delta)$ is known and setting of $c_w = \sigma_w + B_w\sqrt{d_{\mathcal{F},\mathcal{W}}(\delta)/n}$ and $\epsilon = \sqrt{d_{\mathcal{F},\mathcal{W}}(\delta)/n}$.

Since we assume c_w is known, then we can define the estimator (Scaled MRO, or SMRO):

$$\widehat{f}_{\text{SMRO}} := \operatorname*{arginf}_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} (\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w))/c_w, \tag{10}$$

We now present our main results for the statistical properties of $\widehat{f}_{\text{SMRO}}$ under the assumptions of Theorems 2 and 4, before comparing the results and discussing some key consequences.

Theorem 7 *Under assumptions of Theorem 2, consider the estimator $\widehat{f}_{SMRO}$ with $c_w = \widehat{\sigma}_w + B_w/\sqrt{n}$, where $\widehat{\sigma}_w^2 = \frac{1}{n}\sum_{i=1}^n w(z_i)^2$ is the empirical second moment of the importance weights. Then with probability at least $1 - \delta$, we have $\forall w \in \mathcal{W}$:*

$$\text{Regret}_w(\widehat{f}_{SMRO}) \leq c_w \inf_{f\in\mathcal{F}} \sup_{w'\in\mathcal{W}} \frac{\text{Regret}_{w'}(f)}{c_{w'}} + \underbrace{d_{\mathcal{F},\mathcal{W}}(\delta)\ \mathcal{O}\left(\sqrt{\sigma_w^2/n} + B_w/n\right)}_{\epsilon'_w}.$$

We can also state a version for squared loss and convex classes.

Theorem 8 *Under assumptions of Theorem 4, consider the estimator $\widehat{f}_{SMRO}$ with $c_w = B_w$. Then with probability at least $1 - \delta$, we have*

$$\forall w \in \mathcal{W} \ : \ \text{Regret}_w(\widehat{f}_{SMRO}) \leq B_w \left[\text{Regret}_\star + \mathcal{O}\left(\sqrt{\text{Regret}_\star \cdot \frac{d_{\mathcal{F},\mathcal{W}}(\delta)}{n}} + \frac{d_{\mathcal{F},\mathcal{W}}(\delta)}{n} \right) \right],$$

where $\text{Regret}_\star = \inf_{f \in \mathcal{F}} \sup_{w' \in \mathcal{W}} \frac{\text{Regret}_{w'}(f)}{B_{w'}}$.

Proofs for both the theorems are given in Appendix C. Comparing Theorems 7 and 8 with their counterparts of MRO, we notice a subtle but important difference. The finite sample deviation term ϵ'_w in Theorem 7 depends on the weights w for which the bound is stated, while the corresponding term in Theorem 7 is $\sup_{w' \in \mathcal{W}} \epsilon_{w'}$. In other words, the heterogeneous scaling in SMRO allows us to obtain a regret bound for each distribution depending on the deviation properties of that particular distribution, unlike in the basic MRO approach. To see the benefits of this approach, we state a corollary of Theorems 7 and 8 next.

Corollary 9 (Aligned distribution class) *Suppose that P_0 and $\mathcal{W}$ satisfy that $R_w(f_w) = R_w(f^\star)$ for all $w \in \mathcal{W}$, with $f^\star \in \mathcal{F}$. Under conditions of Theorem 7, with probability at least $1 - \delta$:*

$$\forall w \in \mathcal{W} \ : \ R_w(\widehat{f}_{SMRO}) \leq R_w(f^\star) + d_{\mathcal{F},\mathcal{W}}(\delta) \ \mathcal{O}\left(\sqrt{\frac{\sigma_w^2}{n}} + \frac{B_w}{n} \right).$$

In the same setting, under the assumptions of Theorem 8, we have with probability at least $1 - \delta$:

$$\forall w \in \mathcal{W} \ : \ R_w(\widehat{f}_{SMRO}) \leq R_w(f^\star) + \mathcal{O}\left(\frac{B_w d_{\mathcal{F},\mathcal{W}}(\delta)}{n} \right).$$

Both results follow by choosing $f = f_w = f^\star$ in Theorems 7 and 8. This result shows that the rescaling allows our bounds to adapt to the closeness of a target distribution to the data collection distribution, simultaneously for all target distributions. The alignment condition of a shared $f^\star$ is always true in well-specified problems, and we illustrate a consequence of Corollary 9 for well-specified linear regression in Appendix D. Specifically, we observe that in well-specified linear regression, when the distribution P_0 has a well-conditioned covariance, then we can directly estimate the true regression coefficients $\beta^\star$ in ℓ_2 error, and hence predict well on any other distribution. However, a direct application of our earlier results in Theorems 2 or 4 results in a scaling with the importance weight bound even for this setting, which is clearly sub-optimal. On the other hand, whenever the weight class contains the function $w(z) \equiv 1$, so that $P_0 \in \mathcal{P}$, we see that the guarantee of Theorems 8 recovers a similar estimation guarantee for the SMRO predictor as direct regression under P_0. Since regression under P_0 is statistically efficient in this scenario, we do not expect to do better, and the SMRO approach achieves this ideal guarantee (up to an extra $\ln |\mathcal{W}|$ term), while being robust to potential misspecification.

6. Algorithmic considerations

So far our development has focused on the statistical properties of MRO. In this section, we discuss how the MRO estimator can be computed from a finite dataset, given some reasonable computational assumptions on the function class $\mathcal{F}$.

Definition 10 (Empirical Risk Minimization oracle) *Given a dataset of the form $(\omega_i, z_i)_{i=1}^n$, an empirical risk minimization (ERM) oracle for $\mathcal{F}$ solves the weighted empirical risk minimization problem:* $\min_{f\in\mathcal{F}} \omega_i \ell(z_i, f(z_i))$.

While we assume access to an exact ERM oracle, we can weaken the notion to an approximate oracle with an optimization error comparable to the statistical error from finite samples. Given such an oracle, we can now approximately solve the MRO objective (6) (or (10)) by using the well-known strategy of solving minimax problems as two player zero-sum games.

For finite $\mathcal{W}$ which we consider in this paper, it is also possible to find the approximate $\hat{\rho}$ and $\widehat{f}_{\text{MRO}}$ efficiently, using the celebrated result of Freund and Schapire (1996) to solve this minimax problem through no-regret dynamics. We use the best response strategy for the P-player, as finding the best response distribution given a specific ρ is equivalent to finding the function $f \in \mathcal{F}$ that minimizes $\mathbb{E}_{w\sim\rho}\widehat{R}_w(f)$. This minimzation can be done through one call to the ERM oracle. For optimizing ρ, we use the exponentiated gradient algorithm of Kivinen and Warmuth (1997) (closely related to Hedge (Freund and Schapire, 1997)), which is a common no-regret strategy to learn a distribution over a finite collection of experts, with each $w \in \mathcal{W}$ being an "expert" in our setting. More formally, we initialize ρ_1 to be the uniform distribution over $\mathcal{W}$ and repeatedly update:

$$f_t = \operatorname*{arginf}_{f\in\mathcal{F}} \mathbb{E}_{w\sim\rho_t}\widehat{R}_w(f), \quad \rho_{t+1}(w) \propto \rho_t(w)\exp\left(\eta\left(\widehat{R}_w(f_t) - \widehat{R}_w(\widehat{f}_w)\right)\right). \tag{11}$$

Let us denote the distribution $P_t = (f_1 + \ldots + f_t)/t$ for the iterates generated by (11). Then the results of Freund and Schapire (1996) yield the following suboptimality bound on P_t.

Proposition 11 *Assume that the class $\mathcal{F}$ is compact and $|\mathcal{W}| < \infty$. For any T and using $\eta = \sqrt{\frac{\ln|\mathcal{W}|}{B^2T}}$ in the updates (11), the distribution P_T satisfies*

$$\mathbb{E}_{f\sim P_T}\sup_w[\widehat{R}_w(f) - \widehat{R}_w(f_w)] \le \inf_{f\in\mathcal{F}}\sup_{w\in\mathcal{W}}\left[\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)\right] + 2B\sqrt{\frac{\ln|\mathcal{W}|}{T}}.$$

It implies that there exists $t \in [T]$ such that

$$\sup_w[\widehat{R}_w(f_t) - \widehat{R}_w(f_w)] \le \inf_{f\in\mathcal{F}}\sup_{w\in\mathcal{W}}\left[\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)\right] + 2B\sqrt{\frac{\ln|\mathcal{W}|}{T}}.$$

At a high-level, Proposition 11 allows us to choose $T = n^2$ to ensure that the optimization error is no larger than the statistical error, allowing the same bounds to apply up to constants. In particular, this proposition allows us to use the distribution $w \sim \rho_t$, in place of $w \in \mathcal{W}$, and its ERM solution f_t to solve the problem. More generally, for infinity $\mathcal{W}$, we have the following result.

Proposition 12 *Consider any compact $\mathcal{F}$ and $\mathcal{W}$. Assume that Assumption 1 and Assumption 2 hold. Then there exists $\hat{\rho} \in \Delta(\mathcal{W})$ so that the solution of MRO is equivalent to the following solution of the weighted ERM method:*

$$\widehat{f}_{MRO} = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E}_{w\sim\hat{\rho}}\widehat{R}_w(f).$$

Proof Assumption 2 implies that $\mathcal{W}$ is compact, and there exists a finite subset $\mathcal{W}_\epsilon$ of $\mathcal{W}$ so that for all $w \in \mathcal{W}$, there exists $w' \in \mathcal{W}_\epsilon$ such that $\sup_i |w(z_i) - w'(z_i)| \le \epsilon$. Now we apply Proposition 11 to the weight class $\mathcal{W}_\epsilon$ and choose $\widehat{f}_{\text{MRO}}, \hat{\rho}$ as:

$$t_0 = \operatorname*{argmin}_{t\in[T]} \sup_{w\in\mathcal{W}_\epsilon} \widehat{R}_w(f_t) - \widehat{R}_w(\widehat{f}_w), \quad \widehat{f}_{\text{MRO}} = f_{t_0}, \quad \hat{\rho} = \rho_{t_0}.$$

Now Proposition 11 implies that for any $\epsilon' > 2\epsilon$ (choosing T appropriately), we have

$$\begin{aligned}\sup_w [\widehat{R}_w(\widehat{f}_{\text{MRO}}) - \widehat{R}_w(f_w)] &\le \inf_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}_\epsilon} \left[\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)\right] + \epsilon' - 2\epsilon \\ &\le \inf_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \left[\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)\right] + \epsilon'.\end{aligned}$$

By setting $\epsilon' \to 0$, and using the compactness of $\mathcal{F}$, we obtain the result. ■

The optimization strategy used here bears some resemblance to boosting techniques, which also seek to reweight the data in order to ensure a uniformly good performance on all samples, though the reweighting is not constrained to a particular class $\mathcal{W}$ unlike here. Note that while the optimization error scales logarithmically in $|\mathcal{W}|$, the computational complexity is linear in the size of this set, meaning that our strategy is computationally feasible for sets $\mathcal{W}$ of a modest size. On the other hand, our earlier discussion (8) suggests that alternative reformulations of the objective might be computationally preferred in the case of class $\mathcal{W}$ which are continuous. Handling the intermediate regime of a large discrete class $\mathcal{W}$ in a computationally efficient manner is an interesting question for future research.

7. Conclusion

In this paper, we introduce the MRO formulation to address problems with distribution shift, and establish learning theoretic properties of optimizing the MRO criterion from samples. We demonstrate the many benefits of reasoning with *uniform regret* as opposed to uniform risk guarantees, and we expect these observations to have implications beyond the setting of distribution shift.

On a technical side, it remains interesting to further develop scalable algorithms for large datasets and weight classes. Better understanding the statistical scaling with the size of the weight class and refining our techniques for important scenarios such as covariate shift are also important directions for future research.

Acknowledgments

We thank the anonymous reviewers for constructive comments towards improving the manuscript and spotting an error in the initial proof of Proposition 6.

References

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *International Conference on Machine Learning*, pages 145–155. PMLR, 2020a.

Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R Varshney. Empirical or invariant risk minimization? a sample complexity perspective. *arXiv preprint arXiv:2010.16412*, 2020b.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996. ISBN 1886529108.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88, 2007.

Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from data of variable quality. In *NIPS*, pages 219–226, 2005.

Arnak S Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327–2355, 2012.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2019.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

Uriel Feige, Yishay Mansour, and Robert Schapire. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pages 637–657. PMLR, 2015.

Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the ninth annual conference on Computational learning theory*, pages 325–332, 1996.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006.

Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, volume 56. Citeseer, 2000.

Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput.*, 132(1):1–63, 1997.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT 2009*, 2009.

Jiří Matoušek and Jan Vondrák. The probabilistic method. *Lecture Notes, Department of Applied Mathematics, Charles University, Prague*, 2001.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, volume 29, pages 2208–2216, 2016.

Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence. *Dataset Shift in Machine Learning*. Neural Information Processing series. MIT Press, 2009. ISBN 9780262170055. URL https://books.google.com/books?id=7zrhAAAAMAAJ.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2020.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, page 107–194, 2012.

Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32:9134–9144, 2019.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 1998. ISBN 9780262303842. URL https://books.google.com/books?id=U57uDwAAQBAJ.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer, 2003.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *International Conference on Machine Learning*, pages 6586–6595. PMLR, 2019.

Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *International Conference on Machine Learning*, pages 10544–10554. PMLR, 2020.

Jia-Jie Zhu, Wittawat Jitkrittum, Moritz Diehl, and Bernhard Schölkopf. Kernel distributionally robust optimization. *arXiv preprint arXiv:2006.06981*, 2020.

Appendix A. Proof of Theorem 2

Recalling our assumptions that for any $w \in \mathcal{W}$, we have $\max_z w(z) \le B$ and $\mathbb{E}_{z\sim P_0} w(z)^2 \le \sigma_w^2$, we know that for a fixed $w \in \mathcal{W}$ and $f \in \mathcal{F}$, we have, with probability at least $1-\delta$:

$$\left|\widehat{R}_w(f) - R_w(f)\right| = \mathcal{O}\left(\sqrt{\frac{\sigma_w^2 \ln(1/\delta)}{n}} + \frac{B_w \ln(1/\delta)}{n}\right). \tag{12}$$

This is a consequence of Bernstein's inequality applied to the random variable $A = w(z)\ell(z, f(z))$ which is bounded by B_w almost surely, and has a second moment at most σ_w^2 when $z \sim P_0$. Since $\ell(z, f(z))$ is L-Lipschitz in the second argument, if $\|f - f'\|_\infty \le 1/(nLB)$, we have for any $w \in \mathcal{W}$ and any z:

$$|w(z)\ell(z, f(z)) - w(z)\ell(z, f'(z))| \le \frac{B_w L}{nLB} \le \frac{1}{n},$$

where the second inequality follows since $B_w \le B$ for all $w \in \mathcal{W}$ by assumption. Hence, defining $\mathcal{F}'$ to be an ℓ_∞ cover of $\mathcal{F}$ at a scale $1/nLB$, a union bound combined with the bound (12) yields that with probability $1-\delta$, we have for all $f \in \mathcal{F}'$:

$$\left|\widehat{R}_w(f) - R_w(f)\right| = \mathcal{O}\left(\sqrt{\frac{\sigma_w^2 \ln(N(\mathcal{F}, 1/(nLB))/\delta)}{n}} + \frac{B_w \ln(N(\mathcal{F}, 1/(nLB))/\delta)}{n}\right).$$

Using the Lipschitz property of the loss and further taking a union bound over $w \in \mathcal{W}$, this gives for all $f \in \mathcal{F}$ and $w \in \mathcal{W}$, with probability at least $1-\delta$:

$$\begin{aligned}\left|\widehat{R}_w(f) - R_w(f)\right| &= \mathcal{O}\left(\sqrt{\frac{\sigma_w^2 \ln(N(\mathcal{F}, 1/(nLB))|\mathcal{W}|/\delta)}{n}} + \frac{B_w \ln(N(\mathcal{F}, 1/(nLB))|\mathcal{W}|/\delta)}{n}\right) + \frac{1}{n} \\ &= \mathcal{O}\left(\sqrt{\frac{\sigma_w^2 d_{\mathcal{F},\mathcal{W}}(\delta)}{n}} + \frac{B_w d_{\mathcal{F},\mathcal{W}}(\delta)}{n}\right) =: \epsilon_w, \end{aligned} \tag{13}$$

where the second equality recalls our definition of $d_{\mathcal{F},\mathcal{W}} = 1 + \ln\frac{N(1/(nLB),\mathcal{F})}{\delta} + \ln\frac{|\mathcal{W}|}{\delta}$. We now condition on the $1-\delta$ probability event that the bound of Equation 13 holds to avoid stating error probabilities in each bound. In particular, applying the bound above to the empirical minimizer $\widehat{f}_w$ for any $w \in \mathcal{W}$, we observe that

$$R_w(\widehat{f}_w) \le \widehat{R}_w(\widehat{f}_w) + \epsilon_w \le \widehat{R}_w(f_w) + \epsilon_w \le R_w(f_w) + 2\epsilon_w.$$

Hence, for any $f \in \mathcal{F}$:

$$\begin{aligned}
\sup_{w\in\mathcal{W}}\left[R_w(\widehat{f}_{\text{MRO}}) - R_w(f_w)\right] &\le \sup_{w\in\mathcal{W}}[R_w(\widehat{f}_{\text{MRO}}) - \widehat{R}_w(\widehat{f}_w)] + \sup_w \epsilon_w \\
&\le \sup_{w\in\mathcal{W}}[\widehat{R}_w(\widehat{f}_{\text{MRO}}) - \widehat{R}_w(\widehat{f}_w)] + 2\sup_w \epsilon_w \\
&\le \sup_{w\in\mathcal{W}}[\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)] + 2\sup_w \epsilon_w \\
&\qquad \text{(since } \widehat{f}_{\text{MRO}} = \text{arginf}_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)) \\
&\le \sup_{w\in\mathcal{W}}[R_w(f) - R_w(\widehat{f}_w)] + 4\sup_w \epsilon_w \\
&\le \sup_{w\in\mathcal{W}}[R_w(f) - R_w(f_w)] + 4\sup_w \epsilon_w.
\end{aligned}$$

Taking an infimum over $f \in \mathcal{F}$ completes the proof of the theorem.

Appendix B. Proofs for Section 4

We first prove Theorem 4 using Lemma 5, before proceeding to prove the lemma.
Proof [Proof of Theorem 4] We begin with the deviation bound for a fixed $f \in \mathcal{F}$ and $w \in \mathcal{W}$ as before, but use the regret random variable $A = w(z)(\ell(z, f(z)) - \ell(z, f_w(z))$ this time. Then by Lemma 5, we have

$$\begin{aligned}
\mathbb{E}_{P_0}[A^2] &\le B_w \mathbb{E}_{P_0}[w(z)(\ell(z, f(z)) - \ell(z, f_w(z)))^2] = B_w \mathbb{E}_w[(\ell(z, f(z)) - \ell(z, f_w(z)))^2] \\
&= B_w \mathbb{E}_w\left[\left((f(x)-y)^2 - (f_w(x)-y)^2\right)^2\right] \\
&= B_w \mathbb{E}_w\left[(f(x) - f_w(x))^2 (f(x) + f_w(x) - 2y)^2\right] \\
&\le 16 B_w \mathbb{E}_w\left[(f(x) - f_w(x))^2\right] && ((|f(x)|, |f_w(x)|, |y| \le 1) \\
&\le 16 B_w \text{Regret}_w(f). && \text{(Lemma 5)}
\end{aligned}$$

Thus we see that for any fixed $f \in \mathcal{F}$ and $w \in \mathcal{W}$, we have with probability $1-\delta$:

$$\begin{aligned}
|\text{Regret}_w(f) - [\widehat{R}_w(f) - \widehat{R}_w(f_w)]| &= \mathcal{O}\left(\sqrt{\frac{B_w \text{Regret}_w(f)\ln(1/\delta)}{n}} + \frac{B_w \ln(1/\delta)}{n}\right) \\
&\le \gamma \text{Regret}_w(f) + (1+\gamma^{-1})\mathcal{O}\left(\frac{B_w \ln(1/\delta)}{n}\right).
\end{aligned}$$

where $\gamma > 0$ is arbitrary.

Hence, with probability at least $1-\delta$, we have for a fixed f and w:

$$(1-\gamma)\mathrm{Regret}_w(f) \le (\widehat{R}_w(f) - \widehat{R}_w(f_w)) + (1+\gamma^{-1})\mathcal{O}\left(\frac{B_w \ln(1/\delta)}{n}\right), \quad \text{and}$$
$$\widehat{R}_w(f) - \widehat{R}_w(f_w) \le (1+\gamma)\mathrm{Regret}_w(f) + (1+\gamma^{-1})\mathcal{O}\left(\frac{B_w \ln(1/\delta)}{n}\right).$$

Using identical arguments as the proof of Theorem 2 allows us to turn both statements into uniform bounds over all $f \in \mathcal{F}$ and $w \in \mathcal{W}$. Defining $\epsilon' = Bd_{\mathcal{F},\mathcal{W}}(\delta)/n$, we now have with probability at least $1-\delta$:

$$\widehat{R}_w(f_w) \le \widehat{R}_w(\widehat{f}_w) + (1+\gamma^{-1})\epsilon'. \tag{14}$$

Now we reason about the quality of $\widehat{f}_{\text{MRO}}$ as before:

$$\begin{aligned}
&(1-\gamma)\sup_{w\in\mathcal{W}} \mathrm{Regret}(\widehat{f}_{\text{MRO}})\\
&\le \sup_{w\in\mathcal{W}} [\widehat{R}_w(\widehat{f}_{\text{MRO}}) - \widehat{R}_w(f_w)] + (1+\gamma^{-1})\epsilon'\\
&\le \sup_{w\in\mathcal{W}} [\widehat{R}_w(\widehat{f}_{\text{MRO}}) - \widehat{R}_w(\widehat{f}_w)] + (1+\gamma^{-1})\epsilon' && (\widehat{R}_w(f_w) \ge \widehat{R}_w(\widehat{f}_w))\\
&\le \sup_{w\in\mathcal{W}} \widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w) + (1+\gamma^{-1})\epsilon' && (\text{since } \widehat{f}_{\text{MRO}} = \text{arginf}_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w))\\
&\le \sup_{w\in\mathcal{W}} \widehat{R}_w(f) - \widehat{R}_w(f_w) + 2(1+\gamma^{-1})\epsilon' && (\text{Equation 14})\\
&\le (1+\gamma)\sup_{w\in\mathcal{W}} [R_w(f) - R_w(f_w)] + 3(1+\gamma^{-1})\epsilon'\\
&= (1+\gamma)\sup_{w\in\mathcal{W}} \mathrm{Regret}_w(f) + 3(1+\gamma^{-1})\epsilon'.
\end{aligned}$$

Taking an infimum over $f \in \mathcal{F}$ and $\gamma = \min(0.5, \sqrt{\epsilon'/\inf_{f\in\mathcal{F}}\sup_{w\in\mathcal{W}}\mathrm{Regret}_w(f)})$ completes the proof. ■

We now prove Lemma 5.

B.1. Proof of Lemma 5

Since $\mathcal{F}$ is convex, if $f_1 \in \mathcal{F}$ and $f_2 \in \mathcal{F}$, then $\alpha f_1 + (1-\alpha) f_2 \in \mathcal{F}$ for $\alpha \in [0,1]$. Then using $f_P = \text{arginf}_{f\in\mathcal{F}} R_P(f)$, we have for any distribution P and $f \in \mathcal{F}$:

$$\begin{aligned}
0 &\le R_P(\alpha f + (1-\alpha) f_P) - R_P(f_P)\\
&= \alpha \mathbb{E}_P[(\alpha f(x) + (2-\alpha) f_P(x) - 2y)(f(x) - f_P(x))]\\
&= \alpha^2 \mathbb{E}_P[(f(x) - f_P(x))^2] + 2\alpha \mathbb{E}_P[(f_P(x) - y)(f(x) - f_P(x))].
\end{aligned}$$

Since this holds for any $\alpha \in [0,1]$, we take the limit $\alpha \downarrow 0$ to conclude that for all $f \in \mathcal{F}$

$$\mathbb{E}_P[(f_P(x) - y)(f(x) - f_P(x))] \ge 0. \tag{15}$$

This inequality further allows us to conclude for any $f \in \mathcal{F}$

$$\begin{aligned} R_P(f) - R_P(f_P) &= \mathbb{E}_P[(f(x) + f_P(x) - 2y)(f(x) - f_P(x)] \\ &= \mathbb{E}_P[(f(x) - f_P(x))^2] + 2\mathbb{E}_P[(f_P(x) - y)(f(x) - f_P(x))] \\ &\geq \mathbb{E}_P[(f(x) - f_P(x))^2]. \end{aligned}$$

B.2. Proof of Proposition 6

We consider a simple problem where there is a class $\mathcal{P} = \{P_1, P_2\}$ consisting of two distributions that we want to do well under. Each distribution P_i is degenerate, so that $x \sim P_i$ gives the sample $x = \mu_i$ with probability 1. We choose $0 \leq \mu_1 \leq \mu_2 \leq 2$ without loss of generality, for a constant C to be set appropriately. The function class $\mathcal{F} = [0, 2]$. Under these conditions, the loss function $(f - \mu_i)^2$ is 4-Lipschitz, so that $d_{\mathcal{F}} = \mathcal{O}(\ln n)$ by discretizing the interval $[0, 2]$ to a precision $1/4n$. Let $\Delta_\mu = \mu_1 - \mu_2$ be the difference of means. Let us choose P_0 to be supported over $\mathbb{R} \times \{1, 2\}$ so that $z = (x, 1)$ where $x \sim P_1$ with probability 0.5 and $z = (x, 2)$ with $x \sim P_2$ with probability 0.5. We set $w_1(z) = 1$ if $z_2 = 1$ and similarly $w_2(z) = 1$ if $z_2 = 2$. Let us define $g(z) = f(z_1)$ for $f \in \mathcal{F}$. Since $g(z)$ is equivalent to $f(x)$ with $x = z_1$ in this example, we stick to using the notation $f \in \mathcal{F}$ for the rest of the proof for consistency with our notation throughout. Let us use R_i, $\widehat{R}_i$ to denote the empirical and expected risks on samples from P_i (equivalently w_i). Then it is easily seen that

$$R_i(f) = (f - \mu_i)^2, \quad \text{and} \quad \widehat{R}_i(f) = \frac{n_i}{n}(f - \mu_i)^2,$$

where $n_i = \sum_{j=1}^n w_i(z_j)$ is the number of samples observed from the distribution P_i, when sampling from P_0. Then we further have

$$f_{\text{DRO}} = \operatorname*{argmin}_{f \in \mathcal{F}} \max_i R_i(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \max\{(f - \mu_1)^2, (f - \mu_2)^2\} = \frac{\mu_1 + \mu_2}{2},$$

and the best worst case risk is given by $\max_i R_i(f_{\text{DRO}}) = \Delta_\mu^2/4$. Now let us examine the empirical situation. We have

$$\widehat{f}_{\text{DRO}} = \operatorname*{argmin}_{f \in \mathcal{F}} \max_i \widehat{R}_i(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \max\{n_1(f - \mu_1)^2, n_2(f - \mu_2)^2\}.$$

Let us denote $\widehat{f}_i = \mu_i$ as the minimizer of $\widehat{R}_i(f)$. Then we observe that :

$$0 = \widehat{R}_1(\widehat{f}_1) \leq \widehat{R}_2(\widehat{f}_1) \quad \text{and} \quad \widehat{R}_1(\widehat{f}_2) \geq \widehat{R}_2(\widehat{f}_2) = 0. \tag{16}$$

Consequently, since both $\widehat{R}_1(f)$ and $\widehat{R}_2(f)$ are continuous functions of f, they are equal for some $f \in [\mu_2, \mu_1]$. So conditioned on $\mathcal{E}$, we seek a solution to

$$n_1(f - \mu_1)^2 = n_2(f - \mu_2)^2,$$

for which, it suffices to choose

$$\widehat{f}_{\text{DRO}} = \frac{\sqrt{n_1}\mu_1 + \sqrt{n_2}\mu_2}{\sqrt{n_1} + \sqrt{n_2}}.$$

Now we evaluate the population worst-case risk of the empirical minimizer, which is given by:

$$\begin{aligned}\max_i R_i(\widehat{f}_{\text{DRO}}) &= \max\{(\widehat{f}_{\text{DRO}} - \mu_1)^2, (\widehat{f}_{\text{DRO}} - \mu_2)^2\} \\ &= \frac{\Delta_\mu^2}{(\sqrt{n_1} + \sqrt{n_2})^2} \max\{n_1, n_2\} \\ &\geq \frac{\Delta_\mu^2}{2n} \max\{n_1, n_2\},\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz and recalling that $n_1 + n_2 = n$. Now we focus on the $\max(n_1, n_2)$ term, which is at least $n/2 + t$ with probability at least $2P(n_1 \geq n/2 + t)$ for any $t \geq 0$. Now using an anti-concentration estimate for Bernoulli sums (Matoušek and Vondrák, 2001), we see that for any $t \in [0, n/8]$, $P(n_1 \geq n/2 + t) \geq \frac{1}{15}\exp(-16t^2/n)$. Let us choose $t = \sqrt{n}/4$, so that we get

$$P\left(n_1 \geq \frac{n}{2} + t\right) \geq \frac{1}{15}\exp\left(-16 \cdot \frac{1}{16}\right) \geq \frac{1}{45},$$

whenever $n \geq 4$ to ensure that $\sqrt{n}/4 \leq n/8$. Hence, we have with probability at least $2/45$

$$\max_i R_i(\widehat{f}_{\text{DRO}}) \geq \frac{\Delta_\mu^2}{2n}\left(\frac{n}{2} + \frac{\sqrt{n}}{4}\right) = \frac{\Delta_\mu^2}{4} + \frac{\Delta_\mu^2}{8\sqrt{n}} = \max_i R_i(f_{\text{DRO}}) + \frac{1}{8\sqrt{n}},$$

where we choose $\Delta_\mu = 1$ in the last equality. This completes the proof.

Appendix C. Proofs for Section 5

We start with a technical lemma about the concentration of empirical estimates of the second moment of importance weights, which is required for Theorem 7.

Lemma 13 *Under Assumption 2, we have with probability at least $1 - \delta$, for all $w \in \mathcal{W}$:*

$$\mathbb{E}_{P_0} w(z)^2 \leq \frac{2}{n}\sum_{i=1}^n w(z_i)^2 + \mathcal{O}\left(\frac{B_w^2 \ln(|\mathcal{W}|/\delta)}{n}\right).$$

Proof Consider the non-negative random variable $A = w(z)^2$ so that $A \leq B_w^2$ with probability 1 when $z \sim P_0$. Then we have

$$\mathbb{E}_{P_0}[A^2] = \mathbb{E}_{P_0}[w(z)^4] \leq B_w^2 \mathbb{E}_{P_0}[w(z)^2].$$

Then for a fixed $w \in \mathcal{W}$, we have by Bernstein's inequality, with probability at least $1 - \delta$:

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n w(z_i)^2 - \mathbb{E}_{P_0} w(z)^2\right| \\ &= \mathcal{O}\left(\sqrt{\frac{B_w^2 \mathbb{E}_{P_0}[w(z)^2]\ln(1/\delta)}{n}} + \frac{B_w^2 \ln(1/\delta)}{n}\right) \\ &\leq \frac{\mathbb{E}_{P_0}[w(z)^2]}{2} + \mathcal{O}\left(\frac{B_w^2 \ln(1/\delta)}{n}\right).\end{aligned}$$

Rearranging terms and taking a union bound over $w \in \mathcal{W}$ completes the proof. ■

We are now ready to prove Theorem 7.

C.1. Proof of Theorem 7

Recall the definition $\widehat{\sigma}_w^2 = \frac{1}{n}\sum_{i=1}^n w(z_i)^2$. Plugging the result of Lemma 13 in Equation 12, we see that with probability at least $1-\delta$, for a fixed $f \in \mathcal{F}$ and $w \in \mathcal{W}$ we have:

$$\left|\widehat{R}_w(f) - R_w(f)\right| = \mathcal{O}\left(\sqrt{\frac{(2\widehat{\sigma}_w^2 + B_w^2/n)\ln(2/\delta)}{n}} + \frac{B_w \ln(2/\delta)}{n}\right)$$
$$= \mathcal{O}\left(\sqrt{\frac{\widehat{\sigma}_w^2 \ln(2/\delta)}{n}} + \frac{B_w \ln(2/\delta)}{n}\right).$$

Now following the proof of Theorem 2, with probability at least $1-\delta$, we have for all $w \in \mathcal{W}$:

$$\begin{aligned}
&R_w(\widehat{f}_{\text{SMRO}}) - R_w(f_w) \\
&= \widehat{R}_w(\widehat{f}_{\text{SMRO}}) - \widehat{R}_w(\widehat{f}_w) + \mathcal{O}\left(\sqrt{\frac{\widehat{\sigma}_w^2(d_{\mathcal{F}}(\delta) + \log(|\mathcal{W}|/\delta)}{n}} + \frac{B_w(d_{\mathcal{F}}(\delta) + \log(|\mathcal{W}|/\delta)}{n}\right) \\
&= \widehat{R}_w(\widehat{f}_{\text{SMRO}}) - \widehat{R}_w(\widehat{f}_w) + \underbrace{\left(\widehat{\sigma}_w + \frac{B_w}{\sqrt{n}}\right)}_{c_w} \underbrace{\mathcal{O}\left(\frac{d_{\mathcal{F},\mathcal{W}}(\delta)}{\sqrt{n}}\right)}_{\epsilon}.
\end{aligned}$$

Now dividing through by c_w, we see that for all $f \in \mathcal{F}$, we have with probability at least $1-\delta$:

$$\begin{aligned}
\sup_{w\in\mathcal{W}} \frac{R_w(\widehat{f}_{\text{SMRO}}) - R_w(f_w)}{c_w} &\leq \sup_{w\in\mathcal{W}} \frac{\widehat{R}_w(\widehat{f}_{\text{SMRO}}) - \widehat{R}_w(\widehat{f}_w)}{c_w} + \epsilon \\
&\leq \inf_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \frac{\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)}{c_w} + \epsilon \\
&\leq \inf_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \frac{R_w(f) - R_w(\widehat{f}_w)}{c_w} + 2\epsilon \\
&\leq \inf_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \frac{R_w(f) - R_w(f_w)}{c_w} + 2\epsilon,
\end{aligned}$$

where the second inequality follows from the definition of SMRO as the empirical optimizer of the objective (10). As a result, we have for any $w \in \mathcal{W}$, with probability at least $1-\delta$:

$$R_w(\widehat{f}_{\text{SMRO}}) - R_w(f_w) \leq c_w \inf_{f\in\mathcal{F}} \sup_{w'\in\mathcal{W}} \frac{R_{w'}(f) - R_{w'}(f_w)}{c_{w'}} + 2c_w\epsilon.$$

C.2. Proof of Theorem 8

We will be terse as the proof is largely a combination of the proofs of Theorems 4 and 7. Proceeding as in the proof of Theorem 4, we see that with probability at least $1-\delta$, we have for all $w \in \mathcal{W}$ and $f \in \mathcal{F}$:

$$(1-\gamma)\text{Regret}_w(f) \leq [\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)] + (1+\gamma^{-1}) \underbrace{B_w}_{c_w} \underbrace{\mathcal{O}\left(\frac{d_{\mathcal{F},\mathcal{W}}}{n}\right)}_{\epsilon}$$

and

$$[\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)] \leq (1+\gamma)\text{Regret}_w(f) + (1+\gamma^{-1}) \underbrace{B_w}_{c_w} \underbrace{\mathcal{O}\left(\frac{d_{\mathcal{F},\mathcal{W}}}{n}\right)}_{\epsilon}$$

for all $\gamma > 0$.

Dividing through by c_w as before and taking a supremum over $w \in \mathcal{W}$, we obtain with probability at least $1-\delta$, we have:

$$\begin{aligned}
(1-\gamma)\sup_{w\in\mathcal{W}} \frac{\text{Regret}_w(\widehat{f}_{\text{SMRO}})}{c_w} &\leq \sup_{w\in\mathcal{W}} \frac{\widehat{R}_w(\widehat{f}_{\text{SMRO}}) - \widehat{R}_w(\widehat{f}_w)}{c_w} + (1+\gamma^{-1})\epsilon \\
&\leq \inf_{f\in\mathcal{F}}\sup_{w\in\mathcal{W}} \frac{\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)}{c_w} + (1+\gamma^{-1})\epsilon \\
&\leq (1+\gamma)\inf_{f\in\mathcal{F}}\sup_{w\in\mathcal{W}} \frac{\text{Regret}_w(f)}{c_w} + 2(1+\gamma^{-1})\epsilon.
\end{aligned}$$

Now following the remaining proof of Theorem 7 gives the desired bound.

Appendix D. Well-specified linear regression with covariate shift

Let us consider a special case of Theorem 8, where $\mathcal{F} = \{\beta^\top x \;:\; \beta \in \mathbb{R}^d, \|\beta\|_2 \leq 1\}$ is the class of linear prediction functions with unit norm weights and the data satisfies $\|x\|_2 \leq 1$. We further assume that $y = x^\top \beta^\star + \nu$, where $\|\beta^\star\|_2 \leq 1$ and ν is zero-mean noise such that $|y| \leq C$ for some constant C (this just causes additional scaling with C in the bounds of Theorems 4 and 8). Suppose further that the covariance $\Sigma_{P_0} := \mathbb{E}_{x\sim P_0} xx^\top$ is full rank with the smallest eigenvalue equal to λ. Let $\widehat{\beta}_{P_0}$ be the ordinary least squares estimator, given samples from P_0. Then it is well-known that with probability at least $1-\delta$,

$$\|\widehat{\beta}_{P_0} - \beta^\star\|^2_{\Sigma_{P_0}} = \mathcal{O}\left(\frac{d\ln(1/\delta)}{n}\right). \tag{17}$$

Consequently, we have for any other distribution P over (x, y):

$$R_P(\widehat{\beta}_{P_0}) - R_P(\beta^\star) = \mathbb{E}_P[(x^\top\widehat{\beta}_{P_0} - x^\top\beta^\star)^2] = \mathcal{O}\left(\frac{d\ln(1/\delta)}{\lambda n}\right),$$

where the inequality uses $\|x\|_2 \leq 1$ and that the smallest eigenvalue of Σ_{P_0} is at least λ. That is, doing OLS under P_0 yields a strong guarantee for all target distributions P, since we get pointwise accurate predictions in this scenario. However, directly applying the results of Theorem 4 would still have additional scaling with B_w for any target distribution with importance weights w. On the other hand, let us consider the bound of Corollary 9 for any class $\mathcal{W}$ such that $w_0 \equiv 1$ is in $\mathcal{W}$, so that we always include P_0 in our class of target distributions. Since $B_{w_0} = 1$ and $d_{\mathcal{F}} = d\ln(\frac{1}{\delta})$ in this case, the second bound of the corollary yields with probability at least $1 - \delta$:

$$\|\widehat{\beta}_{\text{SMRO}} - \beta^\star\|_{\Sigma_{P_0}}^2 = R_{P_0}(\widehat{\beta}_{\text{SMRO}}) - R_{P_0}(\beta^\star) = \mathcal{O}\left(\frac{(d\ln(1/\delta) + \ln\frac{|\mathcal{W}|}{\delta})}{n}\right).$$

This is comparable to the prediction error bound in (17) for ERM on P_0, only incurring an additional $\ln|\mathcal{W}|$ term compared. Note that ERM is asymptotically efficient in this setting, so we cannot expect to do better and suffer only a small penalty for our worst-case robustness. The approach of learning on P_0 alone is of course not robust to misspecification in the linear regression assumption. The guarantee of Theorem 4, in contrast incurs a bound $\sup_{w\in\mathcal{W}} B_w\mathcal{O}\left(\frac{(d\ln(1/\delta)+\ln\frac{|\mathcal{W}|}{\delta})}{n}\right)$, which can be significantly worse.

Appendix E. Proofs for Section 6

It is clearly seen that the updates of ρ_t correspond to the exponentiated gradient (EG) algorithm applied to the linear objective $-(\mathbb{E}_{w\sim\rho}\widehat{R}_w(f_t) - \widehat{R}_w(\widehat{f}_w))/B$ at round t, using a stepsize of ηB, where the negative sign happens since the EG algorithm is designed for minimization problems, while we apply it to a maximization problem. The regret guarantee for EG, specifically Corollary 2.14 of Shalev-Shwartz (2012) states that for any $w \in \mathcal{W}$

$$-\sum_{t=1}^{T}\mathbb{E}_{w'\sim\rho_t}\frac{\widehat{R}_{w'}(f_t) - \widehat{R}_{w'}(\widehat{f}_{w'})}{B} \le -\sum_{t=1}^{T}\frac{\widehat{R}_w(f_t) - \widehat{R}_w(\widehat{f}_w)}{B} + \frac{\ln|\mathcal{W}|}{\eta B} + \eta BT.$$

Multiplying through by B and substituting $\eta B = \sqrt{\ln|\mathcal{W}|/T}$ gives

$$-\sum_{t=1}^{T}\mathbb{E}_{w'\sim\rho_t}[\widehat{R}_{w'}(f_t) - \widehat{R}_{w'}(\widehat{f}_{w'})] \le -\sum_{t=1}^{T}[\widehat{R}_w(f_t) - \widehat{R}_w(\widehat{f}_w)] + 2B\sqrt{T\ln|\mathcal{W}|}.$$

Now recalling the definition $P_t = (f_1 + \ldots + f_t)/t$, following the proof technique of Freund and Schapire (1996) gives that

$$
\begin{aligned}
&\mathbb{E}_{f\sim P_T} \sup_w \frac{\widehat{R}_w(f) - \widehat{R}_w(f_w)}{B} \\
&= \frac{1}{T}\sum_{t=1}^{T} \sup_w \frac{\widehat{R}_w(f_t) - \widehat{R}_w(f_w)}{B} \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{w\sim\rho_t} \frac{\widehat{R}_w(f_t) - \widehat{R}_w(\widehat{f}_w)}{B} + 2\sqrt{\frac{\ln|\mathcal{W}|}{T}} \\
&\overset{(a)}{\leq} \frac{1}{T}\sum_{t=1}^{T} \inf_{f\in\mathcal{F}} \mathbb{E}_{w\sim\rho_t} \frac{\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)}{B} + 2\sqrt{\frac{\ln|\mathcal{W}|}{T}} \\
&\leq \inf_{f\in\mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{w\sim\rho_t} \frac{\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)}{B} + 2\sqrt{\frac{\ln|\mathcal{W}|}{T}} \\
&\leq \inf_{f\in\mathcal{F}} \sup_{w\in\mathcal{W}} \frac{\widehat{R}_w(f) - \widehat{R}_w(\widehat{f}_w)}{B} + 2\sqrt{\frac{\ln|\mathcal{W}|}{T}},
\end{aligned}
$$

where (a) follows from the best response property of f_t for ρ_t (11).