Does Distributionally Robust Supervised Learning Give Robust Classifiers?

Weihua Hu [1 2] Gang Niu [2] Issei Sato [1 2] Masashi Sugiyama [2 1]

Abstract

Distributionally Robust Supervised Learning (DRSL) is necessary for building reliable machine learning systems. When machine learning is deployed in the real world, its performance can be significantly degraded because test data may follow a different distribution from training data. DRSL with f-divergences explicitly considers the worst-case distribution shift by minimizing the adversarially reweighted training loss. In this paper, we analyze this DRSL, focusing on the classification scenario. Since the DRSL is explicitly formulated for a distribution shift scenario, we naturally expect it to give a robust classifier that can aggressively handle *shifted distributions*. However, surprisingly, we prove that the DRSL just ends up giving a classifier that exactly fits *the given training distribution*, which is too pessimistic. This pessimism comes from two sources: the particular losses used in classification and the fact that the variety of distributions to which the DRSL tries to be robust is too wide. Motivated by our analysis, we propose simple DRSL that overcomes this pessimism and empirically demonstrate its effectiveness.

1. Introduction

Supervised learning has been successful in many application fields. The vast majority of supervised learning research falls into the Empirical Risk Minimization (ERM) framework (Vapnik, 1998) that assumes a test distribution to be the same as a training distribution. However, such an assumption can be easily contradicted in real-world applications due to sample selection bias or non-stationarity of the environment (Quionero-Candela et al., 2009). Once the distribution shift occurs, the performance of the traditional machine learning techniques can be significantly degraded. This makes the traditional techniques unreliable for practitioners to use in the real world.

[1]University of Tokyo, Japan [2]RIKEN, Tokyo, Japan. Correspondence to: Weihua Hu <weihua916@gmail.com>.

Distributionally Robust Supervised Learning (DRSL) is a promising paradigm to tackle this problem by obtaining prediction functions explicitly robust to distribution shift. More specifically, DRSL considers a minimax game between a learner and an adversary: the adversary first shifts the test distribution from the training distribution within a pre-specified uncertainty set so as to maximize the expected loss on the test distribution. The learner then minimizes the adversarial expected loss.

DRSL with f-divergences (Bagnell, 2005; Ben-Tal et al., 2013; Duchi et al., 2016; Namkoong & Duchi, 2016; 2017) is particularly well-studied and lets the uncertainty set for test distributions be an f-divergence ball from a training distribution (see Section 2 for the detail). This DRSL has been mainly studied under the assumption that *the same continuous loss is used for training and testing*. This is not the case in the *classification* scenario, in which we care about the *0-1 loss* (i.e., the mis-classification rate) at test time, while at training time, we use a *surrogate loss* for optimization tractability.

In this paper, we revisit DRSL with f-divergences, providing novel insight for the *classification* scenario. In particular, we prove rather surprising results (Theorems 1–3), showing that when the DRSL is applied to classification, the obtained classifier ends up being optimal for the *training distribution*. This is too pessimistic for DRSL given that DRSL is explicitly formulated for a distribution shift scenario and is naturally expected to give a classifier different from the one that exactly fits the given training distribution. Such pessimism comes from two sources: the particular losses used in classification and the over-flexibility of the uncertainty set used by DRSL with f-divergences.

Motivated by our analysis, we propose simple DRSL that overcomes the pessimism of the previous DRSL by incorporating structural assumptions on distribution shift (Section 4). We establish convergence properties of our proposed DRSL (Theorem 4) and derive efficient optimization algorithms (Section 5). Finally, we demonstrate the effectiveness of our DRSL through experiments (Section 6).

Related work: Besides DRSL with f-divergences, different DRSL considers different classes of uncertainty sets for test distributions. DRSL by Globerson & Roweis (2006) considered the uncertainty of features deletion, while DRSL by Liu & Ziebart (2014) considered the uncer-

tainty of unknown properties of the conditional label distribution. DRSL by Esfahani & Kuhn (2015), Blanchet et al. (2016) and Sinha et al. (2017) lets the uncertainty set of test distributions be a Wasserstein ball from the training distribution. DRSL with the Wasserstein distance can make classifiers robust to adversarial examples (Sinha et al., 2017), while DRSL with f-divergences can make classifiers robust against adversarial reweighting of data points as shown in Section 2. Recently, in the context of fair machine learning, Hashimoto et al. (2018) applied DRSL with f-divergences in an attempt to achieve fairness without demographic information.

2. Review of ERM and DRSL

In this section, we first review the ordinary ERM framework. Then, we explain a general formulation of DRSL and review DRSL with f-divergences.

Suppose training samples, $\{(x_1, y_1), \ldots, (x_N, y_N)\} \equiv \mathcal{D}$, are drawn i.i.d. from an unknown training distribution over $\mathcal{X} \times \mathcal{Y}$ with density $p(x, y)$, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y}$ is an output domain. Let g_θ be a prediction function with parameter θ, mapping $x \in \mathcal{X}$ into a real scaler or vector, and let $\ell(\widehat{y}, y)$ be a loss between y and real-valued prediction $\widehat{y}$.

ERM: The objective of the risk minimization (RM) is

$$\min_\theta \underbrace{\mathbb{E}_{p(x,y)}[\ell(g_\theta(x), y)]}_{\equiv \mathcal{R}(\theta)}, \tag{1}$$

where $\mathcal{R}(\theta)$ is called the risk. In ERM, we approximate the expectation in Eq. (1) by training data $\mathcal{D}$:

$$\min_\theta \underbrace{\frac{1}{N} \sum_{i=1}^{N} \ell(g_\theta(x_i), y_i)}_{\equiv \widehat{\mathcal{R}}(\theta)}, \tag{2}$$

where $\widehat{\mathcal{R}}(\theta)$ is called the empirical risk. To prevent overfitting, we can add regularization term $\Omega(\theta)$ to Eq. (2) and minimize $\widehat{\mathcal{R}}(\theta) + \lambda\Omega(\theta)$, where $\lambda \geq 0$ is a trade-off hyperparameter.

General formulation of DRSL: ERM implicitly assumes the test distribution to be the same as the training distribution, which does not hold in most real-world applications. DRSL is explicitly formulated for a distribution shift scenario, where test density $q(x, y)$ is different from training density $p(x, y)$. Let $\mathcal{Q}_p$ be an uncertainty set for test distributions. In DRSL, the learning objective is

$$\min_\theta \sup_{q \in \mathcal{Q}_p} \mathbb{E}_{q(x,y)}[\ell(g_\theta(x), y)]. \tag{3}$$

We see that Eq. (3) minimizes the risk w.r.t. the *worst-case test distribution* within the uncertainty set $\mathcal{Q}_p$.

DRSL with f-divergences: Let $q \ll p$ denote that q is absolutely continuous w.r.t. p, i.e., $p(x, y) = 0$ implies $q(x, y) = 0$. Bagnell (2005) and Ben-Tal et al. (2013) considered the particular uncertainty set

$$\mathcal{Q}_p = \{q \ll p \mid \mathrm{D}_f[q\|p] \leq \delta\}, \tag{4}$$

where $\mathrm{D}_f[\cdot\|\cdot]$ is an f-divergence defined as $\mathrm{D}_f[q\|p] \equiv \mathbb{E}_p[f(q/p)]$, and $f(\cdot)$ is convex with $f(1) = 0$. The f-divergence (Ciszar, 1967) measures a discrepancy between probability distributions. When $f(x) = x \log x$, we have the well-known Kullback-Leibler divergence as an instance of it. Hyper-parameter $\delta > 0$ in Eq. (4) controls the degree of the distribution shift. Define $r(x, y) \equiv q(x, y)/p(x, y)$. Through some calculations, the objective of DRSL with f-divergences can be rewritten as

$$\min_\theta \underbrace{\sup_{r \in \mathcal{U}_f} \mathbb{E}_{p(x,y)}[r(x, y)\ell(g_\theta(x), y)]}_{\equiv \mathcal{R}_{\mathrm{adv}}(\theta)}, \tag{5}$$

$$\mathcal{U}_f \equiv \{r(x, y) \mid \mathbb{E}_{p(x,y)}[f(r(x, y))] \leq \delta,$$
$$\mathbb{E}_{p(x,y)}[r(x, y)] = 1,$$
$$r(x, y) \geq 0, \ \forall(x, y) \in \mathcal{X} \times \mathcal{Y}\}. \tag{6}$$

We call $\mathcal{R}_{\mathrm{adv}}(\theta)$ the *adversarial risk* and call the minimization problem of Eq. (5) the *adversarial risk minimization* (ARM). In ARM, the density ratio, $r(x, y)$, can be considered as the weight put by the adversary on the loss of data (x, y). Then, Eq. (5) can be regarded as a minimax game between the learner (corresponding to $\min_\theta$) and the adversary (corresponding to $\sup_{r \in \mathcal{U}_f}$): the adversary first reweights the losses using $r(\cdot, \cdot)$ so as to maximize the expected loss; the learner then minimizes the *reweighted* expected loss, i.e., adversarial risk $\mathcal{R}_{\mathrm{adv}}(\theta)$.

For notational convenience, we denote $\ell(g_\theta(x_i), y_i)$ by $\ell_i(\theta)$. Also, let $\boldsymbol{r} \equiv (r_1, \ldots, r_N)$ be a vector of density ratios evaluated at training data points, i.e., $r_i \equiv r(x_i, y_i)$ for $1 \leq i \leq N$. Equations (5) and (6) can be empirically approximated as[1]

$$\min_\theta \underbrace{\sup_{\boldsymbol{r} \in \widehat{\mathcal{U}}_f} \frac{1}{N} \sum_{i=1}^{N} r_i \ell_i(\theta)}_{\equiv \widehat{\mathcal{R}}_{\mathrm{adv}}(\theta)}, \tag{7}$$

$$\widehat{\mathcal{U}}_f = \left\{\boldsymbol{r} \ \middle| \ \frac{1}{N} \sum_{i=1}^{N} f(r_i) \leq \delta, \ \frac{1}{N} \sum_{i=1}^{N} r_i = 1, \ \boldsymbol{r} \geq 0\right\}, \tag{8}$$

---

[1]The formulation in Eqs. (7) and (8) is similar to Duchi et al. (2016), Namkoong & Duchi (2016) and Namkoong & Duchi (2017) except that they decay δ linearly w.r.t. the number of training data N. Different from us, they assume $\delta = 0$ in Eq. (4) (thus, their objective is the ordinary risk) and try to be robust to apparent distribution fluctuations due to the *finiteness of training samples*. On the other hand, we consider using the same $\delta > 0$ for both Eqs. (4) and (8) and try to be robust to the actual distribution change between training and test stages.

where the inequality constraint for a vector is applied in an element-wise fashion. We call $\widehat{\mathcal{R}}_{\mathrm{adv}}(\theta)$ the *adversarial empirical risk* and call the minimization problem of Eq. (7) the *adversarial empirical risk minimization* (AERM). In AERM, the adversary (corresponding to $\sup_{\boldsymbol{r}\in\widehat{\mathcal{U}}_f}$) reweights data losses through $\boldsymbol{r}$ to maximize the empirical loss in Eq. (7). To prevent overfitting, we can add regularization term $\Omega(\theta)$ to Eq. (7).

3. Analysis of DRSL with f-divergences in classification

At first glance, DRSL with f-divergences (which we call ARM and AERM in this paper) is reasonable to give a distributionally robust classifier in the sense that it explicitly minimizes the loss for the shifted worst-case *test distribution*. However, we show rather surprising results, suggesting that the DRSL, when applied to *classification*, still ends up giving a classifier optimal for a *training* distribution. This is too pessimistic for DRSL because it ends up behaving similarly to ordinary ERM-based supervised classification that does *not* explicitly consider distribution shift. To make a long story short, our results hold because of the particular losses used in classification (especially, the 0-1 loss at test time) and the overly flexible uncertainty sets used by ARM and AERM. We will detail these points after we state our main results.

Classification setting: Let us first briefly review classification settings to set up notations. In binary classification, we have $g_\theta(\cdot) : x \mapsto \widehat{y} \in \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$ and $\ell(\cdot, \cdot) : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. In K-class classification for $K \geq 2$, we have $g_\theta(\cdot) : x \mapsto \widehat{y} \in \mathbb{R}^K$, $\mathcal{Y} = \{1, 2, \ldots, K\}$ and $\ell(\cdot, \cdot) : \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. The goal of classification is to learn the prediction function that minimizes the mis-classification rate on the test distribution. The mis-classification rate corresponds to the *use of the 0-1 loss*, i.e., $\ell(\widehat{y}, y) \equiv \mathbf{1}\{\mathrm{sign}(\widehat{y}) \neq y\}$ for binary classification, and $\ell(\widehat{y}, y) \equiv \mathbf{1}\{\mathrm{argmax}_k \widehat{y}_k \neq y\}$ for multi-class classification, where $\mathbf{1}\{\cdot\}$ is the indicator function and $\widehat{y}_k$ is the k-th element of $\widehat{y} \in \mathbb{R}^K$. However, since the 0-1 loss is non-convex and non-continuous, learning with it is difficult in practice. Therefore, at training time, we instead use *surrogate losses* that are easy to optimize, such as the logistic loss and the cross-entropy loss.

In the following, we state our main results, analyzing ARM and AERM in the classification scenario by considering the use of the 0-1 loss and a surrogate loss.

The 0-1 loss case: Theorem 1 establishes the non-trivial relationship between the adversarial risk and the ordinary risk when the 0-1 loss is used.

Theorem 1. *Let $\ell(\widehat{y}, y)$ be the 0-1 loss. Then, there is a monotonic relationship between $\mathcal{R}_{\mathrm{adv}}(\theta)$ and $\mathcal{R}(\theta)$ in the sense that for any pair of parameters θ_1 and θ_2, the followings hold.*

If $\mathcal{R}_{\mathrm{adv}}(\theta_1) < 1$, then

$$\mathcal{R}_{\mathrm{adv}}(\theta_1) < \mathcal{R}_{\mathrm{adv}}(\theta_2) \iff \mathcal{R}(\theta_1) < \mathcal{R}(\theta_2). \quad (9)$$

If $\mathcal{R}_{\mathrm{adv}}(\theta_1) = 1$, then

$$\mathcal{R}(\theta_1) \leq \mathcal{R}(\theta_2) \implies \mathcal{R}_{\mathrm{adv}}(\theta_2) = 1. \quad (10)$$

The same monotonic relationship also holds between their empirical approximations: $\widehat{\mathcal{R}}_{\mathrm{adv}}(\theta)$ and $\widehat{\mathcal{R}}(\theta)$.

See Appendix A for the proof. Theorem 1 shows a surprising result that *when the 0-1 loss is used*, $\mathcal{R}(\theta)$ and $\mathcal{R}_{\mathrm{adv}}(\theta)$ are essentially equivalent objective functions in the sense that the minimization of one objective function results in the minimization of another objective function. This readily implies that $\mathcal{R}(\theta)$ and $\mathcal{R}_{\mathrm{adv}}(\theta)$ have exactly the same set of global minima in the regime of $\mathcal{R}_{\mathrm{adv}}(\theta) < 1$. An immediate practical implication is that if we select hyper-parameters such as λ for regularization according to *the adversarial risk with the 0-1 loss*, we will end up choosing hyper-parameters that attain the minimum mis-classification rate on the *training distribution*.

The surrogate loss case: We now turn our focus on the training stage of classification, where we use a *surrogate loss* instead of the 0-1 loss. In particular, for binary classification, we consider a class of classification calibrated losses (Bartlett et al., 2006) that are margin-based, i.e., $\ell(\widehat{y}, y)$ is a function of product $y\widehat{y}$. For multi-class classification, we consider a class of classification calibrated losses (Tewari & Bartlett, 2007) that are invariant to class permutation, i.e., for any class permutation $\pi : \mathcal{Y} \to \mathcal{Y}$, $\ell(\widehat{y}^\pi, \pi(y)) = \ell(\widehat{y}, y)$ holds, where $\widehat{y}_k^\pi = \widehat{y}_{\pi(k)}$ for $1 \leq k \leq K$. Although we only consider the sub-class of general classification-calibrated losses (Bartlett et al., 2006; Tewari & Bartlett, 2007), we note that ours still includes some of the most widely used losses: the logistic, hinge, and exponential losses for binary classification and the softmax cross entropy loss for multi-class classification.

We first review Proposition 1 by Bartlett et al. (2006) and Tewari & Bartlett (2007) that justifies the use of classification-calibrated losses in ERM for classification. We then show a surprising fact in Theorem 2 that the similar property also holds for ARM using the sub-class of classification-calibrated losses.

Proposition 1 (Bartlett et al. (2006); Tewari & Bartlett (2007))**.** *Let $\ell(\widehat{y}, y)$ be a classification calibrated loss, and assume that the hypothesis class is equal to all measurable functions. Then, the risk minimization (RM) gives the Bayes optimal classifier[2].*

Theorem 2. *Let $f(\cdot)$ be differentiable, the hypothesis class be all measurable functions, and $\ell(\widehat{y}, y)$ be a classification-calibrated loss that is margin-based or invariant to class*

---

[2]The classifier that minimizes the mis-classification rate for the training density $p(x, y)$ (the 0-1 loss is considered), i.e., the classifier whose prediction on x is equal to $\arg\max_{y\in\mathcal{Y}} p(y|x)$.

permutation. Let $g^{(\mathrm{adv})}$ be any solution of ARM[3] under the above setting, and define

$$r^* \equiv \underset{r \in \mathcal{U}_f}{\arg\max} \, \mathbb{E}_{p(x,y)}[r(x,y)\ell(g^{(\mathrm{adv})}(x),y)]. \quad (11)$$

Then, the prediction of $g^{(\mathrm{adv})}$ coincides with that of the Bayes optimal classifier almost surely over $q^*(x) \equiv \sum_{y \in \mathcal{Y}} r^*(x,y)p(x,y)$. Furthermore, among the solutions of ARM, there exists $g^{(\mathrm{adv})}$ whose prediction coincides with that of the Bayes optimal classifier almost surely over $p(x)$.

See Appendix B for the proof. Theorem 2 indicates that ARM, similarly to RM, ends up giving the optimal decision boundary for the *training* distribution, if the hypothesis class is all measurable functions and we have access to true density $p(x,y)$. Even though the assumptions made are strong, Theorem 2 together with Proposition 1 highlight the non-trivial fact that when a certain surrogate loss is used, AERM and ERM demonstrate the similar *asymptotic* behavior in classification.

We proceed to consider a more practical scenario, where we only have a finite amount of training data and the hypothesis class is limited. In the rest of the section, we focus on a differentiable loss and a real-valued scalar output, i.e., $\widehat{y} \in \mathbb{R}$, which includes the scenario of binary classification.

We first define the notion of a *steeper* loss, which will play a central role in our result.

Definition 1 (Steeper loss). Loss function $\ell_{\mathrm{steep}}(\widehat{y}, y)$ is said to be steeper than loss function $\ell(\widehat{y}, y)$, if there exists a non-constant, non-decreasing and non-negative function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that

$$\frac{\partial \ell_{\mathrm{steep}}(\widehat{y}, y)}{\partial \widehat{y}} = h(\ell(\widehat{y}, y)) \frac{\partial \ell(\widehat{y}, y)}{\partial \widehat{y}}. \quad (12)$$

For example, following Definition 1, we can show that the exponential loss is *steeper* than the logistic loss. Intuitively, outlier-sensitive losses are *steeper* than more outlier-robust losses. Lemma 1 shows an important property of a steeper loss in a classification scenario.

Lemma 1. *Let $\ell(\widehat{y}, y)$ be a margin-based convex classification-calibrated loss. Then, its steeper loss defined in Eq. (12) is also convex classification-calibrated if $h(\ell(0, y)) > 0$.*

See Appendix C for the proof.

Now we are ready to state our result in Theorem 3 that considers $\widehat{y} \in \mathbb{R}$. Theorem 3 holds for *any* hypothesis class that is parametrized by θ and sub-differentiable w.r.t. θ, e.g., linear-in-parameter models and deep neural networks.

Theorem 3. *Let θ^* be a stationary point of AERM in Eq. (7) using $\ell(\widehat{y}, y)$. Then, there exists a steeper loss function, $\ell_{\mathrm{DRSL}}(\widehat{y}, y)$, such that θ^* is also a stationary point of the following ERM.*

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell_{\mathrm{DRSL}}(g_\theta(x_i), y_i). \quad (13)$$

See Appendix D for the proof.

Remark 1 (Conditions for convexity). Let $\ell(\widehat{y}, y)$ be convex in $\widehat{y}$, $g_\theta(x)$ be a linear-in-parameter model. Then, both AERM in Eq. (7) and ERM in Eq. (13) become convex in θ. This implies that the stationary point θ^* in Theorem 3 turns out to be the *global optimum* for both Eqs. (7) and (13) in this usual setting.

Note that Theorem 3 holds for general real-valued scalar prediction, i.e., $\widehat{y} \in \mathbb{R}$; thus, the result holds for ordinary regression (using the same loss for training and testing) as well as for binary classification. However, as we discuss in the following, the implication of Theorem 3 is drastically different for the two scenarios.

Implication for classification: Theorem 3 together with Lemma 1 indicate that under a mild condition,[4] AERM using a convex classification-calibrated margin-based loss reduces to Eq. (13), which is ERM using a convex classification-calibrated loss. This implies that AERM, similarly to ordinary ERM using a classification-calibrated loss, will try to give a classifier optimal for the *training distribution*.

Why does the use of the steeper surrogate loss fail to give meaningful robust classifiers? This is because we are dealing with classification tasks, where we care about the performance *in terms of the 0-1 loss* at test time. The use of the steeper *surrogate* loss may make a classifier distributionally robust *in terms of the surrogate loss*,[5] but not necessarily so in terms of the 0-1 loss. Moreover, even if we obtain a classifier that minimizes the adversarial risk *in terms of the 0-1 loss*, the obtained classifier ends up being optimal for the training distribution (see Theorem 1). In any case, the use of the steeper loss does not in general give classifiers that are robust to change from a training distribution.

In summary, in the *classification* scenario, the use of the steeper loss does more harm (making a classifier sensitive

---

[3]There can be multiple solutions that achieve the same minimum adversarial risk.

[4]The condition that $h(\ell(0, y)) > 0$ in Lemma 1. Whether the condition holds or not generally depends on the uncertainty set, the model, the loss function, and training data. Nonetheless, the condition is mild in practice; especially, the condition always holds when the Kullback-Leibler divergence is used. See Appendix C for detailed discussion.

[5]For fixed δ (non-decaying w.r.t. N), whether AERM is consistent with ARM or not is an open problem. Nevertheless, we empirically confirm in Section 6 that AERM achieves lower adversarial risk than other baselines *in terms of the surrogate loss*.

to outliers due to the use of the steeper surrogate loss) than good (making a classifier robust to change from a training distribution).

Implication for ordinary regression: For comparison, let us rethink about the classical regression scenario, in which we use *the same loss*, e.g., the squared loss, during training and testing. In such a case, the use of the steeper loss may indeed make regressors distributionally robust *in terms of the same loss*. Nonetheless, learning can be extremely sensitive to outliers due to the use of the *steeper* loss. Hence, when applying DRSL with f-divergences to real-world regression tasks, we need to pay extra attention to ensure that there are no outliers in datasets.

4. DRSL with Latent Prior Probability Change

In this section, motivated by our theoretical analysis in Section 3, we propose simple yet practical DRSL that overcomes the over pessimism of ARM and AERM in the classification scenario. We then analyze its convergence property and discuss the practical use of our DRSL.

Theoretical motivation: What insight can we get from our theoretical analyses in Section 3? Our key insight from proving the theorems is that the adversary of ARM has too much (non-parametric) freedom to shift the test distribution, and as a result, the learner becomes overly pessimistic. In fact, the proofs of all the theorems rely on the over-flexibility of the uncertainty set $\mathcal{U}_f$ in Eq. (6), i.e., the values of $r(\cdot, \cdot)$ are *not tied together* for different (x, y) within $\mathcal{U}_f$ (see Eqs. (5) and (6)). Consequently, the adversary of ARM simply assigns larger weight $r(x, y)$ to data (x, y) with a larger loss. This fact, combined with the fact that we use the different losses during training and testing in classification (see discussion at the end of Section 3), led to the pessimistic results of Theorems 1–3.

Our theoretical insight suggests that in order to overcome the pessimism of ARM applied to classification, it is crucial to *structurally constrain* $r(\cdot, \cdot)$ in $\mathcal{U}_f$, or equivalently, to impose *structural assumptions* on the distribution shift. To this end, in this section, we propose DRSL that overcomes the limitation of the DRSL by incorporating structural assumptions on distribution shift.

Practical structural assumptions: In practice, there can be a variety of ways to impose structural assumptions on distribution shift. Here, as one possible way, we adopt the *latent prior probability change assumption* (Storkey & Sugiyama, 2007) because this particular class of assumptions enjoys the following two practical advantages.

1. Within the class, users of our DRSL can easily and intuitively model their assumptions on distribution shift (see the discussion at the end of this section).

2. Efficient learning algorithms can be derived (see Section 5).

Let us introduce a latent variable $z \in \mathcal{Z} \equiv \{1, \ldots, S\}$, which we call a *latent category*, where S is a constant. The latent prior probability change assumes

$$p(x, y|z) = q(x, y|z), \quad q(z) \neq p(z), \quad (14)$$

where p and q are the training and test densities, respectively. The intuition is that we assume a two-level *hierarchical* data-generation process: we first sample latent category z from the prior and then sample actual data (x, y) conditioned on z. We then assume that only the prior distribution over the latent categories changes, leaving the conditional distribution intact.

We assume the structural assumption in Eq. (14) to be provided externally by users of our DRSL based on their knowledge of potential distribution shift, rather than something to be inferred from data. As we will see at the end of this section, specifying Eq. (14) amounts to grouping training data points according to their latent categories, which is quite intuitive to do in practice.

Objective function of our DRSL: With the latent prior probability change of Eq. (14), the uncertainty set for test distributions in our DRSL becomes

$$\mathcal{Q}_p = \{q \ll p \mid \mathrm{D}_f[q(x, y, z)||p(x, y, z)] \leq \delta,$$
$$q(x, y|z) = p(x, y|z)\}. \quad (15)$$

Then, corresponding to Eq. (3), the objective of our DRSL can be written as

$$\min_{\theta} \underbrace{\sup_{w \in \mathcal{W}_f} \mathbb{E}_{p(x, y, z)} [w(z)\ell(g_{\theta}(x), y)]}_{\equiv \mathcal{R}_{\text{s-adv}}(\theta)}, \quad (16)$$

$$\mathcal{W}_f \equiv \left\{ w(z) \; \middle| \; \sum_{z \in \mathcal{Z}} p(z)f(w(z)) \leq \delta, \right.$$
$$\left. \sum_{z \in \mathcal{Z}} p(z)w(z) = 1, \; w(z) \geq 0, \; \forall z \in \mathcal{Z} \right\}, \quad (17)$$

where $w(z) \equiv q(z)/p(z) = q(x, y, z)/p(x, y, z)$ because of $q(x, y|z) = p(x, y|z)$. We call $\mathcal{R}_{\text{s-adv}}(\theta)$ the *structural adversarial risk* and call the minimization problem of Eq. (16) the *structural adversarial risk minimization* (structural ARM). Similarly to ARM, structural ARM is a minimax game between the learner and the adversary. Differently from ARM, the adversary of structural ARM (corresponding to $\sup_{w \in \mathcal{W}_f}$) uses $w(\cdot)$ to reweight data; hence, it has much less (only parametric) freedom to shift the test distribution compared to the adversary of ARM that uses non-parametric weight $r(\cdot, \cdot)$ (see Eq. (5)). Because of this limited freedom for the adversary, we can show that Theorems 1–3 do *not* hold for structural ARM, and we can expect to learn meaningful classifiers that are robust to *structurally constrained* distribution shift.

Discussion and proposal of evaluation metric for distributional robustness: Recall from Theorem 1 that when the 0-1 loss is used, the adversarial risk ends up being equivalent to the ordinary risk as an evaluation metric, which is too pessimistic as a metric for the distributional robustness of a classifier. In contrast, we can easily verify that our structural adversarial risk using the 0-1 loss does not suffer from the pessimism. We argue that our structural adversarial risk can be an alternative metric in distributionally robust classification. To better understand its property, inspired by Namkoong & Duchi (2017), we decompose it as[6]

$$\mathcal{R}_{\text{s-adv}}(\theta) = \underbrace{\mathcal{R}(\theta)}_{\text{(a) ordinary risk}} + \sqrt{\delta} \cdot \underbrace{\sqrt{\sum_{z \in \mathcal{Z}} p(z)(\mathcal{R}_z(\theta) - \mathcal{R}(\theta))^2}}_{\text{(b) sensitivity}}, \quad (18)$$

where $\mathcal{R}_z(\theta)(\equiv \mathbb{E}_{p(x,y|z)}[\ell(g_\theta(x), y)])$ is the risk of the classifier on latent category $z \in \mathcal{Z}$. We see that $\mathcal{R}_{\text{s-adv}}(\theta)$ in Eq. (18) contains the *risk variance* term (b). This variance term (b) can be large when the obtained classifier performs extremely poorly on a small number of latent categories. Once a test density concentrates on those poorly-performed latent categories, the test accuracy of the classifier can extremely deteriorate. In this sense, the classifier with large (b) is sensitive to distribution shift. In contrast, the small risk variance (b) indicates that the obtained classifier attains similar accuracy on all the latent categories. In such a case, the test accuracy of the classifier is insensitive to latent category prior change. In this sense, the classifier with small (b) is robust to distribution shift. To sum up, the additional term (b) measures the sensitivity of the classifier to the specified structural distribution shift.

On the basis of the above discussion, we see that $\mathcal{R}_{\text{s-adv}}(\theta)$ in Eq. (18) simultaneously captures (a) the ordinary risk, i.e., the mis-classification rate *when no distribution shift occurs*, and (b) the *sensitivity* to distribution shift. In this sense, our structural adversarial risk is an intuitive and reasonable metric for distributional robustness of a classifier, and we will employ this metric in our experiments in Section 6.

Empirical approximation: We explain how to empirically approximate the objective functions in Eqs. (16) and (17) using training data $\mathcal{D}' \equiv \{(x_1, y_1, z_1), \ldots, (x_N, y_N, z_N)\}$ drawn independently from $p(x, y, z)$.

Define $\mathcal{G}_s \equiv \{i \mid z_i = s, 1 \leq i \leq N\}$ for $1 \leq s \leq S$, which is a set of data indices belonging to latent category s. In our DRSL, users are responsible for specifying the groupings of training data points, i.e., $\{\mathcal{G}_s\}_{s=1}^S$. By specifying these groupings, the users incorporate their structural

---

[6]This particular decomposition holds when the Pearson (PE) divergence is used and δ is not so large. Refer to Appendix E for the derivation. Analogous decomposition can be also derived when other f-divergences are used.

assumptions on distribution shift into our DRSL. We will discuss how this can be done in practice at the end of this section.

For notational convenience, let $w_s \equiv w(s)$, $1 \leq s \leq S$, and define $\boldsymbol{w} \equiv (w_1, \ldots, w_S)$. Equations (16) and (17) can be empirically approximated as follows using $\mathcal{D}'$:

$$\min_\theta \underbrace{\sup_{\boldsymbol{w} \in \widehat{\mathcal{W}}_f} \frac{1}{N} \sum_{s=1}^S n_s w_s \widehat{\mathcal{R}}_s(\theta)}_{\equiv \widehat{\mathcal{R}}_{\text{s-adv}}(\theta)} \quad (19)$$

$$\widehat{\mathcal{W}}_f = \left\{ \boldsymbol{w} \in \mathbb{R}^S \middle| \frac{1}{N} \sum_{s=1}^S n_s f(w_s) \leq \delta, \right.$$
$$\left. \frac{1}{N} \sum_{s=1}^S n_s w_s = 1, \ \boldsymbol{w} \geq 0 \right\}, \quad (20)$$

where n_s is the cardinality of $\mathcal{G}_s$ and $\widehat{\mathcal{R}}_s(\theta)(\equiv \frac{1}{n_s} \sum_{i \in \mathcal{G}_s} \ell_i(\theta))$ is the average loss of all data points in $\mathcal{G}_s$. We call $\widehat{\mathcal{R}}_{\text{s-adv}}(\theta)$ the *structural adversarial empirical risk* and call the minimization problem of Eq. (19) the *structural adversarial empirical risk minimization* (structural AERM). We can add regularization term $\Omega(\theta)$ to Eq. (19) to prevent overfitting.

Convergence rate and estimation error: We establish the convergence rate of the model parameter and the order of the estimation error for structural AERM in terms of the number of training data points N. Due to the limited space, we only present an informal statement here. The formal statement can be found in Appendix G and its proof can be found in Appendix H.

Theorem 4 (Convergence rate and estimation error, informal statement). *Let θ^* be the solution of structural ARM, and $\widehat{\theta}_N$ be the solution of regularized structural AERM given training data of size N. Assume $g_\theta(x)$ is linear in θ, and regularization hyper-parameter λ decreases at a rate of $\mathcal{O}(N^{-1/2})$. Under mild conditions, as $N \to \infty$, we have $\|\widehat{\theta}_N - \theta^*\|_2 = \mathcal{O}(N^{-1/4})$ and consequently, $|\mathcal{R}_{\text{s-adv}}(\widehat{\theta}_N) - \mathcal{R}_{\text{s-adv}}(\theta^*)| = \mathcal{O}(N^{-1/4})$.*

Notice that the convergence rate of $\widehat{\theta}_N$ to θ^* is not the optimal parametric rate $\mathcal{O}(N^{-1/2})$. This is because the inner maximization of Eq. (19) converges in $\mathcal{O}(N^{-1/4})$ that slows down the entire convergence rate. Theorem 4 applies to any f-divergence where $f(t)$ is nonlinear in t, while knowing which f-divergence is used may improve the result to the optimal parametric rate.

Discussion on groupings: In our structural ARM and AERM, users need to incorporate their structural assumptions by grouping training data points. Here, we discuss how this can be done in practice.

Most straightforwardly, a user of our DRSL may assume

class prior change (Saerens et al., 2002) or sub-category[7] prior change. To incorporate such assumptions into our DRSL, the user can simply group training data by class labels or a sub-categories, respectively.

Alternatively, a user of our DRSL can group data by available meta-information of data such as time and places in which data are collected. The intuition is that data collected in the same situations (e.g., time and places) are likely to *"share the same destiny"* in the future distribution shift; hence, it is natural to assume that only the *prior* over the situations changes at test time while the conditionals remain the same.

In any case, it is crucial that the users provide structural assumptions on distribution shift so that we can overcome the pessimism of ARM and AERM for classification raised in Section 3.

5. Efficient Learning Algorithms

In this section, we derive efficient gradient-based learning algorithms for our structural AERM in Eq. (19). Thanks to Danskin's theorem (Danskin, 1966), we can obtain the gradient $\nabla_\theta \widehat{\mathcal{R}}_{\text{s-adv}}(\theta)$ as

$$\nabla_\theta \widehat{\mathcal{R}}_{\text{s-adv}}(\theta) = \frac{1}{N} \sum_{s=1}^{S} n_s w_s^* \nabla_\theta \widehat{\mathcal{R}}_s(\theta), \qquad (21)$$

where $\boldsymbol{w}^* = (w_1^*, \ldots, w_S^*)$ is the solution of inner maximization of AERM in Eq. (19).

In the following, we show that $\boldsymbol{w}^*$ can be obtained very efficiently for two well-known instances of f-divergences.

Kullback-Leibler (KL) divergence: For the KL divergence, $f(x) = x \log x$, we have

$$w_s^* = \frac{N}{Z(\gamma)} \cdot \exp\left(\frac{\widehat{\mathcal{R}}_s(\theta)}{\gamma}\right), \ 1 \leq s \leq S, \qquad (22)$$

where γ is a scalar such that the first constraint of $\widehat{\mathcal{W}}_f$ in Eq. (20) holds with equality, and $Z(\gamma) \equiv \sum_{s=1}^{S} n_s \exp\left(\widehat{\mathcal{R}}_s(\theta)/\gamma\right)$ is a normalizing constant in order to satisfy the second constraint of $\widehat{\mathcal{W}}_f$. To compute γ, we can simply perform a binary search.

Pearson (PE) divergence: For the PE divergence, $f(x) = (x-1)^2$. For small δ, $\boldsymbol{w} \geq 0$ of $\widehat{\mathcal{W}}_f$ is often satisfied in practice. We drop the inequality for simplicity; then, the solution of the inner maximization of Eq. (19) becomes analytic and efficient to obtain:

$$\boldsymbol{w}^* = \sqrt{\frac{N\delta}{\sum_{s=1}^{S} n_s v_s^2}} \boldsymbol{v} + \mathbf{1}_S, \qquad (23)$$

where $\mathbf{1}_S$ is the S-dimensional vector with all the elements equal to 1. $\boldsymbol{v}$ is the S-dimensional vector such that $v_s = \widehat{\mathcal{R}}_s(\theta) - \widehat{\mathcal{R}}(\theta)$, $1 \leq s \leq S$.

Computational complexity: The time complexity for obtaining $\boldsymbol{w}^*$ is: $\mathcal{O}(mS)$ for the KL divergence and $\mathcal{O}(S)$ for the PE divergence, where m is the number of the binary search iterations to compute γ in Eq. (22). Calculating the adversarial weights therefore adds negligible computational overheads to computing $\nabla \ell_i(\theta)$ and $\ell_i(\theta)$ for $1 \leq i \leq N$, which for example requires $\mathcal{O}(Nb)$-time for a b-dimensional linear-in-parameter model.

6. Experiments

In this section, we experimentally analyze our DRSL (structural AERM) in classification by comparing it with ordinary ERM and DRSL with f-divergences (AERM). We empirically demonstrate (i) the undesirability of AERM in classification and (ii) the robustness of structural AERM against specified distribution shift.

Datasets: We obtained six classification datasets from the UCI repository (Blake & Merz, 1998), two of which are for multi-class classification. We also obtained MNIST (LeCun et al., 1998) and 20newsgroups (Lang, 1995). Refer to Appendix I for the details of the datasets.

Evaluation metrics: We evaluated the three methods (ordinary ERM, AERM and structural AERM) with three kinds of metrics: the ordinary risk, the adversarial risk, and the structural adversarial risk, where the *0-1 loss is used* for all the metrics.[8] We did not explicitly report the adversarial risk in our experiments because of Theorem 1.

Both the risk and structural adversarial risk are estimated using held-out test data. In particular, the structural adversarial risk can be estimated similarly to Eqs. (19) and (20), i.e., calculating the mis-classification rate on the held-out test data and *structurally and adversarially reweight* them. See discussion of Eq. (18) for why the structural adversarial risk is a meaningful evaluation metric to measure distributional robustness of classifiers.

Experimental protocols: For our DRSL, we consider learning classifiers robust against (a) the class prior change and (b) the sub-category prior change. This corresponds to grouping training data by (a) class labels and (b) subcategory labels, respectively. In the benchmark datasets, the sub-category labels are not available. Hence, we manually created such labels as follows. First, we converted the original multi-class classification problems into classification problems with fewer classes by integrating some classes together. Then, the original class labels are regarded as the subcategories. In this way, we converted the satimage, letter and MNIST datasets into binary classification problems, and 20newsgroups into a 7-class classifica-

---

[7] A sub-category (Ristin et al., 2015) is a *refined* category of a class label, e.g., a "flu" label contains three *sub-categories*: types A, B, and C flu.

[8] To gain more insight on the methods, we also reported all the metrics in terms of the surrogate loss in Appendix K.

Table 1: Experimental comparisons of the three methods w.r.t. the estimated ordinary risk and the estimated structural adversarial risk *using the 0-1 loss (%)*. The lower these values are, the better the performance of the method is. The KL divergence is used and distribution shift is assumed to be (a) class prior change and (b) sub-category prior change. Mean and standard deviation over 50 random train-test splits were reported. The best method and comparable ones based on the t-test at the significance level 1% are highlighted in boldface.

(a) Class prior change.

Dataset	Estimated ordinary risk			Estimated structural adversarial risk		
	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM
blood	**22.4 (0.7)**	26.7 (5.0)	33.4 (2.0)	62.3 (2.4)	53.5 (10.6)	**36.7 (1.9)**
adult	**15.3 (0.2)**	15.4 (0.2)	18.7 (0.2)	30.4 (0.4)	30.4 (0.5)	**19.1 (0.3)**
fourclass	**24.0 (1.4)**	25.7 (2.6)	27.2 (1.4)	36.9 (2.2)	38.0 (6.2)	**29.5 (2.0)**
phishing	**6.1 (0.2)**	6.4 (0.2)	**6.0 (0.2)**	7.6 (0.4)	8.2 (0.4)	**6.5 (0.3)**
20news	**28.9 (0.4)**	30.6 (0.4)	34.7 (0.4)	44.1 (0.5)	45.0 (0.5)	**40.0 (0.6)**
satimage	**25.2 (0.3)**	30.8 (0.3)	32.2 (0.4)	**39.5 (0.6)**	47.9 (0.5)	**39.6 (0.6)**
letter	**14.3 (0.5)**	15.2 (0.6)	19.3 (0.8)	36.6 (1.5)	34.7 (1.7)	**22.5 (1.0)**
mnist	**10.0 (0.1)**	11.5 (0.1)	12.7 (0.1)	14.4 (0.2)	15.9 (0.2)	**13.8 (0.2)**

(b) Sub-category prior change.

Dataset	Estimated ordinary risk			Estimated structural adversarial risk		
	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM
20news	**19.0 (0.3)**	20.6 (0.4)	23.3 (0.5)	35.6 (0.6)	37.8 (0.9)	**31.1 (0.5)**
satimage	**36.4 (0.3)**	44.2 (2.5)	40.7 (0.9)	62.1 (0.6)	66.2 (4.2)	**50.5 (0.6)**
letter	**17.5 (0.4)**	28.0 (5.4)	34.2 (2.0)	52.1 (0.6)	60.1 (5.8)	**43.0 (1.7)**
mnist	**13.3 (0.1)**	13.7 (0.1)	17.3 (0.2)	22.6 (0.2)	23.2 (0.2)	**19.9 (0.2)**

tion. Appendix J details how we grouped the class labels.

For all the methods, we used linear models with softmax output for the prediction function $g_\theta(x)$. The cross-entropy loss with ℓ_2 regularization was adopted. The regularization hyper-parameter λ was selected from $\{1.0, 0.1, 0.01, 0.001, 0.0001\}$ via 5-fold cross validation.

We used the two f-divergences (the KL and PE divergences) and set $\delta = 0.5$ for AERM and structural AERM. The same δ and f-divergence were used for estimating the structural adversarial risk. At the end of this section, we discuss how we can choose δ in practice.

Results: In Table 1, we report experimental results on the classification tasks when the KL divergence is used. Refer to Appendix L for the results when the PE divergence is used, which showed similar tendency.

We see from the left half of Table 1 that ordinary ERM achieved lower estimated risks as expected. On the other hand, we see from the entire Table 1 that AERM, which does not incorporate any structural assumptions on distribution shift, performed poorly in terms of both of two evaluation metrics; hence, it also performed poorly in terms of the adversarial risk (see Theorem 1). This may be because AERM was excessively sensitive to outliers as implied by Theorem 3. We see from the right half of Table 1 that structural AERM achieved significantly lower estimated structural adversarial risks. Although this was expected, our experiments confirmed that structural AERM indeed obtained classifiers robust against the *structural* distribution shift.[9]

---

[9]When we used the surrogate loss to evaluate the methods

Discussion: Here we provide an insight for users to determine δ in our DRSL (structural ARM and AERM). We see from Eq. (18) that the structural adversarial risk can be decomposed into the sum of the ordinary risk and the sensitivity term, where δ acts as a trade-off hyper-parameter between the two terms. In practice, users of our DRSL may want to have good balance between the two terms, i.e., the learned classifier should achieve high accuracy on the training distribution while being robust to specified distribution shift. Since both terms in Eq. (18) can be estimated by cross validation, the users can adjust δ of AERM at training time to best trade-off the two terms for their purposes, e.g., increasing δ during training to decrease the sensitivity term at the expense of a slight increase of the risk term.

7. Conclusion

In this paper, we theoretically analyzed DRSL with f-divergences applied to classification. We showed that the DRSL ends up giving a classifier optimal for the *training distribution*, which is too pessimistic in terms of the original motivation of distributionally robust classification. To rectify this, we presented simple DRSL that gives a robust classifier based on structural assumptions on distribution shift. We derived efficient optimization algorithms for our DRSL and empirically demonstrated its effectiveness.

(which is not the case in ordinary classification), we confirmed that the methods indeed achieved the best performance in terms of the metrics they optimized for, i.e., ERM, AERM, and structural AERM performed the best in terms of the ordinary risk, adversarial risk and structural adversarial risk, respectively. See Appendix K for the actual experimental results.

Acknowledgement

References

Bagnell, J. A. Robust supervised learning. In *Proceedings of Association for the Advancement of Artificial Intelligence*, volume 20, pp. 714–719, 2005.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

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

Blake, C. and Merz, C. J. UCI repository of machine learning databases. http://archive.ics.uci.edu/ml/index.html, 1998.

Blanchet, J., Kang, Y., and Murthy, K. Robust wasserstein profile inference and applications to machine learning. *arXiv preprint arXiv:1610.05627*, 2016.

Bonnans, J. F. and Shapiro, A. Optimization problems with perturbations, a guided tour. *SIAM Review*, 40(2):228–264, 1998.

Chung, K. L. *A Course in Probability Theory*. Academic Press, 1968.

Ciszar, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Danskin, J. M. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

Duchi, J., Glynn, P., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

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

Globerson, A. and Roweis, S. Nightmare at test time: robust learning by feature deletion. In *Proceedings of International Conference on Machine learning*, pp. 353–360, 2006.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *Proceedings of International Conference on Machine learning*, 2018.

Lang, K. Newsweeder: Learning to filter netnews. In *Proceedings of International Conference on Machine Learning*, pp. 331–339, 1995.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324, 1998.

Liu, A. and Ziebart, B. Robust classification under sample selection bias. In *Advances in Neural Information Processing Systems*, pp. 37–45, 2014.

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

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pp. 2975–2984, 2017.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. The MIT Press, 2009.

Ristin, M., Gall, J., Guillaumin, M., and Van Gool, L. From categories to subcategories: large-scale image classification with partial class label refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 231–239, 2015.

Robinson, S. M. A characterization of stability in linear programming. *Operations Research*, 25:435–447, 1977.

Saerens, M., Latinne, P., and Decaestecker, C. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Sinha, A., Namkoong, H., and Duchi, J. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Storkey, A. J. and Sugiyama, M. Mixture regression for covariate shift. In *Advances in Neural Information Processing Systems*, pp. 1337–1344, 2007.

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8(May):1007–1025, 2007.

Vapnik, V. N. *Statistical Learning Theory*. Wiley New York, 1998.

A. Proof of Theorem 1

Given θ (the parameter of the prediction function), we obtain the adversarial risk by the following optimization:

$$\max_{r \in \mathcal{U}_f} \mathbb{E}_{p(x,y)}[r(x,y)\ell(g_\theta(x),y)] \tag{24}$$

$$\text{where} \quad \mathcal{U}_f \equiv \{r(x,y) \mid \mathbb{E}_{p(x,y)}[f(r(x,y))] \leq \delta, \ \mathbb{E}_{p(x,y)}[r(x,y)] = 1, \ r(x,y) \geq 0 \ (\forall (x,y) \in \mathcal{X} \times \mathcal{Y})\}. \tag{25}$$

Here, we are considering that $\ell(\cdot,\cdot)$ is the 0-1 loss. Let $\Omega_\theta^{(0)} \equiv \{(x,y) \mid \ell(g_\theta(x),y) = 0\} \subseteq \mathcal{X} \times \mathcal{Y}$. Then, we have $\Omega_\theta^{(1)} \equiv \{(x,y) \mid \ell(g_\theta(x),y) = 1\} = \mathcal{X} \times \mathcal{Y} \setminus \Omega_\theta$. Let $r^*(\cdot,\cdot)$ be the optimal solution of Eq. (24).

Note that Eq. (24) is a convex optimization problem. This is because the objective is linear in $r(\cdot,\cdot)$ and the uncertainty set is convex, which follows from the fact that $f(\cdot)$ is convex. Therefore, any local maximum of Eq. (24) is the global maximum. Nonetheless, there can be multiple solutions that attain the same global maxima. Among those solutions, we now show that there exists $r^*(\cdot,\cdot)$ such that it takes the same values within $\Omega_\theta^{(0)}$ and $\Omega_\theta^{(1)}$, respectively, i.e.,

$$r^*(x,y) = r_0^* \ \text{for} \ \forall(x,y) \in \Omega_\theta^{(0)}, \qquad r^*(x,y) = r_1^* \ \text{for} \ \forall(x,y) \in \Omega_\theta^{(1)}, \tag{26}$$

where r_0^* and r_1^* are some constant values. This is because for any given optimal solution $r'^*(\cdot,\cdot)$ of Eq. (24), we can always obtain optimal solution $r^*(\cdot,\cdot)$ that satisfies Eq. (26) by the following transformation:

$$r_0^* = \mathbb{E}_{p(x,y)}[r'^*(x,y) \cdot \mathbf{1}\{(x,y) \in \Omega_\theta^{(0)}\}]/\mathbb{E}_{p(x,y)}[\mathbf{1}\{(x,y) \in \Omega_\theta^{(0)}\}], \tag{27}$$

$$r_1^* = \mathbb{E}_{p(x,y)}[r'^*(x,y) \cdot \mathbf{1}\{(x,y) \in \Omega_\theta^{(1)}\}]/\mathbb{E}_{p(x,y)}[\mathbf{1}\{(x,y) \in \Omega_\theta^{(1)}\}], \tag{28}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function. Eqs. (27) and (28) are simple average operations of $r'^*(\cdot,\cdot)$ on regions $\Omega_\theta^{(0)}$ and $\Omega_\theta^{(1)}$, respectively. Utilizing the convexity of $f(\cdot)$, it is straightforward to see that $r^*(\cdot,\cdot)$ constructed in this way is still in the feasible region of Eq. (25). It also attains exactly the same objective of Eq. (24) as the original solution $r^{*\prime}(\cdot,\cdot)$ does. This concludes our proof that there exists an optimal solution of Eq. (24) such that it takes the same values within $\Omega_\theta^{(0)}$ and $\Omega_\theta^{(1)}$, respectively.

Let $p_{\Omega_\theta^{(i)}} = \mathbb{E}_{p(x,y)}[\mathbf{1}\{(x,y) \in \Omega_\theta^{(i)}\}]$ for $i = 0,1$, which is the proportion of data that is correctly and incorrectly classified, respectively. We note that $p_{\Omega_\theta^{(0)}} + p_{\Omega_\theta^{(1)}} = 1$. Also, we see that $p_{\Omega_\theta^{(1)}}$ is by definition the misclassification rate; thus, it is equal to the ordinary risk, i.e., $\mathbb{E}_{p(x,y)}[\ell(g_\theta(x),y)]$. By using Eq. (26), we can simplify Eq. (24) as

$$\sup_{(r_0,r_1) \in \mathcal{U}_f'} p_{\Omega_\theta^{(1)}} r_1, \tag{29}$$

$$\text{where} \quad \mathcal{U}_f' \equiv \{(r_0,r_1) \mid p_{\Omega_\theta^{(0)}} f(r_0) + p_{\Omega_\theta^{(1)}} f(r_1) \leq \delta, \ p_{\Omega_\theta^{(0)}} r_0 + p_{\Omega_\theta^{(1)}} r_1 = 1, \ r_0 \geq 0, \ r_1 \geq 0\}. \tag{30}$$

In the following, we show that Eq. (29) has monotonic relationship with $p_{\Omega_\theta^{(1)}}$. With a fixed value of $p_{\Omega_\theta^{(1)}}$, we can obtain the optimal (r_0,r_1) by solving Eq. (29). Let $(r_0^*(p_1), r_1^*(p_1))$ be the solution of Eq. (29) when $p_{\Omega_\theta^{(1)}}$ is fixed to p_1 with $0 \leq p_1 \leq 1$.

First, we note that the first inequality constraint in Eq. (30) is a convex set and includes $(1,1)$ in its relative interior because $p_{\Omega_\theta^{(0)}} f(1) + p_{\Omega_\theta^{(1)}} f(1) = 0 < \delta$ and $p_{\Omega_\theta^{(0)}} \cdot 1 + p_{\Omega_\theta^{(1)}} \cdot 1 = 1$. Note that in a two dimensional space, the number of intersections between a line and a boundary of a convex set is at most two. For $\delta > 0$, there are always exactly two different points that satisfies both $p_{\Omega_\theta^{(0)}} f(r_0) + p_{\Omega_\theta^{(1)}} f(r_1) = \delta$ and $p_{\Omega_\theta^{(0)}} r_0 + p_{\Omega_\theta^{(1)}} r_1 = 1$. We further see that the optimal solution of r_1 is always greater than 1 because the objective in Eq. (29) is an increasing function of r_1. Taking these facts into account, we can see that the optimal solution, $(r_0^*(p_1), r_0^*(p_1))$, satisfies either of the following two cases depending on whether the inequality constraint $r_0 \geq 0$ in Eq. (30) is active or not.

Case 1: $\quad p_0 \cdot f(r_0^*(p_1)) + p_1 \cdot f(r_1^*(p_1)) = \delta, \ p_0 \cdot r_0^*(p_1) + p_1 \cdot r_1^*(p_1) = 1, \ 0 < r_0^*(p_1) < 1 < r_1^*(p_1). \tag{31}$

Case 2: $\quad r_0^*(p_1) = 0, \ r_1^*(p_1) = \dfrac{1}{p_1}, \tag{32}$

where $p_0 = 1 - p_1$. We now show that there is the monotonic relation between $p_{\Omega_\theta^{(1)}}$ and Eq. (29) for both the cases. To this end, pick any p_1' such that $p_1 < p_1' \leq 1$, and let $(r_0^*(p_1'), r_1^*(p_1'))$ be the solution of Eq. (29) when $p_{\Omega_\theta^{(1)}}$ is fixed to p_1'.

Regarding Case 2 in Eq. (32), the adversarial risk in Eq. (29) is $p_1 \cdot \frac{1}{p_1} = 1$. On the other hand, it is easy to see that the active equality constraint stays $r_0 \geq 0$ in Eq. (30) for $p_{\Omega_\theta^{(1)}}$ larger p_1. Hence, we can show that $r_0^*(p_1') = 0, r_1^*(p_1') = \frac{1}{p_1'}$. Therefore, the adversarial risk in Eq. (29) stays $p_1' \cdot \frac{1}{p_1'} = 1$. This concludes our proof for Eq. (10) in Theorem 1.

Regarding Case 1 in Eq. (31), we note that both the ordinary risk p_1 and the adversarial risk $p_1 \cdot r_1^*(p_1)$ are strictly less than 1. Our goal is to show Eq. (9) in Theorem 1, which is equivalent to showing

$$p_1 \cdot r_1^*(p_1) < p_1' \cdot r_1^*(p_1'). \tag{33}$$

To do so, we further consider the following two sub-cases of Case 1 in Eq. (31):

$$\textbf{Case 1-a:} \qquad p_1' < p_1 \cdot r_1^*(p_1) \tag{34}$$
$$\textbf{Case 1-b:} \qquad p_1' \geq p_1 \cdot r_1^*(p_1). \tag{35}$$

In Case 1-b, we can straightforwardly show Eq. (33) as follows.

$$p_1 \cdot r_1^*(p_1) \leq p_1' < p_1' \cdot r_1^*(p_1'), \tag{36}$$

where the last inequality follows from $1 < r_1^*(p_1')$.

Now, assume Cases 1 and 1-a in Eqs. (31) and (34). Our goal is to show that $r_1^*(p_1')$ satisfies $r_1^*(p_1') > \frac{p_1}{p_1'} r_1^*(p_1)$ because then Eq. (33) holds. To this end, we show that

$$r_1' = \frac{p_1}{p_1'} \cdot r_1^*(p_1) \tag{37}$$

is contained in the relative interior (excluding the boundary) of Eq. (30) with $p_{\Omega_\theta^{(1)}}$ fixed to p_1'. Then, because our objective in Eq. (29) is linear in r_1, $r_1' < r_1^*(p_1')$ holds in our setting. Then, we arrive at $r_1^*(p_1') > \frac{p_1}{p_1'} \cdot r_1^*(p_1)$. Formally, our goal is to show r_1' in Eq. (37) satisfies

$$p_0' \cdot f(r_0') + p_1' \cdot f(r_1') < \delta, \quad p_0' r_0' + p_1' r_1' = 1, \quad r_0' > 0, \quad r_1' > 0, \tag{38}$$

where $p_0' = 1 - p_1'$. By Eqs. (31), (37) and the second equality of Eq. (38), we have

$$r_0' = \frac{1 - p_1' r_1'}{p_0'} = \frac{1 - p_1 \cdot r_1^*(p_1)}{p_0'} = \frac{p_0}{p_0'} \cdot r_0^*(p_1). \tag{39}$$

The latter two inequalities of Eq. (38), i.e., $r_0' > 0$ and $r_1' > 0$ follow straightforwardly from the assumptions. Combining the assumption in Eq. (34) and the last inequality in Eq. (31), we have the following inequality.

$$0 < r_0^*(p_1) < r_0' < 1 < r_1' < r_1^*(p_1). \tag{40}$$

Thus, we can write r_0' (resp. r_1') as a linear interpolation of $r_0^*(p_1)$ and 1 (resp. 1 and $r_1^*(p_1)$) as follows.

$$r_0' = \alpha \cdot r_0^*(p_1) + (1 - \alpha) \cdot 1, \qquad r_1' = \beta \cdot r_1^*(p_1) + (1 - \beta) \cdot 1, \tag{41}$$

where $0 < \alpha, \beta < 1$. Substituting Eqs. (37) and (39), we have

$$\alpha = \frac{1}{1 - r_0^*(p_1)} \cdot \frac{p_0' - p_0 \cdot r_0^*(p_1)}{p_0'}, \tag{42}$$

$$\beta = \frac{1}{r_1^*(p_1) - 1} \cdot \frac{p_1 \cdot r_1^*(p_1) - p_1'}{p_1'} = \frac{1}{r_1^*(p_1) - 1} \cdot \frac{p_0' - p_0 \cdot r_0^*(p_1)}{p_1'}. \tag{43}$$

Then, we have

$$
\begin{aligned}
p_0' f(r_0') + p_1' f(r_1') &= p_0' f(\alpha \cdot r_0^*(p_1) + (1-\alpha) \cdot 1) + p_1' f(\beta \cdot r_1^*(p_1) + (1-\beta) \cdot 1) \\
&\leq p_0' \cdot \{\alpha \cdot f(r_0^*(p_1)) + (1-\alpha) \cdot f(1)\} + p_1' \cdot \{\beta \cdot f(r_1^*(p_1)) + (1-\beta) \cdot f(1)\} \;\; (\because \text{convexity of } f(\cdot)) \\
&= p_0' \alpha \cdot f(r_0^*(p_1)) + p_1' \beta \cdot f(r_1^*(p_1)) \;\; (\because f(1) = 0) \\
&= (p_0' - p_0 \cdot r_0^*(p_1)) \left(\frac{1}{1 - r_0^*(p_1)} \cdot f(r_0^*(p_1)) + \frac{1}{r_1^*(p_1) - 1} \cdot f(r_1^*(p_1)) \right) \;\; (\because \text{Eqs. (42) and (43)}) \\
&= (p_1 \cdot r_1^*(p_1) - p_1') \left(\frac{1}{1 - r_0^*(p_1)} \cdot f(r_0^*(p_1)) + \frac{1}{r_1^*(p_1) - 1} \cdot f(r_1^*(p_1)) \right) \\
&< (p_1 \cdot r_1^*(p_1) - p_1) \left(\frac{1}{1 - r_0^*(p_1)} \cdot f(r_0^*(p_1)) + \frac{1}{r_1^*(p_1) - 1} \cdot f(r_1^*(p_1)) \right) \;\; (\because p_1' > p_1.) \\
&= p_0 \cdot f(r_0^*(p_1)) + p_1 \cdot f(r_1^*(p_1)) \\
&= \delta. \;\; (\because \text{the first equation of Eq. (31).})
\end{aligned}
$$

This concludes our proof that Eq. (29) has monotonic relationship with $p_{\Omega_\theta^{(1)}}$. Recall that $p_{\Omega_\theta^{(1)}}$ is by definition equal to the ordinary risk, $\mathcal{R}(\theta)$. Therefore, for any pair of parameters θ_1 and θ_2, if $\mathcal{R}_{\mathrm{adv}}(\theta_1) < 1$, we have

$$
\mathcal{R}(\theta_1) < \mathcal{R}(\theta_2) \implies \mathcal{R}_{\mathrm{adv}}(\theta_1) < \mathcal{R}_{\mathrm{adv}}(\theta_2). \tag{44}
$$

To show that the opposite direction of Eq. (44) holds, we need to show that any pair of parameters θ_1 and θ_2, the following holds:

$$
\mathcal{R}(\theta_1) = \mathcal{R}(\theta_2) \implies \mathcal{R}_{\mathrm{adv}}(\theta_1) = \mathcal{R}_{\mathrm{adv}}(\theta_2). \tag{45}
$$

This is obvious from Eq. (29) because the adversarial risk depends on the parameter of the model *only through the risk of the model.* This concludes the proof of Theorem 1, in which the adversarial risk and ordinary risk are compared. For the case of empirical approximations, the same argument can be used by replacing the expectations with empirical averages. $\quad\square$

B. Proof of Theorem 2

We prove by contradiction. Let Ω be the subset of $\mathbb{R}^{|\mathcal{Y}|}$. We consider the multi-class classification and assume that the loss $\ell(\cdot, \cdot) : \Omega \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is classification-calibrated. Although we will mainly focus on a multi-class classification scenario, our proof easily extends to a binary classification scenario, which we will discuss at the end of the proof.

Let $g(\cdot) : \mathcal{X} \to \mathbb{R}^K$ be prediction function, where K is the number of classes. Assume the prediction function, g, can take any measurable functions. Then, g^* that minimizes the ordinary risk using the classification-calibrated loss, $\ell(\cdot, \cdot)$, i.e.,

$$
\begin{aligned}
g^* &= \arg\min_g \mathbb{E}_{p(x,y)} \left[\ell(g(x), y) \right] \\
&= \arg\min_g \mathbb{E}_{p(x)} \left[\sum_{y \in \mathcal{Y}} p(y|x) \ell(g(x), y) \right],
\end{aligned} \tag{46}
$$

is the Bayes optimal classifier[10] (Bartlett et al., 2006; Tewari & Bartlett, 2007).

Our goal is to show that $g^{(\mathrm{adv})}$ that minimizes the adversarial risk using classification-calibrated loss is also Bayes optimal w.r.t. $p(x, y)$. More specifically, we consider

$$
g^{(\mathrm{adv})} = \arg\min_g \sup_{q : D_f[q||p] \leq \delta} \mathbb{E}_{q(x,y)} [\ell(g(x), y)] \tag{47}
$$

$$
= \arg\min_g \sup_{r(\cdot,\cdot) \in \mathcal{U}_f} \mathbb{E}_{p(x,y)} [r(x,y) \ell(g(x), y)]. \tag{48}
$$

---

[10]The classifier that minimizes the mis-classification rate for the training density $p(x, y)$ (the 0-1 loss is considered), i.e., the classifier whose prediction on x is equal to $\arg\max_{y \in \mathcal{Y}} p(y|x)$.

Recall that $q(x, y)$ in Eq. (47) and $r(x, y)$ in Eq. (48) are related by $r(x, y) \equiv q(x, y)/p(x, y)$. In the following, with a slight abuse of notation, we denote $r(x) \equiv q(x)/p(x)$ and $r(y|x) \equiv q(y|x)/p(y|x)$. Obviously, we have $r(x, y) = r(x)r(y|x)$.

Let $r^*(\cdot, \cdot)$ be the solution of the inner maximization of Eq. (48) with $g^{(\mathrm{adv})}$, i.e.,

$$r^*(\cdot, \cdot) = \arg\max_{r(\cdot,\cdot) \in \mathcal{U}_f} \mathbb{E}_{p(x,y)} \left[r(x, y)\ell(g^{(\mathrm{adv})}(x), y) \right]. \tag{49}$$

Then, by Danskin's theorem (Danskin, 1966), Eq. (48) can be rewritten as

$$
\begin{aligned}
g^{(\mathrm{adv})} &= \arg\min_g \mathbb{E}_{p(x,y)} \left[r^*(x, y)\ell(g(x), y) \right] \\
&= \arg\min_g \mathbb{E}_{p(x)} \left[r^*(x)\mathbb{E}_{p(y|x)}[r^*(y|x)\ell(g(x), y)] \right] \\
&= \arg\min_g \mathbb{E}_{p(x)} \left[r^*(x) \sum_{y \in \mathcal{Y}} p(y|x)r^*(y|x)\ell(g(x), y) \right].
\end{aligned}
\tag{50}
$$

Now, suppose that $g^{(\mathrm{adv})}$ is not Bayes optimal almost surely over $q^*(x) \equiv r^*(x)p(x)$. Then, we have

$$\int_{x \in \mathcal{S}} q^*(x)\mathrm{d}x > 0, \tag{51}$$

where

$$\mathcal{S} \equiv \left\{ x \ \middle| \ x \in \mathcal{X}, \ p(x) > 0, \ q^*(x) > 0, \ \arg\max_{y \in \mathcal{Y}} p(y|x) \neq \arg\max_{y \in \mathcal{Y}} g_y^{(\mathrm{adv})}(x) \right\}. \tag{52}$$

In the following, we denote $x \in \mathcal{S}$ by $x^\dagger$, i.e., whenever we denote $x^\dagger$, we implicitly assume $x^\dagger \in \mathcal{S}$. We immediately have $r^*(x^\dagger) = q^*(x^\dagger)/p(x^\dagger) > 0$. We let $y^{(\mathrm{max})}(x^\dagger) \equiv \arg\max_{y \in \mathcal{Y}} p(y|x^\dagger)$ and $y^{(\mathrm{adv})}(x^\dagger) \equiv \arg\max_{y \in \mathcal{Y}} g_y^{(\mathrm{adv})}(x^\dagger)$. Since $\ell(\cdot, \cdot)$ is classification-calibrated, from Eq. (50) and the definition of the classification-calibrated loss (Bartlett et al., 2006; Tewari & Bartlett, 2007), we have

$$y^{(\mathrm{adv})}(x^\dagger) = \arg\max_{y \in \mathcal{Y}} p(y|x^\dagger)r^*(y|x^\dagger). \tag{53}$$

Because we have $x^\dagger \in \mathcal{S}$, $y^{(\mathrm{adv})}(x^\dagger) \neq y^{(\mathrm{max})}(x^\dagger)$ holds. Thus, we have

$$p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger)r^*(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) > p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger)r^*(y^{(\mathrm{max})}(x^\dagger)|x^\dagger). \tag{54}$$

Combining this with $p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) > p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger)$, we have

$$r^*(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) > r^*(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) > 0. \tag{55}$$

We construct a new ratio function, $r_{\mathrm{new}}(\cdot, \cdot)$, by the following operations. We first set $r_{\mathrm{new}}(x, y) \leftarrow r^*(x, y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Then, for $\forall x \in \mathcal{S}$, we let

$$r_{\mathrm{new}}(y^{(\mathrm{max})}(x)|x) \leftarrow r^*(y^{(\mathrm{max})}(x)|x) + \epsilon \cdot p(y^{(\mathrm{adv})}(x)|x), \tag{56}$$

$$r_{\mathrm{new}}(y^{(\mathrm{adv})}(x)|x) \leftarrow r^*(y^{(\mathrm{adv})}(x)|x) - \epsilon \cdot p(y^{(\mathrm{max})}(x)|x), \tag{57}$$

where $\epsilon > 0$ is a sufficiently small number. We show that such $r_{\mathrm{new}}(\cdot, \cdot)$ is still in $\mathcal{U}_f$. As shown in Eqs. (56) and (57), the value of $r_{\mathrm{new}}(\cdot, \cdot)$ changed from $r^*(\cdot, \cdot)$ only in $\mathcal{S}$. Therefore, given that $r^*(\cdot, \cdot) \in \mathcal{U}_f$, in order to show $r_{\mathrm{new}}(\cdot, \cdot) \in \mathcal{U}_f$, it is sufficient to show the following three equality/inequalities:

$$\mathbb{E}_{p(x,y)} \left[f\left(r_{\mathrm{new}}(x, y)\right) \right] \leq \delta, \ \mathbb{E}_{p(x,y)}[r_{\mathrm{new}}(x, y)] = 1, \ r_{\mathrm{new}}(x, y) \geq 0 \ (\forall(x, y) \in \mathcal{X} \times \mathcal{Y}). \tag{58}$$

Because we know $r^*(\cdot, \cdot) \in \mathcal{U}_f$, it is sufficient to show

$$
\begin{aligned}
&\mathbb{E}_{p(x,y)} \left[f\left(r_{\mathrm{new}}(x, y)\right) \right] \leq \mathbb{E}_{p(x,y)} \left[f\left(r^*(x, y)\right) \right], \\
&\mathbb{E}_{p(x,y)}[r_{\mathrm{new}}(x, y)] = \mathbb{E}_{p(x,y)}[r^*(x, y)], \\
&r_{\mathrm{new}}(x, y) \geq 0, \forall(x, y) \in \mathcal{X} \times \mathcal{Y}.
\end{aligned}
\tag{59}
$$

Furthermore, $r_{\mathrm{new}}(\cdot, \cdot)$ only differs from $r^*(\cdot, \cdot)$ in $(x, y^{(\mathrm{max})}) \in \mathcal{S} \times \mathcal{Y}$ and $(x, y^{(\mathrm{adv})}) \in \mathcal{S} \times \mathcal{Y}$. Therefore, to show Eq. (59), all we need to show is

$$\int_{x \in \mathcal{S}} \left\{ p(x, y^{(\mathrm{adv})}(x)) \cdot f(r_{\mathrm{new}}(x, y^{(\mathrm{adv})}(x))) + p(x^\dagger, y^{(\mathrm{max})}(x)) \cdot f(r_{\mathrm{new}}(x, y^{(\mathrm{max})}(x))) \right\} \mathrm{d}x$$

$$\leq \int_{x \in \mathcal{S}} \left\{ p(x, y^{(\mathrm{adv})}(x)) \cdot f(r^*(x, y^{(\mathrm{adv})}(x))) + p(x, y^{(\mathrm{max})}(x)) \cdot f(r^*(x, y^{(\mathrm{max})}(x))) \right\} \mathrm{d}x, \tag{60}$$

$$\int_{x \in \mathcal{S}} \left\{ p(x, y^{(\mathrm{adv})}(x)) \cdot r_{\mathrm{new}}(x, y^{(\mathrm{adv})}(x)) + p(x, y^{(\mathrm{max})}(x)) \cdot r_{\mathrm{new}}(x, y^{(\mathrm{max})}(x)) \right\} \mathrm{d}x$$

$$= \int_{x \in \mathcal{S}} \left\{ p(x, y^{(\mathrm{adv})}(x)) \cdot r^*(x, y^{(\mathrm{adv})}(x)) + p(x, y^{(\mathrm{max})}(x)) \cdot r^*(x, y^{(\mathrm{max})}(x)) \right\} \mathrm{d}x, \tag{61}$$

$$r_{\mathrm{new}}(x, y^{(\mathrm{adv})}(x)) \geq 0, \ \ r_{\mathrm{new}}(x, y^{(\mathrm{max})}(x)) \geq 0, \ \forall x \in \mathcal{S}. \tag{62}$$

First, since $r^*(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) > 0$ holds from Eq. (54), $r_{\mathrm{new}}(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) = r^*(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) - \epsilon \cdot p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \geq 0$ holds for sufficiently small ϵ. Thus, $r_{\mathrm{new}}(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) = r_{\mathrm{new}}(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) r(x^\dagger) \geq 0$. Hence, Eq. (62) holds for sufficiently small ϵ. Also, Eq. (61) follows because

Integrand of L.H.S. of Eq. (61)

$$= p(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \cdot \left[r_{\mathrm{new}}(x^\dagger) \cdot \left\{ r^*(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) - \epsilon \cdot p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \right\} \right]$$

$$+ p(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \cdot \left[r_{\mathrm{new}}(x^\dagger) \cdot \left\{ r^*(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) + \epsilon \cdot p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \right\} \right] \tag{63}$$

$$= p(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \cdot \left[r^*(x^\dagger) \cdot \left\{ r^*(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) - \epsilon \cdot p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \right\} \right]$$

$$+ p(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \cdot \left[r^*(x^\dagger) \cdot \left\{ r^*(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) + \epsilon \cdot p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \right\} \right] \tag{64}$$

$$= p(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \cdot r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) + p(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \cdot r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger))$$

$$+ \epsilon \cdot r^*(x^\dagger) p(x^\dagger) \underbrace{\left\{ p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) - p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \right\}}_{=0} \tag{65}$$

$$= \text{Integrand of R.H.S. of Eq. (61)} \tag{66}$$

Finally, we show Eq. (60). Substituting Eqs. (56) and (57) into the L.H.S. of Eq. (60), we have

Integrand of L.H.S. of Eq. (60)

$$= p(x^\dagger) \cdot \left\{ p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \cdot f\left(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) - \epsilon \cdot r^*(x^\dagger) p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \right) \right.$$

$$\left. + p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \cdot f\left(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) + \epsilon \cdot r^*(x^\dagger) p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \right) \right\}. \tag{67}$$

Because $f(\cdot)$ is differentiable, we can apply the first order Taylor expansion to the two terms involving $f(\cdot)$ in Eq. (67) as

$$f\left(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) - \epsilon \cdot r^*(x^\dagger) p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \right)$$

$$= f\left(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \right) - \epsilon \cdot f'\left(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \right) \cdot r^*(x^\dagger) p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) + \mathcal{O}(\epsilon^2),$$

$$f\left(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) + \epsilon \cdot r^*(x^\dagger) p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \right)$$

$$= f\left(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \right) + \epsilon \cdot f'\left(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \right) \cdot r^*(x^\dagger) p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) + \mathcal{O}(\epsilon^2). \tag{68}$$

Substituting Eq. (68) into Eq. (67), we have

Integrand of L.H.S. of Eq. (60)

$$= p(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \cdot f\left(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \right) + p(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \cdot f\left(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \right)$$

$$+ \epsilon \cdot r^*(x^\dagger) p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \cdot \left\{ f'\left(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \right) - f'\left(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \right) \right\} + \mathcal{O}(\epsilon^2), \tag{69}$$

Since $f(\cdot)$ is convex, its derivative $f'(\cdot)$ is non-decreasing. Also $r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) > r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger))$ holds because of Eq. (55). Therefore, we have $f'(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger))) - f'(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger))) \leq 0$.

First, assume $f'(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger))) - f'(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger))) = 0$. Then, $f(\cdot)$ is exactly linear in the interval of $[r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger)), r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger))]$; hence, $\mathcal{O}(\epsilon^2)$ term in Eq. (69) is exactly 0 for sufficiently small ϵ. Hence, we have

$$
\begin{aligned}
\text{Eq. (69)} &= p(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \cdot f(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger))) + p(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \cdot f(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger))) \\
&= \text{Integrand of R.H.S. of Eq. (60).}
\end{aligned}
\tag{70}
$$

Next, assume $f'(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger))) - f'(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger))) < 0$. In this case, since the coefficient of ϵ in Eq. (71) is negative, there exits sufficiently small $\epsilon > 0$ such that

$$
\epsilon \cdot \underbrace{r^*(x^\dagger)}_{>0} \underbrace{p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger)p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger)}_{>0} \cdot \underbrace{\left\{ f'\left(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger))\right) - f'\left(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger))\right) \right\}}_{<0} + \mathcal{O}(\epsilon^2) < 0 \tag{71}
$$

Thus, we have

$$
\begin{aligned}
\text{Eq. (69)} &< p(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) \cdot f(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger))) + p(x^\dagger, y^{(\mathrm{max})}(x^\dagger)) \cdot f(r^*(x^\dagger, y^{(\mathrm{max})}(x^\dagger))) \\
&= \text{R.H.S. of Eq. (60).}
\end{aligned}
\tag{72}
$$

In both Eqs. (70) and (72), by taking an integral in $\mathcal{S}$, Eq. (60) holds.

In summary, since Eqs. (60), (61) and (62) all hold, the newly constructed, $r_{\mathrm{new}}(\cdot, \cdot)$, is still in $\mathcal{U}_f$.

We now show that $r_{\mathrm{new}}(\cdot, \cdot)$ actually gives larger objective of Eq. (49) than $r^*(\cdot, \cdot)$, which contradicts Eq. (49). Since the value of $r_{\mathrm{new}}(\cdot, \cdot)$ is mostly the same as that of $r^*(\cdot, \cdot)$ except that we have Eqs. (56) and (57), we only need to consider the part they differ. Therefore, it is sufficient to show

$$
\int_{x \in \mathcal{S}} p(x) r^*(x) \underbrace{\left\{ p(y^{(\mathrm{adv})}(x)|x) r_{\mathrm{new}}(y^{(\mathrm{adv})}(x)|x) \ell(g^{(\mathrm{adv})}(x), y^{(\mathrm{adv})}(x)) + p(y^{(\mathrm{max})}(x)|x) r_{\mathrm{new}}(y^{(\mathrm{max})}(x)|x) \ell(g^{(\mathrm{adv})}(x), y^{(\mathrm{max})}(x)) \right\}}_{(a)} \mathrm{d}x
$$

$$
> \int_{x \in \mathcal{S}} p(x) r^*(x) \underbrace{\left\{ p(y^{(\mathrm{adv})}(x)|x) r^*(y^{(\mathrm{adv})}(x)|x) \ell(g^{(\mathrm{adv})}(x), y^{(\mathrm{adv})}(x)) + p(y^{(\mathrm{max})}(x)|x) r^*(y^{(\mathrm{max})}(x)|x) \ell(g^{(\mathrm{adv})}(x), y^{(\mathrm{max})}(x)) \right\}}_{(b)} \mathrm{d}x.
\tag{73}
$$

Subtracting (b) from (a) in Eq. (73), and using Eqs. (56) and (57), we have

$$
\begin{aligned}
& p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger)) \left\{ r_{\mathrm{new}}(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) - r^*(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \right\} \\
& \quad + p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{max})}(x^\dagger)) \left\{ r_{\mathrm{new}}(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) - r^*(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \right\} \\
& \quad = \epsilon \cdot p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \left\{ -\ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger)) + \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{max})}(x^\dagger)) \right\}.
\end{aligned}
\tag{74}
$$

We now show $\ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{max})}(x^\dagger)) > \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger))$. Suppose $\ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{max})}(x^\dagger)) \leq \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger))$. Construct $g' \in \mathbb{R}^K$ by swapping the $y^{(\mathrm{max})}$-th and $y^{(\mathrm{adv})}$-th elements of $g^{(\mathrm{adv})}(x^\dagger)$, while retaining other elements to be exactly the same. Then, because of the assumption that $\ell(\cdot, \cdot)$ is invariant to class permutation, we have $\ell(g', y^{(\mathrm{max})}(x^\dagger)) = \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger))$ and $\ell(g', y^{(\mathrm{adv})}(x^\dagger)) = \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{max})}(x^\dagger))$. Combining this with Eq. (54), we have

$$
\begin{aligned}
& p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) r^*(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \cdot \ell(g', y^{(\mathrm{adv})}(x^\dagger)) + p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) r^*(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \cdot \ell(g', y^{(\mathrm{max})}(x^\dagger)) \\
& < p(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) r^*(y^{(\mathrm{adv})}(x^\dagger)|x^\dagger) \cdot \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger)) + p(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) r^*(y^{(\mathrm{max})}(x^\dagger)|x^\dagger) \cdot \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{max})}(x^\dagger)),
\end{aligned}
\tag{75}
$$

which is in contradiction with the fact that $g^{(\mathrm{adv})}$ achieves the minimal value of Eq. (50). Therefore, we have $\ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{max})}(x^\dagger)) > \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger))$. Hence, Eq. (74) is positive. Multiplying Eq. (74) by $p(x) r^*(x)$ and taking an integral in $\mathcal{S}$, it is still positive because of Eq. (51). Thus, we have Eq. (73).

In conclusion, using $r_{\text{new}}(\cdot, \cdot)$ gives larger objective of Eq. (49) than $r^*(\cdot, \cdot)$. This contradicts the fact that $r^*(\cdot, \cdot)$ is the solution of Eq. (49). Therefore, $g^{(\text{adv})}$, which is obtained via ARM in Eq. (47), is Bayes optimal and coincides with g^* that is obtained via ordinary RM in Eq. (46).

So far, we showed that $g^{(\text{adv})}$, which is any solution of ARM, coincides with the Bayes optimal classifier almost surely over $q^*(x)$. We now turn our focus to the region

$$\mathcal{S}' \equiv \{x \mid x \in \mathcal{X}, \ p(x) > 0, \ q^*(x) = 0\}. \tag{76}$$

Because $g^{(\text{adv})}$ is chosen from all measurable functions, its function value at x is obtained in a point-wise manner:

$$g^{(\text{adv})}(x) = \underset{\widehat{y} \in \mathbb{R}^K}{\arg\min} \sum_{y \in \mathcal{Y}} p(y|x) r^*(y|x) \ell(\widehat{y}, y). \tag{77}$$

For $x \in \mathcal{S}'$, we immediately have $r^*(x) = q^*(x)/p(x) = 0$. Then, for all $x \in \mathcal{S}'$, we have $r^*(x, y) = 0$ for any $r(y|x), y \in \mathcal{Y}$. Therefore, for $x \in \mathcal{S}'$, we can *virtually* set $r^*(y|x) = 1$ for all $y \in \mathcal{Y}$. Substituting this to Eq. (77), we have

$$g^{(\text{adv})}(x) = \underset{\widehat{y} \in \mathbb{R}^K}{\arg\min} \sum_{y \in \mathcal{Y}} p(y|x) \ell(\widehat{y}, y), \ \text{ for } x \in \mathcal{S}'. \tag{78}$$

If follows from Eq. (78) and the use of the classification-calibrated loss that

$$\underset{y}{\arg\max}\, g_y^{(\text{adv})}(x) = \underset{y}{\arg\max}\, p(y|x), \quad \text{for } x \in \mathcal{S}'. \tag{79}$$

This particular $g^{(\text{adv})}$ coincides with the Bayes optimal classifier for all $x \in \mathcal{S}'$. Define

$$\mathcal{S}_{\text{diff}} \equiv \left\{x \ \middle| \ x \in \mathcal{X}, \ p(x) > 0, \ \underset{y \in \mathcal{Y}}{\arg\max}\, p(y|x) \neq \underset{y \in \mathcal{Y}}{\arg\max}\, g_y^{(\text{adv})}(x)\right\}.$$

$$= \mathcal{S} \cup \underbrace{\left\{x \ \middle| \ x \in \mathcal{X}, \ p(x) > 0, \ q^*(x) = 0, \ \underset{y \in \mathcal{Y}}{\arg\max}\, p(y|x) \neq \underset{y \in \mathcal{Y}}{\arg\max}\, g_y^{(\text{adv})}(x)\right\}}_{\equiv \mathcal{S}'_{\text{diff}}}. \tag{80}$$

Then we have

$$\int_{x \in \mathcal{S}_{\text{diff}}} p(x)\mathrm{d}x = \int_{x \in \mathcal{S}} p(x)\mathrm{d}x + \underbrace{\int_{x \in \mathcal{S}'_{\text{diff}}} p(x)\mathrm{d}x}_{=0 \ \because \text{ Eq. (79)}.}$$

$$= \int_{x \in \mathcal{S}} \frac{q^*(x)}{r^*(x)}\mathrm{d}x$$

$$\leq \frac{1}{\min_{x \in \mathcal{S}} r^*(x)} \int_{x \in \mathcal{S}} q^*(x)\mathrm{d}x$$

$$= 0. \tag{81}$$

Therefore, the particular $g^{(\text{adv})}$ coincides with the Bayes optimal classifier almost surely over $p(x)$.

Finally, we consider binary classification, where the key differences to multi-class classification are that the prediction function is $\mathcal{X} \to \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, and the prediction is made based on the sign of the output of the prediction function. Therefore, we need to replace all 'argmax' in the above proof with 'sign'. In addition, to show $\ell(g^{(\text{adv})}(x^\dagger), y^{(\max)}(x^\dagger)) > \ell(g^{(\text{adv})}(x^\dagger), y^{(\text{adv})}(x^\dagger))$, we construct g' in Eq. (75) by $-g^{(\text{adv})}(x^\dagger)$. With these two modifications, all the arguments for multi-class classification hold for binary classification. $\qquad \square$

Remark 2. Here, we show that *if the KL divergence is used* for the f-divergence, the prediction of *any* solution $g^{(\text{adv})}$ of ARM coincides with that of the Bayes optimal classifier almost surely over $p(x)$.

In the following, assume the KL divergence is used. Given prediction function g, the density ratio put by the adversary becomes

$$r^*(x, y) = \frac{1}{Z(\gamma)} \exp\left(\frac{\ell(g(x), y)}{\gamma}\right), \tag{82}$$

where

$$Z(\gamma) = \mathbb{E}_{p(x,y)}\left[\exp\left(\frac{\ell(g(x),y)}{\gamma}\right)\right], \tag{83}$$

and γ is chosen so that the following equality holds:

$$\mathbb{E}_{p(x,y)}\left[r^*(x,y)\log r^*(x,y)\right] = \delta. \tag{84}$$

From Eq. (82), we see that $r^*(\cdot,\cdot)$ is a positive function for any g as long as $\ell(\cdot,\cdot)$ is bounded. Thus, $q^*(x) \equiv \sum_{y\in\mathcal{Y}} r^*(x,y)p(x,y)$ is also positive for $x \in \mathcal{X}$ such that $p(x) > 0$. On the other hand, because we assume $q \ll p$, $p(x) = 0$ implies $q(x) = 0$. Thus, we have $q^*(x) > 0$ iff $p(x) > 0$.

Now, assume that $q^*(x) > 0$ iff $p(x) > 0$, and $g^{(\mathrm{adv})}$ coincides with the Bayes optimal classifier almost surely over $q^*(x)$. Then, we have

$$\int_{x\in\mathcal{S}} q^*(x)\mathrm{d}x = 0, \tag{85}$$

where

$$\mathcal{S}' \equiv \left\{ x \;\middle|\; x \in \mathcal{X},\; p(x) > 0,\; q^*(x) > 0,\; \arg\max_{y\in\mathcal{Y}} p(y|x) \neq \arg\max_{y\in\mathcal{Y}} g_y^{(\mathrm{adv})}(x) \right\}. \tag{86}$$

Because $r^*(x) \equiv q^*(x)/p(x)$ is positive, $\epsilon \equiv \min_{x\in\mathcal{S}} r^*(x)$ is also positive. Then, we have

$$\begin{aligned}
\int_{x\in\mathcal{S}} p(x)\mathrm{d}x &= \int_{x\in\mathcal{S}} \frac{q^*(x)}{r^*(x)}\mathrm{d}x \\
&\leq \frac{1}{\epsilon}\int_{x\in\mathcal{S}} q^*(x)\mathrm{d}x \\
&= 0. \tag{87}
\end{aligned}$$

Therefore, $g^{(\mathrm{adv})}$ coincides with the Bayes optimal classifier almost surely over the training density $p(x)$.

C. Proof of Lemma 1

By assumption, $\ell(\widehat{y}, y)$ is a convex margin loss. Thus, we can let $\ell(\widehat{y}, y) = \phi(y\widehat{y})$, where $\phi(\cdot)$ is a convex function. On the other hand, by Definition 1, for some non-constant, non-decreasing and non-negative function $h(\cdot)$, the steeper loss $\ell_{\mathrm{steep}}(\widehat{y}, y)$, satisfies

$$\begin{aligned}
\frac{\partial \ell_{\mathrm{steep}}(\widehat{y}, y)}{\partial \widehat{y}} &= h(\ell(\widehat{y}, y))\frac{\partial \ell(\widehat{y}, y)}{\partial \widehat{y}} \\
&= h(\phi(y\widehat{y}))\frac{\partial \phi(y\widehat{y})}{\partial \widehat{y}}. \tag{88}
\end{aligned}$$

From Eq. (88), it is easy to see that $\ell_{\mathrm{steep}}(\widehat{y}, y)$ is also a margin loss and can be written as $\ell_{\mathrm{steep}}(\widehat{y}, y) = \phi_{\mathrm{steep}}(y\widehat{y})$. Our first goal is to show that $\phi_{\mathrm{steep}}(\cdot)$ is convex. To this end, it is sufficient to show that $\frac{\partial \phi_{\mathrm{steep}}(y\widehat{y})}{\partial \widehat{y}} = h(\phi(y\widehat{y}))\frac{\partial \phi(y\widehat{y})}{\partial \widehat{y}}$ is non-decreasing in $\widehat{y}$, $\mathbb{Y} = \{+1, -1\}$.

Since $\phi(y\widehat{y})$ is convex in $\widehat{y}$, $\frac{\phi(y\widehat{y})}{\partial \widehat{y}}$ is non-decreasing in $\widehat{y}$. Let $\widehat{y}_\alpha$ be the smallest $\widehat{y}_\alpha$ such that $\frac{\ell(\widehat{y}_\alpha, y_i)}{\partial \widehat{y}} = 0$, if such $\widehat{y}_\alpha$ exists. In the following, we analyze $\phi(y\widehat{y})\frac{\partial \phi(y\widehat{y})}{\partial \widehat{y}}$, considering two cases: 1) $\widehat{y} \leq \widehat{y}_\alpha$ and 2) $\widehat{y}_\alpha \leq \widehat{y}$. Note that $\widehat{y}_\alpha$ may not always exist because $\frac{\phi(y\widehat{y})}{\partial \widehat{y}}$ can be negative for any finite $\widehat{y}$, which is the case for the widely-used classification losses such as the exponential loss and the logistic loss. In such a case, we only consider the first case, letting $\widehat{y}_\alpha$ arbitrarily large.

Case 1 $\widehat{y} \leq \widehat{y}_\alpha$**:** By convexity of $\phi(y\widehat{y})$, for $\widehat{y} \leq \widehat{y}_\alpha$, $\frac{\phi(y\widehat{y})}{\partial \widehat{y}} \leq 0$ holds and therefore, $\phi(y\widehat{y})$ is non-increasing in $\widehat{y}$. Since $h(\cdot)$ is a non-decreasing function, $h(\phi(y\widehat{y}))$ is non-increasing in $\widehat{y}$ for $\widehat{y} \leq \widehat{y}_\alpha$. In summary, for $\widehat{y} \leq \widehat{y}_\alpha$, $\frac{\phi(y\widehat{y})}{\partial \widehat{y}}$ is a non-positive

non-decreasing function of $\widehat{y}$, and $h(\phi(y\widehat{y}))$ is a non-negative non-increasing function of $\widehat{y}$. Thus, for $\widehat{y} \leq \widehat{y}_\alpha$, their product $h(\phi(y\widehat{y}))\frac{\phi(y\widehat{y})}{\partial \widehat{y}}$ is a non-decreasing function of $\widehat{y}$.

Case 2 $\widehat{y}_\alpha \leq \widehat{y}$: By convexity of $\phi(y\widehat{y})$, for $\widehat{y}_\alpha \leq \widehat{y}$, $\frac{\phi(y\widehat{y})}{\partial \widehat{y}} \geq 0$ holds and therefore, $\phi(y\widehat{y})$ is non-decreasing in $\widehat{y}$. Since $h(\cdot)$ is a non-decreasing function, $h(\phi(y\widehat{y}))$ is non-decreasing in $\widehat{y}$ for $\widehat{y} \leq \widehat{y}_\alpha$. In summary, for $\widehat{y}_\alpha \leq \widehat{y}$, $\frac{\phi(y\widehat{y})}{\partial \widehat{y}}$ is a non-negative non-decreasing function of $\widehat{y}$, and $h(\phi(y\widehat{y}))$ is a non-negative non-decreasing function of $\widehat{y}$. Thus, for $\widehat{y} \leq \widehat{y}_\alpha$, their product $h(\phi(y\widehat{y}))\frac{\phi(y\widehat{y})}{\partial \widehat{y}}$ is a non-decreasing function of $\widehat{y}$.

Therefore, *for any $\widehat{y}$, $h(\phi(y\widehat{y}))\frac{\phi(y\widehat{y})}{\partial \widehat{y}}$ is a non-decreasing function of $\widehat{y}$*, which directly indicates that the steeper loss, $\phi_{\text{steep}}(y\widehat{y})$, is convex.

We now utilize the fact that a convex margin loss $\psi(y\widehat{y})$ is classification calibrated iff $\psi'(0) < 0$ [Theorem 6 in (Bartlett et al., 2006)]. Using this fact, because $\phi(y\widehat{y})$ is classification calibrated, we have $\phi'(0) < 0$. Furthermore, from the assumption, we have $h(\phi(0)) > 0$. Therefore, we have $\phi'_{\text{steep}}(0) = h(\phi(0))\phi'(0) < 0$. Using the fact again, we immediately have that $\phi_{\text{steep}}(y\widehat{y})$ is classification calibrated. $\square$

Remark 3. *In the proof, we need to assume $h(\phi(0)) > 0$. From Appendix D, we know that $h(\ell)$ corresponds to the weight put by the adversary to data points with a loss value of ℓ. We see from Eq. (22) that when the KL divergence is used, the adversary will only assign positive weights to data losses. Therefore, Lemma 1 always holds when the KL divergence is used.*

D. Proof of Theorem 3

Let θ^* be the stationary point of Eq. (7). By using a chain rule and Danskin's theorem (Danskin, 1966), θ^* satisfies

$$\frac{1}{N} \sum_{i=1}^{N} r_i^* \left.\frac{\partial \ell(\widehat{y}, y_i)}{\partial \widehat{y}}\right|_{\widehat{y}=g_{\theta^*}(x_i)} \cdot \nabla_\theta g_\theta(x_i)|_{\theta=\theta^*} \in \mathbf{0}, \tag{89}$$

where $\boldsymbol{r}^*$ is the solution of inner maximization of Eq. (7) at the stationary point.

Now, we analyze $\boldsymbol{r}^*$, which is the solution of Eq. (7) at the stationary point θ^*. For notational convenience, for $1 \leq i \leq N$, let us denote $\ell_i(\theta^*)$ by ℓ_i^*. Then, $\boldsymbol{r}^*$ is the solution of the following optimization problem.

$$\max_{\boldsymbol{r}\in\widehat{\mathcal{U}}_f} \frac{1}{N} \sum_{i=1}^{N} r_i \ell_i^*, \tag{90}$$

$$\widehat{\mathcal{U}}_f = \left\{ \boldsymbol{r} \, \middle| \, \frac{1}{N} \sum_{i=1}^{N} f(r_i) \leq \delta, \; \frac{1}{N} \sum_{i=1}^{N} r_i = 1, \; \boldsymbol{r} \geq 0 \right\}, \tag{91}$$

Note that Eq. (90) is a convex optimization problem because it has a linear objective with a convex constraint; thus, any local maximum is the global maximum. Nonetheless, there can be multiple solutions that attain the same global maxima. Among those solutions, we now show that there exists $\boldsymbol{r}^*$ such that elements of $\boldsymbol{r}^*$ has the monotonic relationship to the corresponding data losses, i.e., for any $1 \leq j \neq k \leq N$,

$$\ell_j^* < \ell_k^* \Rightarrow 0 \leq r_j^* \leq r_k^*, \tag{92}$$

$$\ell_j^* = \ell_k^* \Rightarrow 0 \leq r_j^* = r_k^*. \tag{93}$$

To prove this, we assume we obtain one of optimal solutions of Eq. (90), which we denote as $\boldsymbol{r}'^*$. If this $\boldsymbol{r}'^*$ satisfies Eqs. (92) and (93) for any j and k, then we are done. In the following, we assume $\boldsymbol{r}'^*$ does not satisfy either Eqs. (92) or (93).

First, assume that $\boldsymbol{r}'^*$ does not satisfy Eq. (92). Then, there exist $1 \leq j \neq k \leq N$ such that $\ell_j^* < \ell_k^*$ but $r_j'^* > r_k'^*$. Define $\boldsymbol{r}''^*$ such that

$$r_i''^* = \begin{cases} r_i'^* & \text{if } i \neq j, k \\ r_j'^* & \text{if } i = k \\ r_k'^* & \text{if } i = j \end{cases} \quad \text{for } 1 \leq i \leq N. \tag{94}$$

Then, it is easy to see $r''^* \in \widehat{\mathcal{U}}_f$, and the following holds:

$$\frac{1}{N}\sum_{i=1}^{N} r_i''^* \ell_i^* - \frac{1}{N}\sum_{i=1}^{N} r_i'^* \ell_i^* = \frac{1}{N}\left(r_j''^*\ell_j^* + r_k''^*\ell_k^* - r_j'^*\ell_j^* - r_k'^*\ell_k^*\right)$$

$$= \frac{1}{N}\left(r_k'^*\ell_j^* + r_j'^*\ell_k^* - r_j'^*\ell_j^* - r_k'^*\ell_k^*\right)$$

$$= \frac{1}{N}\left(r_j'^* - r_k'^*\right)\left(\ell_k^* - \ell_j^*\right)$$

$$> 0. \tag{95}$$

Therefore, the newly defined r''^* attains the larger objective value of Eq. (90), which contradicts the assumption that r'^* is the optimal solution of Eq. (90). Thus, r'^* always satisfies Eq. (92).

Second, assume that r'^* does not satisfy Eq. (93). Then, there exist $1 \le j \ne k \le N$ such that $\ell_j^* = \ell_k^*$ but $r_j'^* \ne r_k'^*$. Define r''^* such that

$$r_i''^* = \begin{cases} r_i'^* & \text{if } i \ne j, k \\ (r_i'^* + r_j'^*)/2 & \text{if } i = j, k \end{cases} \quad \text{for } 1 \le i \le N. \tag{96}$$

Then, it is easy to see $r''^* \in \widehat{\mathcal{U}}_f$ because

$$\frac{1}{N}\sum_{i=1}^{N} f\left(r_i''^*\right) = \frac{1}{N}\left\{\left(\sum_{i \ne j,k} f\left(r_i''^*\right)\right) + f\left(r_j''^*\right) + f\left(r_k''^*\right)\right\}$$

$$= \frac{1}{N}\left\{\left(\sum_{i \ne j,k} f\left(r_i'^*\right)\right) + f\left((r_j'^* + r_k'^*)/2\right) + f\left((r_j'^* + r_k'^*)/2\right)\right\}$$

$$\le \frac{1}{N}\left\{\left(\sum_{i \ne j,k} f\left(r_i'^*\right)\right) + f\left(r_j'^*\right) + f\left(r_k'^*\right)\right\} \quad (\because \text{convexity of } f(\cdot).)$$

$$= \frac{1}{N}\sum_{i=1}^{N} f\left(r_i'^*\right)$$

$$\le \delta. \quad \left(\because r'^* \in \widehat{\mathcal{U}}_f.\right) \tag{97}$$

Also, it is easy to see that r''^* attain the same maximum value as r'^*; thus, r''^* is the optimal solution of Eq. (90), and notably, we have $r_j''^* = r_k''^*$ for $\ell_j^* = \ell_k^*$. In general, we can start from any $r'^* \in \widehat{\mathcal{U}}_f$ and equally distribute the weights to the same losses to obtain r''^*, which is still in $\widehat{\mathcal{U}}_f$ and attains exactly the same global optimal value in Eq. (90).

In the following, we assume we have r^* that satisfies Eqs (92) and (93) for any $1 \le j \ne k \le N$. Then, there exits a non-decreasing non-negative function $r^*(\cdot) : \mathbb{R} \to \mathbb{R}$, such that

$$r^*(\ell_i^*) = r_i^*, \quad \text{for } 1 \le i \le N. \tag{98}$$

Let us construct a new loss function $\ell_{\mathrm{DRSL}}(\widehat{y}, y)$ by its derivative:

$$\frac{\partial \ell_{\mathrm{DRSL}}(\widehat{y}, y)}{\partial \widehat{y}} \equiv r^*(\ell(\widehat{y}, y))\frac{\partial \ell(\widehat{y}, y)}{\partial \widehat{y}}. \tag{99}$$

Then, from Eqs. (89), (98) and (99), we immediately have

$$\frac{1}{N}\sum_{i=1}^{N} \left.\frac{\partial \ell_{\mathrm{DRSL}}(\widehat{y}, y_i)}{\partial \widehat{y}}\right|_{\widehat{y}=g_{\theta^*}(x_i)} \cdot \left.\nabla_\theta g_\theta(x_i)\right|_{\theta=\theta^*} \in \mathbf{0}. \tag{100}$$

This readily implies that θ^* is a stationary point of Eq. (13), i.e., ERM using $\ell_{\mathrm{DRSL}}(\widehat{y}, y)$. Furthermore, from Eq. (99) and the non-negativeness and non-decreasingness of $r^*(\cdot)$, we see that the newly constructed loss, $\ell_{\mathrm{DRSL}}(\widehat{y}, y)$, is steeper than the original loss, $\ell(\widehat{y}, y)$ (see Definition 1 for the definition of the steeper loss). Here we see that $h(\cdot)$ in Definition 1 exactly corresponds to $r^*(\cdot)$ defined in Eq. (98). $\qquad\square$

E. Derivation of the Decomposition of the Adversarial Risk

Here, we derive Eq. (18) for the PE divergence.

$$
\begin{aligned}
\mathcal{R}_{\text{s-adv}}(\theta) - \mathcal{R}(\theta) &\equiv \sup_{w \in \mathcal{W}_{\text{PE}}} \mathbb{E}_{p(x,y,z)}\left[\{w(z) - 1\}\ell(g_\theta(x), y)\right] \\
&= \sup_{w \in \mathcal{W}_{\text{PE}}} \mathbb{E}_{p(z)}\left[\{w(z) - 1\}\mathbb{E}_{p(x,y|z)}\left[\ell(g_\theta(x), y)\right]\right] \\
&= \sup_{w \in \mathcal{W}_{\text{PE}}} \sum_{z \in \mathcal{Z}} p(z)\{w(z) - 1\}\mathcal{R}_z(\theta).
\end{aligned}
\tag{101}
$$

It follows from Eq. (23) that for $z \in \mathcal{Z}$, we have the adversarial weight as[11]

$$
w^*(z) = \sqrt{\frac{\delta}{\sum_{z' \in \mathcal{Z}} p(z')(\mathcal{R}_{z'}(\theta) - \mathcal{R}(\theta))^2}}(\mathcal{R}_z(\theta) - \mathcal{R}(\theta)) + 1.
\tag{102}
$$

Hence, Eq. (101) becomes

$$
\begin{aligned}
\sum_{z \in \mathcal{Z}} p(z)\{w^*(z) - 1\}\mathcal{R}_z(\theta) &= \sqrt{\frac{\delta}{\sum_{z' \in \mathcal{Z}} p(z')(\mathcal{R}_{z'}(\theta) - \mathcal{R}(\theta))^2}} \sum_{z \in \mathcal{Z}} p(z)(\mathcal{R}_z(\theta) - \mathcal{R}(\theta))\mathcal{R}_z(\theta) \\
&= \sqrt{\frac{\delta}{\sum_{z' \in \mathcal{Z}} p(z')(\mathcal{R}_{z'}(\theta) - \mathcal{R}(\theta))^2}} \sum_{z \in \mathcal{Z}} p(z)(\mathcal{R}_z(\theta) - \mathcal{R}(\theta))^2 \\
&= \sqrt{\delta} \cdot \sqrt{\sum_{z \in \mathcal{Z}} p(z)(\mathcal{R}_z(\theta) - \mathcal{R}(\theta))^2},
\end{aligned}
\tag{103}
$$

which concludes our derivation.

F. Comparison between the Use of Different f-divergences

We qualitatively compare the use of different f-divergences. For $1 \le x$, the f functions for the PE, KL divergences are $(x - 1)^2$, $x \log x$, respectively. The function f in Eq. (20) penalizes the deviation of the adversarial weights from the uniform weights, $\mathbf{1}_S$. With the quadratic penalty of the PE divergence, it is hard for the adversary to concentrate large weights onto a small portion of latent categories. In contrast, when the KL divergence is used, the adversary tends to put large weights to a small portion of latent categories. Hence, users can choose the appropriate divergence depending on their belief on how concentrated the distribution shift occurs.

G. Formal Statement of the Convergence Rate

Denote by $p_z = p(z)$ and $w_z = w(z)$ for $z \in \mathcal{Z}$ and define a set-valued function $\Phi : \mathbb{R}^S \to 2^{\mathbb{R}^S}$ as

$$
\Phi(\boldsymbol{u}) = \left\{\boldsymbol{w} \in \mathbb{R}^S \mid \sum_s (p_s + u_s)f(w_s) \le \delta, \sum_s (p_s + u_s)w_s = 1, w_s \ge 0\right\}.
$$

Then, $\mathcal{W}_f = \Phi(\mathbf{0})$ and $\widehat{\mathcal{W}}_f = \Phi(\boldsymbol{u})$ where $u_s = n_s/N - p_s$ for $s = 1, \ldots, S$. Similarly, denote by $l_z = \mathbb{E}_{p(x,y)}[p(z \mid x, y)\ell(g_\theta(x), y)]$ and define a function $R_\theta : \mathbb{R}^S \to \mathbb{R}$ indexed by θ as

$$
R_\theta(\boldsymbol{u}') = \sum_s w_s(l_s + u_s').
$$

Then, $\mathbb{E}_{p(x,y,z)}[w_z \ell(g_\theta(x), y)] = R_\theta(\mathbf{0})$ and $\widehat{\mathcal{R}}(\boldsymbol{w}, \theta) = R_\theta(\boldsymbol{u}')$ where $u_s' = n_s \overline{\ell_s}(\theta)/N - l_s$ for $s = 1, \ldots, S$. Finally, the perturbed objective function can be defined by

$$
J(\theta, \boldsymbol{u}, \boldsymbol{u}') = \sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} R_\theta(\boldsymbol{u}') + \lambda(\boldsymbol{u}, \boldsymbol{u}')\Omega(\theta),
$$

---

[11]Here, we need to assume that δ is not so large. Then, we can validly drop the non-negativity inequality constraint of $\widehat{\mathcal{W}}_f$, which is needed to obtain the analytic solution in Eq. (23).

where the function $\lambda(\boldsymbol{u}, \boldsymbol{u}') \geq 0$ serves as the regularization parameter, so that the truly optimal θ^* is the minimizer of $J(\theta, \boldsymbol{0}, \boldsymbol{0})$ and the empirically optimal $\widehat{\theta}$ is the minimizer of $J(\theta, \boldsymbol{u}, \boldsymbol{u}')$ with the aforementioned perturbations $\boldsymbol{u}$ and $\boldsymbol{u}'$.

According to the central limit theorem (Chung, 1968), $u_s = \mathcal{O}_p(1/\sqrt{N})$, and $u'_s = \mathcal{O}_p(1/\sqrt{N})$ if the loss ℓ is finite. Therefore, we only consider perturbations $\boldsymbol{u}$ and $\boldsymbol{u}'$ such that $\|\boldsymbol{u}\|_2 \leq \epsilon$ and $\|\boldsymbol{u}'\|_2 \leq \epsilon$ in our analysis, where $0 < \epsilon \leq \delta/(5\sqrt{S}|f'(1)|)$ is a sufficiently small constant.

We make the following assumptions:

(a) $g_\theta(x)$ is linear in θ, and for all θ under consideration, $\|\nabla_\theta g_\theta\|_\infty = \sup_x \|\nabla_\theta g_\theta(x)\|_2 < \infty$, which implies $\|g_\theta\|_\infty = \sup_x |g_\theta(x)| < \infty$;[12]
(b) $\partial\ell(t, y)/\partial t$ is bounded from below and above for all t such that $|t| \leq \|g_\theta\|_\infty$;[13]
(c) $f(t)$ is twice differentiable, and this second derivative is bounded from below by a positive number for all t such that $0 \leq t \leq \sup_{\|\boldsymbol{u}\|_2 \leq \epsilon} \sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} \max_s w_s$;[14]
(d) $\Omega(\theta)$ is Lipschitz continuous, and $\lambda(\boldsymbol{u}, \boldsymbol{u}')$ converges to $\lambda(\boldsymbol{0}, \boldsymbol{0})$ in $\mathcal{O}(\|\boldsymbol{u}\|_2 + \|\boldsymbol{u}'\|_2)$.

We also assume either one of the two conditions holds:

(e1) $\Omega(\theta)$ is strongly convex in θ and $\lambda(\boldsymbol{0}, \boldsymbol{0}) > 0$;
(e2) $\ell(t, y)$ is twice differentiable w.r.t. t, and $\partial^2\ell(t, y)/\partial t^2$ is lower bounded by a positive number for all t such that $|t| \leq \|g_{\theta^*}\|_\infty$. If t is vector-valued, $\partial^2\ell(t, y)/\partial t_i^2$ is lower bounded for all dimensions of t such that $\|t\|_\infty \leq \sup_x \|g_{\theta^*}(x)\|_\infty$.[15]

Theorem 5 (Perturbation analysis). *Assume (a), (b), (c), (d), and (e1) or (e2). Let θ^* be the minimizer of $J(\theta, \boldsymbol{0}, \boldsymbol{0})$ and $\theta_{\boldsymbol{u},\boldsymbol{u}'}$ be the minimizer of $J(\theta, \boldsymbol{u}, \boldsymbol{u}')$. Then, for all $\boldsymbol{u}$ and $\boldsymbol{u}'$ such that $\|\boldsymbol{u}\|_2 \leq \epsilon$ and $\|\boldsymbol{u}'\|_2 \leq \epsilon$,*

$$\|\theta_{\boldsymbol{u},\boldsymbol{u}'} - \theta^*\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2^{1/2} + \|\boldsymbol{u}'\|_2),$$

$$\|J(\theta_{\boldsymbol{u},\boldsymbol{u}'}, \boldsymbol{0}, \boldsymbol{0}) - J(\theta^*, \boldsymbol{0}, \boldsymbol{0})\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2^{1/2} + \|\boldsymbol{u}'\|_2).$$

The convergence rate of the model parameter and the order of the estimation error are immediate corollaries of Theorem 5.

Theorem 6 (Convergence rate and estimation error). *Assume (a), (b), (c), (d), and (e1) or (e2). Let θ^* be the minimizer of the adversarial expected risk and $\widehat{\theta}_N$ be the minimizer of the adversarial empirical risk given some training data of size N. Then, as $N \to \infty$,*

$$\|\widehat{\theta}_N - \theta^*\|_2 = \mathcal{O}(N^{-1/4}),$$

and

$$\left\|\mathcal{R}_{\text{s-adv}}(\widehat{\theta}_N) - \mathcal{R}_{\text{s-adv}}(\theta^*)\right\|_2 = \mathcal{O}(N^{-1/4})$$

H. Proof of the Convergence Rate

We begin with the growth condition of $J(\theta, \boldsymbol{0}, \boldsymbol{0})$ at $\theta = \theta^*$.

Lemma 2 (Second-order growth condition). *There exists a constant $C_{J''} > 0$ such that*

$$J(\theta, \boldsymbol{0}, \boldsymbol{0}) \geq J(\theta^*, \boldsymbol{0}, \boldsymbol{0}) + C_{J''}\|\theta - \theta^*\|_2^2.$$

Proof. First consider the assumption (e1). Let $C_{J''} = (1/2)\lambda(\boldsymbol{0}, \boldsymbol{0})$, so that $J(\theta, \boldsymbol{0}, \boldsymbol{0})$ is strongly convex with parameter $C_{J''}$, i.e.,

$$J(\theta, \boldsymbol{0}, \boldsymbol{0}) \geq J(\theta^*, \boldsymbol{0}, \boldsymbol{0}) + \nabla_\theta J(\theta^*, \boldsymbol{0}, \boldsymbol{0})^\top (\theta - \theta^*) + C_{J''}\|\theta - \theta^*\|_2^2.$$

---

[12] This makes $\ell(g_\theta(x), y)$ convex in θ for all $\ell(t, y)$ convex in t and $\nabla_\theta^2 \ell(g_\theta(x), y)$ easy to handle.

[13] This is a sufficient condition for the Lipschitz continuity of ℓ. In fact, it must be valid given (a) since ℓ is continuously differentiable w.r.t. t.

[14] This is for the Lipschitz continuity of $J(\theta, \boldsymbol{u}, \boldsymbol{u}') - J(\theta, \boldsymbol{0}, \boldsymbol{0})$. It is satisfied by the KL divergence since $f''(t) = 1/t$ and $\Phi(\boldsymbol{u})$ is bounded and then $\sup_{\|\boldsymbol{u}\|_2 \leq \epsilon} \sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} \max_s w_s < \infty$, and by the PE divergence since $f''(t) = 2$

[15] This makes $J(\theta, \boldsymbol{0}, \boldsymbol{0})$ locally strongly convex in θ around θ^*. It is satisfied by the logistic loss with the lower bound as $1/(2 + \exp(\|g_{\theta^*}\|_\infty) + \exp(-\|g_{\theta^*}\|_\infty))$ and the softmax cross-entropy loss with the lower bound as $\min_y \min_{t_i = \pm \sup_x \|g_{\theta^*}\|_\infty} \exp(t_y) \sum_{i \neq y} \exp(t_i)/(\sum_i \exp(t_i))^2$.

The lemma follows from the optimality condition which says $\nabla_\theta J(\theta^*, \mathbf{0}, \mathbf{0}) = \mathbf{0}$.

Second consider the assumption (e2) if (e1) does not hold. Without loss of generality, assume that $\lambda(\mathbf{0}, \mathbf{0}) = 0$. Let $\boldsymbol{w}^* = \arg\sup_{\boldsymbol{w} \in \Phi(\mathbf{0})} R_{\theta^*}(\mathbf{0})$, then according to Danskin's theorem (Danskin, 1966),

$$\nabla_\theta J(\theta^*, \mathbf{0}, \mathbf{0}) = \mathbb{E}_{p(x,y,z)}[w_z^* \nabla_\theta \ell(g_{\theta^*}(x), y)]$$
$$= \mathbb{E}_{p(x,y,z)}[w_z^* \ell'(g_{\theta^*}(x), y) \nabla_\theta g_{\theta^*}(x)]$$

where $\ell'(g_{\theta^*}(x), y)$ means $\partial \ell(t, y)/\partial t|_{t=g_{\theta^*}(x)}$. The assumption (a) guarantees that $\nabla_\theta g_{\theta^*}(x)$ is no longer a function of θ, and thus

$$\nabla_\theta^2 J(\theta^*, \mathbf{0}, \mathbf{0}) = \mathbb{E}_{p(x,y,z)}[w_z^* \nabla_\theta \ell'(g_{\theta^*}(x), y) \nabla_\theta g_{\theta^*}(x)^\top]$$
$$= \mathbb{E}_{p(x,y,z)}[w_z^* \ell''(g_{\theta^*}(x), y) \nabla_\theta g_{\theta^*}(x) \nabla_\theta g_{\theta^*}(x)^\top]$$

where $\ell''(g_{\theta^*}(x), y)$ means $\partial^2 \ell(t, y)/\partial t^2|_{t=g_{\theta^*}(x)}$.

Let $C_{\ell''} = \inf_{|t| \le \|g_{\theta^*}\|_\infty} \min_y \ell''(t, y)$, and by assumption $C_{\ell''} > 0$. Also let $C_{\lambda,z}$ be the smallest eigenvalue of $\mathbb{E}_{p(x|z)}[\nabla_\theta g_\theta(x) \nabla_\theta g_\theta(x)^\top]$ at $\theta = \theta^*$ for $z \in \mathcal{Z}$. Note that $p(x \mid z)$ generates infinite number of x, and $\mathbb{E}_{p(x|z)}[\nabla_\theta g_\theta(x) \nabla_\theta g_\theta(x)^\top]$ as an average of infinitely many independent positive semi-definite matrices $\nabla_\theta g_\theta(x) \nabla_\theta g_\theta(x)^\top$ (they are independent as long as $\nabla_\theta g_\theta(x)$ depends on x) is positive definite. Thus, $C_{\lambda,z} > 0$ for all $z \in \mathcal{Z}$, and subsequently,

$$(\theta - \theta^*)^\top \nabla_\theta^2 J(\theta^*, \mathbf{0}, \mathbf{0})(\theta - \theta^*)$$
$$\ge \left(\inf_{|t| \le \|g_{\theta^*}\|_\infty} \min_y \ell''(t, y) \right) \cdot (\theta - \theta^*)^\top \mathbb{E}_{p(x,y,z)}[w_z^* \nabla_\theta g_{\theta^*}(x) \nabla_\theta g_{\theta^*}(x)^\top](\theta - \theta^*)$$
$$= C_{\ell''} \cdot (\theta - \theta^*)^\top \mathbb{E}_{p(x,z)}[w_z^* \nabla_\theta g_{\theta^*}(x) \nabla_\theta g_{\theta^*}(x)^\top](\theta - \theta^*)$$
$$= C_{\ell''} \cdot (\theta - \theta^*)^\top \left(\sum_{s=1}^S p_s w_s^* \mathbb{E}_{p(x|z=s)}[\nabla_\theta g_{\theta^*}(x) \nabla_\theta g_{\theta^*}(x)^\top] \right) (\theta - \theta^*)$$
$$= C_{\ell''} \left(\sum_{s=1}^S p_s w_s^* (\theta - \theta^*)^\top \mathbb{E}_{p(x|z=s)}[\nabla_\theta g_{\theta^*}(x) \nabla_\theta g_{\theta^*}(x)^\top](\theta - \theta^*) \right)$$
$$\ge C_{\ell''} \left(\sum_{s=1}^S p_s w_s^* C_{\lambda,s} \|\theta - \theta^*\|_2^2 \right)$$
$$\ge C_{\ell''} \min_s C_{\lambda,s} \left(\sum_{s=1}^S p_s w_s^* \right) \|\theta - \theta^*\|_2^2$$
$$= C_{\ell''} \min_s C_{\lambda,s} \|\theta - \theta^*\|_2^2.$$

This completes the proof by letting $C_{J''} = C_{\ell''} \min_s C_{\lambda,s}$. $\qquad\square$

We then study the Lipschitz continuity of $J(\theta, \boldsymbol{u}, \boldsymbol{u}')$.

Lemma 3 (Lipschitz continuity of the perturbed objective). *For all $\boldsymbol{u}$ and $\boldsymbol{u}'$ such that $\|\boldsymbol{u}\|_2 \le \epsilon$ and $\|\boldsymbol{u}'\|_2 \le \epsilon$, $J(\theta, \boldsymbol{u}, \boldsymbol{u}')$ is Lipschitz continuous with a (not necessarily the best) Lipschitz constant independent of $\boldsymbol{u}$ and $\boldsymbol{u}'$.*

Proof. Define $F(\theta, \boldsymbol{u}, \boldsymbol{u}') = \sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} R_\theta(\boldsymbol{u}')$ and let $\boldsymbol{w}^* = \arg\sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} R_\theta(\boldsymbol{u}')$. According to Danskin's theorem (Danskin, 1966), $\nabla_\theta F(\theta, \boldsymbol{u}, \boldsymbol{u}') = \sum_s w_s^* \nabla_\theta l_s$ where

$$\nabla_\theta l_s = \mathbb{E}_{p(x,y)}[p(z = s \mid x, y) \ell'(g_\theta(x), y) \nabla_\theta g_\theta(x)].$$

The assumptions (a) and (b) say that $\|\nabla_\theta g_\theta\|_\infty < \infty$ and $|\ell'(g_\theta(x), y)| < \infty$ so that

$$\|\nabla_\theta l_s\|_2 \leq \|\nabla_\theta g_\theta\|_\infty \left(\sup_{|t| \leq \|g_\theta\|_\infty} \max_y |\ell'(t, y)| \right) \mathbb{E}_{p(x,y)}[p(z = s \mid x, y)]$$

$$= \|\nabla_\theta g_\theta\|_\infty \left(\sup_{|t| \leq \|g_\theta\|_\infty} \max_y |\ell'(t, y)| \right) p_s$$

$$< \infty,$$

and it is clear that $w_s^* < \infty$. Hence,

$$\|\nabla_\theta F(\theta, \boldsymbol{u}, \boldsymbol{u}')\|_2 \leq \sum_{s=1}^S w_s^* \|\nabla_\theta l_s\|_2 < \infty,$$

which means $F(\theta, \boldsymbol{u}, \boldsymbol{u}')$ is Lipschitz continuous with a Lipschitz constant independent of $\boldsymbol{u}$ and $\boldsymbol{u}'$.

By the assumption (d), $\Omega(\theta)$ is Lipschitz continuous and there exists a constant $C_\lambda > 0$ such that

$$\lambda(\boldsymbol{u}, \boldsymbol{u}') \leq \lambda(\boldsymbol{0}, \boldsymbol{0}) + C_\lambda(\|\boldsymbol{u}\|_2 + \|\boldsymbol{u}'\|_2)$$

$$\leq \lambda(\boldsymbol{0}, \boldsymbol{0}) + 2C_\lambda \epsilon$$

$$< \infty.$$

As a result, $\lambda(\boldsymbol{u}, \boldsymbol{u}')\Omega(\theta)$ possesses a Lipschitz constant independent of $\boldsymbol{u}$ and $\boldsymbol{u}'$ as well. $\qquad \square$

From now on, we investigate the Lipschitz continuity of the difference function

$$D(\theta) = J(\theta, \boldsymbol{u}, \boldsymbol{u}') - J(\theta, \boldsymbol{0}, \boldsymbol{0}),$$

which is the most challenging task in our perturbation analysis. Define

$$D_1(\theta) = F(\theta, \boldsymbol{u}, \boldsymbol{u}') - F(\theta, \boldsymbol{u}, \boldsymbol{0}),$$
$$D_2(\theta) = F(\theta, \boldsymbol{u}, \boldsymbol{0}) - F(\theta, \boldsymbol{0}, \boldsymbol{0}),$$

where $F(\theta, \boldsymbol{u}, \boldsymbol{u}') = \sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} R_\theta(\boldsymbol{u}')$ defined in Lemma 3, and then $D(\theta)$ can be decomposed as

$$D(\theta) = D_1(\theta) + D_2(\theta) + (\lambda(\boldsymbol{u}, \boldsymbol{u}') - \lambda(\boldsymbol{0}, \boldsymbol{0}))\Omega(\theta).$$

Given the assumption (d), the third function $(\lambda(\boldsymbol{u}, \boldsymbol{u}') - \lambda(\boldsymbol{0}, \boldsymbol{0}))\Omega(\theta)$ is Lipschitz continuous with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}\|_2 + \|\boldsymbol{u}'\|_2)$. We are going to prove the same property for $D_1(\theta)$ and $D_2(\theta)$ using the assumptions (a), (b) and (c).

Lemma 4 (Lipschitz continuity of the difference function, I)**.** *For any fixed $\boldsymbol{u}$ and all $\boldsymbol{u}'$ such that $\|\boldsymbol{u}'\|_2 \leq \epsilon$, $D_1(\theta)$ is Lipschitz continuous with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}'\|_2)$.*

Proof. According to the chain rule in calculus,

$$\|\nabla_\theta D_1(\theta)\|_2 = \left\| \sum_{s=1}^S \frac{\partial D_1(\theta)}{\partial l_s} \nabla_\theta l_s \right\|_2$$

$$\leq \left| \sum_{s=1}^S \frac{\partial D_1(\theta)}{\partial l_s} \right| \cdot \max_s \|\nabla_\theta l_s\|_2$$

$$= \mathcal{O} \left(\left| \sum_{s=1}^S \frac{\partial D_1(\theta)}{\partial l_s} \right| \right),$$

since we have proven that $\|\nabla_\theta l_s\|_2 < \infty$ given the assumptions (a) and (b) in Lemma 3.

By definition,

$$D_1(\theta) = \sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} R_\theta(\boldsymbol{u}') - \sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} R_\theta(\boldsymbol{0})$$

$$= \sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} \sum_{s=1}^{S} w_s(l_s + u'_s) - \sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} \sum_{s=1}^{S} w_s l_s.$$

Let $\boldsymbol{w}^* = \arg\sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} \sum_s w_s l_s$ and $\boldsymbol{v}^* = \arg\sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} \sum_s w_s(l_s + u'_s)$, then according to Danskin's theorem (Danskin, 1966), $\partial D_1(\theta)/\partial l_s = v_s^* - w_s^*$ and

$$\left| \sum_{s=1}^{S} \frac{\partial D_1(\theta)}{\partial l_s} \right| \le \sum_{s=1}^{S} |v_s^* - w_s^*|$$

$$\le \sqrt{S}\|\boldsymbol{v}^* - \boldsymbol{w}^*\|_2,$$

which means $\mathcal{O}(\|\nabla_\theta D_1(\theta)\|_2) = \mathcal{O}(\|\boldsymbol{v}^* - \boldsymbol{w}^*\|_2)$.

Consider the perturbation analysis of the following optimization problem

$$\min_{\boldsymbol{w}} -\sum_{s=1}^{S} w_s(l_s + u'_s) \quad \text{s.t. } \boldsymbol{w} \in \Phi(\boldsymbol{u}), \tag{104}$$

whose objective is perturbed and feasible region is unperturbed. Let

$$L(\boldsymbol{w}, \alpha, \alpha', \boldsymbol{u}') = -\sum_{s=1}^{S} w_s(l_s + u'_s) + \alpha \left(\sum_{s=1}^{S} (p_s + u_s)f(w_s) - \delta \right)$$

$$+ \alpha' \left(\sum_{s=1}^{S} (p_s + u_s)w_s - 1 \right)$$

be the Lagrangian function, where $\alpha \ge 0$ and α' are Lagrange multipliers, and for simplicity the nonnegative constraints are omitted. Note that given the assumption (c), if $\alpha \ne 0$,

$$\frac{\partial^2}{\partial w_i \partial w_j} L(\boldsymbol{w}, \alpha, \alpha', \boldsymbol{u}') = \begin{cases} \alpha f''(w_i) > 0, & i = j, \\ 0, & i \ne j, \end{cases}$$

namely, $L(\boldsymbol{w}, \alpha, \alpha', \boldsymbol{u}')$ is locally strongly convex in $\boldsymbol{w}$. Thus,

- if $\alpha^* > 0$, the second-order sufficient condition (see Definition 6.2 in (Bonnans & Shapiro, 1998)) holds at $\boldsymbol{w}^*$ that implies the corresponding second-order growth condition according to Theorem 6.3 in (Bonnans & Shapiro, 1998);
- if $\alpha^* = 0$, (104) is locally a standard linear programming around $\boldsymbol{w}^*$ and it is fairly easy to see $\|\boldsymbol{v}^* - \boldsymbol{w}^*\|_2 = \mathcal{O}(\|\boldsymbol{u}'\|_2)$ according to Theorem 1 in (Robinson, 1977).

In the former case, it is obvious that for (104),

- the objective $-\sum_s w_s(l_s + u'_s)$ is Lipschitz continuous with a Lipschitz constant $\|\boldsymbol{l}\|_2 + \epsilon$ independent of $\boldsymbol{u}'$;
- the difference function $-\sum_s w_s u'_s$ is Lipschitz continuous with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}'\|_2)$.

Therefore, $\|\boldsymbol{v}^* - \boldsymbol{w}^*\|_2 = \mathcal{O}(\|\boldsymbol{u}'\|_2)$ by applying Proposition 6.1 in (Bonnans & Shapiro, 1998). □

In order to prove the same property for $D_2(\theta)$, we need several lemmas.

Lemma 5. *Denote by $f'(\boldsymbol{w}) = (f'(w_1), \ldots, f'(w_S))^\top$. There exists a constant $C_{\cos} > 0$, such that $\cos(\boldsymbol{d} \circ f'(\boldsymbol{w}), \boldsymbol{d}) \le 1 - C_{\cos}$ for all $\boldsymbol{u}$ satisfying $\|\boldsymbol{u}\|_2 \le \epsilon$, $\boldsymbol{w}$ satisfying $\sum_s (p_s + u_s)f(w_s) = \delta$ and $\sum_s (p_s + u_s)w_s = 1$, and $\boldsymbol{d} > \boldsymbol{0}$.*

Proof. Suppose the lemma is false, i.e., for any sufficiently large n, there exists some $\boldsymbol{w}_n$ such that $\cos(\boldsymbol{d} \circ f'(\boldsymbol{w}_n), \boldsymbol{d}) =$

$1 - 1/(2n^2)$. Let $\zeta_n = \|\boldsymbol{d} \circ f'(\boldsymbol{w}_n)\|_2$ and $\eta_n = \zeta_n/\|\boldsymbol{d}\|_2$, then

$$
\begin{aligned}
\|\boldsymbol{d} \circ f'(\boldsymbol{w}_n) - \eta_n \boldsymbol{d}\|_2^2 &= \|\boldsymbol{d} \circ f'(\boldsymbol{w}_n)\|_2^2 + \eta_n^2 \|\boldsymbol{d}\|_2^2 - 2\eta_n (\boldsymbol{d} \circ f'(\boldsymbol{w}_n))^\top \boldsymbol{d} \\
&= 2\zeta_n^2 - 2\eta_n \cos(\boldsymbol{d} \circ f'(\boldsymbol{w}_n), \boldsymbol{d}) \|\boldsymbol{d} \circ f'(\boldsymbol{w}_n)\|_2 \|\boldsymbol{d}\|_2 \\
&= 2\zeta_n^2 - 2(1 - 1/(2n^2))\zeta_n^2 \\
&= \zeta_n^2/n^2.
\end{aligned}
$$

In other words, for $s = 1, \ldots, S$,

$$
\begin{aligned}
|f'(w_{n,s}) - \eta_n| &= |d_s f'(w_{n,s}) - \eta_n d_s|/d_s \\
&\leq \|\boldsymbol{d} \circ f'(\boldsymbol{w}_n) - \eta_n \boldsymbol{d}\|_2/d_s \\
&= (\zeta_n/n)/d_s \\
&\leq \zeta_n'/n,
\end{aligned}
$$

where $\zeta_n' = \zeta_n / \min_s d_s$. Consequently, for any $1 \leq i, j \leq S$ and $i \neq j$,

$$
\begin{aligned}
|f'(w_{n,i}) - f'(w_{n,j})| &\leq |f'(w_{n,i}) - \eta_n| + |f'(w_{n,j}) - \eta_n| \\
&\leq 2\zeta_n'/n.
\end{aligned}
$$

Let $C_{f''} > 0$ be the lower bound of $f''(t)$ mentioned in the assumption (c). This assumption also guarantees that $f'(t)$ is continuous, and by the mean value theorem, there is some t between $w_{n,i}$ and $w_{n,j}$ such that

$$
\begin{aligned}
|w_{n,i} - w_{n,j}| &= \left| \frac{f'(w_{n,i}) - f'(w_{n,j})}{f''(t)} \right| \\
&\leq 2\zeta_n'/(C_{f''}n).
\end{aligned}
$$

Let $\eta_n' = \sum_s w_{n,s}/S$, then $|w_{n,s} - \eta_n'| \leq 2\zeta_n'/(C_{f''}n)$ for $s = 1, \ldots, S$.

Recall that $\sum_s (p_s + u_s)w_{n,s} = 1$, then

$$
\begin{aligned}
\left(1 + \sum_{s=1}^S u_s\right)\eta_n' - 1 &= \sum_{s=1}^S (p_s + u_s)\eta_n' - \sum_{s=1}^S (p_s + u_s)w_{n,s} \\
&= \sum_{s=1}^S (p_s + u_s)(\eta_n' - w_{n,s}),
\end{aligned}
$$

and hence

$$
\begin{aligned}
\left| \left(1 + \sum_{s=1}^S u_s\right)\eta_n' - 1 \right| &\leq \frac{2\zeta_n'}{C_{f''}n} \sum_{s=1}^S (p_s + u_s) \\
&= \frac{2\zeta_n'}{C_{f''}n} \left(1 + \sum_{s=1}^S u_s\right).
\end{aligned}
$$

This ensures $|\eta_n' - 1/(1 + \sum_s u_s)| \leq 2\zeta_n'/(C_{f''}n)$ and implies $|w_{n,s} - 1/(1 + \sum_s u_s)| \leq 4\zeta_n'/(C_{f''}n)$ for $s = 1, \ldots, S$. Since $f(t)$ is twice differentiable and $\|\boldsymbol{w}_n\|_2 < \infty$, we must have $\zeta_n' < \infty$ and then $\lim_{n\to\infty} w_{n,s} = 1/(1 + \sum_s u_s)$ for all $s = 1, \ldots, S$.

The Taylor expansion of $f(t)$ at $t = 1$ is $f(t) = f'(1)(t - 1) + \mathcal{O}((t - 1)^2)$ since $f(1) = 0$, and if $t = 1/(1 + \sum_s u_s)$,

$$
f\left(\frac{1}{1 + \sum_{s=1}^S u_s}\right) = -f'(1) \cdot \frac{\sum_{s=1}^S u_s}{1 + \sum_{s=1}^S u_s} + \mathcal{O}\left(\left(\frac{\sum_{s=1}^S u_s}{1 + \sum_{s=1}^S u_s}\right)^2\right).
$$

When $\|\boldsymbol{u}\|_2 \leq \epsilon$, $|\sum_s u_s| \leq \|\boldsymbol{u}\|_1 \leq \sqrt{S}\epsilon$, and

$$f\left(\frac{1}{1+\sum_{s=1}^{S} u_s}\right) \leq |f'(1)|\frac{\sqrt{S}\epsilon}{1-\sqrt{S}\epsilon} + \mathcal{O}(\epsilon^2)$$

$$\leq 2\sqrt{S}|f'(1)|\epsilon.$$

As a result,

$$\lim_{n\to\infty} \sum_{s=1}^{S}(p_s + u_s)f(w_{n,s}) = \sum_{s=1}^{S}(p_s + u_s)f\left(\frac{1}{1+\sum_{s=1}^{S} u_s}\right)$$

$$= \left(1 + \sum_{s=1}^{S} u_s\right) f\left(\frac{1}{1+\sum_{s=1}^{S} u_s}\right)$$

$$\leq (1 + \sqrt{S}\epsilon) \cdot 2\sqrt{S}|f'(1)|\epsilon$$

$$\leq 4\sqrt{S}|f'(1)|\epsilon.$$

However, this is impossible since $\sum_s(p_s + u_s)f(w_{n,s}) = \delta \geq 5\sqrt{S}|f'(1)|\epsilon$. $\qquad\square$

Based on Lemma 5, we derive the convergence rate of $\Phi(\boldsymbol{u})$ to $\Phi(\boldsymbol{0})$.

Lemma 6. *Let $d_{\mathcal{H}}(V, W)$ be the Hausdorff distance between two sets V and W:*

$$d_{\mathcal{H}}(V, W) = \max\left\{\sup_{\boldsymbol{v}\in V}\inf_{\boldsymbol{w}\in W}\|\boldsymbol{v}-\boldsymbol{w}\|_2, \sup_{\boldsymbol{w}\in W}\inf_{\boldsymbol{v}\in V}\|\boldsymbol{v}-\boldsymbol{w}\|_2\right\}.$$

Then $d_{\mathcal{H}}(\Phi(\boldsymbol{u}), \Phi(\boldsymbol{0})) = \mathcal{O}(\|\boldsymbol{u}\|_2)$ for all $\boldsymbol{u}$ satisfying $\|\boldsymbol{u}\|_2 \leq \epsilon$.

Proof. We are going to prove $\sup_{\boldsymbol{w}\in\Phi(\boldsymbol{0})}\inf_{\boldsymbol{v}\in\Phi(\boldsymbol{u})}\|\boldsymbol{v}-\boldsymbol{w}\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2)$, and the other direction can be proven similarly.

Pick an arbitrary $\boldsymbol{w}_0 \in \Phi(\boldsymbol{0})$. Let $\beta = \delta/(\delta + \|f(\boldsymbol{w}_0)\|_2\|\boldsymbol{u}\|_2)$ and consider $\boldsymbol{v}_1 = \beta\boldsymbol{w}_0 + (1-\beta)\boldsymbol{1}$,

$$\|\boldsymbol{v}_1 - \boldsymbol{w}_0\|_2 = \|(\beta-1)\boldsymbol{w}_0 + (1-\beta)\boldsymbol{1}\|_2$$

$$= (1-\beta)\|\boldsymbol{w}_0 - \boldsymbol{1}\|_2$$

$$\leq \|f(\boldsymbol{w}_0)\|_2\|\boldsymbol{u}\|_2\|\boldsymbol{w}_0 - \boldsymbol{1}\|_2/\delta$$

$$= \mathcal{O}(\|\boldsymbol{u}\|_2).$$

Moreover,

$$\sum_{s=1}^{S}(p_s + u_s)f(v_{1,s}) = \sum_{s=1}^{S}(p_s + u_s)f(\beta w_{0,s} + (1-\beta))$$

$$\leq \sum_{s=1}^{S}(p_s + u_s)(\beta f(w_{0,s}) + (1-\beta)f(1))$$

$$= \beta\left(\sum_{s=1}^{S}p_s f(w_{0,s}) + \sum_{s=1}^{S}u_s f(w_{0,s})\right)$$

$$\leq \beta(\delta + \|f(\boldsymbol{w}_0)\|_2\|\boldsymbol{u}\|_2)$$

$$= \delta,$$

where the second line is due to the convexity of $f(t)$, the third line is because $f(1) = 0$, and the fourth line is according to Jensen's inequality. This means $\boldsymbol{v}_1$ belongs to the set $V_1 = \{\boldsymbol{w} \in \mathbb{R}^S \mid \sum_s(p_s + u_s)f(w_s) \leq \delta, w_s \geq 0\}$.

However, v_1 does not belong to the set $V_2 = \{w \in \mathbb{R}^S \mid \sum_s (p_s + u_s) w_s = 1, w_s \geq 0\}$. Since V_2 is a hyperplane, we can easily project v_1 onto V_2 to obtain v_2, and

$$
\begin{aligned}
\|v_2 - v_1\|_2 &= \frac{1}{\|p + u\|_2} \left| \sum_{s=1}^S (p_s + u_s) v_{1,s} - 1 \right| \\
&\leq \frac{2}{\|p\|_2} \left| \sum_{s=1}^S (p_s + u_s)(\beta w_{0,s} + 1 - \beta) - 1 \right| \\
&\leq \sqrt{4S} \left| \beta \sum_{s=1}^S p_s w_{0,s} + \beta \sum_{s=1}^S u_s w_{0,s} + (1 - \beta) \sum_{s=1}^S p_s + (1 - \beta) \sum_{s=1}^S u_s - 1 \right| \\
&= \sqrt{4S} \left| \beta + \beta \sum_{s=1}^S u_s w_{0,s} + (1 - \beta) + (1 - \beta) \sum_{s=1}^S u_s - 1 \right| \\
&\leq \sqrt{4S} \beta \left| \sum_{s=1}^S u_s w_{0,s} \right| + \mathcal{O}(\|u\|_2^2) \\
&= \mathcal{O}(\|u\|_2).
\end{aligned}
$$

After this projection, $v_2 \notin V_1$ again.

Let v_3 be the projection of v_2 onto $\Phi(u) = V_1 \cap V_2$, $T_w V_1(v_3)$ be the tangent hyperplane to V_1 at v_3 of $S - 1$ dimensions, and $T_w(V_1 \cap V_2)(v_3)$ be that to $V_1 \cap V_2$ at v_3 of $S - 2$ dimensions. As a consequence, $v_3 - v_2 \in V_2$ is one of normal vectors to $T_w(V_1 \cap V_2)(v_3)$ at v_3, which is also the projection of the normal vector to $T_w V_1(v_3)$ at v_3 onto V_2. This means the normal vector to $T_w V_1(v_3)$ at v_3 belongs to the 2-dimensional plane determined by $v_3 - v_2$ and $v_1 - v_2$, since the latter is a normal vector to V_2 at v_2.

Consider the triangle $(v_1 - v_2, v_3 - v_2, v_1 - v_3)$. This is a right-angled triangle since $v_1 - v_2 \perp V_2$ and $v_3 - v_2 \in V_2$, so that

$$
\|v_1 - v_3\|_2 = \frac{\|v_1 - v_2\|_2}{\sin(v_2 - v_3, v_1 - v_3)}.
$$

Subsequently, let v_4 be the intersection of $v_1 - v_2$ and $T_w V_1(v_3)$, due to the convexity of V_1,

$$
\begin{aligned}
\sin^2(v_2 - v_3, v_1 - v_3) &\geq \sin^2(v_2 - v_3, v_4 - v_3) \\
&= 1 - \cos^2(v_2 - v_3, v_4 - v_3) \\
&= 1 - \cos^2(p + u, (p + u) \circ f'(v_3)),
\end{aligned}
$$

where $p + u$ is a normal vector to V_2 containing $v_2 - v_3$, $(p + u) \circ f'(v_3)$ is a normal vector to $T_w V_1(v_3)$ containing $v_4 - v_3$, and both of them belong to the 2-dimensional plane determined by $v_3 - v_2$ and $v_1 - v_2$ as we have proven above. By definition, v_3 is on the boundary of V_1 such that $\sum_s (p_s + u_s) f(v_{3,s}) = \delta$, and according to Lemma 5,

$$
\begin{aligned}
1 - \cos^2(p + u, (p + u) \circ f'(v_3)) &\geq 1 - (1 - C_{\cos})^2 \\
&= C_{\cos}(2 - C_{\cos}),
\end{aligned}
$$

which implies

$$
\begin{aligned}
\|v_1 - v_3\|_2 &\leq \frac{\|v_1 - v_2\|_2}{\sqrt{C_{\cos}(2 - C_{\cos})}} \\
&= \mathcal{O}(\|v_1 - v_2\|_2) \\
&= \mathcal{O}(\|u\|_2).
\end{aligned}
$$

Combining $\|v_1 - w_0\|_2 = \mathcal{O}(\|u\|_2)$ and $\|v_1 - v_3\|_2 = \mathcal{O}(\|u\|_2)$ gives us $\|v_3 - w_0\|_2 = \mathcal{O}(\|u\|_2)$, and thus $\inf_{v \in \Phi(u)} \|v - w_0\|_2 \leq \|v_3 - w_0\|_2 = \mathcal{O}(\|u\|_2)$. Since w_0 is arbitrarily picked from $\Phi(0)$, the proof is completed. $\square$

Table 2: Summary of dataset statistics.

Dataset	#Points	#classes	Dimension
blood	748	2	4
adult	32561	2	123
fourclass	862	2	2
phishing	11055	2	68
20news	18040	20	50
satimage	4435	6	36
letter	20000	26	16
mnist	70000	10	50

Lemma 7 (Lipschitz continuity of the difference function, II). *For all $\boldsymbol{u}$ such that $\|\boldsymbol{u}\|_2 \leq \epsilon$, $D_2(\theta)$ is Lipschitz continuous with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}\|_2^{1/2})$.*

Proof. The proof goes along the same line with Lemma 4. Let $\boldsymbol{w}^* = \arg\sup_{\boldsymbol{w} \in \Phi(\boldsymbol{0})} \sum_s w_s l_s$ and $\boldsymbol{v}^* = \arg\sup_{\boldsymbol{w} \in \Phi(\boldsymbol{u})} \sum_s w_s l_s$, and consider the perturbation analysis of the following optimization problem

$$\min_{\boldsymbol{w}} \; -\sum_{s=1}^{S} w_s l_s \quad \text{s.t. } \boldsymbol{w} \in \Phi(\boldsymbol{u}),$$

whose objective is unperturbed and feasible region is perturbed. According to Lemma 6, we have $d_{\mathcal{H}}(\Phi(\boldsymbol{u}), \Phi(\boldsymbol{0})) = \mathcal{O}(\|\boldsymbol{u}\|_2)$, which ensures that the multifunction $\boldsymbol{u} \mapsto \Phi(\boldsymbol{u})$ is upper Lipschitz continuous and that $d_{\mathcal{H}}(\{\boldsymbol{w}^*\}, \Phi(\boldsymbol{u})) = \mathcal{O}(\|\boldsymbol{u}\|_2)$. Hence, $\|\boldsymbol{v}^* - \boldsymbol{w}^*\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2^{1/2})$ by applying Proposition 6.4 in (Bonnans & Shapiro, 1998). $\square$

Let us summarize what we have obtained so far:

- a second-order growth condition of $J(\theta, \boldsymbol{0}, \boldsymbol{0})$ at $\theta = \theta^*$;
- the Lipschitz continuity of $J(\theta, \boldsymbol{u}, \boldsymbol{u}')$ with a Lipschitz constant independent of $\boldsymbol{u}$ and $\boldsymbol{u}'$;
- the Lipschitz continuity of $D(\theta)$ with a Lipschitz constant of order $\mathcal{O}(\|\boldsymbol{u}\|_2^{1/2} + \|\boldsymbol{u}'\|_2)$.

Note that θ is unconstrained, by applying Proposition 6.1 in (Bonnans & Shapiro, 1998), we can obtain

$$\|\theta_{\boldsymbol{u}, \boldsymbol{u}'} - \theta^*\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2^{1/2} + \|\boldsymbol{u}'\|_2).$$

This immediately implies

$$\|J(\theta_{\boldsymbol{u}, \boldsymbol{u}'}, \boldsymbol{0}, \boldsymbol{0}) - J(\theta^*, \boldsymbol{0}, \boldsymbol{0})\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2^{1/2} + \|\boldsymbol{u}'\|_2),$$

due to the Lipschitz continuity of $J(\theta, \boldsymbol{0}, \boldsymbol{0})$. $\square$

I. Datasets

We obtained six classification datasets from the UCI repository[16] and also obtained 20newsgroups[17] and MNIST datasets. We used the raw features for the datasets from the UCI repository. For the 20newgroups dataset, we removed stop words and retained the 2000 most frequent words. We then removed documents with fewer than 10 words. We extracted tf-idf features and applied Principle Component Analysis (PCA) to reduce the dimensionality to 50. For the MNIST dataset, we applied PCA on the raw features to reduce the dimensionality to 50. The dataset statistics are summarized in Table 2.

J. Details of the Subcategory Shift Scenario

In this section, we give details on how we converted the original multi-class classification problems into multi-class classification problems with fewer classes.

---

[16]http://archive.ics.uci.edu/ml/index.html
[17]http://qwone.com/ jason/20Newsgroups/

Table 3: Experimental comparisons of the three methods w.r.t. the estimated ordinary risk and the estimated structural adversarial risk *using the surrogate loss (the logistic loss)*. The lower these values are, the better the performance of the method is. The KL divergence is used and distribution shift is assumed to be (a) class prior change and (b) sub-category prior change. Mean and standard deviation over 50 random train-test splits were reported. The best method and comparable ones based on the t-test at the significance level 1% are highlighted in boldface.

(a) Class prior change.

Dataset	Estimated ordinary risk			Estimated adversarial risk			Estimated structural adversarial risk		
	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM
blood	**0.52 (0.05)**	0.69 (0.0)	0.62 (0.02)	1.04 (0.1)	**0.69 (0.0)**	0.97 (0.03)	0.86 (0.23)	0.69 (0.0)	**0.63 (0.19)**
adult	**0.33 (0.0)**	0.65 (0.03)	0.39 (0.0)	1.28 (0.02)	**0.69 (0.01)**	1.42 (0.01)	0.59 (0.3)	0.67 (0.01)	**0.4 (0.38)**
fourclass	**0.51 (0.05)**	0.69 (0.0)	0.54 (0.05)	0.91 (0.04)	**0.69 (0.0)**	0.88 (0.04)	0.65 (0.13)	0.69 (0.0)	**0.56 (0.13)**
phishing	**0.15 (0.01)**	0.41 (0.08)	**0.15 (0.0)**	0.86 (0.01)	**0.59 (0.08)**	0.85 (0.01)	0.18 (0.06)	0.41 (0.02)	**0.16 (0.05)**
20news	**1.05 (0.01)**	1.49 (0.04)	1.22 (0.02)	3.42 (0.02)	**3.0 (0.04)**	3.58 (0.03)	1.43 (0.1)	1.74 (0.19)	**1.32 (0.13)**
satimage	**1.01 (0.01)**	1.26 (0.02)	1.29 (0.01)	2.54 (0.02)	**2.15 (0.02)**	2.86 (0.02)	1.41 (0.05)	1.59 (0.01)	**1.38 (0.04)**
letter	**0.37 (0.01)**	0.47 (0.03)	0.51 (0.02)	1.65 (0.02)	**1.19 (0.03)**	2.27 (0.03)	0.77 (0.17)	0.77 (0.09)	**0.58 (0.21)**
mnist	**0.35 (0.0)**	0.59 (0.05)	0.45 (0.0)	1.96 (0.01)	**1.38 (0.04)**	1.85 (0.01)	0.49 (0.06)	0.73 (0.02)	**0.47 (0.04)**

(b) Sub-category prior change.

Dataset	Estimated ordinary risk			Estimated adversarial risk			Estimated structural adversarial risk		
	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM
20news	**0.61 (0.01)**	0.84 (0.06)	0.76 (0.05)	2.76 (0.02)	**2.03 (0.05)**	2.91 (0.03)	1.02 (0.14)	1.08 (0.29)	**0.89 (0.21)**
satimage	**0.63 (0.0)**	0.69 (0.0)	0.68 (0.0)	0.99 (0.01)	**0.69 (0.0)**	0.74 (0.0)	0.81 (0.02)	0.69 (0.0)	**0.69 (0.02)**
letter	**0.47 (0.0)**	0.69 (0.0)	0.64 (0.02)	1.05 (0.03)	**0.69 (0.0)**	0.86 (0.01)	0.93 (0.02)	0.69 (0.0)	**0.68 (0.06)**
mnist	**0.31 (0.0)**	0.68 (0.01)	0.39 (0.0)	1.28 (0.01)	**0.69 (0.0)**	1.25 (0.0)	0.49 (0.05)	0.68 (0.0)	**0.42 (0.03)**

For the datasets from the UCI repository, we systematically grouped the class labels into binary categories by the following procedure. First, class labels are sorted by the number of data points in the classes. Then, 1, 3, 5, . . . -th labels are assigned to a positive category and the others are assigned to a negative category. For MNIST, we considered a binary classification between odd and even numbers and set the original classes as subcategories. For 20newsgroups, we converted the original 20-class classification problem into a 7-class one with each class corresponding to a high-level topic: comp, rec, sci, misc, alt, soc, and talk. We then set the original classes as subcategories.

K. Experimental Results measured by Surrogate Loss

In this section, we report the experimental results measured by the surrogate loss (the logistic loss). We used the KL and the PE divergences, where we set $\delta = 0.5$. We used the same f-divergence and the same δ during training and testing. The experimental results using the KL and the PE divergences are reported in Tables 3 and 4, respectively. We empirically confirmed that *in terms of the surrogate loss*, each method indeed achieved the best performance in terms of the metric it optimizes for.

L. Experimental Results with the PE divergence

In this section, we report the experimental results using the PE divergence, where we set $\delta = 0.5$. The experimental results are reported in Table 5.

Table 4: Experimental comparisons of the three methods w.r.t. the estimated ordinary risk and the estimated structural adversarial risk *using the surrogate loss (the logistic loss)*. The lower these values are, the better the performance of the method is. The PE divergence is used and distribution shift is assumed to be (a) class prior change and (b) sub-category prior change. Mean and standard deviation over 50 random train-test splits were reported. The best method and comparable ones based on the t-test at the significance level 1% are highlighted in boldface.

(a) Class prior change.

Dataset	Estimated ordinary risk			Estimated adversarial risk			Estimated structural adversarial risk		
	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM
blood	**0.52 (0.05)**	0.67 (0.02)	0.61 (0.03)	0.77 (0.04)	**0.69 (0.0)**	0.81 (0.02)	0.71 (0.07)	0.69 (0.0)	**0.62 (0.07)**
adult	**0.33 (0.0)**	0.41 (0.02)	0.39 (0.01)	0.69 (0.01)	**0.61 (0.01)**	0.77 (0.01)	0.49 (0.02)	0.51 (0.0)	**0.4 (0.03)**
fourclass	**0.52 (0.05)**	0.66 (0.02)	**0.53 (0.05)**	0.73 (0.02)	**0.69 (0.01)**	0.77 (0.04)	0.6 (0.04)	0.67 (0.0)	**0.54 (0.06)**
phishing	**0.15 (0.01)**	0.2 (0.02)	**0.15 (0.0)**	0.43 (0.01)	**0.4 (0.02)**	0.43 (0.01)	0.17 (0.02)	0.21 (0.01)	**0.15 (0.01)**
20news	**1.04 (0.01)**	1.15 (0.11)	1.17 (0.02)	1.99 (0.01)	**1.95 (0.1)**	2.2 (0.02)	1.29 (0.03)	1.36 (0.06)	**1.24 (0.04)**
satimage	**1.01 (0.01)**	1.1 (0.01)	1.1 (0.01)	1.81 (0.01)	**1.72 (0.02)**	2.0 (0.01)	1.26 (0.02)	1.32 (0.01)	**1.17 (0.02)**
letter	**0.36 (0.01)**	0.4 (0.01)	0.42 (0.02)	0.89 (0.02)	**0.81 (0.02)**	1.0 (0.02)	0.6 (0.04)	0.59 (0.03)	**0.53 (0.05)**
mnist	**0.35 (0.0)**	0.41 (0.01)	0.43 (0.0)	0.96 (0.01)	**0.87 (0.01)**	1.04 (0.01)	0.45 (0.02)	0.5 (0.01)	**0.44 (0.01)**

(b) Sub-category prior change.

Dataset	Estimated ordinary risk			Estimated adversarial risk			Estimated structural adversarial risk		
	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM
20news	**0.61 (0.01)**	0.67 (0.01)	0.68 (0.02)	1.34 (0.01)	**1.22 (0.01)**	1.4 (0.01)	0.86 (0.04)	0.86 (0.03)	**0.81 (0.05)**
satimage	**0.63 (0.0)**	0.69 (0.0)	0.68 (0.0)	0.86 (0.01)	**0.69 (0.0)**	0.73 (0.0)	0.76 (0.01)	0.69 (0.0)	**0.69 (0.01)**
letter	**0.47 (0.0)**	0.67 (0.0)	0.57 (0.04)	0.79 (0.01)	**0.69 (0.0)**	0.73 (0.01)	0.74 (0.01)	0.69 (0.0)	**0.66 (0.03)**
mnist	**0.31 (0.0)**	0.4 (0.02)	0.36 (0.0)	0.72 (0.0)	**0.6 (0.01)**	0.72 (0.0)	0.44 (0.01)	0.48 (0.0)	**0.41 (0.01)**

Table 5: Experimental comparisons of the three methods w.r.t. the estimated ordinary risk and the estimated structural adversarial risk *using the 0-1 loss (%)*. The lower these values are, the better the performance of the method is. The PE divergence is used and distribution shift is assumed to be (a) class prior change and (b) sub-category prior change. Mean and standard deviation over 50 random train-test splits were reported. The best method and comparable ones based on the t-test at the significance level 1% are highlighted in boldface.

(a) Class prior change.

Dataset	Estimated ordinary risk			Estimated structural adversarial risk		
	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM
blood	**22.2 (0.6)**	**22.4 (0.5)**	33.0 (2.0)	47.4 (1.6)	49.1 (1.1)	**36.6 (2.4)**
adult	**15.3 (0.1)**	**15.3 (0.1)**	18.7 (0.2)	24.9 (0.3)	24.8 (0.3)	**19.1 (0.3)**
fourclass	**23.7 (1.2)**	**23.5 (1.2)**	27.0 (1.3)	32.7 (1.7)	32.6 (1.8)	**28.7 (1.7)**
phishing	6.0 (0.2)	6.1 (0.2)	**5.9 (0.2)**	7.1 (0.3)	7.4 (0.3)	**6.4 (0.3)**
20news	**28.8 (0.3)**	29.7 (0.3)	33.7 (0.3)	37.9 (0.3)	38.4 (0.4)	**37.5 (0.4)**
satimage	**25.1 (0.2)**	26.7 (0.3)	28.1 (0.3)	33.7 (0.4)	35.6 (0.4)	**32.0 (0.4)**
letter	**14.2 (0.5)**	**14.5 (0.5)**	15.5 (0.5)	26.5 (0.9)	25.6 (0.9)	**20.8 (0.7)**
mnist	**10.0 (0.1)**	10.1 (0.1)	12.2 (0.1)	13.3 (0.1)	13.2 (0.1)	**13.1 (0.1)**

(b) Sub-category prior change.

Dataset	Estimated ordinary risk			Estimated structural adversarial risk		
	ERM	AERM	Structural AERM	ERM	AERM	Structural AERM
20news	**19.0 (0.3)**	19.6 (0.4)	20.8 (0.4)	29.2 (0.4)	29.7 (0.4)	**27.3 (0.4)**
satimage	**36.4 (0.3)**	41.1 (1.9)	39.6 (0.5)	53.9 (0.4)	56.7 (2.7)	**47.0 (0.5)**
letter	**17.4 (0.4)**	18.5 (0.5)	23.1 (3.3)	**38.0 (0.5)**	39.5 (0.6)	38.9 (1.0)
mnist	**13.3 (0.1)**	13.5 (0.1)	15.6 (0.2)	20.0 (0.2)	20.2 (0.2)	**18.6 (0.2)**