## A. Proof of Theorem 1

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

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

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

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.

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

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

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):

$$\textbf{Case 1-a}: \quad p_1' < p_1 \cdot r_1^*(p_1) \tag{34}$$

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

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 $\theta_1$ and $\theta_2$, if $\mathcal{R}_{\mathrm{adv}}(\theta_1) < 1$, we have

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

To show that the opposite direction of Eq. (44) holds, we need to show that any pair of parameters $\theta_1$ and $\theta_2$, the following holds:

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

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. □

## B. Proof of Theorem 2

We prove by contradiction. Let $\Omega$ 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

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

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

[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.,

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

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

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

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

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

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 $y^{(\max)}(x^\dagger) \equiv \arg\max_{y\in\mathcal{Y}} p(y|x^\dagger)$ 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

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

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

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

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

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

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

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

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

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 $\epsilon$. 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 $\epsilon$. 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 $r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger)) > r^*(x^\dagger, y^{(\max)}(x^\dagger))$ holds because of Eq. (55). Therefore, we have $f'(r^*(x^\dagger, y^{(\max)}(x^\dagger))) - f'(r^*(x^\dagger, y^{(\mathrm{adv})}(x^\dagger))) \leq 0$.

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

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

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

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

Thus, we have

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

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

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

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

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

We now show $\ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\max)}(x^\dagger)) > \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger))$. Suppose $\ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\max)}(x^\dagger)) \leq \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger))$. Construct $g' \in \mathbb{R}^K$ by swapping the $y^{(\max)}$-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 $\ell(g', y^{(\max)}(x^\dagger)) = \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger))$ and $\ell(g', y^{(\mathrm{adv})}(x^\dagger)) = \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\max)}(x^\dagger))$. Combining this with Eq. (54), we have

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

which is in contradiction with the fact that $g^{(\mathrm{adv})}$ achieves the minimal value of Eq. (50). Therefore, we have $\ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\max)}(x^\dagger)) > \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger))$. 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 $r_{\mathrm{new}}(\cdot,\cdot)$ 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, $g^{(\mathrm{adv})}$, 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 $g^{(\mathrm{adv})}$, 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 $g^{(\mathrm{adv})}$ is chosen from all measurable functions, its function value at $x$ is obtained in a point-wise manner:

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

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

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

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

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

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

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

Then we have

$$\begin{aligned}
\int_{x\in\mathcal{S}_{\mathrm{diff}}} p(x)\mathrm{d}x &= \int_{x\in\mathcal{S}} p(x)\mathrm{d}x + \underbrace{\int_{x\in\mathcal{S}'_{\mathrm{diff}}} p(x)\mathrm{d}x}_{=0 \ \because \text{Eq. (79).}} \\
&= \int_{x\in\mathcal{S}} \frac{q^*(x)}{r^*(x)}\mathrm{d}x \\
&\leq \frac{1}{\min_{x\in\mathcal{S}} r^*(x)} \int_{x\in\mathcal{S}} q^*(x)\mathrm{d}x \\
&= 0.
\end{aligned} \tag{81}$$

Therefore, the particular $g^{(\mathrm{adv})}$ 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 $\ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{max})}(x^\dagger)) > \ell(g^{(\mathrm{adv})}(x^\dagger), y^{(\mathrm{adv})}(x^\dagger))$, we construct $g'$ in Eq. (75) by $-g^{(\mathrm{adv})}(x^\dagger)$. With these two modifications, all the arguments for multi-class classification hold for binary classification. □

**Remark 2.** Here, we show that *if the KL divergence is used* for the $f$-divergence, the prediction of *any* solution $g^{(\mathrm{adv})}$ 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 $\gamma$ 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

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

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

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

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, $\phi_{\mathrm{steep}}(y\widehat{y})$, 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 $\phi'_{\mathrm{steep}}(0) = h(\phi(0))\phi'(0) < 0$. Using the fact again, we immediately have that $\phi_{\mathrm{steep}}(y\widehat{y})$ is classification calibrated. □

**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 $\ell$. 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 $\theta^*$ be the stationary point of Eq. (7). By using a chain rule and Danskin's theorem (Danskin, 1966), $\theta^*$ 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 $\theta^*$. For notational convenience, for $1 \leq i \leq N$, let us denote $\ell_i(\theta^*)$ by $\ell_i^*$. Then, $\boldsymbol{r}^*$ is the solution of the following optimization problem.

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

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

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

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

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

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} r_i''^* \ell_i^* - \frac{1}{N}\sum_{i=1}^{N} r_i'^* \ell_i^* &= \frac{1}{N}\left(r_j''^* \ell_j^* + r_k''^* \ell_k^* - r_j'^* \ell_j^* - r_k'^* \ell_k^*\right) \\ &= \frac{1}{N}\left(r_k'^* \ell_j^* + r_j'^* \ell_k^* - r_j'^* \ell_j^* - r_k'^* \ell_k^*\right) \\ &= \frac{1}{N}\left(r_j'^* - r_k'^*\right)\left(\ell_k^* - \ell_j^*\right) \\ &> 0. \end{aligned} \tag{95}$$

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

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

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

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

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} f\left(r_i''^*\right) &= \frac{1}{N}\left\{\left(\sum_{i\neq j,k} f\left(r_i''^*\right)\right) + f\left(r_j''^*\right) + f\left(r_k''^*\right)\right\} \\ &= \frac{1}{N}\left\{\left(\sum_{i\neq j,k} f\left(r_i'^*\right)\right) + f\left((r_j'^* + r_k'^*)/2\right) + f\left((r_j'^* + r_k'^*)/2\right)\right\} \\ &\leq \frac{1}{N}\left\{\left(\sum_{i\neq j,k} f\left(r_i'^*\right)\right) + f\left(r_j'^*\right) + f\left(r_k'^*\right)\right\} \quad (\because \text{convexity of } f(\cdot).) \\ &= \frac{1}{N}\sum_{i=1}^{N} f\left(r_i'^*\right) \\ &\leq \delta. \quad \left(\because \boldsymbol{r}'^* \in \widehat{\mathcal{U}}_f.\right) \end{aligned} \tag{97}$$

Also, it is easy to see that $\boldsymbol{r}''^*$ attain the same maximum value as $\boldsymbol{r}'^*$; thus, $\boldsymbol{r}''^*$ 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 $\boldsymbol{r}'^* \in \widehat{\mathcal{U}}_f$ and equally distribute the weights to the same losses to obtain $\boldsymbol{r}''^*$, 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 $\boldsymbol{r}^*$ that satisfies Eqs (92) and (93) for any $1 \leq j \neq k \leq N$. Then, there exits a non-decreasing non-negative function $r^*(\cdot) : \mathbb{R} \to \mathbb{R}$, such that

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

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

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

This readily implies that $\theta^*$ 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). □

## 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 $1 \leq x$, 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

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

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 $\theta$ as

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

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 $\delta$ 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 $\theta^*$ is the minimizer of $J(\theta, \mathbf{0}, \mathbf{0})$ 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 $\ell$ 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 $\theta$, and for all $\theta$ 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 $\lambda(\mathbf{0}, \mathbf{0})$ 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 $\theta$ and $\lambda(\mathbf{0}, \mathbf{0}) > 0$;
(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 $\theta^*$ be the minimizer of $J(\theta, \mathbf{0}, \mathbf{0})$ 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 $\|\boldsymbol{u}'\|_2 \leq \epsilon$,*

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

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 $\theta^*$ 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 $J(\theta, \mathbf{0}, \mathbf{0})$ at $\theta = \theta^*$.

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

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

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

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

[12] This makes $\ell(g_\theta(x), y)$ convex in $\theta$ 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 $\ell$. In fact, it must be valid given (a) since $\ell$ is continuously differentiable w.r.t. $t$.

[14] This is for the Lipschitz continuity of $J(\theta, \boldsymbol{u}, \boldsymbol{u}') - J(\theta, \mathbf{0}, \mathbf{0})$. 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 $J(\theta, \mathbf{0}, \mathbf{0})$ locally strongly convex in $\theta$ around $\theta^*$. 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),

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

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 $\theta$, and thus

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

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,

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

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

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

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

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

$$\begin{aligned}\lambda(\boldsymbol{u}, \boldsymbol{u}') &\leq \lambda(\mathbf{0}, \mathbf{0}) + C_\lambda(\|\boldsymbol{u}\|_2 + \|\boldsymbol{u}'\|_2) \\ &\leq \lambda(\mathbf{0}, \mathbf{0}) + 2C_\lambda \epsilon \\ &< \infty.\end{aligned}$$

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

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

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

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

$$\begin{aligned}D_1(\theta) &= F(\theta, \boldsymbol{u}, \boldsymbol{u}') - F(\theta, \boldsymbol{u}, \mathbf{0}), \\ D_2(\theta) &= F(\theta, \boldsymbol{u}, \mathbf{0}) - F(\theta, \mathbf{0}, \mathbf{0}),\end{aligned}$$

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

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

Given the assumption (d), the third function $(\lambda(\boldsymbol{u}, \boldsymbol{u}') - \lambda(\mathbf{0}, \mathbf{0}))\Omega(\theta)$ 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,

$$\begin{aligned}\|\nabla_\theta D_1(\theta)\|_2 &= \left\| \sum_{s=1}^{S} \frac{\partial D_1(\theta)}{\partial l_s} \nabla_\theta l_s \right\|_2 \\ &\leq \left| \sum_{s=1}^{S} \frac{\partial D_1(\theta)}{\partial l_s} \right| \cdot \max_s \|\nabla_\theta l_s\|_2 \\ &= \mathcal{O}\left( \left| \sum_{s=1}^{S} \frac{\partial D_1(\theta)}{\partial l_s} \right| \right),\end{aligned}$$

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

By definition,

$$\begin{aligned} D_1(\theta) &= \sup_{\boldsymbol{w}\in\Phi(\boldsymbol{u})} R_\theta(\boldsymbol{u}') - \sup_{\boldsymbol{w}\in\Phi(\boldsymbol{u})} R_\theta(\mathbf{0}) \\ &= \sup_{\boldsymbol{w}\in\Phi(\boldsymbol{u})} \sum_{s=1}^{S} w_s(l_s + u_s') - \sup_{\boldsymbol{w}\in\Phi(\boldsymbol{u})} \sum_{s=1}^{S} w_s l_s. \end{aligned}$$

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

$$\begin{aligned} \left|\sum_{s=1}^{S} \frac{\partial D_1(\theta)}{\partial l_s}\right| &\le \sum_{s=1}^{S} |v_s^* - w_s^*| \\ &\le \sqrt{S}\|\boldsymbol{v}^* - \boldsymbol{w}^*\|_2, \end{aligned}$$

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

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

be the Lagrangian function, where $\alpha \ge 0$ and $\alpha'$ 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 $\boldsymbol{d} > \mathbf{0}$.*

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

$$\begin{aligned} f\left(\frac{1}{1+\sum_{s=1}^S u_s}\right) &\le |f'(1)|\frac{\sqrt{S}\epsilon}{1-\sqrt{S}\epsilon} + \mathcal{O}(\epsilon^2) \\ &\le 2\sqrt{S}|f'(1)|\epsilon. \end{aligned}$$

As a result,

$$\begin{aligned} \lim_{n\to\infty}\sum_{s=1}^S (p_s+u_s)f(w_{n,s}) &= \sum_{s=1}^S (p_s+u_s) f\left(\frac{1}{1+\sum_{s=1}^S u_s}\right) \\ &= \left(1+\sum_{s=1}^S u_s\right) f\left(\frac{1}{1+\sum_{s=1}^S u_s}\right) \\ &\le (1+\sqrt{S}\epsilon)\cdot 2\sqrt{S}|f'(1)|\epsilon \\ &\le 4\sqrt{S}|f'(1)|\epsilon. \end{aligned}$$

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

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

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

*Proof.* We are going to prove $\sup_{\boldsymbol{w}\in\Phi(\mathbf{0})}\inf_{\boldsymbol{v}\in\Phi(\boldsymbol{u})}\|\boldsymbol{v}-\boldsymbol{w}\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2)$, and the other direction can be proven similarly. Pick an arbitrary $\boldsymbol{w}_0 \in \Phi(\mathbf{0})$. Let $\beta = \delta/(\delta + \|f(\boldsymbol{w}_0)\|_2\|\boldsymbol{u}\|_2)$ and consider $\boldsymbol{v}_1 = \beta\boldsymbol{w}_0 + (1-\beta)\mathbf{1}$,

$$\begin{aligned} \|\boldsymbol{v}_1 - \boldsymbol{w}_0\|_2 &= \|(\beta-1)\boldsymbol{w}_0 + (1-\beta)\mathbf{1}\|_2 \\ &= (1-\beta)\|\boldsymbol{w}_0 - \mathbf{1}\|_2 \\ &\le \|f(\boldsymbol{w}_0)\|_2\|\boldsymbol{u}\|_2\|\boldsymbol{w}_0-\mathbf{1}\|_2/\delta \\ &= \mathcal{O}(\|\boldsymbol{u}\|_2). \end{aligned}$$

Moreover,

$$\begin{aligned} \sum_{s=1}^S (p_s+u_s)f(v_{1,s}) &= \sum_{s=1}^S (p_s+u_s)f(\beta w_{0,s} + (1-\beta)) \\ &\le \sum_{s=1}^S (p_s+u_s)(\beta f(w_{0,s}) + (1-\beta)f(1)) \\ &= \beta\left(\sum_{s=1}^S p_s f(w_{0,s}) + \sum_{s=1}^S u_s f(w_{0,s})\right) \\ &\le \beta(\delta + \|f(\boldsymbol{w}_0)\|_2\|\boldsymbol{u}\|_2) \\ &= \delta, \end{aligned}$$

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 $V_1 = \{\boldsymbol{w}\in\mathbb{R}^S \mid \sum_s (p_s+u_s)f(w_s) \le \delta, w_s \ge 0\}$.

However, $\boldsymbol{v}_1$ does not belong to the set $V_2 = \{\boldsymbol{w} \in \mathbb{R}^S \mid \sum_s (p_s + u_s) w_s = 1, w_s \geq 0\}$. Since $V_2$ is a hyperplane, we can easily project $\boldsymbol{v}_1$ onto $V_2$ to obtain $\boldsymbol{v}_2$, and

$$\begin{aligned}
\|\boldsymbol{v}_2 - \boldsymbol{v}_1\|_2 &= \frac{1}{\|\boldsymbol{p} + \boldsymbol{u}\|_2} \left| \sum_{s=1}^{S} (p_s + u_s) v_{1,s} - 1 \right| \\
&\leq \frac{2}{\|\boldsymbol{p}\|_2} \left| \sum_{s=1}^{S} (p_s + u_s)(\beta w_{0,s} + 1 - \beta) - 1 \right| \\
&\leq \sqrt{4S} \left| \beta \sum_{s=1}^{S} p_s w_{0,s} + \beta \sum_{s=1}^{S} u_s w_{0,s} + (1-\beta) \sum_{s=1}^{S} p_s + (1-\beta) \sum_{s=1}^{S} u_s - 1 \right| \\
&= \sqrt{4S} \left| \beta + \beta \sum_{s=1}^{S} u_s w_{0,s} + (1-\beta) + (1-\beta) \sum_{s=1}^{S} u_s - 1 \right| \\
&\leq \sqrt{4S} \beta \left| \sum_{s=1}^{S} u_s w_{0,s} \right| + \mathcal{O}(\|\boldsymbol{u}\|_2^2) \\
&= \mathcal{O}(\|\boldsymbol{u}\|_2).
\end{aligned}$$

After this projection, $\boldsymbol{v}_2 \notin V_1$ again.

Let $\boldsymbol{v}_3$ be the projection of $\boldsymbol{v}_2$ onto $\Phi(\boldsymbol{u}) = V_1 \cap V_2$, $T_w V_1(\boldsymbol{v}_3)$ be the tangent hyperplane to $V_1$ at $\boldsymbol{v}_3$ of $S-1$ dimensions, and $T_w(V_1 \cap V_2)(\boldsymbol{v}_3)$ be that to $V_1 \cap V_2$ at $\boldsymbol{v}_3$ of $S-2$ dimensions. As a consequence, $\boldsymbol{v}_3 - \boldsymbol{v}_2 \in V_2$ is one of normal vectors to $T_w(V_1 \cap V_2)(\boldsymbol{v}_3)$ at $\boldsymbol{v}_3$, which is also the projection of the normal vector to $T_w V_1(\boldsymbol{v}_3)$ at $\boldsymbol{v}_3$ onto $V_2$. This means the normal vector to $T_w V_1(\boldsymbol{v}_3)$ at $\boldsymbol{v}_3$ belongs to the 2-dimensional plane determined by $\boldsymbol{v}_3 - \boldsymbol{v}_2$ and $\boldsymbol{v}_1 - \boldsymbol{v}_2$, since the latter is a normal vector to $V_2$ at $\boldsymbol{v}_2$.

Consider the triangle $(\boldsymbol{v}_1 - \boldsymbol{v}_2, \boldsymbol{v}_3 - \boldsymbol{v}_2, \boldsymbol{v}_1 - \boldsymbol{v}_3)$. This is a right-angled triangle since $\boldsymbol{v}_1 - \boldsymbol{v}_2 \perp V_2$ and $\boldsymbol{v}_3 - \boldsymbol{v}_2 \in V_2$, so that

$$\|\boldsymbol{v}_1 - \boldsymbol{v}_3\|_2 = \frac{\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2}{\sin(\boldsymbol{v}_2 - \boldsymbol{v}_3, \boldsymbol{v}_1 - \boldsymbol{v}_3)}.$$

Subsequently, let $\boldsymbol{v}_4$ be the intersection of $\boldsymbol{v}_1 - \boldsymbol{v}_2$ and $T_w V_1(\boldsymbol{v}_3)$, due to the convexity of $V_1$,

$$\begin{aligned}
\sin^2(\boldsymbol{v}_2 - \boldsymbol{v}_3, \boldsymbol{v}_1 - \boldsymbol{v}_3) &\geq \sin^2(\boldsymbol{v}_2 - \boldsymbol{v}_3, \boldsymbol{v}_4 - \boldsymbol{v}_3) \\
&= 1 - \cos^2(\boldsymbol{v}_2 - \boldsymbol{v}_3, \boldsymbol{v}_4 - \boldsymbol{v}_3) \\
&= 1 - \cos^2(\boldsymbol{p} + \boldsymbol{u}, (\boldsymbol{p} + \boldsymbol{u}) \circ f'(\boldsymbol{v}_3)),
\end{aligned}$$

where $\boldsymbol{p} + \boldsymbol{u}$ is a normal vector to $V_2$ containing $\boldsymbol{v}_2 - \boldsymbol{v}_3$, $(\boldsymbol{p} + \boldsymbol{u}) \circ f'(\boldsymbol{v}_3)$ is a normal vector to $T_w V_1(\boldsymbol{v}_3)$ containing $\boldsymbol{v}_4 - \boldsymbol{v}_3$, and both of them belong to the 2-dimensional plane determined by $\boldsymbol{v}_3 - \boldsymbol{v}_2$ and $\boldsymbol{v}_1 - \boldsymbol{v}_2$ as we have proven above. By definition, $\boldsymbol{v}_3$ 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,

$$\begin{aligned}
1 - \cos^2(\boldsymbol{p} + \boldsymbol{u}, (\boldsymbol{p} + \boldsymbol{u}) \circ f'(\boldsymbol{v}_3)) &\geq 1 - (1 - C_{\cos})^2 \\
&= C_{\cos}(2 - C_{\cos}),
\end{aligned}$$

which implies

$$\begin{aligned}
\|\boldsymbol{v}_1 - \boldsymbol{v}_3\|_2 &\leq \frac{\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2}{\sqrt{C_{\cos}(2 - C_{\cos})}} \\
&= \mathcal{O}(\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2) \\
&= \mathcal{O}(\|\boldsymbol{u}\|_2).
\end{aligned}$$

Combining $\|\boldsymbol{v}_1 - \boldsymbol{w}_0\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2)$ and $\|\boldsymbol{v}_1 - \boldsymbol{v}_3\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2)$ gives us $\|\boldsymbol{v}_3 - \boldsymbol{w}_0\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2)$, and thus $\inf_{\boldsymbol{v} \in \Phi(\boldsymbol{u})} \|\boldsymbol{v} - \boldsymbol{w}_0\|_2 \leq \|\boldsymbol{v}_3 - \boldsymbol{w}_0\|_2 = \mathcal{O}(\|\boldsymbol{u}\|_2)$. Since $\boldsymbol{w}_0$ is arbitrarily picked from $\Phi(\boldsymbol{0})$, the proof is completed. □

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 $\boldsymbol{w}^* = \arg\sup_{\boldsymbol{w}\in\Phi(\mathbf{0})} \sum_s w_s l_s$ 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 $d_{\mathcal{H}}(\Phi(\boldsymbol{u}), \Phi(\mathbf{0})) = \mathcal{O}(\|\boldsymbol{u}\|_2)$, 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). □

Let us summarize what we have obtained so far:

- a second-order growth condition of $J(\theta, \mathbf{0}, \mathbf{0})$ 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 $\theta$ 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

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

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

## 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 $\delta$ 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)** |