Supplement for Class-Weighted Classification: Trade-offs and Robust Approaches

A. Organization

Our appendices contain proofs, all of which are omitted from the main text, and additional details on the weighting approach to imbalanced classification. In Appendix B, we prove our results for plug-in classification. Additionally, we show that a threshold-shifted version of Tsybakov's noise condition implies precise rates for the convergence of expected excess risk. Finally, we briefly discuss the universality of weighting, i.e., the fact that choosing the correct weighting is often the means to optimizing other classification metrics, for a class of classification metrics.

In Appendix C, we show a result analogous to Proposition 3 for empirical risk minimization. However, the result is less illuminating, since it depends on the optimal classifiers f_q^* $f_{q'}^*$ for weights q and q' within the class $\mathcal{F}$, which is difficult to analyze more precisely in any generality.

In Appendix D, we prove our results for robust weighting. This includes both the convergence and duality results. In Appendix E, we prove the analog of Theorem 1 for the conditional sampling model. The only difference to observe is that the bounded differences inequality is used with respect to a different number of variables, which leads to a slightly stronger bound.

In Appendix F, we discuss gradient descent-ascent, which is a standard algorithm for solving robust optimization problems. This may be used in cases where the uncertainty set Q does not lead to LCVaR or LHCVaR. In Appendix G and Appendix H, we provide technical and standard lemmas respectively.

Finally, we include additional experiment details, and an algorithm for analytically deriving dual variables in the empirical LCVaR and LHCVaR formulations in Appendix I.

B. Plug-in Classification Details

In this appendix, we provide additional details surrounding plug-in classification. We first start with the proofs of results from the main text, and then we provide more concrete results based on an additional assumption of that gives us faster rates of convergence. Finally, we provide details on the universality of weighting.

For simplicity, we assume that our density estimator $\widehat{\eta}$ is a local polynomial estimator (Stone, 1982), but the properties that the estimator must have for the following proofs to succeed can also be satisfied by other nonparametric estimators such as kernelized regression (Krzyzak & Pawlak, 1987), and nearest-neighbors regression (Györfi, 1981).

B.1. Proofs

Proof of Lemma 1. By the definition of the q-weighted risk and the tower property, we have

$$\begin{aligned} R_{01,q}(f) &= \mathbb{E}[q_Y R_Y(f)] \\ &= \mathbb{E}\left[q_0(1-\eta(X))\mathbb{E}[\mathbf{1}\left\{f(X)=1\right\}|Y=0] + q_1\eta(X)\mathbb{E}[\mathbf{1}\left\{f(X)=0\right\}|Y=1]\right] \\ &= \mathbb{E}\left[q_0(1-\eta(X))\mathbf{1}\left\{f(X)=1\right\} + q_1\eta(X)\mathbf{1}\left\{f(X)=0\right\}\right]. \end{aligned}$$

By inspection, we observe that the f^* minimizing the q-risk satisfies

$$f^*(x) = \begin{cases} 1 & q_0(1-\eta(x)) < q_1\eta(x) \\ 0 & q_0(1-\eta(x)) > q_1\eta(x). \end{cases}$$

When $q_0(1-\eta(x)) = q_1\eta(x)$, we note that the decision may be arbitrary because it does not affect the risk. So, by simple algebraic manipulation, we have

$$f^*(x) = \mathbf{1}\left\{\eta(x) \geq \frac{q_0}{q_0+q_1}\right\},$$

which completes the proof. □

Now, we turn to Proposition 1, Proposition 2, and Proposition 3. Our proofs rely on the following lemma of (Yang, 1999). First, we introduce a few additional definitions. Denote the ε-entropy of Σ with respect to the L_p norm for $1 \leq p \leq \infty$ by $\mathcal{H}(\varepsilon, \Sigma, L_p)$. We define the norm

$$\|\widehat{\eta} - \eta\|_{L_1(P_X)} = \int |\eta(x) - \widehat{\eta}(x)|\, dP_X$$

.

Lemma 1 (Theorem 1 of Yang 1999)**.** *Let η be an element of Σ where Σ is a class of functions from $\mathbb{R}^d$ to $[0,1]$. Suppose the ε-entropy satisfies*

$$\mathcal{H}(\varepsilon, \Sigma, L_p) \leq C\varepsilon^{-\rho},$$

where $C > 0, \rho > 0$. Then the minimax upper bound on the mean convergence rate of any regression estimator $\widehat{\eta}$ is

$$\min_{\widehat{\eta}} \max_{\eta\in\Sigma} \mathbb{E}\left[\|\eta - \widehat{\eta}\|_{L_1(P_X)}\right] \leq O\left(n^{-\frac{1}{2+\rho}}\right),$$

where the expectation is taken over the samples for estimating $\widehat{\eta}$.

The upper bound converges at a rate of $O\left(n^{-1/(2+\rho)}\right)$ where ρ is a smoothness parameter for η, with standard assumptions on the function class of η. For the class of β-Hölder functions, $\rho = \beta/d$, which is our setting of interest.

Proof of Proposition 1. We start by bounding the excess q-risk for a classifier f by

$$\begin{aligned}
\mathcal{E}_q(f) &= R_q(f) - R_q(f_q^*) \\
&= (q_0+q_1)\int \left|\eta(x) - \frac{q_0}{q_0+q_1}\right| \mathbf{1}\left\{f(x) \neq f_q^*(x)\right\} dP_X \\
&\leq (q_0+q_1)\int |\eta(x) - \widehat{\eta}(x)|\, dP_X,
\end{aligned}$$

where the upper bound follows when $|\eta(x) - q_0/(q_0+q_1)| \leq |\eta(x) - \widehat{\eta}(x)|$ when $f(x) \neq f_q^*(x)$. Finally, applying Lemma 1 for β-Hölder functions as noted above completes the proof. □

Proof of Proposition 2. The proposition follows from basic algebraic manipulations and one common observation in nonparametric classification. We have

$$\begin{aligned}
\text{(IE)} &= \mathbb{E}\left[R_{q'}(f_{q'}^*(X)) - R_{q'}(f_q^*(X))\right] \\
&= \int |q_0'(1-\eta(x)) + q_1'\eta(x)|\, dP_X \mathbf{1}\left\{f_{q'}^*(x) \neq f_q^*(x)\right\} \\
&= (q_0'+q_1')\int |\eta(x) - t_{q'}|\, \mathbf{1}\left\{f_{q'}^*(x) \neq f_q^*(x)\right\} dP_X \\
&\leq (q_0'+q_1')\,|t_q - t_{q'}|\,\mathbb{P}\left(f_{q'}^*(X) \neq f_q^*(X)\right),
\end{aligned}$$

where in the inequality we use the fact that if $f_{q'}^*(X) \neq f_q^*(X)$ then $\eta(X)$ must be in $[\underline{t}_{q,q'}, \overline{t}_{q,q'}]$. Thus, we have $|\eta(x) - t_{q'}| \leq |\overline{t}_{q,q'} - \underline{t}_{q,q'}| = |t_q - t_{q'}|$. □

Proof of Proposition 3. Recall that the expected estimation error is

$$\text{(EE)} = \mathbb{E}\left[R_{q'}(\widehat{f}_q) - R_{q'}(f_q^*)\right]$$

We can upper bound the term inside the expectation by

$$\begin{aligned}
R'_q(\widehat{f}_q) - R_{q'}(f_q^*) &= \int q'_0(1-\eta(x))\mathbf{1}\{\widehat{f}_q(x)=1\} + q'_1\eta(x)\mathbf{1}\{\widehat{f}_q(x)=0\}dP_X \\
&\quad - \int q'_0(1-\eta(x))\mathbf{1}\{f_q^*(x)=1\} + q'_1\eta(x)\mathbf{1}\{f_q^*(x)=0\}dP_X \\
&= \int (q'_0(1-\eta(x)) - q'_1\eta(x))\mathbf{1}\{\widehat{f}_q(x)=1, f_q^*(x)=0\}dP_X \\
&\quad + (q'_1\eta(x) - q'_0(1-\eta(x)))\int \mathbf{1}\{\widehat{f}_q(x)=0, f_q^*(x)=1\}dP_X \\
&= (q'_0+q'_1)\int \left|\eta(x) - \frac{q'_0}{q'_0+q'_1}\right| \mathbf{1}\{\widehat{f}_q(x)\neq f_q^*(x)\}dP_X \\
&\leq (q'_0+q'_1)\int \left(|\eta(x)-t_q| + |t_{q'}-t_q|\right)\mathbf{1}\{\widehat{f}_q(x)\neq f_q^*(x)\}dP_X,
\end{aligned}$$

where we use the triangle inequality in the final line. Next, using the fact that $|\eta(x)-t_q| \leq |\eta(x)-\widehat{\eta}(x)|$ when $f(x)\neq f_q^*(x)$, we have

$$\begin{aligned}
R'_q(\widehat{f}_q) - R_{q'}(f_q^*) \leq\ & (q'_0+q'_1)\left(\int |\eta(x)-t_q|\,\mathbf{1}\left\{\widehat{f}_q(x)\neq f_q^*(x)\right\}dP_X \right. \\
& \left. + |t_{q'}-t_q|\,\mathbb{P}\left(\widehat{f}_q(x)\neq f_q^*(x)\right)\right) \\
\leq\ & (q'_0+q'_1)\left(\int |\eta(x)-\widehat{\eta}(x)|\,dP_X + |t_{q'}-t_q|\,\mathbb{P}\left(\widehat{f}_q(x)\neq f_q^*(x)\right)\right)
\end{aligned}$$

Thus, we obtain the upper bound

$$\text{(EE)} \leq (q'_0+q'_1)\left(\mathbb{E}\left[\int |\eta(x)-\widehat{\eta}(x)|\,dP_X\right] + |t_{q'}-t_q|\,\mathbb{E}\left[\mathbb{P}\left(\widehat{f}_q(x)\neq f_q^*(x)\right)\right]\right)$$

Therefore we have completed the proof. Applying Lemma 1 to the first term also proves Corollary 1. □

B.2. Shifted Margin Assumption

An important tool in nonparametric classification is the Tsybakov margin condition.

Definition 1. *A distribution $P_{X,Y}$ satisfies the (α, C)-margin condition if for all $t>0$, we have*

$$\mathbb{P}\left(0 \leq \left|\eta(X) - \frac{1}{2}\right| \leq t\right) \leq Ct^\alpha.$$

Subsequent works (Audibert & Tsybakov, 2007; Chaudhuri & Dasgupta, 2014) leverage this assumption to provide fast, explicit rates of convergence for expected risk. The margin condition is naturally suited to standard plug-in classification because the decision threshold is $1/2$; for weighted plug-in classification, we need a shifted margin condition.

Definition 2. *A distribution $P_{X,Y}$ satisfies the (q, α, C)-margin condition if for all $t>0$, we have*

$$\mathbb{P}\left(0 \leq |\eta(x) - t_q| \leq t\right) \leq Ct^\alpha.$$

Using the shifted margin condition, we can obtain better results than we presented in the main paper. However, the shifted margin condition may be be less interpretable than the original margin condition. Intuitively, the original margin condition says that there is very little probability mass where distinguishing between $Y = 0$ and $Y = 1$ is difficult, i.e., near $\eta(X) = 1/2$. For other t_q, the decision may not be difficult in that t_q may be far from $1/2$, but we would still require little mass near this point.

Proposition 1. *Suppose the distribution $P_{X,Y}$ satisfies the (q, α, C)-margin condition and X has a density that is lower bounded by some constant $\mu_{\min}$ on its support. Additionally, suppose that η is β-Hölder. Then, the excess expected q'-risk of $\widehat{f}_q$ satisfies the bound*

$$\mathbb{E}\mathcal{E}_{q'}(\widehat{f}_q) \leq (q'_0 + q'_1)\left(O\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} + |t_{q'} - t_q|\, O\left(\frac{\log n}{n}\right)^{\frac{\alpha\beta}{2\beta+d}}\right) + \text{(IE)}$$

Before proving this proposition, we prove a helpful lemma that leverages the shifted margin condition, similar to one from (Audibert & Tsybakov, 2007).

Lemma 2. *For a fixed density estimate $\widehat{\eta}$, if $P_{X,Y}$ satisfies the (q, α, C)-margin condition, then following upper bound is always true:*

$$\mathbb{P}\left(\widehat{f}_q(x) \neq f_q^*(x), \eta(x) \neq t_q\right) \leq C \left\|\eta - \widehat{\eta}\right\|_\infty^\alpha .$$

Proof. We use a simple upper bound on the error probability event and apply the margin condition to obtain

$$\begin{aligned}
\mathbb{P}\left(\widehat{f}_q(x) \neq f_q^*(x), \eta(x) \neq t_q\right) &\leq \mathbb{P}\left(0 \leq |\eta(x) - t_q| \leq |\eta(x) - \widehat{\eta}(x)|\right) \\
&\leq \mathbb{P}\left(0 \leq |\eta(x) - t_q| \leq \left\|\eta - \widehat{\eta}\right\|_\infty\right) \\
&\leq C_0 \left\|\eta - \widehat{\eta}\right\|_\infty^\alpha .
\end{aligned}$$

This completes the proof. □

Since, by Lemma 2, we have proved an upper bound in terms of $\|\eta - \widehat{\eta}\|_\infty^\alpha$, we now cite an upper bound on that quantity that is a property of regression estimator.

Lemma 3 (Theorem 1 of Stone 1982)**.** *Let $\widehat{\eta}$ be a local polynomial regression estimator, and suppose X has a density that is lower bounded by some constant $\mu_{\min} > 0$ on its support. Then, we have the following upper bound:*

$$\mathbb{E}\left[\left\|\eta - \widehat{\eta}\right\|_\infty^\alpha\right] \leq C\left(\frac{\log n}{n}\right)^{\frac{\alpha\beta}{2\beta+d}} . \tag{1}$$

The above bound is the optimal rate of uniform convergence for nonparametric estimators under the regularity conditions shown here, and local polynomial regression achieves this optimal rate (Stone, 1982).

Proof of Proposition 1. It suffices to prove an upper bound on the estimation error. We have

$$\text{(EE)} \leq (q'_0 + q'_1)\left(\mathbb{E}\left[\int |\eta(x) - \widehat{\eta}(x)|\, dP_X\right] + |t_{q'} - t_q|\, \mathbb{E}\left[\mathbb{P}(\widehat{f}_q(x) \neq f_q^*(x))\right]\right)$$

by the final equation of the proof of Proposition 3. Next, we use the fact that for all x in $\mathcal{X}$ we have $\eta(x) - \widehat{\eta}(x) \leq \left\|\eta - \widehat{\eta}\right\|_\infty$ and Lemma 2 to obtain

$$\text{(EE)} \leq (q'_0 + q'_1)\left(\mathbb{E}\left[\left\|\eta - \widehat{\eta}\right\|_\infty\right] + |t_{q'} - t_q|\, C_0 \mathbb{E}\left[\left\|\eta - \widehat{\eta}\right\|_\infty^\alpha\right]\right)$$

Finally, we apply Lemma 3 to obtain

$$\text{(EE)} \leq (q'_0 + q'_1)\left(C\left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} + |t_{q'} - t_q|\, C_0 C\left(\frac{\log n}{n}\right)^{\frac{\alpha\beta}{2\beta+d}}\right),$$

which completes the proof. □

B.3. Universality of Weighting

Since we may be interested in performance in error metrics other than risk, we discuss other classification metrics here. In particular, we simply show that weighting is "universal" in that it can be used to optimize these other classification metrics. The reason for this is that, in plug-in classification, optimizing many classification metrics is equivalent to altering the threshold for the classification, and this has been observed to lead to the optimal decision rule in many cases (Lewis, 1995; Menon et al., 2013; Narasimhan et al., 2014; Koyejo et al., 2014). We examine the specific case of metrics considered in (Koyejo et al., 2014).

Definition 3. *Let f be a classifier over $\mathcal{X}$. Define the true positive, false negative, false positive, and true negative proportions to be*

$$\begin{aligned} \text{TP} &= \mathbb{P}(Y = 1, f(X) = 1) & \text{FP} &= \mathbb{P}(Y = 0, f(X) = 1) \\ \text{FN} &= \mathbb{P}(Y = 1, f(X) = 0) & \text{TN} &= \mathbb{P}(Y = 0, f(X) = 0). \end{aligned}$$

A linear-fractional metric is defined as

$$\mathcal{L}(f, P_X, \eta) = \frac{a_0 + a_{11}\text{TP} + a_{10}\text{FP} + a_{01}\text{FN} + a_{00}\text{TN}}{b_0 + b_{11}\text{TP} + b_{10}\text{FP} + b_{01}\text{FN} + b_{00}\text{TN}}$$

for constants $a_0, a_{11}, a_{10}, a_{01}, a_{00}, b_0, b_{11}, b_{10}, b_{01}, b_{00}$.

(Koyejo et al., 2014) showed that the optimal classifier for any linear-fractional metric is simply a threshold classifier. Specifically, the following theorem is true.

Theorem 1 (Koyejo et al. 2014)**.** *Let $\mathcal{L}$ be a linear-fractional metric, and let P_X be absolutely continuous with respect to the dominating measure ν on $\mathcal{X}$. Define*

$$\mathcal{L}^* = \max_f \mathcal{L}(f, P_X, \eta)$$

and

$$\delta^* = \frac{(b_{10} - b_{00})\mathcal{L}^* - a_{10} + a_{00}}{a_{11} - a_{10} - a_{01} + a_{00} - (b_{11} - b_{10} - b_{01} + b_{00})\mathcal{L}^*}.$$

Then, the optimal classifier for $\mathcal{L}$ is $f^*_{\mathcal{L}}(x) = \mathbf{1}\{\eta(x) > \delta^*\}$ if

$$a_{11} - a_{10} - a_{01} + a_{00} - (b_{11} - b_{10} - b_{01} + b_{00})\mathcal{L}^* > 0$$

and $f^*_{\mathcal{L}}(x) = \mathbf{1}\{\eta(x) < \delta^*\}$ otherwise.

Corollary 1. *We note by Proposition 1 that for an metric $\mathcal{L}$ where*

$$a_{11} - a_{10} - a_{01} + a_{00} - (b_{11} - b_{10} - b_{01} + b_{00})\mathcal{L}^* > 0,$$

if we set define q to be

$$\begin{aligned} q_0 &= (b_{10} - b_{00})\mathcal{L}^* - a_{10} + a_{00} \\ q_1 &= (b_{01} - b_{11})\mathcal{L}^* - a_{01} + a_{11}, \end{aligned}$$

then $f^*_q = f^*_{\mathcal{L}}$.

Performance metrics that are used in evaluating classifiers such as F1 and arithmetic mean satisfy the the conditions of Corollary 1. Thus, we can reformulate optimization of a classifier in these error metrics as a specific weighting the risk.

C. The Fundamental Trade-off in Empirical Risk Minimization

Part of our motivation for the robust weighted problem is the fundamental trade-off under different weightings q and q'. We demonstrated this for plug-in classification in the main text because it elucidates the nature of the problem naturally via

thresholds, but we should also convince ourselves that this is not simply a quirk of plug-in classification. To this end, we provide a brief analysis for empirical risk minimization.

Let $\widehat{f}_q$ and f_q^* denote the empirical risk minimizer and risk minimizer within $\mathcal{F}$. Define the excess risk to be the difference between $R(\widehat{f}_q)$ and $R_q(f_q^*)$. Suppose that we have a uniform convergence guarantee

$$R_q(f) - \widehat{R}_q(f) \leq O\left(n^{-\frac{1}{2}}\right)$$

for all f in $\mathcal{F}$. Then, a standard chaining argument reveals that the excess risk decay rate satisfies

$$\begin{aligned}
\mathcal{E}_q(\widehat{f}_q) &= R_q(\widehat{f}_q) - R(f_q^*) \\
&= R_q(\widehat{f}_q) - \widehat{R}_q(\widehat{f}_q) + \widehat{R}_q(\widehat{f}_q) - \widehat{R}_q(f_q^*) + \widehat{R}_q(f_q^*) - R(f_q^*) \\
&\leq O\left(n^{-\frac{1}{2}}\right) + 0 + O\left(n^{-\frac{1}{2}}\right) \\
&= O\left(n^{-\frac{1}{2}}\right),
\end{aligned}$$

where in the inequality we used our uniform convergence guarantee twice and the fact that $\widehat{f}_q$ is the empirical q-risk minimizer. This mirrors the case of q-weighted plug-in estimation in that the excess q-risk still converges to 0 at the standard rate.

On the other hand, we obtain a constant term when performing a similar analysis for $\mathcal{E}_{q'}(\widehat{f}_q)$. Specifically, we get

$$\begin{aligned}
\mathcal{E}_{q'}(\widehat{f}_q) &= R_{q'}(\widehat{f}_q) - R_{q'}(f_q^*) \\
&= R_q(\widehat{f}_q) - R_q(f_q^*) + R_{q'}(\widehat{f}_q) - R_q(\widehat{f}_q) + R_q(f_q^*) - R_{q'}(f_{q'}^*) \\
&\leq O\left(n^{-\frac{1}{2}}\right) + R_{q'}(\widehat{f}_q) - R_q(\widehat{f}_q) + R_q(f_q^*) - R_{q'}(f_{q'}^*).
\end{aligned}$$

Now, using the prior convergence result for the empirical risk minimizers, we obtain

$$\begin{aligned}
\mathcal{E}_{q'}(\widehat{f}_q) &\leq R_{q'}(\widehat{f}_q) - R_q(\widehat{f}_q) + R_q(f_q^*) - R_{q'}(f_{q'}^*) + O\left(n^{-\frac{1}{2}}\right) \\
&\leq R_{q'}(f_q^*) - R_q(f_q^*) + R_q(f_q^*) - R_{q'}(f_{q'}^*) + O\left(n^{-\frac{1}{2}}\right) \\
&= \underbrace{R_{q'}(f_q^*) - R_{q'}(f_{q'}^*)}_{A} + O\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Since $f_{q'}^*$ minimizes $R_{q'}$ and f_q^* minimizes R_q, we see that $A \geq 0$. Thus, even though there is not a clear threshold interpretation, we do see that there is irreducible error that arises in the empirical risk minimization setting as well.

D. Robust Weighting Proofs

In this section, we prove our results for robust weighting. We start with our generalization and excess risk bounds.

Proof of Theorem 1. Define the risk $R_{i,\mathbf{1}}$ as

$$\widehat{R}_{i,\mathbf{1}}(f) = \widehat{p}_i \widehat{R}_i(f) = \frac{1}{n} \sum_{j=1}^{n} \ell_{\text{mar}}(f, z_j) \mathbf{1}\left\{y_j = i\right\}.$$

Let $R_{i,\mathbf{1}}(f)$ denote $\mathbb{E}\widehat{R}_{i,\mathbf{1}}(f)$. Note that we have

$$R_{i,\mathbf{1}}(f) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left[\ell_{\text{mar}}(f, z_j) \mathbf{1}\left\{y_j = i\right\}\right] = \frac{1}{n} \sum_{j=1}^{n} p_i \mathbb{E}\left[\ell_{\text{mar}}(f, z_j) | y_j = i\right] = p_i R_i(f).$$

By definition, we have

$$R_Q(f) = \sup_{q \in Q} \sum_{i=1}^k q_i p_i R_i(f) = \sup_{q \in Q} \sum_{i=1}^k q_i R_{i,\mathbf{1}}(f),$$

and so for our purposes, it suffices to analyze $\widehat{R}_{i,\mathbf{1}}$. Define the class

$$\mathcal{F}_{i,\mathbf{1}} = \{\ell_{\text{mar}}(f, \cdot)\mathbf{1}\{y_j = i\} : f \in \mathcal{F}\}.$$

By Lemma 8, we have with probability at least $1 - \delta/k$ that

$$R_{i,\mathbf{1}}(f) \leq \widehat{R}_{i,\mathbf{1}}(f) + 2\mathfrak{R}_n(\ell_{\text{mar}} \circ \mathcal{F}_{i,\mathbf{1}}) + \sqrt{\frac{\log \frac{k}{\delta}}{2n}}$$

for each f in $\mathcal{F}$. So, it suffices to analyze the Rademacher complexity term. Let σ_j be iid Rademacher random variables. We condition on the value of $y_1, \ldots, y_n$. Let $\mathcal{H}_Y$ be the sigma-field $\sigma(y_1, \ldots, y_n)$. Suppose without loss of generality that under the conditioning, we have $y_1 = \cdots = y_{N_i} = i$ and $y_j \neq i$ for all $j > N_i$. Then, we have

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{F}_{i,\mathbf{1}}) &= \frac{1}{n}\mathbb{E}\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \sigma_j \ell_{\text{mar}}(f, z_j)\mathbf{1}\{y_j = i\} \middle| \mathcal{H}_Y\right] \\
&= \frac{1}{n}\mathbb{E}\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{j=1}^{N_i} \sigma_j \ell_{\text{mar}}(f, z_j) \middle| \mathcal{H}_Y\right] \\
&= \mathbb{E}\left[\frac{N_i}{n}\widehat{\mathfrak{R}}_{N_i}(\ell_{\text{mar}} \circ \mathcal{F})\right].
\end{aligned}$$

By the proof of Lemma 10, we have

$$\widehat{\mathfrak{R}}_{N_i}(\ell_{\text{mar}} \circ \mathcal{F}) \leq 2k\widehat{\mathfrak{R}}_{N_i}(\Pi_1(\mathcal{F})).$$

Putting everything together completes the proof of the generalization bound; now we turn to the excess $(\mathcal{F}, q)$-risk bound.

Recall that $\widehat{f}_Q$ is the empirical Q-risk minimizer and f_Q^* is the population Q-risk minimizer. By Lemma 9, we have with probability at least $1 - \delta/k$ that

$$R_{i,\mathbf{1}}(\widehat{f}_Q) \leq \widehat{R}_{i,\mathbf{1}}(\widehat{f}_Q) + 4\mathfrak{R}_n(\ell_{\text{mar}} \circ \mathcal{F}_{i,\mathbf{1}}) + \sqrt{\frac{\log \frac{k}{\delta}}{2n}}.$$

Summing, we have

$$\begin{aligned}
R_q(\widehat{f}_Q) = \sum_{i=1}^k q_i p_i R_i(\widehat{f}_Q) &= \sum_{i=1}^k q_i (R_{i,\mathbf{1}}\widehat{f}_Q) \\
&\leq \sum_{i=1}^k q_i \left(\widehat{R}_{i,\mathbf{1}}(\widehat{f}_Q) + 4\mathfrak{R}_n(\ell_{\text{mar}} \circ \mathcal{F}_{i,\mathbf{1}}) + \sqrt{\frac{\log \frac{k}{\delta}}{2n}}\right) \\
&\leq \widehat{R}_q(\widehat{f}_Q) + \sum_{i=1}^k q_i p_i \left(\frac{4}{p_i}\mathfrak{R}_n(\ell_{\text{mar}} \circ \mathcal{F}_{i,\mathbf{1}}) + \sqrt{\frac{\log \frac{k}{\delta}}{2p_i^2 n}}\right).
\end{aligned}$$

Using the proof of Lemma 10 as before, we then obtain

$$R_q(\widehat{f}_Q) \leq \widehat{R}_q(\widehat{f}_Q) + \sum_{i=1}^k q_i p_i \left(8k\mathbb{E}\left[\frac{N_i}{p_i n}\widehat{\mathfrak{R}}_{N_i}(\Pi_1(\mathcal{F}))\right] + \sqrt{\frac{\log \frac{k}{\delta}}{2p_i^2 n}}\right).$$

Thus, by taking supremums, we observe that

$$R_Q(\widehat{f}_Q) \leq \widehat{R}_Q(\widehat{f}_Q) + \sup_{q \in Q} \sum_{i=1}^{k} q_i p_i \left(8k\mathbb{E}\left[\frac{N_i}{p_i n}\widehat{\mathfrak{R}}_{N_i}(\Pi_1(\mathcal{F}))\right] + \sqrt{\frac{\log\frac{k}{\delta}}{2p_i^2 n}} \right). \tag{2}$$

Similarly, by Lemma 9, we have

$$-R_{i,\mathbf{1}}(f_Q^*) \leq -\widehat{R}_{i,\mathbf{1}}(f_Q^*) + 4\mathfrak{R}_n(\ell_{\text{mar}} \circ \mathcal{F}_{i,\mathbf{1}}) + \sqrt{\frac{\log\frac{k}{\delta}}{2n}}.$$

Summing as before and using the proof of Lemma 10, we have

$$-R_q(f_Q^*) \leq -\widehat{R}_q(f_Q^*) + \sum_{i=1}^{k} q_i p_i \left(8k\mathbb{E}\left[\frac{N_i}{p_i n}\widehat{\mathfrak{R}}_{N_i}(\Pi_1(\mathcal{F}))\right] + \sqrt{\frac{\log\frac{k}{\delta}}{2p_i^2 n}} \right).$$

Taking the infimum and using Lemma 7, we have

$$-R_Q(f_Q^*) \leq -\widehat{R}_Q(f_Q^*) + \sup_{q \in Q} \sum_{i=1}^{k} q_i p_i \left(8k\mathbb{E}\left[\frac{N_i}{p_i n}\widehat{\mathfrak{R}}_{N_i}(\Pi_1(\mathcal{F}))\right] + \sqrt{\frac{\log\frac{k}{\delta}}{2p_i^2 n}} \right). \tag{3}$$

Summing equation (2) and equation (3) and noting that $\widehat{f}_Q$ minimizes the empirical robust risk, we have

$$\mathcal{E}_Q(\mathcal{F}) = R_Q(\widehat{f}_Q) - R_Q(f_Q^*) \leq 2\sup_{q \in Q} \sum_{i=1}^{k} q_i p_i \left(8k\mathbb{E}\left[\frac{N_i}{p_i n}\widehat{\mathfrak{R}}_{N_i}(\Pi_1(\mathcal{F}))\right] + \sqrt{\frac{\log\frac{k}{\delta}}{2p_i^2 n}} \right),$$

and this completes the proof. □

Proof of Corollary 2. The only thing we need to do here is calculate the Rademacher complexity term of Theorem 1. Using our assumption and Jensen's inequality, we have

$$\mathbb{E}\left[\frac{N_i}{n}\widehat{\mathfrak{R}}_{N_i}(\Pi_1(\mathcal{F}))\right] \leq \frac{C(\mathcal{F})}{n}\mathbb{E}\left[\sqrt{N_i}\right] \leq \frac{C(\mathcal{F})}{n}\mathbb{E}[N_i]^{1/2} = C(\mathcal{F})\sqrt{\frac{p_i}{n}}.$$

This completes the proof of the corollary. □

Next, we prove our duality results. We start with LCVaR.

Proof of Proposition 4. The Lagrangian of LCVaR is

$$L(q, \lambda) = \mathbb{E}[q_Y R_Y(f)] + \lambda(1 - \mathbb{E}[q_Y]) = \mathbb{E}\left[q_Y(R_Y(f) - \lambda)\right] + \lambda.$$

Our goal is to use the minimax theorem, which we state as Theorem 3, to switch the infimum over λ and the supremum over q. First, we do not need the minimax theorem to obtain

$$\inf_{\lambda \in \mathbb{R}} L(q, \lambda) \leq \inf_{\lambda \in \mathbb{R}} \sup_{q : q(\cdot) \in [0, \alpha^{-1}]} \mathbb{E}\left[q_Y(R_Y(f) - \lambda)\right] + \lambda = \inf_{\lambda \in \mathbb{R}} \left\{ \mathbb{E}\left[\alpha^{-1}\mathbb{E}(R_Y(f) - \lambda)_+ + \lambda\right] \right\}, \tag{4}$$

since the inequality follows the trivial direction of the minimax theorem and we can solve the inner maximization problem by setting

$$q_i = \begin{cases} 0 & R_i(f) - \lambda < 0 \\ \alpha^{-1} & R_i(f) - \lambda \geq 0. \end{cases}$$

Our present goal is to verify the conditions of the minimax theorem. First, we note that $\lambda \mapsto L(q, \lambda)$ is linear and therefore convex for any q, and similarly, $q \mapsto L(q, \lambda)$ is linear and therefore concave for any q. Additionally, the domain of q, in this case $[0, \alpha^{-1}]^k$, is compact and convex by definition; so we only need to prove that it suffices to consider λ on a compact, convex domain.

Denote the right hand side of equation (4) by $\inf_{\lambda \in \mathbb{R}} D(\lambda)$. Let $F_f(\lambda)$ denote the cumulative distribution function of R_Y at λ. By Lemma 5, the derivative of $D(\lambda)$ is given by

$$D'(\lambda) = 1 + \alpha^{-1}(F_f(\lambda) - 1),$$

when F_f is continuous at λ. If it is not, then the same result holds for the left and right limits. Thus by considering signs of the derivative, we see that λ achieves minimizes $D(\lambda)$ for a value in the interval $[\lambda_*(f), \lambda^*(f)]$ where

$$\lambda_*(f) = \inf\{t : F_f(t) \geq 1 - \alpha\} \text{ and } \lambda^*(f) = \sup\{t : F_f(t) \leq 1 - \alpha\}.$$

Note further that when $\mathcal{F}$ is compact in, say, sup norm, then we also have finite $\lambda_* = \inf_{f \in \mathcal{F}} t_*(\lambda)$ and $\lambda^* = \inf_{f \in \mathcal{F}} t^*(\lambda)$. In any case, we see that it suffices to define λ on a compact set $\Lambda = [\lambda_*, \lambda^*]$, and so we may assume without loss of generality that the domain of λ is compact.

This verifies the conditions of the minimax theorem, and so we have

$$\text{LCVaR}_\alpha(f) = \inf_{\lambda \in \mathbb{R}} \sup_{q : q(\cdot) \in [0, \alpha^{-1}]} \mathbb{E}\left[q_Y(R_Y(f) - \lambda)\right] + \lambda = \inf_{\lambda \in \mathbb{R}} \left\{\mathbb{E}\left[\alpha^{-1}\mathbb{E}(R_Y(f) - \lambda)_+ + \lambda\right]\right\},$$

which completes the proof. □

Next, we consider LHCVaR.

Proof of Proposition 5. The proof is similar to that of Proposition 4. The Lagrangian of LHCVaR is

$$L(q, \lambda) = \mathbb{E}[q_Y R_Y(f)] + \lambda\left(1 - \mathbb{E}[q_Y]\right) = \mathbb{E}\left[q_Y(R_Y(f) - \lambda)\right] + \lambda.$$

Next, by the trivial direction of the minimax theorem, we have

$$\text{LHCVaR}_\alpha(f) \leq \inf_{\lambda \in \mathbb{R}} \sup_{q : q_Y \in [0, \alpha_Y^{-1}]} L(q, \lambda) = \inf_{\lambda \in \mathbb{R}} \mathbb{E}\left[\alpha_Y^{-1}(R_Y(f) - \lambda)_+\right] + \lambda. \tag{5}$$

So, now our goal is to verify the conditions of the minimax theorem. As with LCVaR, the Lagrangian L is linear and therefore concave in q; is linear and therefore convex in λ; and is defined over a compact domain of values of q given by $[0, \alpha^{-1}]^k$. Thus, the only difficulty, as with LCVaR, is showing that it suffices to define λ over a compact interval. To this end, define the right hand side of equation (5) to be $\inf_{\lambda \in \mathbb{R}} H(\lambda)$. It suffices to show that $D(\lambda)$ achieves its infimum on a closed interval, in which case we can restrict the domain of λ to this compact, convex set.

To prove such an interval exists, we wish to show that there exist constants λ_* and λ^* such that H is decreasing for all $\lambda < \lambda_*$ and increasing for all $\lambda > \lambda^*$. By Lemma 6, we see that the derivative of H is

$$H'(\lambda) = 1 - \mathbb{E}[\alpha_Y^{-1}\mathbf{1}\{R_Y(f) > \lambda\}] = 1 - \sum_{i=1}^k \alpha_i^{-1} p_i \mathbf{1}\{R_i(f) > \lambda\}$$

when H' exists; otherwise the result holds for the left and right derivatives. Let $\lambda_*(f) = \min_{i=1,\ldots,k} R_i(f)$. Then, for $\lambda \leq \lambda_*(f)$, we have

$$H'(\lambda) = 1 - \sum_{i=1}^k \alpha_i^{-1} p_i \leq 0.$$

Next, pick $\lambda^*(f) = \max_{i=1,\ldots,k} R_i(f) + 1$. Then, for all $\lambda \geq \lambda^*(f)$, we have

$$H'(\lambda) = 1 \geq 0.$$

If ℓ is continuous, then each $R_i(f)$ is continuous in f. Moreover, when $\mathcal{F}$ is compact on $\mathcal{X}$ in the supremum norm, then we can define finite constants $\lambda_* = \inf_{f \in \mathcal{F}} \lambda_*(f)$ and $\lambda^*(f) = \sup_{f \in \mathcal{F}} \lambda^*(f)$.

Thus, we may restrict the domain of λ to $[\lambda_*, \lambda^*]$ without loss of generality. The minimax theorem now implies that equation (5) holds with equality, which completes the proof. □

E. Results for the Conditional Sampling Model

Now, we present the alternative result for the conditional sampling model. Recall that n_i is the number of samples of class i, which is assumed to be fixed.

Theorem 2. *Let ℓ be the multiclass margin loss. With probability at least $1-\delta$, for every f in $\mathcal{F}$ we have*

$$R_Q \leq \max_{q \in Q} \left\{ \widehat{R}_q(f) + \sum_{i=1}^{k} q_i \widehat{p}_i \left(2k\mathfrak{R}_{n_i}(\mathcal{F}) + \sqrt{\frac{\log \frac{k}{\delta}}{2n_i}} \right) \right\}.$$

Proof. The proof is similar to that of (Cao et al., 2019). We apply Lemma 8 and Lemma 10 to obtain

$$R_i(f) \leq \widehat{R}_i(f) + 2k\mathfrak{R}_{n_i}(\mathcal{F}) + \sqrt{\frac{\log \frac{k}{\delta}}{2n_i}}.$$

Multiplying by $q_i \widehat{p}_i$, summing over i, and taking a supremum over Q completes the proof. □

F. Gradient Descent-Ascent

In general, the robust classification problem is a saddle-point problem. For our purposes, define a saddle-point problem to be an optimization problem of the form

$$\inf_{a \in \mathcal{A}} \sup_{b \in \mathcal{B}} f(a, b). \tag{6}$$

One of the seminal results in game theory is that the minimax problem is equivalent to the maximin problem.

Theorem 3 (minimax theorem)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be compact convex sets. Let $f : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ be a function such that $a \mapsto f(a, b)$ is convex and $b \mapsto f(a, b)$ is concave. Then, we have*

$$\inf_{a \in \mathcal{A}} \sup_{b \in \mathcal{B}} f(a, b) = \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} f(a, b).$$

Algorithm 1: Online Gradient Descent

Input : Convex domain $\mathcal{A}$, $a_1 \in \mathcal{A}$, step sizes η_t, number of rounds T

for $t = 1, \ldots, T$ **do**

Play a_t and observe cost $f_t(a_t)$.

Update and project

$$x_{t+1} = a_t - \eta_t \nabla f_t(a_t)$$
$$a_{t+1} = \Pi_{\mathcal{A}}(x_{t+1}).$$

end

Output : The average iterate $\bar{a}_T = \frac{1}{T} \sum_{t=1}^{T} a_t$.

Lemma 4 (Theorem 3.1 of Hazan 2016)**.** *Let $f_1, \ldots, f_T : \mathcal{A} \to \mathbb{R}$ be a sequence of L-Lipschitz convex functions. If the step size for online gradient descent is chosen to be*

$$\eta_t = \frac{D}{L\sqrt{t}},$$

then we have

$$\sum_{t=1}^{T} f_t(a_t) - \min_{a^* \in \mathcal{A}} \sum_{t=1}^{T} f_t(a^*) \leq \frac{3}{2} DL\sqrt{T}.$$

Algorithm 2: Gradient Descent-Ascent

Input : Convex-concave function f, step sizes $\eta_{a,t}$ and $\eta_{b,t}$, number of rounds T

for $t = 1, \ldots, T$ **do**

Play (a_t, b_t) and observe cost $f(a_t, b_t)$.

Update and project

$$x_{t+1} = a_t - \eta_t \nabla_a f(a_t, b_t)$$
$$a_{t+1} = \Pi_{\mathcal{A}}(x_{t+1}).$$

Update and project

$$y_{t+1} = b_t + \eta_t \nabla_b f(a_t, b_t)$$
$$b_{t+1} = \Pi_{\mathcal{A}}(y_{t+1}).$$

end

Output : The average iterates $\bar{a}_T = \frac{1}{T}\sum_{t=1}^{T} a_t$ and $\bar{b}_T = \frac{1}{T}\sum_{t=1}^{T} b_t$.

Now we return to the saddle-point problem. We give the gradient descent-ascent algorithm in Algorithm 2 and the convergence result in Proposition 2.

Proposition 2. *Let $\mathcal{A}$ and $\mathcal{B}$ be convex, compact sets. Suppose that $\mathcal{A}$ has diameter D_a and $\mathcal{B}$ has diameter D_b. Let $f : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ be convex-concave, L_a-Lipschitz in its first argument, and L_b-Lipschitz in its second argument. Let (a^*, b^*) denote the solution to the saddle-point problem of equation (6). If $(\bar{a}_T, \bar{b}_T)$ is the output of Algorithm 2, then we have*

$$f(a^*, b^*) - \frac{3(L_a D_a + L_b D_b)}{2\sqrt{T}} \leq f(\bar{a}_T, \bar{b}_T) \leq f(a^*, b^*) + \frac{3(L_a D_a + L_b D_b)}{2\sqrt{T}}.$$

First, we want to use a lemma from online convex optimization. For this, we also state the standard online gradient descent algorithm. Here, we use $\Pi_{\mathcal{A}}$ to denote projection onto the set $\mathcal{A}$.

Proof. The proof is fairly straightforward from pre-existing results on online gradient descent; so we state it here. We start first with the upper bound. Define the "regret" to be

$$R_T = \sum_{t=1}^{T} [f(a_t, b_t) - f(a^*, b^*)]$$

where (a^*, b^*) is a solution to the saddle-point problem. Then, we have the decomposition

$$R_T = \sum_{t=1}^{T} [f(a_t, b_t) - f(a^*, b_t)] + \sum_{t=1}^{T} [f(a^*, b_t) - f(a^*, b^*)] \leq \frac{3}{2} L_a D_a \sqrt{T} + 0, \tag{7}$$

where the inequality follows from applying Lemma 4 and noting that the second summand is nonpositive by the definition of b^*. Similarly, we have

$$-R_T = \sum_{t=1}^{T} [f(a^*, b^*) - f(a_t, b_t)] + \sum_{t=1}^{T} [f(a_t, b_t) - f(a_t, b_t)] \leq 0 + \frac{3}{2} L_b D_b \sqrt{T}. \tag{8}$$

So, now we consider the averaged iterates. We have

$$\begin{aligned} f(\bar{a}_T, \bar{b}_T) &\leq \max_{b \in \mathcal{B}} f(\bar{a}_T, b) \\ &\leq \frac{1}{T} \max_{b \in \mathcal{B}} \sum_{t=1}^{T} f(a_t, b) \\ &= f(a^*, b^*) + \frac{1}{T} \max_{b \in \mathcal{B}} \sum_{t=1}^{T} [f(a_t, b) - f(a_t, b_t)] + \frac{1}{T} \sum_{t=1}^{T} [f(a_t, b_t) - f(a^*, b^*)] \\ &\leq f(a^*, b^*) + \frac{3L_b D_b}{2\sqrt{T}} + \frac{3L_a D_a}{2\sqrt{T}}. \end{aligned}$$

Note that the second inequality is due to convexity, and the third is due to Lemma 4 and equation (7).

Similarly, we have

$$\begin{aligned} f(\bar{a_T}, \bar{b}_T) &\geq \min_{a \in \mathcal{A}} f(a, \bar{b}_T) \\ &\geq \frac{1}{T} \min_{a \in \mathcal{A}} \sum_{t=1}^{T} f(a, b_t) \\ &= f(a^*, b^*) + \frac{1}{T} \min_{a \in \mathcal{A}} \sum_{t=1}^{T} [f(a, b_t) - f(a_t, b_t)] + \frac{1}{T} \sum_{t=1}^{T} [f(a_t, b_t) - f(a^*, b^*)] \\ &\geq f(a^*, b^*) - \frac{3L_a D_a}{2\sqrt{T}} - \frac{3L_b D_b}{2\sqrt{T}}. \end{aligned}$$

The second inequality follows from concavity, and the final inequality is a result of Lemma 4 applied to the sequence a_t and equation (8). This completes the proof. □

G. Additional Lemmas

Lemma 5. *Define* $D(\lambda) = \alpha^{-1}\mathbb{E}(R_Y(f) - \lambda)_+ + \lambda$, *and let* F_f *denote the cumulative distribution function of* $R_Y(f)$. *Then, we have*

$$D'(\lambda) = 1 + \alpha^{-1}(F_f(\lambda) - 1).$$

Proof. We compute the derivative directly. We obtain

$$\begin{aligned} D'(\lambda) &= 1 + \alpha^{-1} \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left\{ \mathbb{E}\left[(R_Y(f) - \lambda - \varepsilon)_+ - (R_Y(f) - \lambda)_+ \right] \right\} \\ &= 1 + \alpha^{-1} \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left\{ \mathbb{E}\left[-\varepsilon \mathbf{1}\left\{ R_Y(f) - \lambda > 0 \right\} \right] \right\} \\ &= 1 - \alpha^{-1} \mathbb{E}\mathbf{1}\left\{ R_Y(f) > \lambda \right\} \\ &= 1 + \alpha^{-1}(F_f(\lambda) - 1). \end{aligned}$$

This completes the proof. □

Lemma 6. *Define* $H(\lambda) = \mathbb{E}\left[\alpha^{-1}(R_Y - \lambda)_+\right] + \lambda$. *Then, the derivative of* $H(\lambda)$ *is*

$$H'(\lambda) = 1 - \mathbb{E}\left[\alpha_Y^{-1} \mathbf{1}\left\{ R_Y(f) > \lambda \right\}\right].$$

Proof. We again compute directly, obtaining

$$\begin{aligned} H'(\lambda) &= 1 + \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{E}\left[\alpha_Y^{-1}(R_Y(f) - \lambda - \varepsilon)_+ - \alpha_Y^{-1}(R_Y(f) - \lambda)_+\right] \\ &= 1 + \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \mathbb{E}\left[\alpha_Y^{-1}(-\varepsilon)\mathbf{1}\left\{R_Y(f) > \lambda\right\}\right] \\ &= 1 - \mathbb{E}\left[\alpha_Y^{-1}\mathbf{1}\left\{R_Y(f) > \lambda\right\}\right], \end{aligned}$$

as desired. □

Lemma 7. *We have the inequality*

$$\inf_{q \in Q}\left\{A(q) + B(q)\right\} \leq \inf_{q \in Q} A(q) + \sup_{q \in Q} B(q).$$

Proof. We have the inequality $A(q) + B(q) \leq A(q) + \sup_{q' \in Q} B(q')$, and taking infimums completes the proof. □

H. Standard Lemmas

Lemma 8 (Theorem 3.1 of Mohri et al. 2012)**.** *Let G be a family of functions mapping from $\mathbb{R}$ to $[0, 1]$. Then for $\delta > 0$ and all g in G, with probability at least $1 - \delta$, we have*

$$\mathbb{E}g(Z) \leq \frac{1}{n}\sum_{i=1}^{n} g(Z_i) + 2\mathfrak{R}_n(G) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

For our excess $(\mathcal{F}, q)$-risk bounds, we also use a slight variant, the proof of which is nearly identical to that of Lemma 8.

Lemma 9. *Let G be a family of functions mapping from $\mathbb{R}$ to $[0, 1]$. Then for $\delta > 0$ and all g in G, with probability at least $1 - \delta$, we have*

$$\left|\mathbb{E}g(Z) - \frac{1}{n}\sum_{i=1}^{n} g(Z_i)\right| \leq 4\mathfrak{R}_n(G) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

The following learning bound handles the multi-class margin loss more effectively in the number of classes (Kuznetsov et al., 2015).

Lemma 10. *Let $\mathcal{F}$ be a set of $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Recall that*

$$\Pi_1(\mathcal{F}) = \left\{x \mapsto f_y(x) : y \in \mathcal{Y}, f \in \mathcal{F}\right\}.$$

Then, under the margin loss, we have the bound

$$R(f) \leq \widehat{R}(f) + 4k\mathfrak{R}_n(\Pi_1(\mathcal{F})) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

for all f in $\mathcal{F}$ with probability at least $1 - \delta$.

I. Additional Experiment Details

For all methods and datasets, we optimized a logistic regression model with gradient descent over the entire data.

For all datasets, we chose a learning rate of 0.01 that was linearly annealed to 0.0001 over 2000 epochs.

I.1. Optimizing LCVaR/LHCVaR formulation

Note that in the formulation for LHCVaR described in Eq. (3), despite its convexity, the optimization is over a non-smooth loss. Thus, λ can be explicitly calculated given the classes of each risk. Let $R_{(i)}$ be the ith largest class risk.

$$\lambda = \min \left(\left\{ R_{(i)} : i \in [k], \sum_{j=1}^{i} \widehat{p}_i \alpha_i^{-1} \leq 1 \right\} \cup \{0\} \right)$$

An algorithm for computing this can be akin to water filling in order from largest to smallest class risk. When optimizing by some form of gradient descent the parameters of the classifier, this analytic form of the LHCVaR formulation can be quickly computed and avoid gradient computations on λ itself. Empirically, we used this formulation to speed up our experiments and leads to faster convergence than performing gradient descent on λ in addition to the model parameters. This algorithm is also applicable when optimizing LCVaR as well.

References

Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pp. 3437–3445, 2014.

Györfi, L. The Rate of Convergence of kn-NN Regression Estimates and Classification Rule. *IEEE Transactions on Information Theory*, 27(3):357–362, 1981. ISSN 0018-9448.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., and Dhillon, I. S. Consistent Binary Classification with Generalized Performance Metrics. In *Advances in Neural Information Processing Systems 27*, pp. 2744–2752. Curran Associates, Inc., 2014.

Krzyzak, A. and Pawlak, M. The pointwise rate of convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, 16:159–166, 1987.

Kuznetsov, V., Mohri, M., and Syed, U. Rademacher complexity margin bounds for learning with a large number of classes. In *ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels*, 2015.

Lewis, D. D. Evaluating and optimizing autonomous text classification systems. In *SIGIR*, volume 95, pp. 246–254. Citeseer, 1995.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pp. 603–611, 2013.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2012.

Narasimhan, H., Vaish, R., and Agarwal, S. On the statistical consistency of plug-in classifiers for non-decomposable performance measures. In *Advances in Neural Information Processing Systems*, pp. 1493–1501, 2014.

Stone, C. J. Optimal Global Rates of Convergence for Nonparametric Regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Yang, Y. Minimax nonparametric classification. i. rates of convergence. *IEEE Transactions on Information Theory*, 45(7): 2271–2284, 1999.