A Tunable Loss Function for Binary Classification

Tyler Sypherd*, Mario Diaz*†, Lalitha Sankar*, and Peter Kairouz‡
*Arizona State University, {tsypherd,mdiaztor,lsankar}@asu.edu
†Centro de Investigación en Matemáticas A.C., diaztorres@cimat.mx
‡Google AI, kairouz@google.com

***Abstract*—We present α-loss, $\alpha \in [1,\infty]$, a tunable loss function for binary classification that bridges log-loss ($\alpha = 1$) and 0-1 loss ($\alpha = \infty$). We prove that α-loss has an equivalent margin-based form and is classification-calibrated, two desirable properties for a good surrogate loss function for the ideal yet intractable 0-1 loss. For logistic regression-based classification, we provide an upper bound on the difference between the empirical and expected risk for α-loss at the critical points of the empirical risk by exploiting its Lipschitzianity along with recent results on the landscape features of empirical risk functions. Finally, we show that α-loss with $\alpha = 2$ performs better than log-loss on MNIST for logistic regression.**

I. Introduction

In learning theory, the performance of a classification algorithm in terms of accuracy, tractability, and convergence guarantees is contingent on the choice of a loss function. Consider a feature vector $X \in \mathcal{X}$, an unknown finite label $Y \in \mathcal{Y}$, and a hypothesis test $h : \mathcal{X} \to \mathcal{Y}$. The canonical 0-1 loss, given by $\mathbb{1}[h(X) \neq Y]$, is considered an ideal loss function that captures the probability of incorrectly guessing the true label Y using $h(X)$. However, since the 0-1 loss is neither continuous nor differentiable, its practical application is intractable with state-of-the-art learning algorithms. As a result, there has been much interest in identifying surrogate loss functions that best approximate the 0-1 loss. Common surrogate loss functions include logistic loss, squared loss, and hinge loss.

For binary classification tasks, a hypothesis test $h : \mathcal{X} \to \{-1, 1\}$ is typically replaced by a classification function $f : \mathcal{X} \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. In this context, loss functions are often written in terms of a *margin*, defined as the product of the label, $Y \in \{-1, 1\}$, and the value of the classification function $f(X)$ (see, [1]–[4]). In [1], Lin defines a margin-based loss function as Fisher consistent if, for any x and a given posterior $P_{Y|X=x}$, its population minimizer has the same sign as the optimal Bayes classifier. In [2], Bartlett *et al.* introduce a stronger surrogate requirement of *classification-calibration* wherein the loss function is Fisher consistent for any $P_{Y|X=x}$.

Yet another property for a good surrogate loss function is captured by the effectiveness of the empirical risk minimizers in approximating the true risk minimizers, a property studied through the *empirical landscape*. In [5], Mei *et al.* prove that for general non-convex loss functions which satisfy certain regularity conditions, all critical features of the landscape including local minimizers/maximizers and saddle points of the empirical risk and the true risk are one-to-one, with the distance between corresponding features decreasing as $O\left(\sqrt{\log n/n}\right)$ for n samples.

In [6], Liao *et al.* introduce α-loss as a new loss function to model information leakage under different adversarial threat models. We consider a more general learning setting and apply α-loss for binary classification. We prove that α-loss has an equivalent margin-based form which is classification-calibrated. For a family of logistic regression based classifiers, we use the Lipschitzianity of α-loss and results in [5] to upper bound the difference between the empirical and expected risk under α-loss at the critical points of the empirical risk. Finally, for the MNIST dataset, we focus on a low capacity learning model using logistic regression (such models are desirable when tuning deep neural networks is challenging) to illustrate the higher classification accuracy of α-loss ($\alpha > 1$) relative to the oft-used cross entropy (log-loss).

II. Preliminaries

A. α-loss

Let $\mathcal{P}(\mathcal{Y})$ be the set of probability distributions over $\mathcal{Y}$. For $\alpha \in [1,\infty]$, Liao et al. [6] define α-loss $l^\alpha : \mathcal{Y} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_+$ as

$$l^\alpha(y, P_Y) := \begin{cases} -\log P_Y(y) & \alpha = 1, \\ \frac{\alpha}{\alpha-1}[1 - P_Y(y)^{1-1/\alpha}] & \alpha \in (1,\infty), \\ 1 - P_Y(y) & \alpha = \infty. \end{cases} \tag{1}$$

Note that for (y, P_Y) fixed, $l^\alpha(y, P_Y)$ is continuous in α.

Consider random variables $(X,Y) \sim P_{X,Y}$. Observing X, one can construct an estimate $\hat{Y}$ of Y such that $Y - X - \hat{Y}$ form a Markov chain. One can use expected α-loss to quantify the effectiveness of the estimated posterior $P_{\hat{Y}|X}$ as $\mathbb{E}_{X,Y}[l^\alpha(Y, P_{\hat{Y}|X})]$. In particular,

$$\mathbb{E}_{X,Y}\left[l^1(Y, P_{\hat{Y}|X})\right] = \mathbb{E}_X\left[H(P_{Y|X=x}, P_{\hat{Y}|X=x})\right], \tag{2}$$

where $H(P,Q) := H(P) + D_{\mathrm{KL}}(P\|Q)$ is the cross-entropy between P and Q. Similarly,

$$\mathbb{E}_{X,Y}[l^\infty(Y, P_{\hat{Y}|X})] = \mathbb{P}[Y \neq \hat{Y}], \tag{3}$$

i.e., the expected α-loss for $\alpha = \infty$ equals the probability of error. It can be shown that the expected α-loss is continuous

This material is based upon work supported by the National Science Foundation under Grant Nos. CCF-1350914 and CIF-1815261.

in α, i.e., (2) and (3) result from the continuous extensions for $\alpha = 1$ and $\alpha = \infty$, respectively. Thus, we see that the extremal points of expected α-loss are expected log-loss and probability of error.

B. Binary Classification in Learning

Let $S_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$ be a training dataset where, for each i, $X_i \in \mathcal{X} \subset \mathbb{R}^d$ is the feature vector and $Y_i \in \mathcal{Y} = \{-1, 1\}$ is the class label. We assume that the samples $\{(X_i, Y_i) : i = 1, \ldots, n\}$ are independently drawn from an unknown distribution $P_{X,Y}$. There are multiple approaches (and nomenclatures) to classification [1]–[4]; in particular, we consider two alternative approaches, namely, using *soft classifiers* and using *classification functions*.

Soft classifier: In this approach, the objective of the learner is to construct, based on the training dataset S_n, a soft classifier $g : \mathcal{X} \to [0, 1]$ capable of predicting the likelihood of a label of previously unseen feature vectors. More specifically, for each $x \in \mathcal{X}$, $g(x)$ estimates the probability of the event $\{Y = 1\}$ given $\{X = x\}$. Usually, the learner selects a soft classifier by minimizing a loss function over a family of soft classifiers. Note that every soft classifier determines a set of beliefs and vice versa. Indeed, given a soft classifier g, we can define $P_{\hat{Y}|X}$ by taking $P_{\hat{Y}|X}(1|x) := g(x)$. Conversely, given a set of beliefs $P_{\hat{Y}|X}$, we can define a soft classifier $g(x) = P_{\hat{Y}|X}(1|x)$.

Observe that the soft classification construct defined above makes α-loss in (1) a natural fit as a loss function. Indeed, one can define the expected α-loss (true risk) of a soft classifier as

$$R_{l^\alpha}(g) = \mathbb{E}_{X,Y}[l^\alpha(Y, P_{\hat{Y}|X})], \tag{4}$$

where $P_{\hat{Y}|X}$ is the set of beliefs associated to g. Analogously, we define the empirical α-loss as

$$\hat{R}_{l^\alpha}(g) = \frac{1}{n}\sum_{i=1}^{n} l^\alpha(y_i, P_{\hat{Y}|X=x_i}). \tag{5}$$

Finally, we denote the conditional risk of the α-loss by

$$C_{l^\alpha}(g) = \mathbb{E}_{Y|X}[l^\alpha(Y, P_{\hat{Y}|X=x})]. \tag{6}$$

Observe that $R_{l^\alpha}(g) = \mathbb{E}_X[C_{l^\alpha}(g)]$.

Classification function: As an alternative approach, a learner can select a classification function $f : \mathcal{X} \to \overline{\mathbb{R}}$ by minimizing a loss function over a given family of classification functions. Observe that any such f can yield a (hard decision) hypothesis $h(X) = \text{sign}(f(X))$. The value $f(x)$ can be regarded as the confidence on the value of Y given $\{X = x\}$; a large value of $f(x)$ corresponds to a high confidence on the event $\{Y = 1\}$ given $\{X = x\}$, while a large value of $-f(x)$ corresponds to a high confidence on the event $\{Y = -1\}$.

For this setting, margin-based loss functions have been proposed as a meaningful family of loss functions. A loss function is said to be margin-based if, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the risk associated to a pair $(y, f(x))$ is given by $\tilde{l}(yf(x))$ for some function $\tilde{l} : \overline{\mathbb{R}} \to \mathbb{R}_+$. In this case, the risk of the pair $(y, f(x))$ only depends on the product $yf(x)$, where the product $yf(x)$ is called the margin. Observe that a negative margin corresponds to a mismatch between the signs of $f(x)$ and y, i.e., a classification error by f. Similarly, a positive margin corresponds to a match between the signs of $f(x)$ and y, i.e., a correct classification by f. Hence, most margin-based losses have a graph similar to those depicted in Figure 1(a). Since margin-based loss functions synthesize two quantities (Y and f) into a single margin, they are commonly found in the binary classification literature [1], [2], [7]. The risk of a classification function f with respect to (w.r.t.) a margin-based loss function $\tilde{l}$ is defined as

$$R_{\tilde{l}}(f) = \mathbb{E}_{X,Y}[\tilde{l}(Yf(X))]. \tag{7}$$

For notational convenience, the risk of the 0-1 loss is denoted by $R(f)$, i.e.,

$$R(f) = \mathbb{E}[\mathbb{1}(\text{sign}(f(X)) \neq Y)]. \tag{8}$$

We now introduce a margin-based α-loss. Let $\sigma : \overline{\mathbb{R}} \to [0, 1]$ be the sigmoid function, i.e.,

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \tag{9}$$

Observe that σ is invertible and $\sigma^{-1} : [0, 1] \to \overline{\mathbb{R}}$ is given by

$$\sigma^{-1}(z) = \log\left(\frac{z}{1 - z}\right). \tag{10}$$

Definition 1. *We define the margin α-loss $\tilde{l}^\alpha : \overline{\mathbb{R}} \to \mathbb{R}_+$ as*

$$\tilde{l}^\alpha(z) := \begin{cases} -\log(\sigma(z)) & \alpha = 1, \\ \frac{\alpha}{\alpha-1}\left(1 - \sigma(z)^{1-1/\alpha}\right) & \alpha \in (1, \infty), \\ 1 - \sigma(z) & \alpha = \infty. \end{cases} \tag{11}$$

In Figure 1(a), we plot the margin-based α-loss for different values of α. Observe that, on the one hand, the penalty assigned to misclassified examples decreases as α increases. In practice, this decrease is desirable as the classification error only depends on the prediction itself and not in the particular confidence (margin). On the other hand, the absolute value of the derivative of $\tilde{l}^\alpha$ decreases as α increases. This behavior makes the computation of the optimal classification function more challenging as α increases (as evidenced by the intractability of 0-1 loss).

C. Classification-Calibration

An important concept in the analysis and design of margin-based losses is that of classification-calibration. To define this, we begin by defining the true posterior $\eta : \mathcal{X} \to [0, 1]$ as $\eta(x) = P_{Y|X}(y = 1|x)$. As in [2], we abbreviate $\eta(x)$ as η, making implicit the dependence on x.

Definition 2 ([2, Definition 1])**.** *A margin-based loss function $\tilde{l}$ is said to be classification-calibrated if, for every $\eta \neq 1/2$,*

$$\inf_{f : f(2\eta-1) \leq 0} (\eta\tilde{l}(f) + (1-\eta)\tilde{l}(-f)) > \inf_{f \in \mathbb{R}} (\eta\tilde{l}(f) + (1-\eta)\tilde{l}(-f)). \tag{12}$$

The conditional risk of f given $\{X = x\}$ is given by

$$\mathbb{E}_{Y|X=x}[\tilde{l}(Yf(x))] = \eta(x)\tilde{l}(f(x)) + (1-\eta(x))\tilde{l}(-f(x)). \tag{13}$$

If $\tilde{l}$ is a classification-calibrated margin-based loss function, then the minimum conditional risk given $\{X = x\}$ is attained by a z_x^* such that $\operatorname{sign}(z_x^*) = \operatorname{sign}(2\eta(x)-1)$. Thus, assuming that the posterior distribution η is known, the optimal classification function for $\tilde{l}$, namely $f^*(x) := z_x^*$, gives rise to the optimal classification function for the 0-1 loss, namely the Bayes decision rule $\operatorname{sign}(2\eta(x) - 1)$.

The following proposition establishes another important consequence of classification-calibration; we will use it in the sequel.

Proposition 1 ([2, Theorem 3])**.** *Assume that $\tilde{l}$ is a classification-calibrated margin-based loss function. Then, for every sequence of measurable functions $(f_i)_{i=1}^{\infty}$ and every probability distribution on $\mathcal{X} \times \mathcal{Y}$,*

$$\lim_{i\to\infty} R_{\tilde{l}}(f_i) = R_{\tilde{l}}^* \text{ implies that } \lim_{i\to\infty} R(f_i) = R^*, \tag{14}$$

where $R_{\tilde{l}}^* := \min_f R_{\tilde{l}}(f)$ and $R^* := \min_f R(f)$.

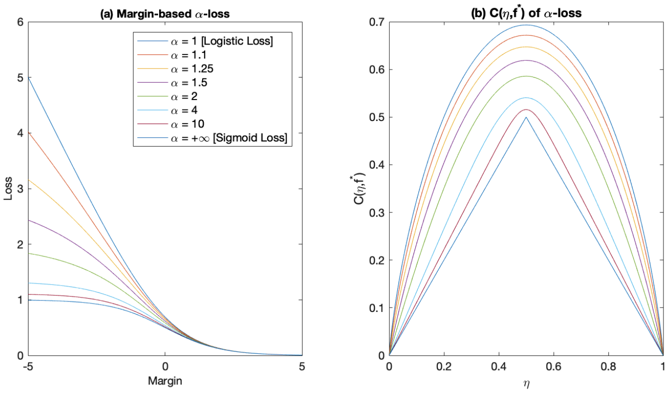

Figure 1. (a) Margin-based α-loss, as a function of the margin $z = yf(x)$; (b) minimum conditional risk for different values of α.

III. Results

A. Relation Between α-loss and its Margin Form

The following proposition shows an important relation between α-loss and its margin form in the context of binary classification. For reasons of brevity, we refer the reader to the full version of the paper for the complete proof.

Proposition 2. *Consider a soft classifier g and let $P_{\hat{Y}|X}$ be the set of beliefs associated to it. If $f(x) = \sigma^{-1}(g(x))$, then, for every $\alpha \in [1, \infty]$,*

$$l^\alpha(y, P_{\hat{Y}|X=x}) = \tilde{l}^\alpha(yf(x)). \tag{15}$$

Conversely, if f is a classification function, then the set of beliefs $P_{\hat{Y}|X}$ associated to $g(x) := \sigma(f(x))$ satisfies (15)*. In particular, for every $\alpha \in [1, \infty]$,*

$$\min_{P_{\hat{Y}|X}} \mathbb{E}_{X,Y}(l^\alpha(Y, P_{\hat{Y}|X})) = \min_f \mathbb{E}_{X,Y}(\tilde{l}^\alpha(Yf(X))). \tag{16}$$

This proposition unifies the probabilistic and margin settings. It also illustrates that the choice of the sigmoid function as the "change of variable" between soft classifiers and classification functions is sensible as the values of the minimization are the same. Furthermore, the minimizers are one-to-one by construction.

B. Statistical Guarantees

Now we establish some statistical properties of the margin-based α-loss that guarantee its appropriateness for classification tasks.

Theorem 1. *For every $\alpha \in [1, \infty]$, the margin-based α-loss $\tilde{l}^\alpha$ is classification-calibrated. In addition, its optimal classification function is given by*

$$f^*(\alpha, \eta) = \alpha \cdot \sigma^{-1}(\eta). \tag{17}$$

Furthermore, its minimum conditional risk is given by

$$C_{\tilde{l}^\alpha}(\eta, f^*) = \begin{cases} -\eta \log \eta - (1-\eta)\log 1 - \eta & \alpha = 1, \\ \frac{\alpha}{\alpha-1}\left[1 - Q(\eta) - Q(1-\eta)\right] & \alpha \in (1, +\infty), \\ \min\{\eta, 1-\eta\} & \alpha \to +\infty, \end{cases} \tag{18}$$

where $Q(z) = \left(\dfrac{z^{\alpha+1-1/\alpha}}{z^\alpha + (1-z)^\alpha}\right)^{1-1/\alpha}$.

Proof. If $\alpha = 1$, then $\tilde{l}^\alpha$ becomes logistic loss which is classification-calibrated, as is shown in [2]. Its optimal classifier and minimum conditional risk are given in [4]. If $\alpha = +\infty$, then $\tilde{l}^\alpha$ becomes sigmoid loss which is known to be classification-calibrated [2]. It can be verified that the optimal classifier for sigmoid loss is degenerated, i.e.,

$$f^*(+\infty, \eta) = \begin{cases} +\infty & \eta > 1/2, \\ -\infty & \eta < 1/2, \end{cases} \tag{19}$$

and $C_{\tilde{l}^\infty}^* = \min\{\eta, 1-\eta\}$.

Let $\alpha \in (1, +\infty)$. By definition of classification-calibration, we have to show that, for every $\eta \neq 1/2$,

$$\inf_{f : f(2\eta-1)\leq 0} (\eta\tilde{l}(f) + (1-\eta)\tilde{l}(-f)) > \inf_{f\in\mathbb{R}} (\eta\tilde{l}(f) + (1-\eta)\tilde{l}(-f)). \tag{20}$$

First we assume that $\eta > 1/2$. In this case, the strategy of proof is to show that the optimization in the right-hand-side of (20) has a unique minimizer f^* and that $f^* > 0$, which means that the right-hand-side of (20) is strictly smaller than the left-hand-side. Indeed, with some straightforward algebra, we can show that $f^* = \alpha \log\left(\dfrac{\eta}{1-\eta}\right)$, which trivially implies that $f^* > 0$. The value of $C_{\tilde{l}^\alpha}^*$ can be obtained by substituting f^* in (6). The case $\eta < 1/2$ can be proved *mutatis mutandis*. □

Proposition 3. *The margin-based α-loss $\tilde{l}^\alpha : \overline{\mathbb{R}} \to \mathbb{R}_+$ is convex for $\alpha = 1$ and quasi-convex for $\alpha > 1$. Furthermore, for every $\alpha \in [1, \infty]$, the minimum conditional risk $C_{\tilde{l}^\alpha}(\eta, f^*)$ is concave as a function η.*

Proof. Since $\tilde{l}^1$ is logistic loss, it is convex with respect to the margin as can be seen by observing its second derivative.

For $\alpha > 1$, it can be shown that $\tilde{l}^{\alpha}$ is monotone, so it is quasi-convex. However, $\tilde{l}^{\alpha}$ is not convex for $\alpha > 1$ since its second derivative is negative for negative values of the margin. Similarly, using a second-derivative argument it can be shown that $C_{\tilde{l}^{\alpha}}(\eta, f^*)$ is concave for every $\alpha \in [1, +\infty]$. □

Many commonly used loss functions in binary classification are convex. Despite the advantages of convex losses in terms of numerical optimization, non-convex loss functions can provide practical benefits as well. For instance, Mei *et al.* [5] state that non-convex loss functions "demonstrate superior robustness and classification accuracy in contrast to convex loss functions". In essence, non-convex loss functions assign less weight to misclassified training examples and therefore algorithms using such losses are less perturbed by outliers. The desirability of non-convex losses is further evidenced by other empirical studies, see, for example, [8]–[10].

Another perspective on the convexity of loss functions is presented in [4] where the authors argue that, for classification tasks, the convexity of a margin-based loss function is non-essential, as long as its minimum conditional risk is concave as a function of η. With regards to α-loss, this is amply observed in Figure 1(b). Since the margin-based α-loss is classification-calibrated and its minimum conditional risk is concave as a function of η, it is a reasonable loss function for binary classification problems.

C. Empirical Landscape of α-loss under Logistic Regression

In this section we consider a setting in which logistic regression is used to perform binary classification. Namely, for a given $\Theta \subset \mathbb{R}^d$, the family of soft classifiers under consideration has the form

$$g_\theta(x) = \sigma(\theta \cdot x), \tag{21}$$

where $\theta \in \Theta$ and σ is the sigmoid function given in (9). This in turn results in α-loss taking the form

$$l^{\alpha}(y, g_\theta(x)) = \frac{\alpha}{\alpha - 1}\Big[1 - \frac{1+y}{2} g_\theta(x)^{1-1/\alpha} - \frac{1-y}{2}(1 - g_\theta(x))^{1-1/\alpha}\Big]. \tag{22}$$

A straightforward computation shows that

$$\frac{\partial}{\partial \theta_i} l^{\alpha}(y, g_\theta(x)) = \Big[\frac{1-y}{2} g_\theta(x)(1 - g_\theta(x))^{1-1/\alpha} - \frac{1+y}{2} g_\theta(x)^{1-1/\alpha}(1 - g_\theta(x))\Big] x_i, \tag{23}$$

where $\theta = (\theta_1, \ldots, \theta_d)$ and $x = (x_1, \ldots, x_d)$. Hence,

$$\nabla_\theta l^{\alpha}(Y, g_\theta(X)) = F_1(\alpha, \theta, X, Y) X, \tag{24}$$

where $F_1(\alpha, \theta, x, y)$ is the expression within brackets in (23).

Recently, Mei *et al.* [5] prove that for non-convex loss functions satisfying certain regularity conditions, there exists a bijection between the critical points of the empirical risk and the critical points of true risk such that the distance between corresponding points decreases at a rate $O\left(\sqrt{\log n / n}\right)$, where n is the sample size. Building upon their work, we establish generalization bounds for logistic regression under α-loss.

Theorem 2. *Let $B_d(r)$ denote the ball of radius r in d-dimensional Euclidean space. Assume that, for some $r > 0$, X is supported over $B_d(r)$ and $\theta \in \Theta \subset B_d(r)$. For each $y \in \{-1, 1\}$, let $X^{[y]}$ be a random variable having the distribution of X conditioned on $Y = y$. We further assume that $X^{[1]} \overset{\mathrm{d}}{=} -X^{[-1]}$, $\mathbb{E}[X^{[1]}] \neq 0$, and $1 - \sigma(-r^2)^2 < \frac{\|\mathbb{E}(X^{[1]})\|}{\mathbb{E}(\|X^{[1]}\|)}$. Let $\hat{\theta}_n$ denote a local minimizer of the empirical risk function $\theta \mapsto \hat{R}_{l^\alpha}(g_\theta)$. If the sample size n is large enough, then, with probability at least $1 - \delta$,*

$$|R_{l^\alpha}(g_{\hat{\theta}_n}) - \hat{R}_{l^\alpha}(g_{\hat{\theta}_n})| \leq C_\alpha \left(\sqrt{\frac{\log(n)}{n}} + \sqrt{\frac{\log(4m/\delta)}{2n}}\right), \tag{25}$$

where C_α is a constant independent of n and m is the number of critical points.

Proof. In Appendices V-D and V-E we show that l^α satisfies the regularity conditions[1] in [5, Thm. 2] and, as a result, the expected risk has finitely many critical points $\{\theta_1, \ldots, \theta_m\}$ and for n large enough, with probability at least $1 - \delta/2$, there exists $\hat{\theta} := \theta_i$ for some $i \in [m]$ such that,

$$\|\hat{\theta}_n - \hat{\theta}\| \leq C\sqrt{\frac{\log(n)}{n}}, \tag{26}$$

where C is a constant independent of n. By the triangle inequality,

$$|R_{l^\alpha}(g_{\hat{\theta}_n}) - \hat{R}_{l^\alpha}(g_{\hat{\theta}_n})| \leq \mathrm{I} + \mathrm{II} + \mathrm{III}, \tag{27}$$

where $\mathrm{I} = |R_{l^\alpha}(g_{\hat{\theta}_n}) - R_{l^\alpha}(g_{\hat{\theta}})|$, $\mathrm{II} = |R_{l^\alpha}(g_{\hat{\theta}}) - \hat{R}_{l^\alpha}(g_{\hat{\theta}})|$, and $\mathrm{III} = |\hat{R}_{l^\alpha}(g_{\hat{\theta}}) - \hat{R}_{l^\alpha}(g_{\hat{\theta}_n})|$.

Observe that, $\mathrm{II} \leq \max_{i=1,\ldots,m} |R_{l^\alpha}(g_{\theta_i}) - \hat{R}_{l^\alpha}(g_{\theta_i})|$. By Hoeffding's inequality and the union bound, see, e.g., [11, Chapter 4], it can be shown that, for any $\epsilon > 0$,

$$\Pr\left(\max_{i=1,\ldots,m} |R_{l^\alpha}(g_{\theta_i}) - \hat{R}_{l^\alpha}(g_{\theta_i})| > \epsilon\right) \leq 2m \exp\left(-\frac{2n(\alpha-1)^2\epsilon^2}{\alpha^2}\right). \tag{28}$$

By taking $\delta = 4m\exp\left(-2n(\alpha-1)^2\epsilon^2/\alpha^2\right)$, we conclude that, with probability at least $1 - \delta/2$,

$$\mathrm{II} \leq \max_i |R_{l^\alpha}(g_{\theta_i}) - \hat{R}_{l^\alpha}(g_{\theta_i})| \leq \frac{\alpha}{\alpha - 1}\sqrt{\frac{\log(4m/\delta)}{2n}}. \tag{29}$$

By the boundedness of X and θ, the derivative in (24) is bounded for all X and θ. Therefore, independently of the training dataset, the empirical risk function $\hat{R}_{l^\alpha}$ is C''_α-Lipschitz for some $C''_\alpha \geq 0$. Hence,

$$\mathrm{III} \leq C''_\alpha \|\hat{\theta}_n - \hat{\theta}\|. \tag{30}$$

[1] These conditions are sub-Gaussian gradient, sub-exponential Hessian, Lipschitz Hessian, and strongly Morse expected risk.

The last inequality and (26) imply that

$$\mathrm{III} \leq C'_\alpha \sqrt{\frac{\log(n)}{n}}, \tag{31}$$

where $C'_\alpha := CC''_\alpha$.

A differentiation under the integral sign argument shows that R_{l^α} is also C''_α-Lipschitz. Thus,

$$|R_{l^\alpha}(g_{\hat{\theta}_n}) - R_{l^\alpha}(g_{\hat{\theta}})| \leq C''_\alpha \|\theta_n - \theta\|. \tag{32}$$

As before, (26) leads to

$$\mathrm{I} = |R_{l^\alpha}(g_{\hat{\theta}_n}) - R_{l^\alpha}(g_{\hat{\theta}})| \leq C'_\alpha \sqrt{\frac{\log(n)}{n}}. \tag{33}$$

The result follows from (29), (31) and (33). □

The following corollary follows as a natural addendum to our main results and establishes that an algorithm perfectly trained using the α-loss converges, with the number of samples n, to an optimal hypothesis w.r.t. the 0-1 loss.

Corollary 1. *For each $n \in \mathbb{N}$, let S_n be a training dataset of size n and $\hat{\theta}_n$ be a global minimizer of the associated empirical risk function $\theta \mapsto \hat{R}_{l^\alpha}(g_\theta)$. Under the assumptions of Theorem 2, the sequence $(\hat{\theta}_n)_{n=1}^{\infty}$ is asymptotically optimal for the* 0-1 *risk, i.e., almost surely,*

$$\lim_{n \to \infty} R(\hat{\theta}_n) = R^*. \tag{34}$$

The Proof of Corollary 1 is given in Appendix F.

D. Simulation Results

We perform simulations on a logistic regression model with randomly initialized weights using a portion of the MNIST dataset. In order to have a binary dataset, we partition the MNIST dataset into the images of 1's and 7's which yields a training set of $12,500$ samples and a test set of $2,050$ samples (evenly divided between the two labels for both train and test data). Of the $12,500$ training samples, we use $11,500$ for training and the remaining $1,000$ for cross-validation.

Since cross entropy (log-loss, i.e. $\alpha = 1$) is the most commonly used loss function for practical implementation in classification [7], we use it as our benchmark for accuracy. In this way, we compare cross entropy and α-loss in terms of accuracy for $\alpha \in \{1.1, 1.2, 1.5, 2.0\}$. In order to have a level playing field, we tune the learning rate during cross-validation, so as to compare the optimal performance of each loss function.

α	Learning Rate	Testing Accuracy
1.0	1.0	**85.3805%**
1.1	1.3	85.4005%
1.2	1.0	85.8527%
1.5	1.9	87.3044%
2.0	2.0	**87.3302%**

Table I
PERFORMANCE REGRESSION

As shown in Table I, for the simple logistic regression model under consideration, α-loss with $\alpha = 2$ exhibits a testing accuracy about $\sim 2\%$ higher than cross entropy. While this is a simple model, the performance of α-loss is encouraging and suggests that further work is needed.

It ought to be mentioned that, with large-capacity models, MNIST data can be classified with an accuracy above 99% [12]. The goal of our numerical experiments with low capacity models (such models are desirable when tuning deep neural networks is challenging) is to show that α-loss can perform better than cross entropy in some situations. Further simulations using state-of-the-art datasets is the subject of ongoing research.

IV. Concluding Remarks

We have proved theoretical properties and highlighted practical preliminary results for α-loss under binary classification. Beyond generalization to multi-hypothesis testing, the optimal choice of α is another important problem and will require exploring the trade-off between the magnitude of the gradients (convergence) and the gradient noise induced by finite samples. Yet another challenging problem to explore is the robustness of α-loss for $\alpha > 1$ against adversarial examples; one approach to doing so is by quantifying its generalization properties by building upon the work in [13].

References

[1] Y. Lin, "A note on margin-based loss functions in classification," *Statistical & Probability Letters*, vol. 68, no. 1, pp. 73–82, 2004.

[2] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[3] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On surrogate loss functions and f-divergences," *AOS*, vol. 37, no. 2, pp. 876–904, 04 2009.

[4] H. Masnadi-Shirazi and N. Vasconcelos, "On the design of loss functions for classification: theory, robustness to outliers, and SavageBoost," in *Advances in neural information processing systems*, 2009, pp. 1049–1056.

[5] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for nonconvex losses," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2747–2774, 2018.

[6] J. Liao, O. Kosut, L. Sankar, and F. P. Calmon, "A tunable measure for information leakage," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 701–705.

[7] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *arXiv preprint arXiv:1702.05659*, 2017.

[8] Y. Wu and Y. Liu, "Robust truncated hinge loss support vector machines," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 974–983, 2007.

[9] O. Chapelle, C. B. Do, C. H. Teo, Q. V. Le, and A. J. Smola, "Tighter bounds for structured estimation," in *Advances in neural information processing systems*, 2009, pp. 281–288.

[10] T. Nguyen and S. Sanner, "Algorithms for direct 0–1 loss optimization in binary classification," in *International Conference on Machine Learning*, 2013, pp. 1085–1093.

[11] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[12] Y. LeCun, C. Cortes, and C. J. C. Burges, "The MNIST database of handwritten digits," http://yann.lecun.com/exdb/mnist/index.html.

[13] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems*, 2017, pp. 2524–2533.

[14] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[15] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

V. Appendix

A. Proof of Proposition 2

Consider a soft classifier g and let $P_{\hat{Y}|X}$ be the set of beliefs associated to it. Suppose $f(x) = \sigma^{-1}(g(x))$, where $g(x) = P_{\hat{Y}|X}(1|x)$. We want to show that

$$l^{\alpha}(y, P_{\hat{Y}|X=x}) = \tilde{l}^{\alpha}(yf(x)). \tag{35}$$

We assume that $\alpha \in (1, \infty)$. Note that the cases where $\alpha = 1$ and $\alpha = \infty$ follow similarly.

Suppose that $g(x) = P_{\hat{Y}|X}(1|x) = \sigma(f(x))$. If $y = 1$, then

$$\begin{aligned} l^{\alpha}(1, P_{\hat{Y}|X}(1|x)) &= l^{\alpha}(1, \sigma(f(x))) && (36) \\ &= \frac{\alpha}{\alpha-1}[1 - \sigma(f(x))^{1-1/\alpha}] && (37) \\ &= \tilde{l}^{\alpha}(f(x)). && (38) \end{aligned}$$

If $y = -1$, then

$$\begin{aligned} l^{\alpha}(-1, P_{\hat{Y}|X}(-1|x)) &= l^{\alpha}(-1, 1 - P_{\hat{Y}|X}(1|x)) && (39) \\ &= l^{\alpha}(-1, 1 - \sigma(f(x))) && (40) \\ &= l^{\alpha}(-1, \sigma(-f(x))) && (41) \\ &= \frac{\alpha}{\alpha-1}[1 - \sigma(-f(x))^{1-1/\alpha}] && (42) \\ &= \tilde{l}^{\alpha}(-f(x)), && (43) \end{aligned}$$

where (41) follows from

$$\sigma(x) + \sigma(-x) = 1, \tag{44}$$

which can be observed by (9). To show the reverse direction of (35) we substitute

$$f(x) = \sigma^{-1}(g(x)) = \sigma^{-1}(P_{\hat{Y}|X}(1|x)) \tag{45}$$

in $\tilde{l}^{\alpha}(yf(x))$. For $y = 1$,

$$\begin{aligned} \tilde{l}^{\alpha}(f(x)) &= \tilde{l}^{\alpha}(\sigma^{-1}(P_{\hat{Y}|X}(1|x))) && (46) \\ &= \frac{\alpha}{\alpha-1}[1 - (\sigma(\sigma^{-1}(P_{\hat{Y}|X}(1|x))))^{1-1/\alpha}] && (47) \\ &= \frac{\alpha}{\alpha-1}[1 - P_{\hat{Y}|X}(1|x)^{1-1/\alpha}] && (48) \\ &= l^{\alpha}(1, P_{\hat{Y}|X}(1|x)). && (49) \end{aligned}$$

For $y = -1$,

$$\begin{aligned} \tilde{l}^{\alpha}(-f(x)) &= \tilde{l}^{\alpha}(-\sigma^{-1}(P_{\hat{Y}|X}(1|x))) && (50) \\ &= \frac{\alpha}{\alpha-1}[1 - \sigma(-\sigma^{-1}(P_{\hat{Y}|X}(1|x)))^{1-1/\alpha}] && (51) \\ &= \frac{\alpha}{\alpha-1}[1 - (1 - \sigma(\sigma^{-1}(P_{\hat{Y}|X}(1|x))))^{1-1/\alpha}] && (52) \\ &= \frac{\alpha}{\alpha-1}[1 - P_{\hat{Y}|X}(-1|x)^{1-1/\alpha}] && (53) \\ &= l^{\alpha}(-1, P_{\hat{Y}|X}(-1|x)), && (54) \end{aligned}$$

where (52) follows from (44).

The equality in the results of the minimization procedures follows from the equality between l^{α} and $\tilde{l}^{\alpha}$. As was shown in [6], the minimizer of the left-hand-side is

$$P^*_{\hat{Y}|X}(y|x) = \frac{P_{Y|X}(y|x)^{\alpha}}{\sum\limits_{y} P_{Y|X}(y|x)^{\alpha}}. \tag{55}$$

Using $f(x) = \sigma^{-1}(P_{\hat{Y}|X}(1|x))$, $f^*(x) = \sigma^{-1}(P^*_{\hat{Y}|X}(1|x))$.

B. Proof of Theorem 1

Suppose $\alpha = 1$, then $\tilde{l}^{\alpha}$ becomes

$$\tilde{l}^{1}(z) = -\log(\sigma(z)) = \log(1 + e^{-z}), \tag{56}$$

which is logistic loss. By solving the minimization procedure in (12), it can be shown as in [2] that $\tilde{l}^1$ is classification-calibrated. Further, the optimal classifier and minimum conditional risk of logistic loss are given in [4].

Suppose $\alpha = +\infty$, then $\tilde{l}^{\alpha}$ becomes

$$\tilde{l}^{\infty}(z) = 1 - \sigma(z) = \frac{e^z}{1 + e^z}, \tag{57}$$

which is sigmoid loss. Similarly, sigmoid loss can be shown to be classification-calibrated as is given in [2]. It can be verified by calculating the minimization procedure in (12) that the optimal classifier for sigmoid loss is degenerate. That is,

$$f^*(+\infty, \eta) = \begin{cases} +\infty & \eta > 1/2 \\ -\infty & \eta < 1/2. \end{cases} \tag{58}$$

Therefore, $C_{\tilde{l}^{\infty}}(\eta, f^*) = \min\{\eta, 1 - \eta\}$. Note that sigmoid loss and 0-1 loss have the same minimum conditional risk. Thus, sigmoid loss can be viewed as a smoothed version of 0-1 loss and will similarly suffer from vanishing gradients for most values of the margin.

Now consider $\alpha \in (1, +\infty)$. Since classification calibration requires proving (12), we begin by expanding the inequality in (12) using $\tilde{\ell}$ in (11) to show that $\forall \eta \neq 1/2$,

$$\inf_{f: f(2\eta-1) \leq 0} (\eta \tilde{l}(f) + (1-\eta)\tilde{l}(-f)) > \inf_{f \in \mathbb{R}} (\eta \tilde{l}(f) + (1-\eta)\tilde{l}(-f)). \tag{59}$$

Without loss of generality, we assume that $\eta > 1/2$. The strategy of the proof is to demonstrate that for $\eta > 1/2$, $f^* > 0$, which means that the right-hand-side of (59) is smaller than the left-hand-side because the attainer of the infimum is not in the search-space of the left-side's infimum. We rearrange the right-hand-side of (59) to obtain

$$\frac{\alpha}{\alpha-1}\left[1 - \sup_{f \in \mathbb{R}} \left[\eta\Big(\frac{1}{1+e^{-f}}\Big)^{1-1/\alpha} + (1-\eta)\Big(\frac{1}{1+e^{f}}\Big)^{1-1/\alpha}\right]\right]. \tag{60}$$

We take the derivative of the expression inside the supremum, which we denote $g(\eta, \alpha, f)$, and obtain

$$\frac{d}{df} g(\eta, \alpha, f) = \Big(1 - \frac{1}{\alpha}\Big)\Big(\frac{1}{e^f + 2 + e^{-f}}\Big)\Big[\eta\Big(1 + e^{-f}\Big)^{1/\alpha} - (1-\eta)\Big(1 + e^{f}\Big)^{1/\alpha}\Big]. \tag{61}$$

One can then obtain the f_0 minimizing (60) by setting $\dfrac{d}{df}g(\eta,\alpha,f)=0$, i.e.,

$$\eta\Big(1+e^{-f_0}\Big)^{1/\alpha}=(1-\eta)\Big(1+e^{f_0}\Big)^{1/\alpha}. \tag{62}$$

Note that the derivative $dg(\eta,\alpha,f)/df$ in (61) approaches zero for both $f\to+\infty$ and $f\to-\infty$ for which g simplifies to η and $(1-\eta)$, respectively. Since $\eta>1/2$, to show that $f_0\in(-\infty,\infty)$ is the point at which $g(\eta,\alpha,f)$ is maximized, we must demonstrate that $g(\eta,\alpha,f_0)>\eta>1/2$. We solve (62) for $(1-\eta)$ and substitute it into $g(\eta,\alpha,f_0)$. Further simplifying, we obtain $\eta(1+e^{-f_0})^{1/\alpha}$ which is always greater than η. Therefore, f_0 is the maximizer of $g(\eta,\alpha,f)$. Solving (62) for f_0, we obtain

$$f_0=f^*(\alpha,\eta)=\alpha\log\Big(\frac{\eta}{1-\eta}\Big)>0, \tag{63}$$

i.e., $\tilde{l}^\alpha$ is classification-calibrated. Since (63) minimizes the right side of (59), it is the optimal classifier for $\tilde{l}^\alpha$ where $\alpha\in(1,+\infty)$. Accordingly, $C_{\tilde{l}^\alpha}(\eta,f^*)$ is obtained by substituting (63) into (60).

C. Proof of Proposition 3

For $\alpha=1$, $\tilde{l}^1(z)=-\log\sigma(z)$. Further,

$$\frac{d^2}{dz^2}\tilde{l}^1(z)=\frac{e^{-z}}{(1+e^{-z})^2}\geq 0, \tag{64}$$

$\forall z\in\mathbb{R}$, so $\tilde{l}^1$ is convex.

For $\alpha\in(1,\infty)$,

$$\frac{d^2}{dz^2}\tilde{l}^\alpha(z)=\frac{(e^{-z}+1)^{1/\alpha}e^z(\alpha e^z-\alpha+1)}{\alpha(e^z+1)^3}. \tag{65}$$

As can be observed in the numerator for $\alpha>1$, there exists some z_0 for which $\alpha e^{z_0}-\alpha+1<0$. Thus $\tilde{l}^\alpha$ is not convex for $\alpha\in(1,\infty)$. Similarly as can be seen in (65) by letting $\alpha\to\infty$, that $\dfrac{d^2}{dz^2}\tilde{l}^\infty(z)=\dfrac{e^z(e^z-1)}{(e^z+1)^3}$, which is less than zero for $z<0$. Thus, $\tilde{l}^\infty$ is also not convex.

It can be shown that, for all $\alpha\in[1,\infty]$, $\tilde{l}^\alpha$ is monotonically decreasing since

$$\frac{d}{dz}\tilde{l}^\alpha(z)=\frac{-(e^{-z}+1)^{1/\alpha}e^z}{(1+e^z)^2}<0, \tag{66}$$

$\forall z\in\mathbb{R}$. Since monotonic functions are quasi-convex [14], we have that $\tilde{l}^\alpha$ is quasi-convex for $\alpha>1$.

With regards to the minimum conditional risk, for $\alpha=1$, it can be shown that $\dfrac{d^2}{d\eta^2}C_{\tilde{l}^1}(\eta,f^*)=\dfrac{1}{(\eta-1)\eta}<0$ since $\eta\in(0,1)$. Despite a cumbersome expression, one can similarly verify that, for $\alpha\in(1,\infty)$, $C_{\tilde{l}^{\tilde{\alpha}}}(\eta,f^*)$ is concave. For $\alpha=\infty$, $C_{\tilde{l}^{\tilde{\infty}}}(\eta,f^*)=\min\{\eta,1-\eta\}$ can be easily verified to be concave as a function of η.

D. Background for Theorem 2

The proof of Theorem 2 relies on a result by Mei *et al.* [5] stated at the end of this section. We start by providing the necessary background.

Definition 3. *A random vector $X\in\mathbb{R}^d$ is σ^2-sub-Gaussian if, for every $\lambda\in\mathbb{R}^d$,*

$$\mathbb{E}[e^{\langle\lambda,X-\mathbb{E}[X]\rangle}]\leq e^{\sigma^2\|\lambda\|_2^2/2}, \tag{67}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product.

Gaussian and bounded random variables are examples of sub-Gaussian random variables, see, for example, [15]. It can be shown that if the components of a random vector are sub-Gaussian, then the random vector itself is sub-Gaussian [15].

Definition 4. *A random matrix Z is τ^2-sub-exponential if, for every $\lambda\in B_d(1/\tau)$,*

$$\mathbb{E}\left[e^{|Z_\lambda-\mathbb{E}[Z_\lambda]|}\right]\leq 2, \tag{68}$$

where $Z_\lambda:=\langle\lambda,Z\lambda\rangle$ and $B_d(r)$ denotes the ball of radius r in d-dimensional Euclidean space.

We now recall the definition of a regularity property known as strongly Morse. Let $[d]:=\{1,2,\ldots,d\}$.

Definition 5. *We say that a twice differentiable function $F:B_d(r)\to\mathbb{R}$ is (ϵ,η)-strongly Morse if $\|\nabla F(x)\|_2>\epsilon$ for $\|x\|_2=r$ and, for any $x\in\mathbb{R}^d$, $\|x\|_2<r$, the following holds:*

$$\|\nabla F(x)\|_2\leq\epsilon\implies\min_{i\in[d]}|\lambda_i(\nabla^2F(x))|\geq\eta, \tag{69}$$

where $\{\lambda_i(\nabla^2F(x)):i\in[d]\}$ are the eigenvalues of $\nabla^2F(x)$.

Now we are in position to state Mei *et al.* result.

Proposition 4 ([5, Thm. 2]). *Let l be a given loss function. Assume that*

1) *the gradient $\nabla_\theta l(\theta)$ is sub-Gaussian;*
2) *the Hessian $\nabla_\theta^2 l(\theta)$ is sub-exponential;*
3) *the Hessian $\nabla_\theta^2 R_l(\theta)$ is bounded at a point and Lipschitz continuous with integrable Lipschitz constant, i.e, there exists J_* such that*

$$J(\mathbf{z})\equiv\sup_{\theta_1\neq\theta_2\in B^p(r)}\frac{\|\nabla^2 l(\theta_1;\mathbf{z})-\nabla^2 l(\theta_2;\mathbf{z})\|_{op}}{\|\theta_1-\theta_2\|_2}, \tag{70}$$

   where $\mathbb{E}[J(\mathbf{Z})]\leq J_*$;
4) *$R_l(\theta)$ is (ϵ,η)-strongly Morse.*

Let $\hat{\theta}_n$ denote a local minimizer of the empirical risk function $\theta\mapsto\hat{R}_l(\theta)$. If the sample size n is large enough, then there exists a critical point $\hat{\theta}$ of the true risk function $\theta\mapsto R_l(\theta)$ such that, with probability at least $1-\delta$,

$$\|\hat{\theta}_n-\hat{\theta}\|_2\leq C\sqrt{\frac{\log n}{n}}, \tag{71}$$

where $C=C(\sigma,\alpha,\epsilon,\eta,d)$ is a positive constant. Further, $\hat{R}_l(\theta)$ is $(\epsilon/2,\eta/2)$-strongly Morse.

$$F_2(\alpha,\theta,x,y) = \frac{1-y}{2}\Big[g_\theta(1-g_\theta)^{2-1/\alpha} - \left(1-\frac{1}{\alpha}\right) g_\theta^2(1-g_\theta)^{1-1/\alpha}\Big] + \frac{1+y}{2}\Big[g_\theta^{2-1/\alpha}(1-g_\theta) - \left(1-\frac{1}{\alpha}\right) g_\theta^{1-1/\alpha}(1-g_\theta)^2\Big] \tag{75}$$

E. Proof that l^α satisfies the Assumptions of Proposition 4

Here, we prove the assumptions stipulated by Proposition 4 hold for α-loss. We restrict ourselves to the setting of logistic regression. Thus, $\hat{R}_{l^\alpha}(\theta) = \hat{R}_{l^\alpha}(g_\theta)$ and $R_{l^\alpha}(\theta) = R_{l^\alpha}(g_\theta)$, where $g_\theta(x) = \sigma(\theta \cdot x)$ and $g_\theta(x)$ is often abbreviated g_θ for convenience.

Proof of Assumption 1: The first assumption requires the gradient of the loss function to be sub-Gaussian. The gradient of α-loss is given by (24). That is,

$$\nabla_\theta l^\alpha(Y, g_\theta(X)) = F_1(\alpha,\theta,X,Y)X, \tag{72}$$

where

$$\begin{aligned} F_1(\alpha,\theta,x,y) = \frac{1-y}{2} & g_\theta(x)(1-g_\theta(x))^{1-1/\alpha} \\ & - \frac{1+y}{2} g_\theta(x)^{1-1/\alpha}(1-g_\theta(x)). \end{aligned} \tag{73}$$

In order to prove (73), we used that

$$\frac{\partial}{\partial\theta}\sigma(\theta\cdot x) = \sigma(\theta\cdot x)(1-\sigma(\theta\cdot x)). \tag{74}$$

By the boundedness of the sigmoid function, we have that $|F_1(\alpha,\theta,X,Y)| \leq 1$. Since $X \in B_d(r)$ by assumption, each component of $\nabla_\theta l^\alpha$ is bounded and, as a consequence, sub-Gaussian. Therefore, the gradient of α-loss is sub-Gaussian.

Proof of Assumption 2: The second assumption requires the Hessian of the loss function to be sub-exponential. It can be shown that the Hessian has the form

$$\nabla_\theta^2 l^\alpha(Y, g_\theta(X)) = F_2(\alpha,\theta,X,Y)XX^T, \tag{76}$$

where $F_2(\alpha,\theta,X,Y)$ is defined on (75). It is straightforward to verify that $|F_2(\alpha,\theta,X,Y)| \leq \frac{1}{4}$. Notice that the product of $\nabla_\theta^2 l^\alpha$ with $\lambda \in B^p(1)$ becomes

$$\langle\lambda, \nabla_\theta^2 l^\alpha \lambda\rangle = \left(F_2(\alpha,\theta,X,Y)^{1/2}\sum_{i=1}^{d}\lambda_i X_i\right)^2. \tag{77}$$

Since both θ and X are assumed to be bounded, $\langle\lambda, \nabla_\theta^2 l^\alpha\lambda\rangle$ is the square of a bounded random variable. Since the square of a sub-gaussian random variable is sub-exponential, we conclude that the Hessian is sub-exponential.

Proof of Assumption 3: The third required assumption is that the Hessian of the loss function is Lipschitz and the Hessian of the population risk is bounded above at a point. The former can be observed by calculating the third derivative of α-loss and showing that it is bounded. The third derivative has the form

$$\frac{\partial}{\partial\theta_i}\nabla_\theta^2 l^\alpha(Y, g_\theta(X)) = F_3(\alpha,\theta,X,Y)XX^TX_i, \tag{79}$$

where $F_3(\alpha,\theta,X,Y)$ is defined in (78). Observe that $|F_3(\alpha,\theta,X,Y)| \leq 2$. Since $\theta, X \in B_d(r)$ by assumption, the derivative of the Hessian is bounded with constant $L = 2r^3$. Therefore, the Hessian is Lipschtiz continuous, in the sense of (70), with integrable Lipschitz constant L. Using similar arguments, it is straightforward to verify that the Hessian of the population risk is bounded at a point.

Proof of Assumption 4: The final assumption requires the population risk to be strongly Morse. Recall that, for each $y \in \{-1, 1\}$, $X^{[y]}$ has the same distribution as X conditioned on $Y = y$. Since $X^{[1]} \overset{\text{d}}{=} -X^{[-1]}$ by assumption, conditioning on Y we obtain that

$$\nabla_\theta R(\theta) = -\mathbb{E}\left[g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]})X^{[1]}\right]. \tag{80}$$

Observe that

$$\begin{aligned} & \|\mathbb{E}[g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]})X^{[1]}] - \mathbb{E}[X^{[1]}]\| && (81) \\ & = \|\mathbb{E}[(g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - 1)X^{[1]}]\| && (82) \\ & \leq \mathbb{E}[|g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - 1|\|X^{[1]}\|], && (83) \end{aligned}$$

where we used the convexity of the norm and Jensen's inequality. Since $\theta, X \in B^d(r)$, it can be verified that

$$\sigma(-r^2)^2 \leq g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) \leq 1. \tag{84}$$

Hence,

$$\begin{aligned} & \|\mathbb{E}[g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]})X^{[1]}] - \mathbb{E}[X^{[1]}]\| && (85) \\ & \leq (1-\sigma(-r^2)^2)\mathbb{E}[\|X^{[1]}\|]. && (86) \end{aligned}$$

By the triangle inequality, we obtain that

$$\sigma(-r^2)^2\mathbb{E}[\|X^{[1]}\|] \leq \|\nabla_\theta R(\theta)\|. \tag{87}$$

By assumption, $\|\mathbb{E}(X)\| \neq 0$, hence $R(\theta) > \epsilon$ for all θ, where

$$\epsilon := \sigma(-r^2)^2\mathbb{E}[\|X^{[1]}\|]. \tag{88}$$

Therefore, by vacuity, $R(\theta)$ satisfies (69) for every $\eta > 0$, i.e., $R(\theta)$ is (ϵ, η)-strongly Morse.

$$\begin{aligned} F_3(\alpha,\theta,x,y) = \frac{1-y}{2} & \Big[g_\theta(1-g_\theta)^{3-\frac{1}{\alpha}} - \left(4+\frac{1}{\alpha}\right)g_\theta^2(1-g_\theta)^{2-\frac{1}{\alpha}} + \left(1-\frac{1}{\alpha}\right)^2 g_\theta^3(1-g_\theta)^{1-\frac{1}{\alpha}}\Big] \\ & - \frac{1+y}{2}\Big[g_\theta^{3-\frac{1}{\alpha}}(1-g_\theta) - \left(4+\frac{1}{\alpha}\right)(1-g_\theta)^2 g_\theta^{2-\frac{1}{\alpha}} + \left(1-\frac{1}{\alpha}\right)^2(1-g_\theta)^3 g_\theta^{1-\frac{1}{\alpha}}\Big] \end{aligned} \tag{78}$$

F. Proof of Corollary 1

We start by proving that, almost surely,

$$\lim_{n\to\infty} R_{l^\alpha}(g_{\hat{\theta}_n}) = \min_{\theta\in\Theta} R_{l^\alpha}(g_\theta). \tag{89}$$

Let θ^* be a minimizer of the expected risk, i.e.,

$$R_{l^\alpha}(g_{\theta^*}) = \min_{\theta\in\Theta} R_{l^\alpha}(g_\theta). \tag{90}$$

Observe that

$$0 \leq R_{l^\alpha}(g_{\hat{\theta}_n}) - R_{l^\alpha}(g_{\theta^*}) = \mathrm{I}_n + \mathrm{II}_n, \tag{91}$$

where $\mathrm{I}_n := R_{l^\alpha}(g_{\hat{\theta}_n}) - \hat{R}_{l^\alpha}(g_{\hat{\theta}_n})$ and $\mathrm{II}_n := \hat{R}_{l^\alpha}(g_{\hat{\theta}_n}) - R_{l^\alpha}(g_{\theta^*})$. After some straightforward manipulations, (25) implies that, for every $\epsilon > 0$,

$$\mathbb{P}\left(|R_{l^\alpha}(g_{\hat{\theta}_n}) - \hat{R}_{l^\alpha}(g_{\hat{\theta}_n})| > \epsilon\right) < 4mne^{-n\epsilon^2/(2C_\alpha^2)}, \tag{92}$$

whenever n is large enough. A routine application of the Borel-Cantelli lemma shows that, almost surely,

$$\lim_{n\to\infty} \mathrm{I}_n = \lim_{n\to\infty} R_{l^\alpha}(g_{\hat{\theta}_n}) - \hat{R}_{l^\alpha}(g_{\hat{\theta}_n}) = 0. \tag{93}$$

Since $\hat{\theta}_n$ is a minimizer of the empirical risk $\hat{R}_{l^\alpha}$,

$$\mathrm{II}_n = \hat{R}_{l^\alpha}(g_{\hat{\theta}_n}) - R_{l^\alpha}(g_{\theta^*}) \leq \hat{R}_{l^\alpha}(g_{\theta^*}) - R_{l^\alpha}(g_{\theta^*}). \tag{94}$$

By Hoeffding's inequality, for every $\epsilon > 0$,

$$\mathbb{P}\left(|\hat{R}_{l^\alpha}(g_{\theta^*}) - R_{l^\alpha}(g_{\theta^*})| > \epsilon\right) \leq 2e^{-2n(\alpha-1)^2\epsilon^2/\alpha^2}. \tag{95}$$

Hence, the Borel-Cantelli lemma implies that, almost surely,

$$\lim_{n\to\infty} |\hat{R}_{l^\alpha}(g_{\theta^*}) - R_{l^\alpha}(g_{\theta^*})| = 0. \tag{96}$$

In particular, we have that, almost surely,

$$\limsup_{n\to\infty} \mathrm{II}_n \leq 0. \tag{97}$$

By plugging (93) and (97) in (91), we obtain that, almost surely,

$$0 \leq \limsup_{n\to\infty} \left[R_{l^\alpha}(g_{\hat{\theta}_n}) - R_{l^\alpha}(g_{\theta^*})\right] \leq 0, \tag{98}$$

from which (89) follows.

For each $n \in \mathbb{N}$, let $f_n : \mathcal{X} \to \overline{\mathbb{R}}$ be given by $f_n(x) = \hat{\theta}_n \cdot x$. Since $f_n(x) = \sigma^{-1}(\sigma(\hat{\theta}_n \cdot x)) = \sigma^{-1}(g_{\hat{\theta}_n}(x))$, Proposition 2 and (89) imply that

$$\lim_{n\to\infty} R_{\tilde{l}^\alpha}(f_{\hat{\theta}_n}) = \min_{\theta\in\Theta} R_{\tilde{l}^\alpha}(f_\theta) =: R^*_{\tilde{l}^\alpha}. \tag{99}$$

Since $\tilde{l}^\alpha$ is classification-calibrated, as established in Theorem 1, Proposition 1 and (99) imply that

$$\lim_{n\to\infty} R(\hat{\theta}_n) = R^*, \tag{100}$$

as required.