A Tunable Loss Function for Robust Classification: Calibration, Landscape, and Generalization

Tyler Sypherd, *Graduate Student Member, IEEE*, Mario Diaz, *Member, IEEE*,
John Kevin Cava, *Graduate Student Member, IEEE*, Gautam Dasarathy, *Senior Member, IEEE*,
Peter Kairouz, *Member, IEEE*, and Lalitha Sankar, *Senior Member, IEEE*

Abstract—We introduce a tunable loss function called α-loss, parameterized by $\alpha \in (0, \infty]$, which interpolates between the exponential loss ($\alpha = 1/2$), the log-loss ($\alpha = 1$), and the 0-1 loss ($\alpha = \infty$), for the machine learning setting of classification. Theoretically, we illustrate a fundamental connection between α-loss and Arimoto conditional entropy, verify the classification-calibration of α-loss in order to demonstrate asymptotic optimality via Rademacher complexity generalization techniques, and build-upon a notion called strictly local quasi-convexity in order to quantitatively characterize the optimization landscape of α-loss. Practically, we perform class imbalance, robustness, and classification experiments on benchmark image datasets using convolutional-neural-networks. Our main practical conclusion is that certain tasks may benefit from tuning α-loss away from log-loss ($\alpha = 1$), and to this end we provide simple heuristics for the practitioner. In particular, navigating the α hyperparameter can readily provide superior model robustness to label flips ($\alpha > 1$) and sensitivity to imbalanced classes ($\alpha < 1$).

Index Terms—α-loss, Arimoto conditional entropy, robustness, classification-calibration, strictly local quasi-convexity, generalization.

I. Introduction

IN THE context of machine learning, the performance of a classification algorithm, in terms of accuracy, tractability, and convergence guarantees crucially depends on the choice of the loss function during training [3], [4]. Consider a feature vector $X \in \mathcal{X}$, an unknown finite-valued label $Y \in \mathcal{Y}$, and a hypothesis $h : \mathcal{X} \rightarrow \mathcal{Y}$. The canonical 0-1 loss, given by $\mathbb{1}[h(X) \neq Y]$, is considered an ideal loss function in the classification setting 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 applicability in state-of-the-art learning algorithms is highly restricted [5]. As a consequence, *surrogate* loss functions that approximate the 0-1 loss such as log-loss, exponential loss, sigmoid loss, etc. have garnered much interest [6]–[16].

In the field of information-theoretic privacy, Liao *et al.* recently introduced a tunable loss function called α-loss for $\alpha \in [1, \infty]$ to model the inferential capacity of an adversary to obtain private attributes [17]–[19]. For $\alpha = 1$, α-loss reduces to log-loss which models a belief-refining adversary; for $\alpha = \infty$, α-loss reduces to the probability of error which models an adversary that makes hard decisions. Using α-loss, Liao *et al.* in [17] derived a new privacy measure called α-*leakage* which continuously interpolates between Shannon's mutual information [20] and maximal leakage introduced by Isaa *et al.* [21]; indeed, Liao *et al.* showed that α-leakage is equivalent to the Arimoto mutual information [22]. In this paper, we extend α-loss to the range $\alpha \in (0, \infty]$ and propose it as a tunable *surrogate* loss function for the ideal 0-1 loss in the machine learning setting of classification. Through our extensive analysis, we argue that: 1) since α-loss continuously interpolates between the exponential ($\alpha = 1/2$), log ($\alpha = 1$), and 0-1 ($\alpha = \infty$) losses and is related to the Arimoto conditional entropy, it is theoretically an object of interest in its own right; 2) navigating the convexity/robustness trade-offs inherent in the α hyperparameter offers significant practical improvements over log-loss, which is a canonical loss function in classification, and can be done quickly and effectively.

A. Related Work

The study and implementation of tunable utility (or loss) metrics which continuously interpolate between useful quantities is a persistent theme in information theory, networking, and machine learning. In information theory, Rényi entropy generalized the Shannon entropy [23], and Arimoto extended the Rényi entropy to conditional distributions [24]. This led to the α-mutual information [22], [25], which is directly related

to a recently introduced privacy measure called α-leakage [17]. More recently in networking, Mo *et al.* introduced α-fairness in [26], which is a tunable utility metric that alters the value of different edge users; similar ideas have recently been studied in the federated learning setting [27]. Even more recently in machine learning, Barron in [15] presented a tunable extension of the l_2 loss function, which interpolates between several known l_2-type losses and has similar convexity/robustness themes as this work. Presently, there is a need in the machine learning setting of *classification* for alternative losses to the cross-entropy loss (one-hot encoded log-loss) [28]. We propose α-loss, which continuously interpolates between the exponential, log, and 0-1 losses, as a viable solution.

In order to evaluate the statistical efficacy of loss functions in the learning setting of classification, Bartlett *et al.* proposed the notion of *classification-calibration* in a seminal paper [6]. *Classification-calibration* is analogous to pointwise Fisher consistency in that it requires that the minimizer of the conditional expectation of a loss function agrees in sign with the Bayes predictor for every value of the feature vector. A more restrictive notion called *properness* requires that the minimizer of the conditional expectation of a loss function exactly replicates the true posterior [29]–[31]. *Properness* of a loss function is a necessary condition for efficacy in the class probability estimation setting (see, e.g., [31]), but for the classification setting which is the focus of this work, the notion of *classification-calibration* is sufficient. In the sequel, we find that the margin-based form of α-loss is classification-calibrated for all $\alpha \in (0, \infty]$ and thus satisfies this necessary condition for efficacy in binary classification.

While early research was predominantly focused on convex losses [6], [8]–[10], more recent works propose the use of non-convex losses as a means to moderate the behavior of an algorithm [7], [11], [15], [32]. This is due to the increased robustness non-convex losses offer over convex losses [15], [32], [33] and the fact that modern learning models (e.g., deep learning) are inherently non-convex as they involve vast functional compositions [34]. There have been numerous theoretical attempts to capture the non-convexity of the optimization landscape which is the loss surface induced by the learning model, underlying distribution, and the surrogate loss function itself [32], [35]–[41]. To this end, Hazan *et al.* [35] introduce the notion of *strictly local quasi-convexity* (SLQC) to parametrically quantify approximately quasi-convex functions, and provide convergence guarantees for the Normalized Gradient Descent (NGD) algorithm (originally introduced in [42]) for such functions. Through a quantification of the SLQC parameters of the expected α-loss, we provide some estimates that strongly suggest that the degree of convexity increases as α decreases less than 1 (log-loss); conversely, the degree of convexity decreases as α increases greater than 1. Thus, we find that there exists a trade-off inherent in the choice of $\alpha \in (0, \infty]$, i.e., trade convexity (and hence optimization speed) for robustness and vice-versa. Since increasing the degree of convexity of the optimization landscape is conducive to faster optimization, our approach could serve as an alternative to other approaches whose objective is to accelerate

the optimization process, e.g., the activation function tuning in [43]–[45] and references therein.

Understanding the generalization capabilities of learning algorithms stands as one of the key problems in theoretical machine learning. A classical approach to this problem consists in deriving algorithm independent generalization bounds, mainly relying on the notion of Rademacher complexity [4, Ch. 26]. A recent line of research, initiated by the works of Russo and Zou [46] and Xu and Raginsky [47], aims to improve generalization bounds by considering the statistical dependency between the input and the output of a given learning algorithm. While there are many extensions and refinements, e.g., [48]–[54], these results are inherently algorithm dependent which makes them hard to instantiate and obfuscates the role of the loss function. Hence, in this work we rely on classical Rademacher complexity tools to provide algorithm independent generalization bounds that lead to the asymptotic optimality of α-loss w.r.t. the 0-1 loss.

There are a few proposed tunable loss functions for the classification setting in the literature [11], [55]–[57]. Notably, the symmetric cross entropy loss introduced by Wang *et al.* in [55] proposes the tunable linear combination of the usual cross entropy loss with the so-called reverse cross entropy loss, which essentially reverses the roles of the one-hot encoded labels and soft prediction of the model. Wang *et al.* report gains under symmetric and asymmetric noisy labels, particularly in the very high noise regime. Another approach introduced by Amid *et al.* in [56] is a bi-tempered logistic loss, which is based on Bregman divergences. As the name suggests, the bi-tempered logistic loss depends on two temperature hyperparameters, which Amid *et al.* show improvements over vanilla cross-entropy loss again on noisy data. Recently, Li *et al.* introduced tilted empirical risk minimization [57], a framework which parametrically generalizes empirical risk minimization using a log-exponential transformation to induce fairness or robustness in the model. Contrasting with this work, we note that our study is exclusively focused on α-loss acting within empirical risk minimization. Summing up, the main distinctions that differentiate this work from related work are that α-loss has a fundamental relationship to the Arimoto conditional entropy, continuously interpolates between the exponential, log, and 0-1 losses, and provides robustness to noisy labels *and* sensitivity to imbalanced classes. Lastly, we note that α-loss has also been recently studied in the context of generative adversarial networks [58] and boosting [59].

B. Contributions

The following are the main contributions of this paper:

- We formulate α-loss in the classification setting, extending it to $\alpha \in (0, 1)$, and we thereby extend the result of Liao *et al.* in [17] which characterizes the relationship between α-loss and the Arimoto conditional entropy.
- For binary classification, we define a margin-based form of α-loss and demonstrate its equivalence to α-loss for all $\alpha \in (0, \infty]$. We then characterize convexity and verify statistical calibration of the margin-based α-loss for $\alpha \in (0, \infty]$. We next derive the minimum conditional

risk of the margin-based α-loss, which we show recovers the relationship between α-loss and the Arimoto conditional entropy for all $\alpha \in (0, \infty]$. Lastly, we provide synthetic experiments on a two-dimensional Gaussian mixture model with asymmetric label flips and class imbalances, where we train linear predictors with α-loss for several values of α.

- For the logistic model in binary classification, we show that the expected α-loss is convex in the logistic parameter for $\alpha \leq 1$ (strongly-convex when the covariance matrix is positive definite), and we show that it retains convexity as α increases greater than 1 provided that the radius of the parameter space is small enough. We provide a point-wise extension of *strictly local quasi-convexity* (SLQC) by Hazan *et al.*, and we reformulate SLQC into a more tractable inequality using a geometric inequality which may be of independent interest. Using a bootstrapping technique which also may be of independent interest, we provide bounds in order to quantify the evolution of the SLQC parameters as α increases.

- Also for the logistic model in binary classification, we characterize the generalization capabilities of α-loss. To this end, we employ standard Rademacher complexity generalization techniques to derive a uniform generalization bound for the logistic model trained with α-loss for $\alpha \in (0, \infty]$. We then combine a result by Bartlett *et al.* and our uniform generalization bound to show (under standard distributional assumptions) that the minimizer of the empirical α-loss is asymptotically optimal with respect to the expected 0-1 loss (probability of error), which is the ideal metric in classification problems.

- Finally, we perform symmetric noisy label and class imbalance experiments on MNIST, FMNIST, and CIFAR-10 using convolutional-neural-networks. We show that models trained with α-loss can either be more robust or sensitive to outliers (depending on the application) over models trained with log-loss ($\alpha = 1$). Following some of our theoretical intuitions, we demonstrate the "Goldilocks zone" of $\alpha \in (0, \infty]$, i.e., for most applications $\alpha^* \in [.8, 8]$. Thus, we argue that α-loss is an effective generalization of log-loss (cross-entropy loss) for classification problems in modern machine learning.

Different subsets of the authors published portions of this paper as conference proceedings in [1] and [2]. Specifically, results provided in [1] primarily comprise a subset of the second bullet in the list above, however, this work extends those published results to $\alpha \in (0, 1)$, clarifies the relationship to Arimoto conditional entropy, and provides synthetic experiments; in addition, results in [2] primarily comprise a subset of the third bullet in the list above, however, this work provides a new convexity result for $\alpha > 1$, provides SLQC background material including a point-wise statement and proof of Lemma 1, and utilizes a bootstrapping argument which significantly improves the bounds in [2]. The remaining three bullets are all comprised of unpublished work.
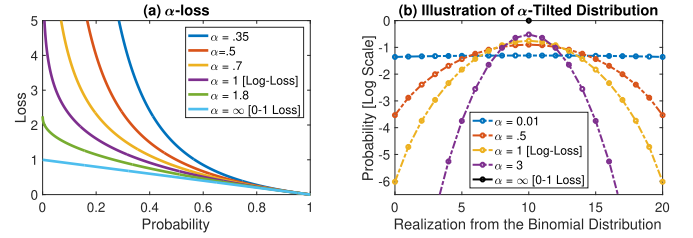

Fig. 1. **(a)** α-loss (2) as a function of the probability for several values of α; **(b)** α-tilted posterior (6) for several values of α where the true underlying distribution is the (20,0.5)-binomial distribution.

II. INFORMATION-THEORETIC MOTIVATIONS OF α-LOSS

Consider a pair of discrete random variables denoted $(X, Y) \sim P_{X,Y}$. Observing X, one can construct an estimate $\hat{Y}$ of Y such that $Y - X - \hat{Y}$ form a Markov chain. It is possible to evaluate the fitness of a given estimate $\hat{Y}$ using a loss function $\ell : \mathcal{Y} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_+$ via the expectation

$$\mathbb{E}_{X,Y}\left[\ell(Y, P_{\hat{Y}|X})\right], \qquad (1)$$

where $\hat{Y}|X \sim P_{\hat{Y}|X}$ is the *learner's* posterior estimate of Y given knowledge of X; for simplicity we sometimes abbreviate $P_{\hat{Y}|X=x}$ as $\hat{P}$ when the context is clear. In [17], Liao *et al.* proposed the definition of α-loss for $\alpha \in [1, \infty]$ in order to quantify adversarial action in the information leakage context. We adapt and extend the definition of α-loss to $\alpha \in (0, \infty]$ in order to study the efficacy of the loss function in the machine learning setting.

Definition 1: Let $\mathcal{P}(\mathcal{Y})$ be the set of probability distributions over $\mathcal{Y}$. For $\alpha \in (0, 1) \cup (1, \infty)$, we define α-loss, denoted by $l^\alpha : \mathcal{Y} \times \mathcal{P}(\mathcal{Y}) \to \mathbb{R}_+$, as

$$l^\alpha(y, \hat{P}) := \frac{\alpha}{\alpha - 1}\left(1 - \hat{P}(y)^{1-1/\alpha}\right), \qquad (2)$$

and, by continuous extension, $l^1(y, \hat{P}) := -\log \hat{P}(y)$ and $l^\infty(y, \hat{P}) := 1 - \hat{P}(y)$.

Note that for $(y, \hat{P})$ fixed, $l^\alpha(y, \hat{P})$ is continuous and monotonically decreasing in α. Also note that l^1 recovers log-loss, and plugging in $\alpha = 1/2$ yields $l^{1/2}(y, \hat{P}) := \hat{P}^{-1}(y) - 1$. One can use expected α-loss $\mathbb{E}_{X,Y}[l^\alpha(Y, P_{\hat{Y}|X})]$, hence called α-risk, to quantify the effectiveness of the estimated posterior $P_{\hat{Y}|X}$. In particular,

$$\mathbb{E}_{X,Y}\left[l^1(Y, P_{\hat{Y}|X})\right] = \mathbb{E}_X\left[H(P_{Y|X=x}, P_{\hat{Y}|X=x})\right], \qquad (3)$$

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

$$\mathbb{E}_{X,Y}[l^\infty(Y, P_{\hat{Y}|X})] = \mathbb{P}[Y \neq \hat{Y}], \qquad (4)$$

i.e., the expected α-loss for $\alpha = \infty$ equals the probability of error. Recall that the expectation of the canonical 0-1 loss, $\mathbb{E}_{X,Y}[\mathbb{1}[Y \neq \hat{Y}]]$, also recovers the probability of error [4]. For this reason, we sometimes refer to l^∞ as the 0-1 loss.

Observe that α-loss presents a tunable class of loss functions that value the probabilistic estimate of the label differently as a function of α; see Fig. 1(a). In the sequel, we find that, when composed with a sigmoid, $l^{1/2}, l^1, l^\infty$ become the

exponential, logistic, and sigmoid (smooth 0-1) losses, respectively. While we note that there may be infinitely many ways to continuously interpolate between the exponential, log, and 0-1 losses, we observe that the interpolation introduced by α-*loss* is monotonic in α, seems to provide an information-theoretic interpretation (Proposition 1), and also appears to be apt for the classification setting which will be further elaborated in the sequel. The following result was shown by Liao *et al.* in [17] for $\alpha \in [1, \infty]$ and provides an explicit characterization of the optimal risk-minimizing posterior under α-loss. We extend the result to $\alpha \in (0, 1)$.

Proposition 1: For each $\alpha \in (0, \infty]$, the minimal α-risk is

$$\min_{P_{\hat{Y}|X}} \mathbb{E}_{X,Y}\left[l^\alpha(Y, P_{\hat{Y}|X})\right] = \frac{\alpha}{\alpha-1}\left(1 - e^{\frac{1-\alpha}{\alpha}H_\alpha^A(Y|X)}\right),$$
(5)

where $H_\alpha^A(Y|X) := \frac{\alpha}{1-\alpha} \log \sum_x \left(\sum_y P_{X,Y}(x,y)^\alpha\right)^{1/\alpha}$ is the Arimoto conditional entropy of order α [60]. The resulting unique minimizer, $\hat{P}_\alpha^*$, is the α-tilted true posterior

$$\hat{P}_\alpha^*(y|x) = \frac{P_{Y|X}(y|x)^\alpha}{\sum_y P_{Y|X}(y|x)^\alpha}.$$
(6)

The proof of Proposition 1 for $\alpha \in [1, \infty]$ can be found in [17] and is readily extended to the case where $\alpha \in (0, 1)$ with similar techniques. Through Proposition 1, we note that α-loss exhibits different operating conditions through the choice of α. Observe that the minimizer of (5) given by the α-tilted distribution in (6) recovers the true posterior only if $\alpha = 1$, i.e., for log-loss. Further, as α decreases from 1 towards 0, α-loss places increasingly higher weights on the low probability outcomes; on the other hand as α increases from 1 to ∞, α-loss increasingly limits the effect of the low probability outcomes. Ultimately, we find that for $\alpha = \infty$, minimizing the corresponding risk leads to making a single guess on the most likely label, i.e., MAP decoding. See Fig. 1(b) for an illustration of the α-tilted distribution on a (20,0.5)-Binomial distribution. Intuitively, empirically minimizing α-loss for $\alpha \neq 1$ could be a boon for learning the minority class ($\alpha < 1$) or ignoring label noise ($\alpha > 1$); see Section VI for experimental consideration of such class imbalance and noisy label trade-offs.

With the information-theoretic motivations of α-loss behind us, we now consider the setting of binary classification, where we study the optimization, statistical, and robustness properties of α-loss.

III. α-LOSS IN BINARY CLASSIFICATION

In this section, we study the role of α-loss in binary classification. First, we provide its margin-based form, which we show is intimately related to the original α-loss formulation in Definition 1; next, we analyze the optimization characteristics and statistical properties of the margin-based α-loss where we notably recover the relationship between α-loss and the Arimoto conditional entropy in the margin setting; finally, we comment on the robustness and sensitivity trade-offs which are inherent in the choice of α through theoretical discussion

and experimental considerations. First, however, we formally discuss the binary classification setting through the role of classification functions and surrogate loss functions.

In binary classification, the learner ideally wants to obtain a classifier $h : \mathcal{X} \to \{-1, +1\}$ that minimizes the probability of error, or the risk (expectation) of the 0-1 loss, given by

$$R(h) = \mathbb{P}[h(X) \neq Y],$$
(7)

where the true 0-1 loss given by $\mathbb{1}[h(X) \neq Y]$. Unfortunately, this optimization problem is NP-hard [5]. Therefore, the problem is typically relaxed by imposing restrictions on the space of possible classifiers and by choosing surrogate loss functions with desirable properties. Thus during the training phase, it is common to optimize a surrogate loss function over classification functions of the form $f : \mathcal{X} \to \overline{\mathbb{R}}, \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, whose output captures the certainty of a model's prediction of the true underlying binary label $Y \in \{-1, 1\}$ associated with X [1], [3], [4], [6]–[9], [61]. Once a suitable classification function has been chosen, the classifier is obtained by making a hard decision, i.e., the model outputs the classification $h(X) = \text{sign}(f(X))$, in order to predict the true underlying binary label $Y \in \{-1, 1\}$ associated with the feature vector $X \in \mathcal{X}$. Examples of learning algorithms which optimize surrogate losses over classification functions include SVM (hinge loss), logistic regression (logistic loss), and AdaBoost (exponential loss), to name a few [3]. With the notions of classification functions and surrogate loss functions in hand, we now turn our attention to an important family of surrogate loss functions in binary classification.

A. Margin-Based α-Loss

Here, we provide the definition of α-loss in binary classification and characterize its relationship to the form presented in Definition 1. First, we discuss an important family of loss functions in binary classification called *margin-based* losses.

A loss function is said to be margin-based if, for all $x \in \mathcal{X}$ and $y \in \{-1, +1\}$, the loss associated to a pair $(y, f(x))$ is given by $\tilde{l}(yf(x))$ for some function $\tilde{l} : \overline{\mathbb{R}} \to \mathbb{R}_+$ [6]–[9], [28]. In this case, the loss of the pair $(y, f(x))$ only depends on the product $z := yf(x)$, the (unnormalized) margin [61]. 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. We now provide the margin-based form of α-loss, which is illustrated in Fig. 2(a).

Definition 2: For $\alpha \in (0, 1) \cup (1, \infty)$, we define the margin-based α-loss, $\tilde{l}^\alpha : \overline{\mathbb{R}} \to \mathbb{R}_+$, as

$$\tilde{l}^\alpha(z) := \frac{\alpha}{\alpha-1}\left(1 - \left(1 + e^{-z}\right)^{1/\alpha-1}\right),$$
(8)

and, by continuous extension, $\tilde{l}^1(z) = \log(1 + e^{-z})$ and $\tilde{l}^\infty(z) = (1 + e^z)^{-1}$.

Note that $\tilde{l}^{1/2}(z) = e^{-z}$. Thus, $\tilde{l}^{1/2}$, $\tilde{l}^1$, and $\tilde{l}^\infty$ recover the exponential, logistic, and sigmoid losses, respectively. Navigating the various regimes of α induces different optimization, statistical, and robustness characteristics for the margin-based
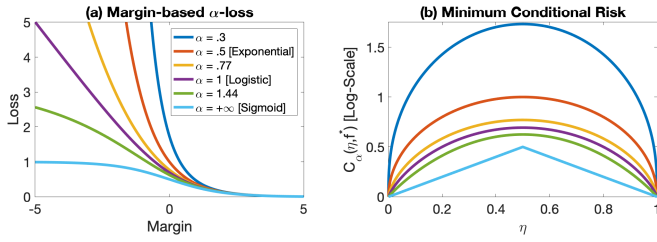
Fig. 2. **(a)** Margin-based α-loss (8) as a function of the margin ($z := yf(x)$) for $\alpha \in \{0.3, 0.5, 0.77, 1, 1.44, \infty\}$; **(b)** Minimum conditional risk (14) for the same values of α.

α-loss; this is elaborated in the sequel. First, we discuss its relationship to the original form in Definition 1, which requires alternative prediction functions to classification functions called soft classifiers.

In binary classification, it is also common to use soft classifiers $g : \mathcal{X} \rightarrow [0,1]$ which encode the conditional distribution, namely, $g(x) := P_{\hat{Y}|X}(1|x)$. In essence, soft classifiers capture a model's *belief* of $Y|X$ [1], [4], [34]. Similar to the classification function setting, the hard decision of a soft classifier is obtained by $h(x) = \text{sign}(g(x) - 1/2)$. Log-loss, and by extension α-loss as given in Definition 1, are examples of loss functions which act on soft classifiers. In practice, a soft classifier can be obtained by composing a classification function with the logistic sigmoid function $\sigma : \overline{\mathbb{R}} \rightarrow [0,1]$ given by

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

which is generalized by the softmax function in the multiclass setting [34]. Observe that σ is invertible and $\sigma^{-1} : [0,1] \rightarrow \overline{\mathbb{R}}$ is given by

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

which is often referred to as the logistic link [31].

With these two transformations, one is able to map classification functions to soft classifiers and vice-versa. Thus, a loss function in one domain is readily transformed into a loss function in the other domain. In particular, we are now in a position to derive the correspondence between α-loss in Defintion 1 and the margin-based α-loss in Definition 2, which generalizes our previous proof in [1].

Proposition 2: Consider a soft classifier $g(x) = P_{\hat{Y}|X}(1|x)$. If $f(x) = \sigma^{-1}(g(x))$, then, for every $\alpha \in (0, \infty]$,

$$l^{\alpha}(y, g(x)) = \tilde{l}^{\alpha}(yf(x)). \qquad (11)$$

Conversely, if f is a classification function, then the soft classifier $g(x) := \sigma(f(x))$ satisfies (11). In particular, for every $\alpha \in (0, \infty]$,

$$\min_g \mathbb{E}_{X,Y}(l^{\alpha}(Y, g(x))) = \min_f \mathbb{E}_{X,Y}(\tilde{l}^{\alpha}(Yf(X))). \qquad (12)$$

Therefore, there is a direct correspondence between α-loss in Definition 1 and the margin-based α-loss which is used in binary classification.

Remark 1: Instead of the fixed inverse link function (9), it is also possible to use any other fixed inverse link function,

or even inverse link functions dependent on α; indeed, it is possible to derive many such tunable margin-based losses this way. However, the margin-based α-loss as given in Definition 2 allows for continuous interpolation between the exponential, logistic, and sigmoid losses, and thus motivates our choice of the fixed sigmoid in (9) as the inverse link.

The following result, which quantifies the convexity of the margin-based α-loss, will be useful in characterizing the convexity of the average loss, or *landscape*, in the sequel.

Proposition 3: As a function of the margin, $\tilde{l}^{\alpha} : \overline{\mathbb{R}} \rightarrow \mathbb{R}_+$ is convex for $\alpha \leq 1$ and quasi-convex for $\alpha > 1$.

Recall that a real-valued function $f : \mathbb{R} \rightarrow \mathbb{R}$ is quasi-convex if, for all $x, y \in \mathbb{R}$ and $\lambda \in [0,1]$, we have that $f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}$, and also recall that any monotonic function is quasi-convex (see e.g., [62]). Intuitively through Fig. 2(a), we find that the quasi-convexity of the margin-based α-loss for $\alpha > 1$ reduces the penalty induced during training for examples which have a negative margin; this has implications for robustness that will also be investigated in the sequel.

B. Calibration of Margin-Based α-Loss

With the definition and basic properties of the margin-based α-loss in hand, we now discuss a statistical property of the margin-based α-loss that highlights its suitability in binary classification. Bartlett *et al.* in [6] introduce *classification-calibration* as a means to compare the performance of a margin-based loss function relative to the 0-1 loss by inspecting the minimizer of its conditional risk. Formally, let $\phi : \mathbb{R} \rightarrow \mathbb{R}_+$ denote a margin-based loss function and let $C_{\phi}(\eta(x), f(x)) = \mathbb{E}[\phi(Yf(X))|X = x]$ denote its conditional expectation (risk), where $\eta(x) = P_{Y|X}(1|x)$ is the true posterior and $f : \mathcal{X} \rightarrow \overline{\mathbb{R}}$ is a classification function. Thus, the conditional risk of the margin-based α-loss for $\alpha \in (0, \infty]$ is given by

$$C_{\alpha}(\eta(x), f(x)) = \mathbb{E}_Y[\tilde{l}^{\alpha}(Yf(X))|X = x]. \qquad (13)$$

We say that $\phi : \overline{\mathbb{R}} \rightarrow \mathbb{R}_+$ is classification-calibrated if, for all $x \in \mathcal{X}$, its minimum conditional risk

$$\inf_{f:\mathcal{X}\rightarrow\mathbb{R}} C_{\phi}(\eta(x), f(x))$$
$$= \inf_{f:\mathcal{X}\rightarrow\mathbb{R}} \eta(x)\phi(f(x)) + (1-\eta(x))\phi(-f(x)), \qquad (14)$$

is attained by a $f^* : \mathcal{X} \rightarrow \overline{\mathbb{R}}$ such that

$$\text{sign}(f^*(x)) = \text{sign}(2\eta(x) - 1). \qquad (15)$$

In words, a margin-based loss function is classification-calibrated if for each feature vector, the minimizer of its minimum conditional risk agrees in sign with the Bayes optimal predictor. Note that this is a pointwise form of Fisher consistency [6], [8].

The expectation of the loss function ϕ, or the ϕ-risk, is denoted $R_{\phi}(f) = \mathbb{E}_X[C_{\phi}(\eta(X), f(X))]$; this notation will be useful in the sequel when we quantify the asymptotic behavior of α-loss. Finally, as is common in the literature [6], [7], we omit the dependence of η and f on x, and we also let $C_{\phi}^*(\eta) = C_{\phi}(\eta, f^*)$ for notional convenience.

With the necessary background on classification-calibrated loss functions in hand, we are now in a position to show that $\tilde{l}^\alpha$ is classification-calibrated for all $\alpha \in (0, \infty]$.

Theorem 1: For $\alpha \in (0, \infty]$, the margin-based α-loss $\tilde{l}^\alpha$ is classification-calibrated. In addition, its optimal classification function is given by

$$f_\alpha^*(\eta) = \alpha \cdot \sigma^{-1}(\eta) = \alpha \log\left(\frac{\eta}{1-\eta}\right). \tag{16}$$

See Appendix A for full proof details. Examining the optimal classification function in (16) more closely, we observe that this expression is readily derived from the α-tilted distribution for a binary label set in Proposition 2. Thus, analogous to the intuitions regarding the α-tilted distribution in (6), the optimal classification function in (16) suggests that $\alpha > 1$ is more robust to slight fluctuations in η and $\alpha < 1$ is more sensitive to slight fluctuations in η. In the sequel, we find that this has practical implications for noisy labels and class imbalances.

Upon plugging (16) into (13), we get the next result which specifies the minimum conditional risk of $\tilde{l}^\alpha$ for $\alpha \in (0, \infty]$.

Corollary 1: For $\alpha \in (0, \infty]$, the minimum conditional risk $C_\alpha^*(\eta)$ of $\tilde{l}^\alpha$ is equal to

$$\begin{cases} \frac{\alpha}{\alpha-1}\left(1 - (\eta^\alpha + (1-\eta)^\alpha)^{1/\alpha}\right) & \alpha \in (0,1) \cup (1,+\infty), \\ -\eta \log \eta - (1-\eta) \log (1-\eta) & \alpha = 1, \\ \min\{\eta, 1-\eta\} & \alpha \to +\infty. \end{cases} \tag{17}$$

Remark 2: Observe that in (17) for $\alpha = 1$, the minimum conditional risk can be rewritten as

$$C_1^*(\eta) = -\eta \log \eta - (1-\eta) \log (1-\eta) \tag{18}$$
$$= H(Y|X = x), \tag{19}$$

where $H(Y|X = x)$ is the Shannon conditional entropy for a Y given $X = x$ [63]. For $\alpha \in (0,1) \cup (1,+\infty)$, also note that in (17), the minimum conditional risk can be rewritten as

$$C_\alpha^*(\eta) = \frac{\alpha}{\alpha-1}\left[1 - (\eta^\alpha + (1-\eta)^\alpha)^{1/\alpha}\right] \tag{20}$$
$$= \frac{\alpha}{\alpha-1}\left[1 - e^{\frac{1-\alpha}{\alpha} H_\alpha^A(Y|X=x)}\right], \tag{21}$$

where $H_\alpha^A(Y|X = x) = \frac{1}{1-\alpha} \log\left(\sum_y P_{Y|X}(y|x)^\alpha\right)$ is the Arimoto conditional entropy of order α [60]. Finally, observe that $\mathbb{E}_X[C_\alpha^*(\eta(X))]$ recovers (5) in Proposition 1.

Finally, note that the minimum conditional risk of the margin-based α-loss is concave for all $\alpha \in (0, \infty]$ (see Fig. 2(b)); indeed, this is known to be a useful property for classification problems [7]. Therefore, since the margin-based α-loss is classification-calibrated and its minimum conditional risk is concave for all $\alpha \in (0, \infty]$, it seems to have reasonable statistical behavior for binary classification problems. We now turn our attention to the robustness and sensitivity tradeoffs induced by traversing the different regimes of α for the margin-based α-loss.
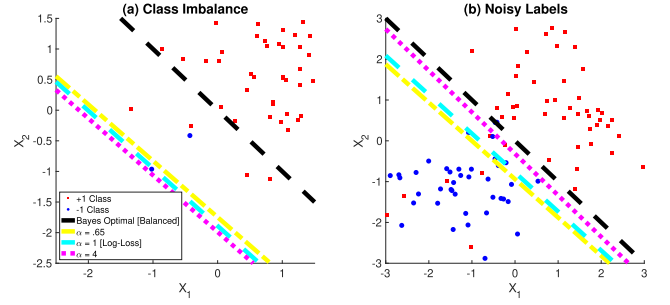

Fig. 3. Two synthetic experiments each averaged over 100 runs highlighting the differences in trained linear predictors of α-loss for $\alpha \in \{0.65, 1, 4\}$ on imbalanced and noisy data, which are compared with the Bayes optimal predictor for the clean, balanced distribution. Training data present in both figures is obtained from the last run in each experiment, respectively. **(a)** Averaged linear predictors trained using α-loss on imbalanced data with 2 examples from $Y = -1$ class per run. Averaged linear predictors for smaller values of α are closer to the Bayes predictor for the balanced distribution, which highlights the sensitivity of α-loss to the minority class for $\alpha < 1$. **(b)** Averaged linear predictors trained using α-loss on noisy data, which is obtained by flipping the labels of the $Y = -1$ class with probability 0.2. Averaged linear predictor for $\alpha = 4$ is closer to the Bayes predictor for the balanced distribution, which highlights the robustness of α-loss to noise for $\alpha > 1$.

C. Robustness and Sensitivity of Margin-Based α-Loss

Despite the advantages of convex losses in terms of numerical optimization and theoretical tractability, non-convex loss functions often provide superior model robustness and classification accuracy [1], [7], [11], [15], [32], [33], [61], [64], [65]. In essence, non-convex loss functions tend to assign less weight to misclassified training examples[1] and therefore algorithms optimizing such losses are often less perturbed by outliers, i.e., examples which induce large negative margins. More concretely, consider Fig. 2(a) for $\alpha = 1/2$ (convex) and $\alpha = 1.44$ (quasi-convex), and suppose that $z_1 = -1$ and $z_2 = -5$. Plugging these parameters into Definition 2, we find that $\tilde{l}^{1/2}(z_1) = e^1 \approx 2.7$, $\tilde{l}^{1/2}(z_2) = e^5 \approx 148.4$, $\tilde{l}^{1.44}(z_1) \approx 1.1$, and $\tilde{l}^{1.44}(z_2) \approx 2.6$. In words, the difference in these loss evaluations for a negative value of the margin, which is representative of a misclassified training example, is approximately exponential versus sub-linear. Indeed, this difference appears to be most relevant for outliers (e.g., noisy or imbalanced training examples) [7], [61].

We explore these ideas with the following synthetic experiment presented in Fig. 3. We assume the practitioner has access to modified training data which approximates the true underlying distribution given by a two-dimensional Gaussian Mixture Model (2D-GMM) with equal mixing probability $\mathbb{P}[Y = -1] = \mathbb{P}[Y = +1]$, symmetric means

$$\mu_{X|Y=-1} = (-1, -1)^\mathsf{T} = -\mu_{X|Y=1}, \tag{22}$$

and shared identity covariance matrix $\Sigma = \mathbb{I}_2$. The first experiment considers the scenario where the training data suffers from a class imbalance; specifically, the number of training examples for the $Y = -1$ class is 2 and the number of training examples for the $Y = +1$ class is 98 for every run.

---

[1]Convex losses grow at least linearly with respect to the negative margin which results in an increased sensitivity to outliers. See Fig. 2(a) for $\alpha = 1$ as an example of this phenomenon.

The second experiment considers the scenario where the training data suffers from noisy labels; specifically, the labels of the $Y = -1$ class are flipped with probability 0.2 and the labels of the $Y = +1$ class are kept fixed. For both experiments we train α-loss on the logistic model, which is the generalization of logistic regression with α-loss and is formally described in the next section. Specifically, we minimize α-loss using gradient descent with the fixed *learning rate* $= 0.01$ for each $\alpha \in \{0.65, 1, 4\}$. Note that $\alpha = 0.65$ (lower limit) and $\alpha = 4$ (upper limit) were both chosen for computational feasibility in the logistic model; in practice, the range of $\alpha \in (0, \infty]$, while usually contracted as in this experiment, is dependent on the model - this is elaborated in the sequel. Training is allowed to progress until convergence as specified by the *optimality parameter* $= 10^{-4}$. The linear predictors presented in Fig. 3 are averaged over 100 runs of randomly generated data according to the parameters for each experiment.

Ideally, the practitioner would like to generate a linear predictor which is invariant to noisy or imbalanced training data and tends to align with the Bayes optimal predictor for the balanced distribution. Indeed, when the training data is balanced (and clean), all averaged linear predictors generated by α-loss collapse to the Bayes predictor; see Fig. 11 in Appendix D.2. However, training on noisy or imbalanced data affects the linear predictors of α-loss in different ways. In the class imbalance experiment in Fig. 3(a), we find that the averaged linear predictor for the smaller values of α more closely approximate the Bayes predictor for the balanced distribution, which suggests that the smaller values of α are more sensitive to the minority class. Similarly in the class noise experiment in Fig. 3(b), we find that the averaged linear predictor for $\alpha = 4$ more closely approximates the Bayes predictor for the balanced distribution, which suggests that the larger values of α are less sensitive to noise in the training data. Both results suggest that $\alpha = 1$ (log-loss) can be improved with the use of α-loss in these scenarios. For quantitative results of this experiment, including a wider range of α's, additional imbalances and noise levels, and results using the F_1 score, see Tables VII, VIII, and IX in Appendix D.2.

In summary, we find that navigating the convexity regimes of α-loss induces different robustness and sensitivity characteristics. We explore these themes in more detail on canonical image datasets in Section VI; theoretical investigations of the robustness of α-loss can be found in [59]. We now turn our attention to theoretically characterizing the optimization complexity of α-loss for the different regimes of α in the logistic model.

IV. OPTIMIZATION GUARANTEES FOR α-LOSS IN THE LOGISTIC MODEL

In this section, we analyze the optimization complexity of α-loss in the logistic model as we vary α by quantifying the convexity of the optimization landscape. First, we show that the α-risk is convex (indeed, strongly-convex if a certain correlation matrix is positive definite) in the logistic model for $\alpha \leq 1$; next, we provide a brief summary of a notion known as *strictly local quasi-convexity* (SLQC); then, we provide a

more tractable reformulation of SLQC which is instrumental for our theory; finally, we study the convexity of the α-risk in the logistic model through SLQC for a range of $\alpha > 1$, which we argue is sufficient due to the rapid saturation effect of α-loss as $\alpha \to \infty$. Notably, our main result depends on a bootstrapping argument that might be of independent interest. Our main conclusion of this section is that there exists a "Goldilocks zone" of $\alpha \in (0, \infty]$ which drastically reduces the hyperparameter search induced by α for the practitioner. Finally, note that all proofs and background material can be found in Appendix B.

A. α-Loss in the Logistic Model

Prior to stating our main results, we clarify the setting and provide necessary definitions. Let $X \in [0,1]^d$ be the normalized feature where $d \in \mathbb{N}$ is the number of dimensions, $Y \in \{-1, +1\}$ the label and we assume that the pair is distributed according to an unknown distribution $P_{X,Y}$, i.e., $(X, Y) \sim P_{X,Y}$. For $\tilde{\theta} \in \mathbb{R}^d$ and $r > 0$, we let $\mathbb{B}_d(\tilde{\theta}, r) := \{\theta \in \mathbb{R}^d : \|\theta - \tilde{\theta}\| \leq r\}$. For simplicity, we let $\mathbb{B}_d(r) = \mathbb{B}_d(\mathbf{0}, r)$ when $\tilde{\theta} = \mathbf{0}$; also note that all norms are Euclidean. Given $r > 0$, we consider the logistic model and its associated hypothesis class $\mathcal{G} = \{g_\theta : \theta \in \mathbb{B}_d(r)\}$, composed of parameterized soft classifiers g_θ such that

$$g_\theta(x) = \sigma(\langle \theta, x \rangle), \qquad (23)$$

with $\sigma : \mathbb{R} \to [0, 1]$ being the sigmoid function given by (9). For convenience, we present the following short form of α-loss in the logistic model which is equivalent to the expanded expression in [1]. For $\alpha \in (0, \infty]$, α-loss is given by

$$l^\alpha(y, g_\theta(x)) = \frac{\alpha}{\alpha - 1} \left[1 - g_\theta(yx)^{1-1/\alpha} \right]. \qquad (24)$$

For $\alpha = 1$, l^1 is the logistic loss and we recover logistic regression by optimizing this loss. Note that in this setting $\langle yx, \theta \rangle$ is the margin, and recall from Proposition 3 that (24) is convex for $\alpha \in (0, 1]$ and quasi-convex for $\alpha > 1$ in $\langle yx, \theta \rangle$. For $\theta \in \mathbb{B}_d(r)$, we define the α-risk R_α as the risk of the loss in (24),

$$R_\alpha(\theta) := \mathbb{E}_{X,Y}[l^\alpha(Y, g_\theta(X))]. \qquad (25)$$

The α-risk (25) is plotted for several values of α in a two-dimensional Gaussian Mixture Model (GMM) in Fig. 4. Further, observe that, for all $\theta \in \mathbb{B}_d(r)$,

$$R_\infty(\theta) := \mathbb{E}_{X,Y}[l^\infty(Y, g_\theta(X))] = \mathbb{P}[Y \neq \hat{Y}_\theta], \qquad (26)$$

where $\hat{Y}_\theta$ is a random variable such that for all $x \in \mathbb{B}_d(1)$, $\mathbb{P}[\hat{Y}_\theta = 1 | X = x] = g_\theta(x)$.

In order to study the landscape of the α-risk, we compute the gradient and Hessian of (24), by employing the following useful properties of the sigmoid

$$\sigma(-z) = 1 - \sigma(z) \quad \text{and} \quad \frac{d}{dz}\sigma(z) = \sigma(z)(1 - \sigma(z)). \qquad (27)$$

Indeed, a straightforward computation shows that

$$\frac{\partial}{\partial \theta^j} l^\alpha(y, g_\theta(x)) = \left[-y g_\theta(yx)^{1-1/\alpha}(1 - g_\theta(yx)) \right] x^j, \qquad (28)$$
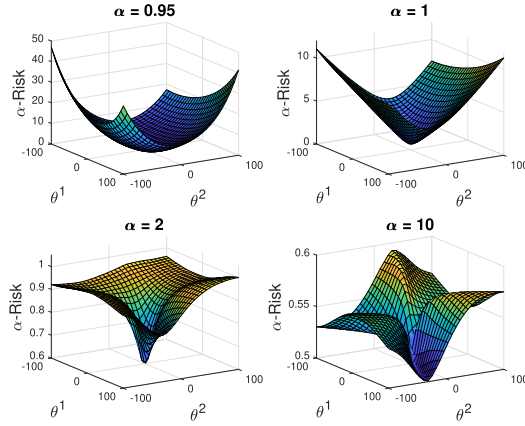
Fig. 4. The landscape of α-loss (R_α for $\alpha = 0.95, 1, 2, 10$) in the logistic model, where features are normalized, for a 2D-GMM with $\mathbb{P}[Y = -1] = 0.12$, $\mu_{X|Y=-1} = (-0.18, 1.49)^\intercal$, $\mu_{X|Y=1} = (-0.01, 0.16)^\intercal$, $\Sigma_{-1} = [3.20, -2.02; -2.02, 2.71]$, and $\Sigma_1 = [4.19, 1.27; 1.27, 0.90]$.

where θ^j, x^j denote the j-th components of θ and x, respectively. Thus, the gradient of α-loss in (24) is

$$\nabla_\theta l^\alpha(Y, g_\theta(X)) = F_1(\alpha, \theta, X, Y)X, \qquad (29)$$

where $F_1(\alpha, \theta, x, y)$ is defined as the expression within brackets in (28). Another straightforward computation yields

$$\nabla^2_\theta l^\alpha(Y, g_\theta(X)) = F_2(\alpha, \theta, X, Y)XX^\intercal, \qquad (30)$$

where F_2 is defined as

$$F_2(\alpha, \theta, x, y) := g_\theta(yx)^{1-1/\alpha} g_\theta(-yx)$$
$$\times \left(g_\theta(yx) - \left(1 - \frac{1}{\alpha}\right) g_\theta(-yx) \right). \qquad (31)$$

B. Convexity of the α-Risk

We now turn our attention to the case where $\alpha \in (0, 1]$; we find that for this regime, R_α is strongly convex; see Fig. 4 for an example. Prior to stating the result, for two matrices $A, B \in \mathbb{R}^{d \times d}$, we let $\succeq$ denote the Loewner (partial) order in the positive semi-definite cone. That is, we write $A \succeq B$ when $A - B$ is a positive semi-definite matrix. For a matrix $A \in \mathbb{R}^{d \times d}$, let $\lambda_1(A), \ldots, \lambda_d(A)$ be its eigenvalues. Finally, we recall that a function is m-strongly convex if and only if its Hessian has minimum eigenvalue $m \geq 0$ [62].

Theorem 2: Let $\Sigma := \mathbb{E}[XX^\intercal]$. If $\alpha \in (0, 1]$, then $R_\alpha(\theta)$ is $\Lambda(\alpha, r\sqrt{d}) \min_{i \in [d]} \lambda_i(\Sigma)$-strongly convex in $\theta \in \mathbb{B}_d(r)$, where

$$\Lambda(\alpha, r\sqrt{d}) := \sigma(r\sqrt{d})^{1-1/\alpha}$$
$$\times \left(\sigma'(r\sqrt{d}) - \left(1 - \frac{1}{\alpha}\right) \sigma(-r\sqrt{d})^2 \right). \qquad (32)$$

Observe that if $\min_{i \in [d]} \lambda_i(\Sigma) = 0$, then the α-risk is merely convex for $\alpha \leq 1$. Also observe that for $r\sqrt{d} > 0$ fixed, $\Lambda(\alpha, r\sqrt{d})$ is monotonically decreasing in α. Thus, R_α becomes more strongly convex as α approaches zero.

While Theorem 2 states that the α-risk is strongly-convex for all $\alpha \leq 1$ and for any $r\sqrt{d} > 0$, the following corollary, which is proved with similar techniques as Theorem 2, states

that the α-risk is strongly-convex for some range of $\alpha > 1$, provided that $r\sqrt{d} > 0$ is small enough.

Corollary 2: Let $\Sigma := \mathbb{E}[XX^T]$. If $r\sqrt{d} \leq \operatorname{arcsinh}(1/2)$, then $R_\alpha(\theta)$ is $\tilde{\Lambda}(\alpha, r\sqrt{d}) \min_{i \in [d]} \lambda_i(\Sigma)$-strongly convex in $\theta \in \mathbb{B}_d(r)$ for $\alpha \in \left(0, (e^{2r\sqrt{d}} - e^{r\sqrt{d}})^{-1}\right]$, where

$$\tilde{\Lambda}(\alpha, r\sqrt{d}) := \sigma(-r\sqrt{d})^{2-1/\alpha} \sigma(r\sqrt{d})$$
$$\times \left(1 - e^{r\sqrt{d}} + \frac{e^{-r\sqrt{d}}}{\alpha} \right). \qquad (33)$$

It could be verified that $(e^{2r\sqrt{d}} - e^{r\sqrt{d}})^{-1} > 1$ whenever $r\sqrt{d} < \operatorname{arcsinh}(1/2)$. By inspecting the relationship between convexity and its dependence on $r\sqrt{d}$, Corollary 2 seems to suggest that as α increases slightly greater than 1, convexity is lost faster nearer to the boundary of the parameter space. Indeed, refer to Fig. 4 to observe an example of this effect for α increasing from $\alpha = 1$ to $\alpha = 2$, and note that convexity is preserved in the small radius about $\mathbf{0}$ for $\alpha = 2$.

Examining the α-risk in Fig. 4 for $\alpha = 2$ more closely, we see that it is reminiscent of a quasi-convex function. Recall that (e.g., Chapter 3.4 in [62]) a function $f : \mathbb{R}^d \to \mathbb{R}$ is quasi-convex if for all $\theta, \theta_0 \in \mathbb{R}^d$, such that $f(\theta_0) \leq f(\theta)$, it follows that

$$\langle -\nabla f(\theta), \theta_0 - \theta \rangle \geq 0. \qquad (34)$$

In other words, the negative gradient of a quasi-convex function always points in the direction of descent. While α-loss (24) is quasi-convex for $\alpha > 1$, this does not imply that the α-*risk* (25) is quasi-convex for $\alpha > 1$ since the sum of quasi-convex functions is not guaranteed to be quasi-convex [62]. Thus, we need a new tool in order to quantify the optimization complexity of the α-risk for $\alpha > 1$ in the large radius regime.

C. Strictly Local Quasi-Convexity and Its Extensions

We use a framework developed by Hazan *et al.* in [35] called *strictly local quasi-convexity* (SLQC), which is a generalization of quasi-convexity. Intuitively, SLQC functions allow for multiple local minima below an ϵ-controlled region while stipulating (strict) quasi-convex functional behavior outside the same region. Formally, we recall the following parameteric definition of SLQC functions provided in [35].

Definition 3 (Definition 3.1, [35]): Let $\epsilon, \kappa > 0$ and $\theta_0 \in \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called $(\epsilon, \kappa, \theta_0)$-strictly locally quasi-convex (SLQC) at $\theta \in \mathbb{R}^d$ if at least one of the following conditions apply:
  1. $f(\theta) - f(\theta_0) \leq \epsilon$,
  2. $\|\nabla f(\theta)\| > 0$ and, for every $\theta' \in \mathbb{B}(\theta_0, \epsilon/\kappa)$,

$$\langle -\nabla f(\theta), \theta' - \theta \rangle \geq 0. \qquad (35)$$

Briefly, in [35] Hazan *et al.* refer to a function as SLQC *in* θ, whereas for the purposes of our analysis we refer to a function as SLQC *at* θ. We recover the uniform SLQC notion of Hazan *et al.* by articulating a function is SLQC *at* θ for *every* θ. Our later analysis of the α-risk in the logistic model benefits from this pointwise consideration.

Observe that where Condition 1 of Definition 3 does not hold, Condition 2 implies quasi-convexity about $\mathbb{B}(\theta_0, \epsilon/\kappa)$
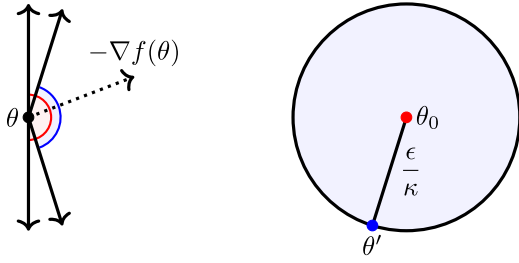
Fig. 5. An illustration highlighting the difference between quasi-convexity as given in (34) and the second SLQC condition of Definition 3. If f is quasi-convex, the red angle describes the possible negative gradients of f at θ with respect to θ_0. If f is SLQC, the blue angle describes the possible negative gradients of f at θ with respect to θ_0 and the given ϵ/κ-radius ball.

as evidence through (34); see Fig. 5 for an illustration of the difference between classical quasi-convexity and SLQC in this regime. We now present the following lemma, which is a structural result for general differentiable functions that provides an alternative formulation of the second requirement of SLQC functions in Definition 3; proof details can be found in Appendix B.1.

Lemma 1: Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, $\theta_0 \in \mathbb{R}^d$ and $\rho > 0$. If $\theta \in \mathbb{R}^d$ is such that $\|\theta - \theta_0\| > \rho$, then the following are equivalent:

1. $\langle -\nabla f(\theta), \theta' - \theta \rangle \geq 0$ for all $\theta' \in \mathbb{B}_d(\theta_0, \rho)$,
2. $\langle -\nabla f(\theta), \theta_0 - \theta \rangle \geq \rho \|\nabla f(\theta)\|$.

Intuitively, the equivalence presented by Condition 2 of Lemma 1 is easier to manipulate in proving SLQC properties of the α-risk as we merely need to control $\langle -\nabla f(\theta), \theta_0 - \theta \rangle$ rather than $\langle -\nabla f(\theta), \theta' - \theta \rangle$ for every $\theta' \in \mathbb{B}(\theta_0, \epsilon/\kappa)$.

In [35], Hazan *et al.* measure the optimization complexity of SLQC functions through the normalized gradient descent (NGD) algorithm, which is almost canonical gradient descent (see, e.g., Chapter 14 in [4]) except gradients are normalized such that the algorithm applies uniform-size directional updates given by a fixed learning rate $\eta > 0$. While NGD may not be the most appropriate optimization algorithm in some applications, we use it as a theoretical benchmark which allows us to understand optimization complexity; further details regarding NGD can be found in Appendix B.1. Indeed, the convergence guarantees of NGD for SLQC functions are similar to those of Gradient Descent for convex functions.

Proposition 4 (Thm. 4.1, [35]): Let $f : \mathbb{R}^d \to \mathbb{R}$, $\theta_1 \in \mathbb{R}^d$, and $\theta^* = \arg\min_{\theta \in \mathbb{R}^d} f(\theta)$. If f is $(\epsilon, \kappa, \theta^*)$-SLQC at θ for every $\theta \in \mathbb{R}^d$, then running the NGD algorithm with learning rate $\eta = \epsilon/\kappa$ for number of iterations $T \geq \kappa^2 \|\theta_1 - \theta^*\|^2/\epsilon^2$ achieves $\min_{t=1,\ldots,T} f(\theta_t) - f(\theta^*) \leq \epsilon$.

For an $(\epsilon, \kappa, \theta_0)$-SLQC function, a smaller ϵ provides better optimality guarantees. Given $\epsilon > 0$, smaller κ leads to faster optimization as the number of required iterations increases with κ^2. Finally, by using projections, NGD can be easily adapted to work over convex and closed sets (e.g., $\mathbb{B}(\theta_0, r)$ for some $\theta_0 \in \mathbb{R}^d$ and $r > 0$).

D. SLQC Parameters of the α-Risk

With the above SLQC preliminaries in hand, we start quantifying the SLQC parameters of the α-risk, R_α. It can be shown that for $\alpha \in (0, \infty]$, R_α is $C_d(r, \alpha)$-Lipschitz in $\theta \in \mathbb{B}_d(r)$ where, for $\alpha \in (0, 1]$,

$$C_d(r, \alpha) := \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha}; \tag{36}$$

and, for $\alpha \in (1, \infty]$,

$$C_d(r, \alpha) := \begin{cases} \sqrt{d}\left(\frac{\alpha-1}{2\alpha-1}\right)^{1-1/\alpha}\left(\frac{\alpha}{2\alpha-1}\right) & e^{r\sqrt{d}} \geq \frac{\alpha-1}{\alpha}, \\ \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha} & e^{r\sqrt{d}} < \frac{\alpha-1}{\alpha}. \end{cases} \tag{37}$$

Thus, in conjunction with Theorem 2, Corollary 2, and a result by Hazan *et al.* in [35] (after Definition 3), we provide the following result that explicitly characterizes the SLQC parameters of the α-risk R_α for two separate ranges of α near 1.

Proposition 5: Suppose that $\Sigma \succ 0$ and $\theta_0 \in \mathbb{B}_d(r)$ is fixed. We have one of the following:

- If $r\sqrt{d} < \text{arcsinh}(1/2)$, then, for every $\epsilon > 0$, R_α is $(\epsilon, C_d(r, \alpha), \theta_0)$-SLQC at θ for every $\theta \in \mathbb{B}_d(r)$ when $\alpha \in \left(0, (e^{2r\sqrt{d}} - e^{r\sqrt{d}})^{-1}\right]$ where $C_d(r, \alpha)$ is given in (36) and (37);
- Otherwise, for every $\epsilon > 0$, R_α is $(\epsilon, C_d(r, \alpha), \theta_0)$-SLQC at θ for every $\theta \in \mathbb{B}_d(r)$ for $\alpha \in (0, 1]$.

Thus, by Proposition 4 and (36), the number of iterations of NGD, T_α, tends to infinity as α tends to zero. This consequence of the result seems somewhat counterintuitive because one would expect that increasing convexity (R_α becomes "more" strongly convex in θ as α decreases, see Theorem 2 and Fig. 4) would improve the convergence rate. However, the number of iterations of NGD tends to infinity as α tends to zero because the Lipschitz constant of R_α, $C_d(r, \alpha) = \kappa$ blows up. This phenomenon of the Lipschitz constant worsening the convergence rate is not merely a feature of the SLQC theory surrounding NGD. It is also present in convergence rates for SGD optimizing convex functions, e.g., see Theorem 14.8 in [4]. Therefore, we find that there exists a trade-off between the desired strong-convexity of R_α and the optimization complexity of NGD.

Next, we quantify the evolution of the SLQC parameters of R_α both in the small radius regime and in the large radius regime. Since R_α tends more towards the probability of error (expectation of 0-1 loss) as α approaches infinity, we find that the SLQC parameters deteriorate and the optimization complexity of NGD increases as we increase α. Fortunately, in the logistic model, α-loss exhibits a saturation effect whereby relatively small values of α resemble the landscape induced by $\alpha = \infty$. In order to quantify this effect, we state the following two Lipschitz inequalities which will also be instrumental for our main SLQC result.
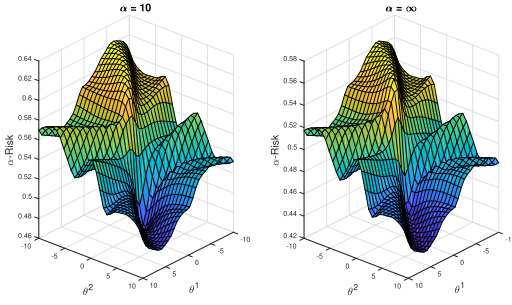
Fig. 6. An illustration of the saturation phenomenon of α-loss (R_α for $\alpha = 10, \infty$) in the logistic model for a 2D-GMM with $\mathbb{P}[Y = 1] = \mathbb{P}[Y = -1]$, $\mu_{X|Y=-1} = (-.91, .50)^\mathsf{T}$, $\mu_{X|Y=1} = (-.27, .20)^\mathsf{T}$, $\Sigma = [1.38, .55; .55, 2.18]$. Note the small difference, uniformly over the parameter space, between R_{10} and R_∞.

Lemma 2: If $\alpha, \alpha' \in [1, \infty]$, then, for all $\theta \in \mathbb{B}_d(r)$,

$$|R_\alpha(\theta) - R_{\alpha'}(\theta)| \leq L_d(\theta) \left| \frac{\alpha - \alpha'}{\alpha \alpha'} \right|, \tag{38a}$$

$$\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\| \leq J_d(\theta) \left| \frac{\alpha - \alpha'}{\alpha \alpha'} \right|, \tag{38b}$$

where

$$L_d(\theta) := \frac{\left(\log \left(1 + e^{\|\theta\| \sqrt{d}} \right) \right)^2}{2}, \tag{39a}$$

$$J_d(\theta) := \sqrt{d} \log \left(1 + e^{\|\theta\| \sqrt{d}} \right) \sigma(\|\theta\| \sqrt{d}). \tag{39b}$$

This result is proved in Appendix B.2, and it can be applied to illustrate a saturation effect of α-loss in the logistic model. That is, let $\alpha = 10$ and $\alpha' = \infty$, then for all $\theta \in \mathbb{B}_d(r)$, we have that

$$|R_{10}(\theta) - R_\infty(\theta)| \leq \frac{L_d(\theta)}{10}, \tag{40a}$$

$$|\nabla R_{10}(\theta) - \nabla R_\infty(\theta)| \leq \frac{J_d(\theta)}{10}, \tag{40b}$$

where $L_d(\theta)$ and $J_d(\theta)$ are both given in (39). In words, the pointwise distance between the $\alpha = 10$ landscape and the $\alpha = \infty$ landscape decreases geometrically; for a visual representation see Fig. 6.

The saturation effect of α-loss suggests that it is unnecessary to work with large values of α. In particular, this motivates us to study the evolution of the SLQC parameters of the α-risk as we increase $\alpha > 1$.

Theorem 3: Let $\alpha_0 \in [1, \infty]$, $\epsilon_0, \kappa_0 > 0$, and $\theta_0, \theta \in \mathbb{B}_d(r)$. If R_{α_0} is $(\epsilon_0, \kappa_0, \theta_0)$-SLQC at θ and

$$0 \leq \alpha - \alpha_0 < \frac{\alpha_0^2 \|\nabla R_{\alpha_0}(\theta)\|}{2 J_d(\theta) \left(1 + r \frac{\kappa_0}{\epsilon_0} \right)}, \tag{41}$$

then R_α is $(\epsilon, \kappa, \theta_0)$-SLQC at θ with

$$\epsilon = \epsilon_0 + 2 L_d(\theta) \left(\frac{\alpha - \alpha_0}{\alpha \alpha_0} \right), \tag{42}$$

$$\frac{\epsilon}{\kappa} = \frac{\epsilon_0}{\kappa_0} \left(1 - \frac{\left(1 + 2r \frac{\kappa_0}{\epsilon_0} \right) J_d(\theta)(\alpha - \alpha_0)}{\alpha \alpha_0 \|\nabla R_{\alpha_0}(\theta)\| - J_d(\theta)(\alpha - \alpha_0)} \right). \tag{43}$$

The proof of Theorem 3 can be found in Appendix B.2. The crux of the proof is a consideration of two cases, dependent on the location of $\theta \in \mathbb{B}_d(r)$ relative to the ϵ_0-plane. The first case considers $\theta \in \mathbb{B}_d(r)$ such that $R_{\alpha_0}(\theta) - R_{\alpha_0}(\theta_0) \leq \epsilon_0$ and provides the required increase for ϵ to capture such points as α increases. The second case considers $\theta \in \mathbb{B}_d(r)$ such that $R_{\alpha_0}(\theta) - R_{\alpha_0}(\theta_0) > \epsilon_0$ and provides the required decrease for ϵ/κ to capture such points as α increases. The second case is far more geometric than the first one, as it makes use of finer gradient information. As a result, the decrease in ϵ/κ is more closely related to the landscape evolution of R_α than the corresponding increase in ϵ. From a numerical point of view, Proposition 4 implies that reducing the radius of the ϵ/κ ball about θ_0 increases the required number of iterations (for optimality), and thus reflects the intuition that increasing $\alpha > 1$ more closely approximates the intractable 0-1 loss. While on the contrary, Proposition 4 implies that increasing the value of ϵ reduces the optimality guarantee itself.

We note that the bounds provided in Theorem 3 are pessimistic, but fortunately, we can improve them by employing a bootstrapping technique - we take infinitesimal steps in α and repeatedly apply the bounds in Theorem 3 to derive improved bounds on α, ϵ, and κ. The following result is the culmination of our analysis regarding the SLQC parameters of the α-risk in the logistic model. The proof can be found in Appendix B.3.

Theorem 4: Let $\alpha_0 \in [1, \infty)$, $\epsilon_0, \kappa_0 > 0$, and $\theta_0, \theta \in \mathbb{B}_d(r)$. Suppose that R_{α_0} is $(\epsilon_0, \kappa_0, \theta_0)$-SLQC at $\theta \in \mathbb{B}_d(r)$ and that there exists $g_\theta > 0$ such that $\|\nabla R_{\alpha'}(\theta)\| > g_\theta$ for every $\alpha' \in [\alpha_0, \infty]$. Then, for every $\lambda \in (0, 1)$, R_{α_λ} is $(\epsilon_\lambda, \kappa_\lambda, \theta_0)$-SLQC at θ where

$$\alpha_\lambda := \alpha_0 + \lambda \frac{\alpha_0^2 g_\theta}{J_d(\theta) \left(1 + 2r \frac{\kappa_0}{\epsilon_0} \right)}, \tag{44}$$

$$\epsilon_\lambda := \epsilon_0 + 2\lambda L_d(\theta) \left(\frac{\alpha_\lambda - \alpha_0}{\alpha_\lambda \alpha_0} \right) \frac{\alpha_0^2 g_\theta}{J_d(\theta) \left(1 + r \frac{\kappa_0}{\epsilon_0} \right)}, \tag{45}$$

$$\frac{\epsilon_\lambda}{\kappa_\lambda} > \frac{\epsilon_0}{\kappa_0}(1 - \lambda). \tag{46}$$

We now provide three different interpretations and comments regarding the previous result. First regarding the SLQC parameters themselves, observe from (44) that the bound on α is improved over Theorem 3 as the factor of 2 in the denominator in (41) is moved into the parentheses; next, it can be observed (upon plugging in α_λ) that ϵ_λ in (45) is linear in λ, which is again an improvement over the first equation in (42); finally, note that the bound on $\epsilon_\lambda/\kappa_\lambda$ in (45) is vastly more tractable and informative than the second expression in (42). Thus, bootstrapping the bounds of Theorem 3 provides strong improvements for all three relevant quantities, α, ϵ, and κ. Next, regarding the extra assumption for Theorem 4 over Theorem 3, i.e., the existence of a lowerbound g_θ on the norm of the gradient $\|\nabla R_{\alpha'}(\theta)\|$ for all $\alpha' \geq \alpha_0$, observe that this is equivalent to the requirement that the landscape at θ does not become "flat" for any $\alpha' \geq \alpha_0$. In essence, this is a distributional assumption in disguise, and it should be addressed in a case-by-case basis. Finally, regarding the effect of the dimensionality of the feature space, d, on the bounds, we observe that for $\theta \in \mathbb{B}_d(r)$ and $d \in \mathbb{N}$ large enough,

$J_d(\theta) \approx d\|\theta\|$ as given in (39). Thus in the high-dimensional regime, the bound on α, i.e., α_λ, is dominated by $1/d$. This implies that the convexity of the landscape worsens as the dimensionality of the feature/parameter vectors d increases.

While a practitioner would ultimately like to approximate the 0-1 loss (captured by $\alpha = \infty$), the bounds presented in Theorem 4 suggest that the optimization complexity of NGD increases as α increases. Fortunately, α-loss exhibits a saturation effect as exemplified in (40) and Fig. 6 whereby smaller values of α quickly resemble the landscape induced by $\alpha = \infty$. Thus, while the optimization complexity increases as α increases (and increases even more rapidly in the high-dimensional regime), the saturation effect suggests that the practitioner need not increase α too much in order to reap the benefits of the ∞-risk. Therefore, for the logistic model, we ultimately posit that there is a narrow range of α useful to the practitioner and we dub this the "Goldilocks zone"; we explore this theme in the experiments in Section VI.

Before this however, we conclude the theoretical analysis of α-loss with a study of the empirical α-risk, and we provide generalization and optimality guarantees for all $\alpha \in (0, \infty]$.

V. GENERALIZATION AND ASYMPTOTIC OPTIMALITY

In this section, we provide generalization and asymptotic optimality guarantees for α-loss for $\alpha \in (0, \infty]$ in the logistic model by utilizing classical Rademacher complexity tools and the notion of classification-calibration introduced by Bartlett *et al.* in [6]. We invoke the same setting and definitions provided in Section IV. In addition, we consider the evaluation of α-loss in the finite sample regime. Formally, let $X \in [0, 1]^d$ be the normalized feature and $Y \in \{-1, +1\}$ the label as before, *and* let $S_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$ be the training dataset where, for each $i \in \{1, \ldots, n\}$, the samples (X_i, Y_i) are independently and identically drawn according to an unknown distribution $P_{X,Y}$. Finally, we let $\hat{R}_\alpha$ denote the empirical α-risk of (24), i.e., for each $\theta \in \mathbb{B}_d(r)$ we have

$$\hat{R}_\alpha(\theta) = \frac{1}{n} \sum_{i=1}^{n} l^\alpha(Y_i, g_\theta(X_i)). \quad (47)$$

In the following sections, we consider the generalization capabilities and asymptotic optimality of a predictor $\theta \in \mathbb{B}_d(r)$ which is learned through empirical evaluation of α-loss (47). First, we recall classical results in Rademacher complexity generalization bounds.

A. Rademacher Complexity Preliminaries

In this section, we provide the main tools we use to derive generalization bounds for α-loss in the sequel. The techniques are standard; see Chapter 26 in [4] for a complete discussion. First, we recall that the Rademacher distribution is the uniform distribution on the set $\{-1, +1\}$. The Rademacher complexity of a set is as follows.

Definition 4: The Rademacher complexity of a nonempty set $A \subset \mathbb{R}^n$ is defined as

$$\mathcal{R}(A) := \mathbb{E}\left(\sup_{a \in A} \frac{1}{n} \langle \sigma, a \rangle\right), \quad (48)$$

where $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_1, \sigma_2, \ldots, \sigma_n$ i.i.d. Rademacher random variables.

In words, the Rademacher complexity of a set approximately measures the richness of the set through the maximal correlation of its elements with uniformly distributed Rademacher vectors. The notion of Rademacher complexity can be used to measure the richness of a hypothesis class as established in the following proposition.

Proposition 6 (Thm. 26.5, [4]): Let $\mathcal{H}$ be a hypothesis class. Assume that $l : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}_+$ is a bounded loss function, i.e., there exists $D > 0$ such that for all $h \in \mathcal{H}$ and for all $(x, y) \in (\mathcal{X}, \mathcal{Y})$ we have that $|l(h, (x, y))| \leq D$. Then, with probability at least $1 - \delta$, for all $h \in \mathcal{H}$,

$$\left|R_l(h) - \hat{R}_l(h)\right| \leq 2\mathcal{R}(l \circ \mathcal{H} \circ S_n) + 4D\sqrt{\frac{2\ln(4/\delta)}{n}}, \quad (49)$$

where $R_l(h)$ and $\hat{R}_l(h)$ denote the true risk and empirical risk of l, respectively, and[2] $l \circ \mathcal{H} \circ S_n \subset \mathbb{R}^n$ which is equal to

$$\{(l(h, (x_1, y_1)), \ldots, l(h, (x_n, y_n))) : h \in \mathcal{H}\}. \quad (50)$$

For linear predictors, obtaining a bound on $\mathcal{R}(l \circ \mathcal{H} \circ S_n)$ is feasible; we now provide two results (in conjunction with Proposition 6) necessary to derive a generalization bound for α-loss in the logistic model.

Lemma 3 (Lemma 26.9, [4]): Suppose $\tilde{l}_1, \ldots, \tilde{l}_n : \mathbb{R} \to \mathbb{R}$ are r_0-Lipschitz functions with common constant $r_0 \geq 0$. If $\tilde{l} = (\tilde{l}_1, \ldots, \tilde{l}_n)$ and $A \subset \mathbb{R}^n$, then $\mathcal{R}(\tilde{l}(A)) \leq r_0 \mathcal{R}(A)$, where $\tilde{l}(A) := \{(\tilde{l}_1(a_1), \ldots, \tilde{l}_n(a_n)) : a \in A\}$.

The previous result, known as the Contraction Lemma, provides an upperbound on the Rademacher complexity of the composition of a function acting on a set. For our purposes, one can think of $\tilde{l} = (\tilde{l}_1, \ldots, \tilde{l}_n)$ as a margin-based loss function acting on a training set with n samples - this will be further elaborated in the sequel. The following result provides an upperbound on the Rademacher complexity of the set comprised of inner products between a given parameter vector drawn from a bounded space and the n-sample training set.

Lemma 4 (Lemma 26.10, [4]): Let $x_{1:n} = \{x_1, \ldots, x_n\}$ be a set of vectors each in $\mathbb{R}^d$, and define the following composition $\mathcal{H} \circ x_{1:n} = \{(\langle \theta, x_1 \rangle, \ldots, \langle \theta, x_n \rangle) : \|\theta\|_2 \leq r\}$. Then,

$$\mathcal{R}(\mathcal{H} \circ x_{1:n}) \leq \frac{r \max_{i \in [n]} \|x_i\|_2}{\sqrt{n}}. \quad (51)$$

With the above Rademacher complexity preliminaries in hand, we now apply these results to derive a generalization bound for α-loss in the logistic model.

B. Generalization and Asymptotic Optimality of α-Loss

We now present the following Lipschitz inequality for the margin-based α-loss (Definition 2) and will be useful in applying Proposition 6. It can readily be shown that the margin-based α-loss, $\tilde{l}^\alpha$ is $C_{r_0}(\alpha)$-Lipschitz in $z \in [-r_0, r_0]$ for every $r_0 > 0$, where for $\alpha \in (0, 1]$,

$$C_{r_0}(\alpha) := \sigma(r_0)\sigma(-r_0)^{1-1/\alpha}; \quad (52)$$

---

[2]In (49) we present the two-sided version of Theorem 26.5 in [4], which can be readily obtained via the symmetrization technique.

and, for $\alpha \in (1, \infty]$,

$$C_{r_0}(\alpha) := \begin{cases} \left(\frac{\alpha-1}{2\alpha-1}\right)^{1-\frac{1}{\alpha}} \left(\frac{\alpha}{2\alpha-1}\right) & e^{r_0} \geq \frac{\alpha-1}{\alpha}, \\ \sigma(r_0)\sigma(-r_0)^{1-\frac{1}{\alpha}} & e^{r_0} < \frac{\alpha-1}{\alpha}. \end{cases} \quad (53)$$

That is, for $\alpha \in (0, \infty]$ and $z, z' \in [-r_0, r_0]$, we have that $|\tilde{l}^\alpha(z) - \tilde{l}^\alpha(z')| \leq C_{r_0}(\alpha)|z - z'|$; see Lemma 6 in Appendix C for the proof. Lastly, note that for any fixed $r_0 > 0$, $C_{r_0}(\alpha)$ is monotonically decreasing in α.

With the Lipschitz inequality for $\tilde{l}^\alpha$ in hand, we are now in a position to state a generalization bound for α-loss in the logistic model.

Theorem 5: If $\alpha \in (0, \infty]$, then, with probability at least $1 - \delta$, for all $\theta \in \mathbb{B}_d(r)$,

$$\left| R_\alpha(\theta) - \hat{R}_\alpha(\theta) \right| \leq C_{r\sqrt{d}}(\alpha) \frac{r\sqrt{d}}{\sqrt{n}} + D_{r\sqrt{d}}(\alpha) \sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{n}}, \quad (54)$$

where $C_{r\sqrt{d}}(\alpha)$ is given in (52) and (53) and where $D_{r\sqrt{d}}(\alpha)$ is given by $D_{r\sqrt{d}}(\alpha) := 4\sqrt{2}\frac{\alpha}{\alpha-1}\left(1 - \sigma(-r\sqrt{d})^{1-1/\alpha}\right)$.

Note that $D_{r\sqrt{d}}(\alpha)$ is also monotonically decreasing in α for fixed $r\sqrt{d} > 0$. Thus, Theorem 5 seems to suggest that generalization improves as $\alpha \to \infty$. However, because R_α and $\hat{R}_\alpha$ also monotonically decrease in α, it is difficult to reach such a conclusion. Nonetheless, Corollary 3 in Appendix C offers an attempt at providing a unifying comparison between the ∞-risk, R_∞, and the empirical α-risk, $\hat{R}_\alpha$.

Lastly, observe that for the generalization result in Theorem 5, we make no distributional assumptions such as those by Tsybakov, *et. al* in [66], where they assume the posterior satisfies a *margin* condition. Under such an assumption, we observe that faster rates could be achieved, but optimal rates are not the focus of this work. Nonetheless, the next theorem relies on the assumption that the minimum α-risk is attained by the logistic model, i.e., given $\alpha \in (0, \infty]$, suppose that

$$\min_{\theta \in \mathbb{B}_d(r)} R_\alpha(\theta) = \min_{f: \mathcal{X} \to \mathbb{R}} R_\alpha(f), \quad (55)$$

where $R_\alpha(\theta)$ is given in (25) and $R_\alpha(f) = \mathbb{E}[\tilde{l}^\alpha(Yf(X))]$ for all measurable f.

Theorem 6: Assume that the minimum α-risk is attained by the logistic model, i.e., (55) holds. Let S_n be a training dataset with $n \in \mathbb{N}$ samples as before. If for each $n \in \mathbb{N}$, $\hat{\theta}_n^\alpha$ is a global minimizer of the associated empirical α-risk $\theta \mapsto \hat{R}_\alpha(\theta)$, then the sequence $(\hat{\theta}_n^\alpha)_{n=1}^\infty$ is asymptotically optimal for the 0-1 risk, i.e., almost surely,

$$\lim_{n \to \infty} R(f_{\hat{\theta}_n^\alpha}) = R^*, \quad (56)$$

where $f_{\hat{\theta}_n^\alpha}(x) = \langle \hat{\theta}_n^\alpha, x \rangle$ for each $n \in \mathbb{N}$ and the Bayes risk R^* is given by $R^* := \min_{f: \mathcal{X} \to \mathbb{R}} \mathbb{P}[Y \neq \text{sign}(f(X))]$.

In words, setting the optimization procedure aside, utilizing α-loss for a given $\alpha \in (0, \infty]$ is asymptotically optimal with respect to the probability of error (expectation of the 0-1 loss). Observe that the assumption in (55) is a stipulation for the underlying data-generating distribution, $P_{X,Y}$,

in disguise. That is, we assume that $P_{X,Y}$ is separable by a linear predictor, which is a global minimizer for the α-risk. In essence, Theorem 6 is a combination of Theorem 5 and classification-calibration.

With the statistical, optimization, and generalization considerations of α-loss behind us, we now provide experimental results in two canonical settings for α-loss in logistic and convolutional-neural-network models.

VI. EXPERIMENTAL RESULTS

As was first introduced in Section III-C, in this section we further experimentally evaluate the efficacy of α-loss in the following two canonical scenarios:

(i) *Noisy Labels:* the classification algorithm is trained on a binary-labeled dataset that suffers from symmetric noisy labels, and it attempts to produce a model which achieves strong performance on the clean test data.

(ii) *Class Imbalance:* the classification algorithm is trained on a binary-labeled dataset that suffers from a class imbalance, and it attempts to produce a model which achieves strong performance on the balanced test data.

Our hypotheses are as follows: for setting (i), tuning $\alpha > 1$ (away from log-loss) improves the robustness of the trained model to symmetric noisy labels; for setting (ii), tuning $\alpha < 1$ (again away from log-loss) improves the sensitivity of the trained model to the minority class. In general, we experimentally validate both hypotheses.

In our experimental procedure, we use the following image datasets: MNIST [67], Fashion MNIST (FMNIST) [68], and CIFAR-10 [69]. While these datasets have predefined training and test sets, we present binary partitions of these datasets for both settings in the main text, in alignment with our theoretical investigations of α-loss for binary classification problems; in Appendix D.4, we present multiclass symmetric noise experiments for the MNIST and FMNIST datasets. Regarding the binary partitions themselves, we chose classes which are visually similar in order to increase the difficulty of the classification task. Specifically, for MNIST we used a binary partition on the *1* and *7* classes, for FMNIST we used a binary partition on the *T-Shirt* and *Shirt* classes, and finally for our binary experiments on CIFAR-10 we used a binary partition on the *Cat* and *Dog* classes.

All code is written in PyTorch, version 1.30 [70]. Architectures learning CIFAR are trained with GPUs, while the architectures learning MNIST and FMNIST are both trained with CPUs. Throughout, we consider two broad classes of architectures: logistic regression (LR) and convolutional neural networks (CNNs) with one or two fully connected layers preceded by varying convolutional layer depths (2, 3, 4, and 6) such that we obtain the shorthand CNN X+Y where X is one of 2, 3, 4, or 6 and Y is one of 1 or 2. For all architectures learning CIFAR, we additionally use a sigmoid at the last layer for smoothing. For each set of experiments, we randomly fix a seed, and for each iteration we reinitialize a new architecture with randomly selected weights. We use softmax activation to generate probabilities over the labels, and we evaluate the model's soft belief using α-loss on a one-hot-encoding of the training data.

All (dataset, architecture) tuples were trained with the same optimizer, vanilla SGD, with fixed learning rates. In order to provide the fairest comparison to log-loss ($\alpha = 1$), for each (dataset, architecture) tuple we select a fixed learning rate from the set $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}\}$ which provides the highest validation accuracy for a model trained with log-loss. Then for the chosen (dataset, architecture) tuple, we train α-loss for each value of α using this fixed learning rate. Regarding the optimization of α-loss itself which is parameterized by $\alpha \in (0, \infty]$, in general we find that searching over $\alpha \in [.8, 8]$ for noisy labels and $\alpha \in [.8, 4]$ for class imbalances is sufficient, and we typically do so in step-sizes of 0.1 or 0.05 (near $\alpha = 1$) or a step-size of 1 (when $\alpha > 1$). This is in line with our earlier theoretical discussions regarding the "Goldilocks zone" of α-loss, i.e., the gradient explosion for very small values of α, the increased difficulty of optimization for large values of α, and the fact that relatively small values of α closely approximate the ∞-loss.

For all experiments, we employ a training batch size of 128 examples. For all experiments on the MNIST and FMNIST datasets, training was allowed to progress for 50 epochs; for all experiments on the CIFAR-10 dataset, training was allowed to progress for 120 epochs - convergence for all values of α was ensured for both choices. Lastly, for each architecture we re-run each experiment 10 times and report the average test accuracies calculated according to the *relative accuracy gain*, which we rewrite for our experimental setting as

$$\text{rel acc gain } \% = \frac{|\alpha\text{-loss acc} - \text{log-loss acc}|}{\text{log-loss acc}} \times 100, \quad (57)$$

where we use acc to denote test accuracy. Also note that α^* is chosen as the α over the search range which maximizes the average test accuracy of its trained models. For more details regarding architecture configurations (i.e., CNN channel sizes, kernel size, etc) and general experiment details, we refer the reader to the code for all of our experiments (including the implementation of α-loss), which can be found at [71].

A. Noisy Labels

For the first set of experiments, we evaluate the robustness of α-loss to symmetric noisy labels, and we generate symmetric noisy labels in the binary training data as follows:

1) For each run of an experiment, we randomly select 0-40% of the training data in increments of 10%.
2) For each training example in the randomly selected group, we flip the label of the selected training example.

Note that for all symmetric noisy label experiments we keep the test data clean, i.e., we do not perform label flips on the test data. Thus, these experiments address the scenario where training data is noisy and test data is clean. Also note that during our 10-iteration averaging for each accuracy value presented in each table, we are also randomizing over the symmetric noisy labels in the training data.

The results on the binary MNIST dataset (composed of classes *1* and *7*), binary FMINIST dataset (composed of classes *T-Shirt* and *Shirt*), and binary CIFAR-10 (composed

TABLE I

SYMMETRIC BINARY NOISY LABEL EXPERIMENT ON **MNIST** CLASSES *1* AND *7*. NOTE THAT ARCH STANDS FOR ARCHITECTURE, LF FOR LABEL FLIP, LL ACC AND α ACC STAND FOR LOG-LOSS ACCURACY AND α-LOSS ACCURACY FOR α^*, RESPECTIVELY, AND THAT GAIN % IS CALCULATED ACCORDING TO (57). ALSO NOTE THAT EACH REPORTED ACCURACY IS AVERAGED OVER 10 RUNS

Arch	LF %	LL Acc %	α^* Acc %	α^*	Gain %
LR	0	99.26	99.26	0.95,1	0.00
	10	99.03	99.13	6	0.10
	20	98.65	99.03	7	0.39
	30	97.89	98.96	3.5	1.10
	40	92.10	98.53	8	6.98
CNN 2+2	0	99.83	99.84	4-8	0.01
	10	95.27	99.68	6,7	4.63
	20	87.41	98.72	8	12.94
	30	77.56	87.86	8	13.28
	40	62.89	66.10	8	5.12

TABLE II

SYMMETRIC BINARY NOISY LABEL EXPERIMENT ON CLASSES *T-Shirt* AND *Shirt* OF THE **FMNIST** DATASET

Arch	LF %	LL Acc %	α^* Acc %	α^*	Gain %
LR	0	84.51	84.78	1.5	0.32
	10	83.80	84.41	2	0.72
	20	83.11	83.94	2.5	1.01
	30	81.29	83.43	3	2.63
	40	74.39	92.02	8	23.69
CNN 2+2	0	86.96	87.19	1.1	0.27
	10	81.14	83.74	5	3.20
	20	72.96	78.00	8	6.93
	30	66.17	69.21	8	4.59
	40	57.90	58.56	3	1.15

of classes *Cat* and *Dog*) are presented in Tables I, II, and III, respectively. As stated previously, in order to report the fairest comparison between log-loss and α-loss, we first find the optimal fixed learning rate for log-loss from our set of learning rates (given above), then we train each chosen architecture with α-loss for all values of α also with this found fixed learning rate. Following this procedure, for the binary MNIST dataset, we trained both the LR and CNN 2+2 architectures with a fixed learning rate of 10^{-2}; for the binary FMNIST dataset, we trained the LR and CNN 2+2 architectures with fixed learning rates of 10^{-4} and 5×10^{-3}, respectively; for the binary CIFAR-10 dataset, we trained the CNN 2+1, 3+2, 4+2, and 6+2 architectures with fixed learning rates of 10^{-2}, 10^{-1}, 5×10^{-2}, and 10^{-1}, respectively.

Regarding the results presented in Tables I, II, and III, in general we find for 0% label flips (from now on referred to as baseline) the extra α hyperparameter does not offer significant gains over log-loss in the test results for each (dataset, architecture) tuple. However once we start to increase the percentage of label flips, we immediately find that α^* increases greater than 1 (log-loss). Indeed for each (dataset, architecture) tuple, we find that as the number of symmetric label flips increases, training with α-loss for a value of $\alpha > 1$ increases the test accuracy on clean data, often significantly outperforming log-loss. Note that this performance increase induced by the new α hyperparameter is not monotonic as the number of label flips increases, i.e.,

TABLE III
SYMMETRIC BINARY NOISY LABEL EXPERIMENT
ON **CIFAR-10** CLASSES *Cat* AND *Dog*

Arch	LF %	LL Acc %	α^* Acc %	α^*	Gain %
	0	80.59	80.68	0.99	0.11
	10	79.61	79.89	1.1	0.35
CNN 2+1	20	77.01	77.15	0.99	0.19
	30	73.67	74.78	2.5	1.51
	40	63.54	68.12	4	7.21
	0	85.80	85.80	1	0.00
	10	82.92	83.15	0.99	0.28
CNN 3+2	20	77.61	80.88	3	4.21
	30	69.53	76.72	5	10.34
	40	59.44	67.19	6	13.04
	0	87.49	87.59	0.9	0.12
	10	83.65	84.69	1.2	1.25
CNN 4+2	20	78.96	81.39	3.5	3.07
	30	69.24	75.56	6	9.13
	40	59.12	64.53	8	9.15
	0	87.31	87.93	1.2	0.70
	10	84.91	85.33	2	0.49
CNN 6+2	20	78.92	81.80	6	3.64
	30	68.88	77.20	7	12.09
	40	58.54	65.16	7	11.32

there appears to be a noise threshold past which the performance of all losses decays, but this occurs for very high noise levels, which are not usually present in practice. Recalling Section III-C, the strong performance of α-loss for $\alpha > 1$ on binary symmetric noisy training labels can intuitively be accounted for by the quasi-convexity of α-loss in this regime, i.e., the reduced sensitivity to outliers. Thus, we conclude that the results in Tables I, II, and III on binary MNIST, FMNIST, and CIFAR-10, respectively, indicate that practitioners should employ α-loss for $\alpha > 1$ when training robust architectures to combat against binary noisy training labels. Lastly, we report two experiments for multiclass symmetric noisy training labels in Appendix D.4. In short, we find similar robustness to noisy labels for $\alpha > 1$, but we acknowledge that further empirical study of α-loss on multiclass datasets is needed.

B. Class Imbalance

For the second set of experiments, we evaluate the sensitivity of α-loss to class imbalances, and we generate binary class imbalances in the training data as follows:

1) Given a dataset, select two classes, Class 1 and Class 2, and generate baseline $50/50$ (balanced) data, i.e., such that $|\text{Class } 1| = |\text{Class } 2| = 2500$ training examples. For all experiments ensure that $|\text{Class } 1| + |\text{Class } 2| = 5000$ randomly drawn training examples.
2) Starting at the baseline $(2500/2500)$ and drawing from the available training examples in each dataset when necessary, increase the number of training examples of Class 1 by 500, 1000, 1500, 2000, and 2250 and reduce the number of training examples of Class 2 by the same amounts in order to generate training example splits of $60/40$, $70/30$, $80/20$, $90/10$, and $95/5$, respectively.
3) Repeat the previous step where the roles of Class 1 and Class 2 are reversed.

Note that the test set is balanced for all experiments with 2000 test examples (1000 for each class). Thus, these experiments address the scenario where training data is imbalanced and the test data is balanced. Also note that during our 10-iteration averaging for each accuracy value presented in each table, we are also randomizing over the training examples present in each class imbalance split, according to the procedure above.

The results on binary FMNIST (composed of classes *T-Shirt* and *Shirt*) and binary CIFAR-10 (composed of classes *Cat* and *Dog*) are presented in Tables IV, V, and VI. For this set of experiments, note that α^* is the optimal $\alpha \in [0.8, 4]$ (in our search set) which maximizes the average test accuracy of the minority class, and also note that there are slight test accuracy discrepancies between the baselines in the symmetric noisy labels and class imbalance experiments because of the reduced training and test set size for the class imbalance experiments. For the binary FMNIST dataset, we trained the LR and CNN 2+2 architectures with fixed learning rates of 10^{-4} and 5×10^{-3}, respectively; for the binary CIFAR-10 dataset, we trained the CNN 2+1, 3+2, 4+2, and 6+2 architectures with fixed learning rates of 10^{-2}, 10^{-1}, 5×10^{-2}, and 10^{-1}, respectively.

In general, we find that the minority class is almost always favored by the smaller values of α, i.e., we typically have that $\alpha^* < 1$. Further, we observe that as the percentage of class imbalance increases, the relative accuracy gain on the minority class typically increases through training with α-loss. This aligns with our intuitions articulated in Section III-C regarding the benefits of "stronger" convexity of α-loss when $\alpha < 1$ over log-loss ($\alpha = 1$), particularly when the practitioner desires models which are more sensitive to outliers. Nonetheless, sometimes there does appear to exist a trade-off between how well learning the majority class influences predictions on the minority class, see e.g., recent work in the area of *stiffness* by Fort *et al.* [72]. This is a possible explanation for why $\alpha < 1$ is not always preferred for the minority class, e.g., 30% and 40% imbalance in Table V when *Dog* is the minority class. Thus we conclude that the results in Tables IV, V, and VI, on binary FMNIST and CIFAR-10, respectively, indicate that practitioners should employ α-loss (typically) for $\alpha < 1$ when training architectures to be sensitive to the minority class in the training data.

C. Key Takeaways

We conclude this section by highlighting the key takeaways from our experimental results.

1) Overall Performance Relative to Log-Loss: The experimental results as evidenced through Tables I to VI suggest that α-loss, more often than not, yields models with improvements in test accuracy over models trained with log-loss, with more prominent gains in the canonical settings of noisy labels and class imbalances in the training data. In order to remedy the extra hyperparameter tuning induced by the seemingly daunting task of searching over $\alpha \in (0, \infty)$, we find that searching over $\alpha \in [.8, 8]$ in the noisy label experiments or $\alpha \in [.8, 4]$ in the class imbalance experiments is sufficient. This aligns with

TABLE IV

BINARY **FMNIST** LOGISTIC REGRESSION IMBALANCE EXPERIMENTS ON THE *T-Shirt* AND *Shirt* CLASSES. NOTE THAT LL-F_1 CORRESPONDS TO THE F_1 SCORE OF LOG-LOSS ON THE IMBALANCED CLASS; SIMILARLY α^*-F_1 CORRESPONDS TO THE F_1 SCORE OF α^*-LOSS ON THE IMBALANCED CLASS. SEE APPENDIX D.1 FOR A BRIEF REVIEW OF THE DEFINITION OF THE F_1 SCORE. THE RELATIVE % GAIN IS DEFINED AS THE RELATIVE PERCENT GAIN (57) ON THE AVERAGE MINORITY CLASS ACCURACY (ON TEST DATA) OF MODELS TRAINED WITH LOG-LOSS VS. THE AVERAGE MINORITY CLASS ACCURACY OF MODELS TRAINED WITH α-LOSS. NOTE THAT OV = OVERALL. LASTLY, OBSERVE THAT FOR THE BASELINE (50% IMBALANCE) EXPERIMENTS, WE PRESENT THE ACCURACY AND α^* FOR BOTH CLASSES

		Log-Loss			α-Loss				
Imb %	Min	Min Acc %	Ov Acc %	LL-F_1	Min Acc %	Ov Acc %	α^*-F_1	α^*	Rel Gain %
50	T-Shirt	85.4	84.31	0.8448	85.7	84.17	0.8441	1.5	0.35
	Shirt	83.2	84.31	0.8413	83.4.	84.33	0.8418	0.85	0.24
40	T-Shirt	80.0	83.68	0.8306	80.2.	83.73	0.8313	1.1	0.25
	Shirt	77.7	83.88	0.8282	77.7	83.90	0.8284	0.99	0.00
30	T-Shirt	72.9	81.89	0.8010	73.0	81.88	0.8011	0.99	0.14
	Shirt	70.8	82.04	0.7977	72.3	82.52	0.8053	0.8	2.12
20	T-Shirt	60.9	77.97	0.7344	61.7	78.20	0.7389	0.8	1.31
	Shirt	63.1	79.81	0.7576	64.5	80.40	0.7669	0.8	2.22
10	T-Shirt	43.0	70.50	0.5931	45.2	71.50	0.6133	0.8	5.12
	Shirt	55.2	76.97	0.7056	56.0	77.25	0.7111	0.8	1.45
5	T-Shirt	24.6	61.85	0.3920	26.0	62.54	0.4097	0.8	5.69
	Shirt	47.5	73.52	0.6421	47.6	73.48	0.6422	0.8	0.21

TABLE V

BINARY **CIFAR-10** CNN 4+2 IMBALANCE EXPERIMENTS ON *Cat* AND *Dog* CLASSES. NOTE THAT LL-F_1 CORRESPONDS TO THE F_1 SCORE OF LOG-LOSS ON THE IMBALANCED CLASS; SIMILARLY α^*-F_1 CORRESPONDS TO THE F_1 SCORE OF α^*-LOSS ON THE IMBALANCED CLASS. NOTE THAT DUE TO OUR CALCULATION OF REL % GAIN THAT DIVISION BY 0 IS ∞, AND THUS ABSOLUTE % GAIN FOR THE MINORITY CLASS *Cat* AT A 5% IMBALANCE IS 9.6%

		Log-Loss			α-Loss				
Imb %	Min	Min Acc %	Ov Acc %	LL-F_1	Min Acc %	Ov Acc %	α^*-F_1	α^*	Rel Gain %
50	Cat	83.7	83.48	0.8352	87.2	83.86	0.8438	1.1	4.18
	Dog	83.3	83.48	0.8345	86.1	84.06	0.8438	0.99	3.36
40	Cat	79.8	83.34	0.8273	82.7	83.39	0.8327	0.95	3.63
	Dog	78.4	83.85	0.8292	82.4	83.20	0.8306	2.5	5.10
30	Cat	73.0	81.98	0.8020	74.6	82.40	0.8000	0.99	2.19
	Dog	72.0	82.00	0.8091	74.9	83.18	0.8166	1.2	4.03
20	Cat	64.6	78.96	0.7543	66.2	78.85	0.7579	0.8	2.48
	Dog	63.1	78.94	0.7498	65.0	79.79	0.7628	0.8	3.01
10	Cat	39.1	68.04	0.5502	41.6	68.88	0.5721	0.9	6.39
	Dog	42.1	70.03	0.5842	48.5	72.53	0.6384	0.8	15.20
5	Cat	0.0	50.00	0.0000	9.6	54.48	0.1742	0.8	∞
	Dog	10.0	54.94	0.1816	23.2	61.31	0.3749	0.8	132.00

TABLE VI

BINARY **CIFAR-10** CNN 6+2 IMBALANCE EXPERIMENTS ON *Cat* AND *Dog* CLASSES

		Log-Loss			α-Loss				
Imb %	Min	Min Acc %	Ov Acc %	LL-F_1	Min Acc %	Ov Acc %	α^*-F_1	α^*	Rel Gain %
50	Cat	84.4	84.30	0.8432	85.2	84.93	0.8497	0.99	0.95
	Dog	84.1	84.30	0.8427	87.0	83.91	0.8439	2	3.45
40	Cat	80.3	83.79	0.8320	82.4	84.87	0.8449	0.8	2.62
	Dog	81.2	84.91	0.8433	84.0	84.83	0.8470	0.9	3.45
30	Cat	74.2	82.72	0.8111	78.2	83.32	0.8242	0.8	5.39
	Dog	73.0	82.92	0.8104	77.2	83.60	0.8248	0.9	5.75
20	Cat	64.6	78.98	0.7545	64.6	78.98	0.7545	1	0.00
	Dog	67.4	81.02	0.7803	70.2	81.75	0.7937	0.99	4.15
10	Cat	38.0	67.69	0.5405	41.8	69.34	0.5769	0.85	10.00
	Dog	46.4	72.14	0.6248	50.1	73.53	0.6543	0.9	7.97
5	Cat	1.7	50.80	0.0334	13.6	56.26	0.2372	0.8	700.00
	Dog	23.7	61.44	0.3807	31.0	64.90	0.4690	0.8	30.80

our earlier theoretical investigations (Section IV) regarding the so-called "Goldilocks zone", i.e., most of the meaningful action induced by α occurs in a narrow region. Notably in the class imbalance experiments, we find that the relevant region is even narrower than our initial choice, i.e., $\alpha^* \in [.8, 2.5]$

(in our search set) for all imbalances. For the noisy label experiments, we always find that $\alpha^* > 1$ and usually α is not too large, and for the class imbalance experiments, we almost always find that $\alpha^* < 1$. These two heuristics enable the practitioner to readily determine a very good α in

these two canonical scenarios. Consequently, α-loss seems to be a principled generalization of log-loss for the practitioner, and it perhaps remedies the concern of Janocha *et al.* in [28] regarding the lack of canonical alternatives to log-loss (cross-entropy loss) in modern machine learning.

VII. CONCLUSION

In this work, we introduced a tunable loss function called α-loss, $\alpha \in (0, \infty]$, which interpolates between the exponential loss ($\alpha = 1/2$), the log-loss ($\alpha = 1$), and the 0-1 loss ($\alpha = \infty$), for the machine learning setting of classification. We illustrated the connection between α-loss and Arimoto conditional entropy (Section II), and then we studied the statistical calibration (Section III), optimization landscape (Section IV), and generalization capabilities (Section V) of α-loss induced by navigating the α hyperparameter. Regarding our main theoretical results, we showed that α-loss is classification-calibrated for all $\alpha \in (0, \infty]$; we also showed that in the logistic model there is a "Goldilocks zone", such that most of the meaningful action induced by α occurs in a narrow region (usually $\alpha \in [.8, 8]$); finally, we showed (under standard distributional assumptions) that empirical minimizers of α-loss for all $\alpha \in (0, \infty]$ are asymptotically optimal with respect to the true 0-1 loss. Practically, following intuitions developed in Section III-C, we performed noisy label and class imbalance experiments on MNIST, FMNIST, and CIFAR-10 using logistic regression and convolutional neural networks (Section VI). Furthermore, we showed that models trained with α-loss can be more robust or sensitive to outliers (depending on the practitioner's choice) over models trained with log-loss ($\alpha = 1$). Therefore, we argue that α-loss seems to be a principled generalization of log-loss for classification algorithms in modern machine learning. Regarding promising avenues to further explore the role of α-loss in machine learning, the robustness of neural-networks to adversarial influence has recently drawn much attention [73]–[75] in addition to learning censored and fair representations that ensure statistical fairness for all downstream learning tasks [76].

APPENDIX

A. α-Loss in Binary Classification

Proposition 2: Consider a soft classifier $g(x) = P_{\hat{Y}|X}(1|x)$. If $f(x) = \sigma^{-1}(g(x))$, then, for every $\alpha \in (0, \infty]$,

$$l^\alpha(y, g(x)) = \tilde{l}^\alpha(yf(x)). \tag{58}$$

Conversely, if f is a classification function, then the set of beliefs $P_{\hat{Y}|X}$ associated to $g(x) := \sigma(f(x))$ satisfies (11). In particular, for every $\alpha \in (0, \infty]$,

$$\min_g \mathbb{E}_{X,Y}(l^\alpha(Y, g(x))) = \min_f \mathbb{E}_{X,Y}(\tilde{l}^\alpha(Yf(X))). \tag{59}$$

Proof: 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

$$l^\alpha(y, P_{\hat{Y}|X=x}) = \tilde{l}^\alpha(yf(x)). \tag{60}$$

We assume that $\alpha \in (0, 1) \cup (1, \infty)$. 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

$$l^\alpha(1, P_{\hat{Y}|X}(1|x)) = l^\alpha(1, \sigma(f(x))) \tag{61}$$

$$= \frac{\alpha}{\alpha - 1}\left[1 - \sigma(f(x))^{1-1/\alpha}\right] \tag{62}$$

$$= \tilde{l}^\alpha(f(x)). \tag{63}$$

If $y = -1$, then

$$l^\alpha(-1, P_{\hat{Y}|X}(-1|x)) = l^\alpha(-1, 1 - P_{\hat{Y}|X}(1|x)) \tag{64}$$

$$= l^\alpha(-1, 1 - \sigma(f(x))) \tag{65}$$

$$= l^\alpha(-1, \sigma(-f(x))) \tag{66}$$

$$= \frac{\alpha}{\alpha - 1}[1 - \sigma(-f(x))^{1-1/\alpha}] \tag{67}$$

$$= \tilde{l}^\alpha(-f(x)), \tag{68}$$

where (66) follows from

$$\sigma(x) + \sigma(-x) = 1, \tag{69}$$

which can be observed by (9). To show the reverse direction of (60) we substitute

$$f(x) = \sigma^{-1}(g(x)) = \sigma^{-1}(P_{\hat{Y}|X}(1|x)), \tag{70}$$

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

$$\tilde{l}^\alpha(f(x)) = \tilde{l}^\alpha(\sigma^{-1}(P_{\hat{Y}|X}(1|x))) \tag{71}$$

$$= \frac{\alpha}{\alpha - 1}[1 - (\sigma(\sigma^{-1}(P_{\hat{Y}|X}(1|x))))^{1-1/\alpha}] \tag{72}$$

$$= \frac{\alpha}{\alpha - 1}[1 - P_{\hat{Y}|X}(1|x)^{1-1/\alpha}] \tag{73}$$

$$= l^\alpha(1, P_{\hat{Y}|X}(1|x)). \tag{74}$$

For $y = -1$,

$$\tilde{l}^\alpha(-f(x)) = \tilde{l}^\alpha(-\sigma^{-1}(P_{\hat{Y}|X}(1|x))) \tag{75}$$

$$= \frac{\alpha}{\alpha - 1}[1 - \sigma(-\sigma^{-1}(P_{\hat{Y}|X}(1|x)))^{1-1/\alpha}] \tag{76}$$

$$= \frac{\alpha}{\alpha - 1}[1 - (1 - \sigma(\sigma^{-1}(P_{\hat{Y}|X}(1|x))))^{1-1/\alpha}] \tag{77}$$

$$= \frac{\alpha}{\alpha - 1}[1 - P_{\hat{Y}|X}(-1|x)^{1-1/\alpha}] \tag{78}$$

$$= l^\alpha(-1, P_{\hat{Y}|X}(-1|x)), \tag{79}$$

where (77) follows from (69).

The equality in the results of the minimization procedures follows from the equality between l^α and $\tilde{l}^\alpha$. As was shown in [17], the minimizer of the left-hand-side is

$$P^*_{\hat{Y}|X}(y|x) = \frac{P_{Y|X}(y|x)^\alpha}{\sum_y P_{Y|X}(y|x)^\alpha}. \tag{80}$$

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

Proposition 3: As a function of the margin, $\tilde{l}^\alpha : \overline{\mathbb{R}} \to \mathbb{R}_+$ is convex for $\alpha \leq 1$ and quasi-convex for $\alpha > 1$.

Proof: The second derivative of the margin-based α-loss for $\alpha \in (0, \infty]$ with respect to the margin is given by

$$\frac{d^2}{dz^2}\tilde{l}^\alpha(z) = \frac{(e^{-z} + 1)^{1/\alpha}e^z(\alpha e^z - \alpha + 1)}{\alpha(e^z + 1)^3}. \tag{81}$$

Observe that if $\alpha \in (0, 1]$, then we have that, for all $z \in \overline{\mathbb{R}}$, $\frac{d^2}{dz^2}\tilde{l}^\alpha(z) \geq 0$, which implies that $\tilde{l}^\alpha$ is convex [62]. If we have $\alpha \in (1, \infty]$, then note that $\alpha e^z - \alpha + 1 < 0$ for all $z \in \overline{\mathbb{R}}$ such that $z < \log\left(1 - \alpha^{-1}\right)$. Thus, the margin-based α-loss, $\tilde{l}^\alpha$, is not convex for $\alpha \in (1, \infty]$. However, observe that

$$\frac{d}{dz}\tilde{l}^\alpha(z) = \frac{-(e^{-z} + 1)^{1/\alpha}e^z}{(1 + e^z)^2}. \tag{82}$$

Since $\frac{d}{dz}\tilde{l}^\alpha(z) < 0$ for $\alpha \in [1, \infty]$ and for all $z \in \overline{\mathbb{R}}$, $\tilde{l}^\alpha$ is monotonically decreasing. Furthermore, since monotonic functions are quasi-convex [62], we have that $\tilde{l}^\alpha$ is quasi-convex for $\alpha > 1$. $\square$

Theorem 1: For $\alpha \in (0, \infty]$, the margin-based α-loss $\tilde{l}^\alpha$ is classification-calibrated. In addition, its optimal classification function is given by

$$f_\alpha^*(\eta) = \alpha \cdot \sigma^{-1}(\eta). \tag{83}$$

Proof: We first show that $\tilde{l}^\alpha$ is classification-calibrated for all $\alpha \in (0, \infty]$. Suppose that $\alpha \in (0, 1]$; we rely on the following result by Bartlett *et al.* in [6].

Proposition 7 (Thm. 6, [6]): Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function in the margin. Then ϕ is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$.

Observe that $\tilde{l}^\alpha$ is smooth and monotonically decreasing for all $\alpha \in (0, \infty]$, and for $\alpha \in (0, 1]$, $\tilde{l}^\alpha$ is convex by Proposition 3. Thus, $\tilde{l}^\alpha$ satisfies Proposition 7, which implies that $\tilde{l}^\alpha$ is classification-calibrated for $\alpha \in (0, 1)$.

Now consider $\alpha \in (1, \infty)$. Since classification-calibration requires proving that the minimizer of (14) agrees in sign with the Bayes predictor, we first obtain the minimizer of the conditional risk for all $\eta \neq 1/2$. We have that

$$\inf_{f \in \mathbb{R}} C_{\tilde{l}^\alpha}(\eta, f) = \inf_{f \in \mathbb{R}} \eta \tilde{l}^\alpha(f) + (1 - \eta)\tilde{l}^\alpha(-f) \tag{84}$$

$$= \frac{\alpha}{\alpha - 1}\left(1 - \sup_{f \in \mathbb{R}}\left[\eta\sigma(f)^{1-1/\alpha} + (1 - \eta)\sigma(-f)^{1-1/\alpha}\right]\right), \tag{85}$$

where we substituted $\tilde{l}^\alpha$ into (84) and pulled the infimum through. We take the derivative of the expression inside the supremum, which we denote $g(\eta, \alpha, f)$, and obtain

$$\frac{d}{df}g(\eta, \alpha, f) = \left(1 - \frac{1}{\alpha}\right)\left(\frac{1}{e^f + 2 + e^{-f}}\right)$$
$$\times \left[\eta\left(1 + e^{-f}\right)^{\frac{1}{\alpha}} - (1 - \eta)\left(1 + e^f\right)^{\frac{1}{\alpha}}\right]. \tag{86}$$

One can then obtain the f^* minimizing (84) by setting $\frac{d}{df}g(\eta, \alpha, f) = 0$, i.e.,

$$\eta\left(1 + e^{-f^*}\right)^{1/\alpha} = (1 - \eta)\left(1 + e^{f^*}\right)^{1/\alpha}, \tag{87}$$

and solving for f^* we have

$$f_\alpha^*(\eta) = \alpha\log\left(\frac{\eta}{1 - \eta}\right) = \alpha \cdot \sigma^{-1}(\eta). \tag{88}$$

Recall that the Bayes predictor, which is optimal, is given by $h_{\text{Bayes}}(\eta) = \text{sign}(2\eta - 1)$, and notice that the classification function representation is simply $f_{\text{Bayes}}(\eta) = 2\eta - 1$. Observe that for all $\eta \neq 1/2$ and for $\alpha \in [1, \infty)$ (indeed $\alpha < 1$ as well), we have that $\text{sign}(f_{\text{Bayes}}(\eta)) = \text{sign}(f_\alpha^*(\eta))$. Thus, $\tilde{l}^\alpha$ is classification-calibrated for $\alpha \in (0, \infty)$. Lastly, if $\alpha = +\infty$, then $\tilde{l}^\alpha$ becomes

$$\tilde{l}^\infty(z) = 1 - \sigma(z) = \frac{e^z}{1 + e^z}, \tag{89}$$

which is sigmoid loss. Similarly, sigmoid loss can be shown to be classification-calibrated as is given in [6]. Therefore, $\tilde{l}^\alpha$ is classification-calibrated for all $\alpha \in (0, \infty]$.

Finally, note that the proof of classification-calibration yielded the optimal classification function given in (88) for all $\alpha \in (0, \infty]$. Alternatively, the optimal classification function can be obtained from Proposition 1 by Liao *et al.* Specifically, substitute the α-tilted distribution (6) for a binary label $\mathcal{Y} = \{-1, +1\}$ into (10) as stated by Proposition 2. Indeed, we have that

$$f^*(x) = \sigma^{-1}(P_{\hat{Y}|X}^*(1|x)) \tag{90}$$

$$= \log\left(\frac{P_{Y|X}(1|x)^\alpha}{P_{Y|X}(-1|x)^\alpha}\right) \tag{91}$$

$$= \alpha\log\left(\frac{\eta(x)}{1 - \eta(x)}\right), \tag{92}$$

which aligns with (16). $\square$

Corollary 1: For $\alpha \in (0, \infty]$, the minimum conditional risk $C_\alpha^*(\eta)$ of $\tilde{l}^\alpha$ is equal to

$$\begin{cases} \frac{\alpha}{\alpha - 1}\left(1 - (\eta^\alpha + (1 - \eta)^\alpha)^{1/\alpha}\right) & \alpha \in (0, 1) \cup (1, +\infty), \\ -\eta\log\eta - (1 - \eta)\log(1 - \eta) & \alpha = 1, \\ \min\{\eta, 1 - \eta\} & \alpha \to +\infty. \end{cases} \tag{93}$$

Proof: For $\alpha = 1$, we recover logistic loss and we know from [7] and [1] that the minimum conditional risk is given by

$$C_1^*(\eta) = -\eta\log\eta - (1 - \eta)\log(1 - \eta). \tag{94}$$

Similarly, for $\alpha = \infty$, we recover the sigmoid loss and we know from [6] and [1] that the minimum conditional risk is given by

$$C_\infty^*(\eta) = \min\{\eta, 1 - \eta\}. \tag{95}$$

Thus, we now consider the case where $\alpha \in (0, \infty) \setminus \{1\}$. The conditional risk of $\tilde{l}^\alpha$ is given by

$$C_\alpha(\eta, f) = \eta\tilde{l}^\alpha(f) + (1 - \eta)\tilde{l}^\alpha(-f) \tag{96}$$

$$= \frac{\alpha}{\alpha - 1}\left[1 - \eta\sigma(f)^{1-1/\alpha} - (1 - \eta)\sigma(-f)^{1-1/\alpha}\right], \tag{97}$$

where we substituted (8) into (96). We can obtain the minimum conditional risk upon substituting (16) into (97) which yields

$$C_\alpha^*(\eta) = \frac{\alpha}{\alpha - 1} - \frac{\alpha}{\alpha - 1}(1 - \eta)\left(\frac{(1 - \eta)^\alpha}{\eta^\alpha + (1 - \eta)^\alpha}\right)^{1 - 1/\alpha}$$
$$- \frac{\alpha}{\alpha - 1}\eta\left(\frac{\eta^\alpha}{\eta^\alpha + (1 - \eta)^\alpha}\right)^{1 - 1/\alpha} \quad (98)$$
$$= \frac{\alpha}{\alpha - 1}\left[1 - (\eta^\alpha + (1 - \eta)^\alpha)^{1/\alpha}\right], \quad (99)$$

where the last equation is obtained after some algebra. Finally, observe that $C_{1/2}^*(\eta) = 2\sqrt{\eta(1 - \eta)}$, which aligns with [7]. □

B. Optimization Guarantees for α-Loss in the Logistic Model

Theorem 2: Let $\Sigma := \mathbb{E}[XX^\intercal]$. If $\alpha \in (0, 1]$, then $R_\alpha(\theta)$ is $\Lambda(\alpha, r\sqrt{d})\min_{i \in [d]} \lambda_i(\Sigma)$-strongly convex in $\theta \in \mathbb{B}_d(r)$, where

$$\Lambda(\alpha, r\sqrt{d}) := \sigma(r\sqrt{d})^{1 - 1/\alpha}$$
$$\times \left(\sigma'(r\sqrt{d}) - \left(1 - \frac{1}{\alpha}\right)\sigma(-r\sqrt{d})^2\right). \quad (100)$$

Proof: For each $\alpha \in (0, 1]$, it can readily be shown that each component of $F_2(\alpha, \theta, x, y)$ is positive and monotonic in $\langle \theta, x \rangle$, which implies that $F_2(\alpha, \theta, x, y) \geq \Lambda(\alpha, r\sqrt{d}) > 0$. Now, consider $R_\alpha(\theta) = \mathbb{E}[l^\alpha(Y, g_\theta(X))]$. We have

$$\nabla_\theta^2 R_\alpha(\theta) = \mathbb{E}_{X,Y}[\nabla_\theta^2 l^\alpha(Y, g_\theta(X))] \quad (101)$$
$$= \mathbb{E}_{X,Y}[F_2(\alpha, \theta, X, Y)XX^\intercal] \quad (102)$$
$$\succeq \Lambda(\alpha, r\sqrt{d})\mathbb{E}[XX^\intercal] \quad (103)$$
$$= \Lambda(\alpha, r\sqrt{d})\Sigma \succeq 0, \quad (104)$$

where we used an identity of positive semi-definite matrices for (103) (see, e.g., [77, Ch. 7]); for (104), we used the fact that $\Lambda(\alpha, r\sqrt{d}) \geq 0$ and we recognize that Σ is positive semi-definite as it is the correlation of the random vector $X \in [0, 1]^d$ (see, e.g., [78, Ch. 7]). We also note that $\min_{i \in [d]} \lambda_i(\Sigma) \geq 0$ (see, e.g., [77, Ch. 7]). Thus, $\nabla_\theta^2 R_\alpha(\theta)$ is positive semi-definite for every $\theta \in \mathbb{B}_d(r)$. Therefore, since $\lambda_{\min}(\nabla^2 R_\alpha(\theta)) \geq \Lambda(\alpha, r\sqrt{d})\min_{i \in [d]} \lambda_i(\Sigma) \geq 0$ for every $\theta \in \mathbb{B}_d(r)$, which follows by the Courant-Fischer min-max theorem [77, Theorem 4.2.6], we have that R_α is $\Lambda(\alpha, r\sqrt{d})\min_{i \in [d]} \lambda_i(\Sigma)$-strongly convex for $\alpha \in (0, 1]$. □

Corollary 2: Let $\Sigma := \mathbb{E}[XX^\intercal]$. If $r\sqrt{d} < \operatorname{arcsinh}(1/2)$, then $R_\alpha(\theta)$ is $\tilde{\Lambda}(\alpha, r\sqrt{d})\min_{i \in [d]} \lambda_i(\Sigma)$-strongly convex in $\theta \in \mathbb{B}_d(r)$ for $\alpha \in \left(0, (e^{2r\sqrt{d}} - e^{r\sqrt{d}})^{-1}\right]$, where

$$\tilde{\Lambda}(\alpha, r\sqrt{d}) := \sigma(-r\sqrt{d})^{2 - 1/\alpha}$$
$$\times \sigma(r\sqrt{d})\left(1 - e^{r\sqrt{d}} + \frac{1}{\alpha}e^{-r\sqrt{d}}\right). \quad (105)$$

Proof: Let $\theta \in \mathbb{B}_d(r)$ be arbitrary. We similarly have that

$$\nabla_\theta^2 R_\alpha(\theta) = \mathbb{E}_{X,Y}[\nabla_\theta^2 l^\alpha(Y, g_\theta(X))] \quad (106)$$
$$= \mathbb{E}_{X,Y}[g_\theta(YX)^{1 - 1/\alpha}(g_\theta'(YX)$$
$$- \left(1 - \frac{1}{\alpha}\right)g_\theta(-YX)^2)XX^\intercal] \quad (107)$$
$$= \mathbb{E}_{X,Y}[g_\theta(YX)^{1 - 1/\alpha}g_\theta(-YX)(g_\theta(YX)$$
$$- \left(1 - \frac{1}{\alpha}\right)g_\theta(-YX))XX^\intercal], \quad (108)$$

where we recall (102) and factored out $g_\theta(-YX)$. Considering the expression in parentheses in (108), we note that this is the only part of the Hessian which can become negative. Examining this term more closely, we find that

$$g_\theta(YX) - \left(1 - \frac{1}{\alpha}\right)g_\theta(-YX)$$
$$= \frac{1}{1 + e^{-\langle \theta, YX \rangle}} - \left(1 - \frac{1}{\alpha}\right)\frac{1}{1 + e^{\langle \theta, YX \rangle}} \quad (109)$$
$$= g_\theta(YX)\left[1 - \left(1 - \frac{1}{\alpha}\right)\frac{1 + e^{-\langle \theta, YX \rangle}}{1 + e^{\langle \theta, YX \rangle}}\right] \quad (110)$$
$$= g_\theta(YX)\left[1 - \left(1 - \frac{1}{\alpha}\right)e^{-\langle \theta, YX \rangle}\right]. \quad (111)$$

Continuing, observe that

$$1 - \left(1 - \frac{1}{\alpha}\right)e^{-\langle \theta, YX \rangle} = 1 - e^{-\langle \theta, YX \rangle} + \frac{e^{-\langle \theta, YX \rangle}}{\alpha} \quad (112)$$
$$\geq 1 - e^{r\sqrt{d}} + \frac{e^{-r\sqrt{d}}}{\alpha} \geq 0, \quad (113)$$

where we lowerbound using the radius of the balls (Cauchy-Schwarz), i.e., $\langle \theta, YX \rangle \leq |Y|\|\theta\|\|X\| \leq r\sqrt{d}$ and the last inequality in (113) holds if $\alpha \leq e^{-r\sqrt{d}}(e^{r\sqrt{d}} - 1)^{-1}$. Thus, returning to (108), we have that

$$\nabla_\theta^2 R_\alpha(\theta)$$
$$= \mathbb{E}_{X,Y}[g_\theta(YX)^{1 - \frac{1}{\alpha}}g_\theta'(YX)(1 - (1 - \frac{1}{\alpha})e^{-\langle \theta, YX \rangle})XX^\intercal] \quad (114)$$
$$\succeq \sigma(-r\sqrt{d})^{2 - \frac{1}{\alpha}}\sigma(r\sqrt{d})\left(1 - e^{r\sqrt{d}} + \frac{e^{-r\sqrt{d}}}{\alpha}\right)\mathbb{E}[XX^\intercal] \quad (115)$$
$$= \sigma(-r\sqrt{d})^{2 - \frac{1}{\alpha}}\sigma(r\sqrt{d})\left(1 - e^{r\sqrt{d}} + \frac{e^{-r\sqrt{d}}}{\alpha}\right)\Sigma \succeq 0, \quad (116)$$

where in (114) we used (111) and the fact as given in (27) that $\sigma'(z) = \sigma(z)\sigma(-z)$, and in (115) and (116) we use the upper-bound derived above and the same arguments as Theorem 2, *mutatis mutandis*. Thus, if we have the following bound $\alpha \leq e^{-r\sqrt{d}}(e^{r\sqrt{d}} - 1)^{-1}$, then we have that $R_\alpha(\theta)$ is $\tilde{\Lambda}(\alpha, r\sqrt{d})\min_{i \in [d]} \lambda_i(\Sigma)$-strongly convex in $\theta \in \mathbb{B}_d(r)$,

$$\tilde{\Lambda}(\alpha, r\sqrt{d}) := \sigma(-r\sqrt{d})^{2 - 1/\alpha}$$
$$\times \sigma(r\sqrt{d})\left(1 - e^{r\sqrt{d}} + \alpha^{-1}e^{-r\sqrt{d}}\right). \quad (117)$$

Finally, recall that $\sinh(x) = (e^x - e^{-x})/2$ and that $\operatorname{arcsinh} x = \log\left(x + \sqrt{x^2 + 1}\right)$. Observe that if we have $r\sqrt{d} \leq \operatorname{arcsinh}(1/2)$, then $e^{-r\sqrt{d}}(e^{r\sqrt{d}} - 1)^{-1} \geq 1$. Also note that $e^{-r\sqrt{d}}(e^{r\sqrt{d}} - 1)^{-1}$ is monotonically decreasing in $r\sqrt{d}$ and that $\operatorname{arcsinh}(1/2) \approx 0.48$. $\qquad\square$

Proposition 5: Suppose that $\Sigma \succ 0$ and $\theta_0 \in \mathbb{B}_d(r)$ is fixed. We have one of the following:

- If $r\sqrt{d} < \operatorname{arcsinh}(1/2)$, then, for every $\epsilon > 0$, R_α is $(\epsilon, C_d(r, \alpha), \theta_0)$-SLQC at θ for every $\theta \in \mathbb{B}_d(r)$ when $\alpha \in \left(0, (e^{2r\sqrt{d}} - e^{r\sqrt{d}})^{-1}\right]$ where $C_d(r, \alpha)$ is given in (36) and (37);
- Otherwise, for every $\epsilon > 0$, R_α is $(\epsilon, C_d(r, \alpha), \theta_0)$-SLQC at θ for every $\theta \in \mathbb{B}_d(r)$ for $\alpha \in (0, 1]$.

Proof: In order to prove the result, we apply a result by Hazan, *et al.* [35] where they show that if a function f is G-Lipschitz and strictly-quasi-convex, then for all $\epsilon > 0$, f is (ϵ, G, θ_0)-SLQC in θ. Thus, one may view κ as approximately quantifying the growth of the gradients of general functions.

First, we show that R_α is $C_d(r, \alpha)$-Lipschitz in $\theta \in \mathbb{B}_d(r)$ where for $\alpha \in (0, 1]$,

$$C_d(r, \alpha) := \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha}; \qquad (118)$$

and, for $\alpha \in (1, \infty]$,

$$C_d(r, \alpha) := \begin{cases} \sqrt{d}\left(\frac{\alpha-1}{2\alpha-1}\right)^{1-1/\alpha}\left(\frac{\alpha}{2\alpha-1}\right) & e^{r\sqrt{d}} \geq \frac{\alpha-1}{\alpha}, \\ \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha} & e^{r\sqrt{d}} < \frac{\alpha-1}{\alpha}. \end{cases} \qquad (119)$$

Formally, we want to show that for all $\theta, \theta' \in \mathbb{B}_d(r)$,

$$|R_\alpha(\theta) - R_\alpha(\theta')| \leq C\|\theta - \theta'\|, \qquad (120)$$

where $C := \sup_{\theta \in \mathbb{B}_d(r)} \|\nabla R_\alpha(\theta)\|$. Recall from (29) that

$$\nabla_\theta R_\alpha(\theta) = \mathbb{E}[\nabla_\theta l^\alpha(Y, g_\theta(X)] \qquad (121)$$
$$= \mathbb{E}[F_1(\alpha, \theta, X, Y)X], \qquad (122)$$

where from (28) we have

$$F_1(\alpha, \theta, x, y) = -y g_\theta(yx)^{1-1/\alpha}(1 - g_\theta(yx)). \qquad (123)$$

It can be shown that for $\alpha \leq 1$,

$$|F_1(\alpha, \theta, x, y)| = g_\theta(yx)^{1-1/\alpha}(1 - g_\theta(yx)), \qquad (124)$$

is monotonically decreasing in $\langle \theta, x \rangle$. Thus for $\alpha \leq 1$,

$$C = \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha}. \qquad (125)$$

It can also be shown that for $\alpha > 1$, $|F_1(\alpha, \theta, x, y)|$ is unimodal and quasi-concave with the maximum obtained at $\langle \theta, x \rangle^* = \log(1 - 1/\alpha)$. If $r\sqrt{d} \geq \log(1 - 1/\alpha)$, we obtain upon plugging in $\langle \theta, x \rangle^*$ for $\alpha > 1$,

$$C = \sqrt{d}\left(\frac{\alpha-1}{2\alpha-1}\right)^{1-1/\alpha}\left(\frac{\alpha}{2\alpha-1}\right). \qquad (126)$$

Otherwise, if $r\sqrt{d} < \log(1 - 1/\alpha)$, then, using the local monotonicity of $|F_1(\alpha, \theta, x, y)|$, we obtain for $\alpha > 1$,

$$C = \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha}, \qquad (127)$$

which mirrors the $\alpha < 1$ case. Thus, combining the two regimes of α we have that R_α is $C_d(r, \alpha)$-Lipschitz in $\theta \in \mathbb{B}_d(r)$ for $\alpha \in (0, \infty]$ where $C_d(r, \alpha)$ is given in (36) and (37).

Finally when R_α is strongly-convex, this implies that R_α is strictly-quasi-convex. That is, since $\Sigma \succ 0$, we merely apply Corollary 2 to obtain strong-convexity of R_α when $\alpha \in (0, (e^{2r\sqrt{d}} - e^{r\sqrt{d}})^{-1}]$ for $r\sqrt{d} < \operatorname{arcsinh}(1/2)$. Similarly, we apply Theorem 2 to obtain strong-convexity of R_α for $\alpha \in (0, 1]$, otherwise. $\qquad\square$

1) Fundamentals of SLQC and Reformulation: In this subsection, we briefly review *strictly locally quasi-convexity* (SLQC) which was introduced by Hazan *et al.* in [35]. Recall that in [35] Hazan *et al.* refer to a function as SLQC *in* θ, whereas for the purposes of our analysis we refer to a function as SLQC *at* θ. We recover the uniform SLQC notion of Hazan *et al.* by articulating a function is SLQC *at* θ *for every* θ. Our later analysis of the α-risk in the logistic model benefits from this pointwise consideration. Intuitively, the notion of SLQC functions extends quasi-convex functions in a parameterized manner. Regarding notation, for $\theta_0 \in \mathbb{R}^d$ and $r > 0$, we let $\mathbb{B}(\theta_0, r) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\| \leq r\}$.

Definition 3 (Definition 3.1, [35]): Let $\epsilon, \kappa > 0$ and $\theta_0 \in \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called $(\epsilon, \kappa, \theta_0)$-strictly locally quasi-convex (SLQC) at $\theta \in \mathbb{R}^d$ if at least one of the following conditions apply:
1. $f(\theta) - f(\theta_0) \leq \epsilon$,
2. $\|\nabla f(\theta)\| > 0$ and, for every $\theta' \in \mathbb{B}(\theta_0, \epsilon/\kappa)$,

$$\langle -\nabla f(\theta), \theta' - \theta \rangle \geq 0. \qquad (128)$$

Observe that the notion of SLQC implies quasi-convexity about $\mathbb{B}(\theta_0, \epsilon/\kappa)$ on $\{\theta \in \Theta : f(\theta) - f(\theta_0) > \epsilon\}$; see Fig. 5 for an illustration of the difference between classical quasi-convexity and SLQC in this regime. In [35], Hazan *et al.* note that if a function f is G-Lipschitz and strictly-quasi-convex, then for all $\tilde{\theta}_1, \tilde{\theta}_2 \in \mathbb{R}^d$, for all $\epsilon > 0$, it holds that f is $(\epsilon, G, \tilde{\theta}_1)$-SLQC at $\tilde{\theta}_2$ for every $\tilde{\theta}_2 \in \mathbb{R}^d$; this will be useful in the sequel.

As shown by Hazan *et al.* in [35], the convergence guarantees of Normalized Gradient Descent (NGD, given in Algorithm 1) for SLQC functions are similar to those of Gradient Descent for convex functions.

Algorithm 1 Normalized Gradient Descent (NGD)

1: **Input:** $T \in \mathbb{N}$ no. of iterations, $\theta_0 \in \mathbb{R}^d$ initial parameter, $\eta > 0$ learning rate
2: **for** $t = 0, 1, \dots, T - 1$ **do**
3:     Update: $\theta_{t+1} = \theta_t - \eta \dfrac{\nabla f(\theta_t)}{\|\nabla f(\theta_t)\|}$
4: **Return** $\bar{\theta}_T = \underset{\theta_1, \dots, \theta_T}{\arg\min} f(\theta_t)$

Proposition 4 (Thm. 4.1, [35]): Let $f : \mathbb{R}^d \to \mathbb{R}$, $\theta_1 \in \mathbb{R}^d$, and $\theta^* = \arg\min_{\theta \in \mathbb{R}^d} f(\theta)$. If f is $(\epsilon, \kappa, \theta^*)$-SLQC at θ for every $\theta \in \mathbb{R}^d$, then running the NGD algorithm with learning rate $\eta = \epsilon/\kappa$ for number of iterations $T \geq \kappa^2 \|\theta_1 - \theta^*\|^2/\epsilon^2$ achieves $\min_{t=1,\dots,T} f(\theta_t) - f(\theta^*) \leq \epsilon$.
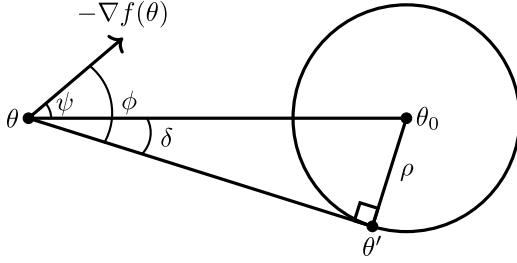
Fig. 7. A companion illustration for Lemma 1 which depicts the relevant quantities involved. Note that there are three different configurations of the angles δ, ϕ and ψ. Refer to Fig. 8 for this illustration.

For an $(\epsilon, \kappa, \theta_0)$-SLQC function, a smaller ϵ provides better optimality guarantees. Given $\epsilon > 0$, smaller κ leads to faster optimization as the number of required iterations increases with κ^2. Hazan, *et al.* [35] show that if a function f is G-Lipschitz and strictly-quasi-convex, then for all $\epsilon > 0$, f is (ϵ, G, θ_0)-SLQC in θ. Thus, one may view κ as approximately quantifying the growth of the gradients of general functions. Finally, by using projections, NGD can be easily adapted to work over convex and closed sets (e.g., $\mathbb{B}(\theta_0, r)$ for some $\theta_0 \in \mathbb{R}^d$ and $r > 0$).

We conclude this subsection by studying the behavior of $(\epsilon, \kappa, \theta_0)$-SLQC functions on the ball $\overline{\mathbb{B}_d(\theta_0, \epsilon/\kappa)}$, which is articulated by the following novel result.

Proposition 8: Let $\epsilon, \kappa > 0$ and $\theta_0 \in \mathbb{R}^d$. Assume f is $(\epsilon, \kappa, \theta_0)$-SLQC at $\theta \in \mathbb{R}^d$. If $\theta \in \mathbb{B}_d(\theta_0, \epsilon/\kappa)$, then $f(\theta) - f(\theta_0) \leq \epsilon$. Indeed, if f is $(\epsilon, \kappa, \theta_0)$-SLQC on Θ, then

$$\overline{\mathbb{B}_d(\theta_0, \epsilon/\kappa) \cap \Theta} \subset \{\theta \in \Theta : f(\theta) - f(\theta_0) \leq \epsilon\}.$$

Proof: Since f is $(\epsilon, \kappa, \theta_0)$-SLQC at $\theta \in \mathbb{R}^d$ we have that at least one condition of Definition 3 holds. Suppose that Condition 2 holds. In this case, we have that $\|\nabla f(\theta)\| > 0$ and $\langle -\nabla f(\theta), \theta' - \theta \rangle \geq 0$ for every $\theta' \in \mathbb{B}(\theta_0, \epsilon/\kappa)$. Since $\|\theta - \theta_0\| < \epsilon/\kappa$, choose $\delta > 0$ small enough such that

$$\theta' := \theta + \delta \nabla f(\theta) \in \mathbb{B}(\theta_0, \epsilon/\kappa). \tag{129}$$

Thus, we have that

$$0 \leq \langle -\nabla f(\theta), \theta' - \theta \rangle \tag{130}$$
$$= \langle -\nabla f(\theta), \theta + \delta \nabla f(\theta) - \theta \rangle \tag{131}$$
$$= -\delta \langle \nabla f(\theta), \nabla f(\theta) \rangle \tag{132}$$
$$= -\delta \|\nabla f(\theta)\|^2, \tag{133}$$

which is a contradiction since $\delta > 0$ and $\|\nabla f(\theta)\| > 0$. Therefore, we must have that Condition 1 of Definition 3 holds, i.e., $f(\theta) - f(\theta_0) \leq \epsilon$. Finally, a continuity argument shows that $f(\theta) - f(\theta_0) \leq \epsilon$ whenever $\theta \in \overline{\mathbb{B}_d(\theta_0, \epsilon/\kappa) \cap \Theta}$. $\square$

The following is the formal statement and proof of Lemma 1, which provides a useful characterization of the gradient of $(\epsilon, \kappa, \theta_0)$-SLQC functions outside the set $\overline{\mathbb{B}_d(\theta_0, \epsilon/\kappa)}$. Refer to Fig. 7 for a picture of the relevant quantities.

Lemma 1: Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, $\theta_0 \in \mathbb{R}^d$ and $\rho > 0$. If $\theta \in \mathbb{R}^d$ is such that $\|\theta - \theta_0\| > \rho$ and $\|\nabla f(\theta)\| > 0$, then the following are equivalent:

(1) $\langle -\nabla f(\theta), \theta' - \theta \rangle > 0$ for all $\theta' \in \mathbb{B}_d(\theta_0, \rho)$;
(2) $\langle -\nabla f(\theta), \theta' - \theta \rangle \geq 0$ for all $\theta' \in \mathbb{B}_d(\theta_0, \rho)$;
(3) $\langle -\nabla f(\theta), \theta_0 - \theta \rangle \geq \rho \|\nabla f(\theta)\|$.

Proof: Clearly (1) $\Rightarrow$ (2). (2) $\Rightarrow$ (3): Let θ' be the point of tangency of a line tangent to $\overline{\mathbb{B}_d(\theta_0, \rho)}$ passing through θ, as depicted in Fig. 7. We define

δ: the angle between $\theta_0 - \theta$ and $\theta' - \theta$;
ϕ: the angle between $-\nabla f(\theta)$ and $\theta' - \theta$;
ψ: the angle between $-\nabla f(\theta)$ and $\theta_0 - \theta$.

Recall that the inner product satisfies that

$$\langle u, v \rangle = \|u\| \|v\| \cos(\varphi_{u,v}), \tag{134}$$

where $\varphi_{u,v} \in [0, \pi]$ is the angle between u and v. By continuity and Condition (2),

$$\|\nabla f(\theta)\| \|\theta' - \theta\| \cos(\phi) = \langle -\nabla f(\theta), \theta' - \theta \rangle \geq 0, \tag{135}$$

which implies that $\phi \leq \frac{\pi}{2}$. Observe that, by construction, we have $\phi = \psi + \delta$. In particular, we have that $\psi \leq \frac{\pi}{2} - \delta$. Since $\cos(\cdot)$ is decreasing over $[0, \pi]$, we have that

$$\cos(\psi) \geq \cos\left(\frac{\pi}{2} - \delta\right) = \sin(\delta). \tag{136}$$

Since the triangle $\triangle \theta \theta' \theta_0$ is a right triangle, we have that $\sin(\delta) = \frac{\rho}{\|\theta_0 - \theta\|}$ and thus

$$\cos(\psi) \geq \frac{\rho}{\|\theta_0 - \theta\|}. \tag{137}$$

Therefore, we conclude that

$$\langle -\nabla f(\theta), \theta_0 - \theta \rangle = \|\nabla f(\theta)\| \|\theta_0 - \theta\| \cos(\psi) \tag{138}$$
$$\geq \rho \|\nabla f(\theta)\|, \tag{139}$$

as we wanted to prove.

(3) $\Rightarrow$ (1): For a given $\theta' \in \mathbb{B}_d(\theta_0, \rho)$, we define ψ, ϕ and δ as above. By assumption,

$$\|\nabla f(\theta)\| \|\theta_0 - \theta\| \cos(\psi) = \langle -\nabla f(\theta), \theta_0 - \theta \rangle \tag{140}$$
$$\geq \rho \|\nabla f(\theta)\| \geq 0. \tag{141}$$

Since $\cos^{-1}(\cdot)$ is decreasing over $[-1, 1]$, (140) implies that

$$\psi \leq \cos^{-1}\left(\frac{\rho}{\|\theta_0 - \theta\|}\right). \tag{142}$$

Also, an immediate application of the law of cosines yields

$$\delta = \cos^{-1}\left(\frac{\|\theta_0 - \theta\|^2 + \|\theta' - \theta\|^2 - \|\theta' - \theta_0\|^2}{2\|\theta_0 - \theta\| \|\theta' - \theta\|}\right). \tag{143}$$

Since $\|\theta' - \theta_0\| < \rho$, we have that

$$\delta < \cos^{-1}\left(\frac{\|\theta_0 - \theta\|^2 + \|\theta' - \theta\|^2 - \rho^2}{2\|\theta_0 - \theta\| \|\theta' - \theta\|}\right). \tag{144}$$

A routine minimization argument further implies that

$$\delta < \cos^{-1}\left(\sqrt{1 - \left(\frac{\rho}{\|\theta_0 - \theta\|}\right)^2}\right) = \sin^{-1}\left(\frac{\rho}{\|\theta_0 - \theta\|}\right), \tag{145}$$

where the equality follows from the trigonometric identity $\cos(\sin^{-1}(x)) = \sqrt{1 - x^2}$. Observe that, in order to prove

$$\langle -\nabla f(\theta), \theta' - \theta \rangle = \|\nabla f(\theta)\| \|\theta' - \theta\| \cos(\phi) > 0, \tag{146}$$
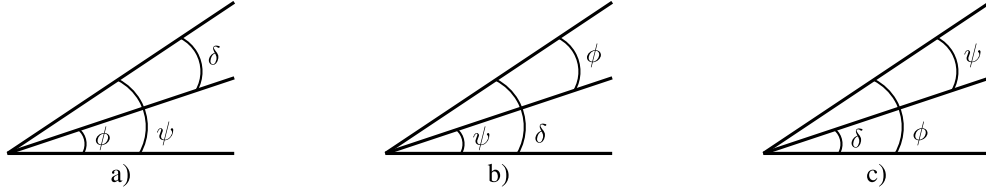
Fig. 8. Three different configurations of the angles δ, ϕ and ψ.

it is enough to show that $\phi < \frac{\pi}{2}$. Depending on the position of θ', the angles δ, ϕ and ψ can be arranged in three different configurations, as depicted in Fig. 8.

a) Since $\frac{\rho}{\|\theta_0 - \theta\|} > 0$, (142) implies that $\psi < \frac{\pi}{2}$. Therefore, $\phi < \frac{\pi}{2}$ as $\phi \leq \psi$.

b) Since $\frac{\rho}{\|\theta_0 - \theta\|} < 1$, (145) implies that $\delta < \frac{\pi}{2}$. Therefore, $\phi < \frac{\pi}{2}$ as $\phi \leq \delta$.

c) Since $\sin^{-1}(x) + \cos^{-1}(x) = \frac{\pi}{2}$, (142) and (145) imply that $\phi = \psi + \delta < \frac{\pi}{2}$.

Since in all cases $\phi < \frac{\pi}{2}$, the result follows. $\square$

2) Lipschitz Inequalities in α^{-1} and Main SLQC Result for the α-Risk:

Lemma 2: If $\alpha, \alpha' \in [1, \infty]$, then, for all $\theta \in \mathbb{B}_d(r)$,

$$|R_\alpha(\theta) - R_{\alpha'}(\theta)| \leq L_d(\theta) \left| \frac{\alpha - \alpha'}{\alpha \alpha'} \right|, \tag{147a}$$

$$\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\| \leq J_d(\theta) \left| \frac{\alpha - \alpha'}{\alpha \alpha'} \right|, \tag{147b}$$

where,

$$L_d(\theta) := \frac{\left(\log \left(1 + e^{\|\theta\|\sqrt{d}} \right) \right)^2}{2}, \tag{148a}$$

$$J_d(\theta) := \sqrt{d} \log \left(1 + e^{\|\theta\|\sqrt{d}} \right) \sigma(\|\theta\|\sqrt{d}). \tag{148b}$$

Proof: Here, we present proofs for both Lipschitz inequalities.

a) Proof of First Inequality: For ease of notation, we denote $\beta = 1/\alpha$. Thus, we have that for $\alpha \in [1, \infty]$, i.e., $\beta \in [0, 1]$,

$$R_\alpha(\theta) = \mathbb{E}[l^\alpha(Y, g_\theta(X))] \tag{149}$$

$$= \mathbb{E} \left[\frac{1}{1 - \beta} \left(1 - g_\theta(yx)^{1-\beta} \right) \right] \tag{150}$$

$$= R_\beta(\theta). \tag{151}$$

To show that R_α is Lipschitz in $\alpha^{-1} = \beta \in [0, 1]$, it suffices to show $\frac{d}{d\beta} R_\beta(\theta) \leq L$ for some $L > 0$. Observe that

$$\frac{d}{d\beta} R_\beta(\theta) = \mathbb{E} \left[\frac{d}{d\beta} \frac{1}{1 - \beta} \left(1 - g_\theta(yx)^{1-\beta} \right) \right], \tag{152}$$

where the equality follows since we assume well-behaved integrals. Consider without loss of generality the expression in the brackets; we denote this expression as

$$f(\beta, \theta, yx) = \frac{d}{d\beta} \frac{1}{1 - \beta} \left(1 - g_\theta(yx)^{1-\beta} \right). \tag{153}$$

It can be shown that

$$f(\beta, \theta, yx) = \frac{g_\theta(yx)^{1-\beta} \log \left(g_\theta(yx) \right)}{1 - \beta} + \frac{1 - g_\theta(yx)^{1-\beta}}{(1 - \beta)^2}, \tag{154}$$

and

$$f(1, \theta, yx) = \frac{\left(\log g_\theta(yx) \right)^2}{2}. \tag{155}$$

In addition, it can be shown that for any $y \in \{-1, +1\}$, $x \in [0, 1]^d$, and $\theta \in \mathbb{B}_d(r)$ that $f(\beta, \theta, yx)$ is monotonically increasing in $\beta \in [0, 1]$. Therefore, for any $\beta \in [0, 1]$, $y \in \{-1, +1\}$, $x \in [0, 1]^d$, and $\theta \in \mathbb{B}_d(r)$,

$$f(\beta, \theta, yx) \leq f(1, \theta, yx) \tag{156}$$

$$= \frac{\left(\log g_\theta(yx) \right)^2}{2} \tag{157}$$

$$\leq \frac{\left(\log \sigma(-\|\theta\|\sqrt{d}) \right)^2}{2}. \tag{158}$$

b) Proof of Second Inequality: For ease of notation, we let $\beta = 1/\alpha$. Since $\alpha \in [1, \infty]$, $\beta \in [0, 1]$. Thus, we have that for $\alpha \in [1, \infty]$, i.e., $\beta \in [0, 1]$,

$$\nabla R_\alpha(\theta) = \mathbb{E}[F_1(\alpha, \theta, X, Y)X] \tag{159}$$

$$= \mathbb{E}[-Y g_\theta(YX)^{1-\beta}(1 - g_\theta(YX))X], \tag{160}$$

and we let $\tilde{F}_1(\beta, \theta, X, Y) := -Y g_\theta(YX)^{1-\beta}(1 - g_\theta(YX))$. For any $\theta \in \mathbb{B}_d(r)$ we have

$$\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\|$$

$$= \|\mathbb{E}[(\tilde{F}_1(\beta, \theta, X, Y) - \tilde{F}_1(\beta', \theta, X, Y))X]\| \tag{161}$$

$$\leq \mathbb{E}[|(\tilde{F}_1(\beta, \theta, X, Y) - \tilde{F}_1(\beta', \theta, X, Y))|\|X\|] \tag{162}$$

$$\leq \sqrt{d} \mathbb{E}[|(\tilde{F}_1(\beta, \theta, X, Y) - \tilde{F}_1(\beta', \theta, X, Y))|], \tag{163}$$

where we used the fact that X has support $[0, 1]^d$ for the second inequality. Here, we obtain a Lipschitz inequality on $\tilde{F}_1$ by considering the variation of $\tilde{F}_1$ with respect to β for any $\theta \in \mathbb{B}_d(r)$, $x \in [0, 1]^d$, and $y \in \{-1, +1\}$. Taking the derivative of $\tilde{F}_1(\beta, \theta, x, y)$ with respect to β we obtain

$$\frac{d}{d\beta} \tilde{F}_1(\beta, \theta, x, y) = \frac{d}{d\beta} -y g_\theta(yx)^{1-\beta}(1 - g_\theta(yx)) \tag{164}$$

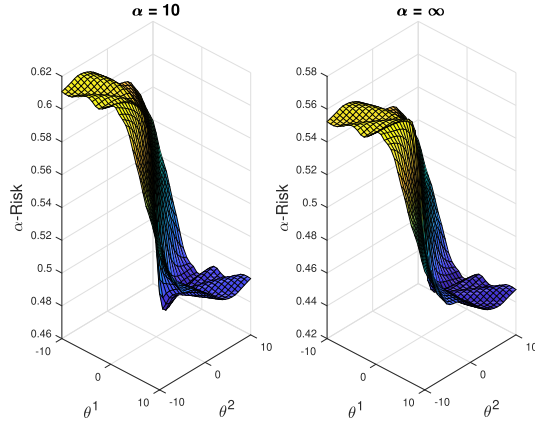$$= y(1 - g_\theta(yx)) g_\theta(yx)^{1-\beta} \log g_\theta(yx), \tag{165}$$

Fig. 9. Another illustration highlighting the saturation of α-loss (R_α for $\alpha = 10, \infty$) in the logistic model for a 2D-GMM with $\mathbb{P}[Y = 1] = .5$, $\mu_{X|Y=-1} = [.5, .5]$, $\mu_{X|Y=1} = [1, 1]$, and shared covariance matrix $\Sigma = [1, .5; .5, 3]$.

where we used the fact that $\dfrac{d}{dx} a^{1-x} = -a^{1-x} \log a$. Continuing, we have

$$y(1 - g_\theta(yx)) g_\theta(yx)^{1-\beta} \log g_\theta(yx)$$
$$\leq \log\left(1 + e^{\|\theta\|\sqrt{d}}\right) \sigma(\|\theta\|\sqrt{d}) \sigma(\|\theta\|\sqrt{d})^{1-\beta} \quad (166)$$
$$= \log\left(1 + e^{\|\theta\|\sqrt{d}}\right) \sigma(\|\theta\|\sqrt{d})^{2-\beta} \quad (167)$$
$$\leq \log\left(1 + e^{\|\theta\|\sqrt{d}}\right) \sigma(\|\theta\|\sqrt{d}). \quad (168)$$

Thus, we have that, for any $\theta \in \mathbb{B}_d(r)$,

$$\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\| \leq J_d(\theta)|\beta - \beta'|, \quad (169)$$

where $\beta, \beta' \in [0, 1]$ ($\alpha, \alpha' \in [1, \infty]$). Therefore, we have that, for any $\theta \in \mathbb{B}_d(r)$,

$$\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\| \leq J_d(\theta)\left|\frac{1}{\alpha} - \frac{1}{\alpha'}\right|, \quad (170)$$

where $\alpha, \alpha' \in [1, \infty]$. □

Theorem 3: Let $\alpha_0 \in [1, \infty]$, $\epsilon_0, \kappa_0 > 0$, and $\theta_0, \theta \in \mathbb{B}_d(r)$. If R_{α_0} is $(\epsilon_0, \kappa_0, \theta_0)$-SLQC at θ and

$$0 \leq \alpha - \alpha_0 < \frac{\alpha_0^2 \|\nabla R_{\alpha_0}(\theta)\|}{2J_d(\theta)\left(1 + r\frac{\kappa_0}{\epsilon_0}\right)}, \quad (171)$$

then R_α is $(\epsilon, \kappa, \theta_0)$-SLQC at θ with

$$\epsilon = \epsilon_0 + 2L_d(\theta)\left(\frac{\alpha - \alpha_0}{\alpha \alpha_0}\right), \quad (172)$$

$$\frac{\epsilon}{\kappa} = \frac{\epsilon_0}{\kappa_0}\left(1 - \frac{\left(1 + 2r\frac{\kappa_0}{\epsilon_0}\right)J_d(\theta)(\alpha - \alpha_0)}{\alpha \alpha_0 \|\nabla R_{\alpha_0}(\theta)\| - J_d(\theta)(\alpha - \alpha_0)}\right). \quad (173)$$

Proof: For ease of notation let $\rho_0 = \dfrac{\epsilon_0}{\kappa_0}$ and $\rho = \dfrac{\epsilon}{\kappa}$, and consider the following two cases.

(a) $\alpha = .9$ loss landscape   (b) $\alpha = 1$ loss landscape

(c) $\alpha = 2$ loss landscape   (d) $\alpha = 10$ loss landscape
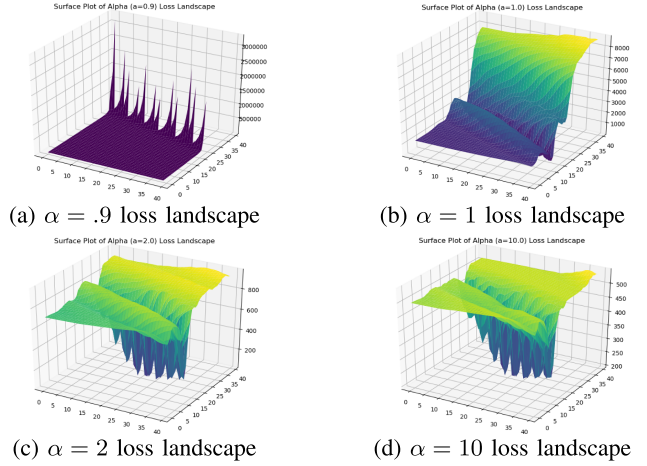
Fig. 10. Loss landscape visualizations obtained using [79] for $\alpha \in \{.9, 1, 2, 10\}$ training a ResNet-18 on the MNIST dataset. The visualization technique finds two "principal directions" of the model to allow for a 3D plot. We note that similar themes as theoretically articulated in Section IV for the simpler logistic model are also evident here; i.e., exploding gradients for α too small, a loss of convexity (and increasing "flatness") as α increases greater than 1, and also a saturation effect as exhibited by the visual similarity between the $\alpha = 2$ and $\alpha = 10$ loss landscapes. This hints at the generality of the theory presented in Section IV.

Case 1: Assume that $R_{\alpha_0}(\theta) - R_{\alpha_0}(\theta_0) \leq \epsilon_0$. Then,

$$R_\alpha(\theta) - R_\alpha(\theta_0)$$
$$= R_\alpha(\theta) - R_{\alpha_0}(\theta) + R_{\alpha_0}(\theta)$$
$$\quad - R_{\alpha_0}(\theta_0) + R_{\alpha_0}(\theta_0) - R_\alpha(\theta_0) \quad (174)$$
$$\leq L_d(\theta)\left(\frac{\alpha - \alpha_0}{\alpha \alpha_0}\right) + \epsilon_0 + L_d(\theta)\left(\frac{\alpha - \alpha_0}{\alpha \alpha_0}\right). \quad (175)$$

Since $\epsilon_0 + 2L_d(\theta)\left(\frac{\alpha - \alpha_0}{\alpha \alpha_0}\right) = \epsilon$, we have $R_\alpha(\theta) - R_\alpha(\theta_0) \leq \epsilon$.

Case 2: Assume that $R_{\alpha_0}(\theta) - R_{\alpha_0}(\theta_0) > \epsilon_0$. Since R_{α_0} is $(\epsilon_0, \kappa_0, \theta_0)$-SLQC at θ by assumption, we have that $\|\nabla R_{\alpha_0}(\theta)\| > 0$ and $\langle -\nabla R_{\alpha_0}(\theta), \theta' - \theta \rangle \geq 0$ for every $\theta' \in \mathbb{B}(\theta_0, \rho_0)$.

Let $\rho = \epsilon/\kappa$ be given as in (42). If $\|\theta - \theta_0\| > \rho$, $\|\nabla R_\alpha(\theta)\| > 0$ and

$$\langle -\nabla R_\alpha(\theta), \theta_0 - \theta \rangle \geq \rho \|\nabla R_\alpha(\theta)\|, \quad (176)$$

then Lemma 1 would imply that R_α is $(\epsilon, \kappa, \theta_0)$-SLQC at θ. In order to show these three expressions, we make ample use of the following three inequalities: The first is the reverse triangle inequality associated with ∇R_α and ∇R_{α_0}, i.e.,

$$\|\nabla R_{\alpha_0}(\theta) - \nabla R_\alpha(\theta)\| \geq |\|\nabla R_\alpha(\theta)\| - \|\nabla R_{\alpha_0}(\theta)\||. \quad (177)$$

The second is that $\nabla R_\alpha(\theta)$ is $J_d(\theta)$-Lipschitz in α^{-1}, i.e.,

$$\left|\frac{1}{\alpha_0} - \frac{1}{\alpha}\right| J_d(\theta) \geq \|\nabla R_{\alpha_0}(\theta) - \nabla R_\alpha(\theta)\|. \quad (178)$$

The third follows from a manipulation of (41), i.e.,

$$\|\nabla R_{\alpha_0}(\theta)\| > 2J_d(\theta)\left(1 + r\rho_0^{-1}\right)\left(\alpha_0^{-1} - \alpha^{-1}\right) \quad (179)$$
$$> J_d(\theta)(\alpha_0^{-1} - \alpha^{-1}), \quad (180)$$

using the fact that $\alpha_0^2 \le \alpha\alpha_0$ and since $r\rho_0^{-1} \ge 1$. With these inequalities in hand, we are now in a position to complete the three steps required to show that R_α is $(\epsilon, \kappa, \theta_0)$-SLQC at θ.

First, we show that $\|\theta - \theta_0\| > \rho$. Since R_{α_0} is $(\epsilon_0, \kappa_0, \theta_0)$-SLQC at θ and $R_{\alpha_0}(\theta) - R_{\alpha_0}(\theta_0) > \epsilon_0$ by assumption, we have by the contrapositive of Proposition 8 that $\theta \notin \mathbb{B}_d(\theta_0, \rho_0)$. Thus, we have that $\|\theta - \theta_0\| > \rho_0$. Next, note that ρ is related to ρ_0 by (42). If we can show that $\rho_0 > \rho$, then we have the desired conclusion. Rearranging the left-hand-side of (179), we have that

$$\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} > 2J_d(\theta)(1 + r\rho_0^{-1}), \quad (181)$$

which can be rewritten to obtain

$$\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_d(\theta) > J_d(\theta)(1 + 2r\rho_0^{-1}). \quad (182)$$

Since by the right-hand-side of (179) we have that

$$\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_d(\theta) > 0, \quad (183)$$

it follows by (182) that

$$1 > \frac{J_d(\theta)(1 + 2r\rho_0^{-1})}{\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_d(\theta)}. \quad (184)$$

Thus examining (42) in light of (184), we have that $\rho_0 > \rho$, which implies that $\|\theta - \theta_0\| > \rho$, as desired.

Second, we show that $\|\nabla R_\alpha(\theta)\| > 0$. Applying (177) to (178) we obtain

$$\|\nabla R_\alpha(\theta)\| \ge \|\nabla R_{\alpha_0}(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1}) > 0, \quad (185)$$

where the right-hand-side inequality again follows by (179). Thus, we have that $\|\nabla R_\alpha(\theta)\| > 0$, as desired.

Finally, we show the expression in (176), i.e., $\langle -\nabla R_\alpha(\theta), \theta_0 - \theta \rangle \ge \rho\|\nabla R_\alpha(\theta)\|$. By the Cauchy-Schwarz inequality, we have

$$\langle -\nabla R_\alpha(\theta), \theta_0 - \theta \rangle$$
$$\ge \langle -\nabla R_{\alpha_0}(\theta), \theta_0 - \theta \rangle - \|\nabla R_\alpha(\theta) - \nabla R_{\alpha_0}(\theta)\|\|\theta_0 - \theta\| \quad (186)$$
$$\ge \rho_0\|\nabla R_{\alpha_0}(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})2r, \quad (187)$$

where in (186) we apply Lemma 1 for the first term; for the second term we use the fact that ∇R_α is $J_d(\theta)$-Lipschitz in α^{-1} as given by (178) and the fact that $\theta_0 - \theta \in \mathbb{B}_d(2r)$. Continuing from (187), we have that

$$\langle -\nabla R_\alpha(\theta), \theta_0 - \theta \rangle$$
$$\ge \rho_0\|\nabla R_\alpha(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})2r$$
$$\quad - \rho_0\|\nabla R_{\alpha_0}(\theta) - \nabla R_\alpha(\theta)\| \quad (188)$$
$$\ge \rho_0\|\nabla R_\alpha(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})(\rho_0 + 2r), \quad (189)$$

where we first apply the reverse triangle inequality in (177) and then we use the fact that $\nabla R_\alpha(\theta)$ is $J_d(\theta)$-Lipschitz in α^{-1}, i.e., the expression in (178). Rearranging the expression in (189), we obtain

$$\rho_0\|\nabla R_\alpha(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})(\rho_0 + 2r)$$
$$= \|\nabla R_\alpha(\theta)\|\left(\rho_0 - \frac{J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})(\rho_0 + 2r)}{\|\nabla R_\alpha(\theta)\|}\right) \quad (190)$$
$$\ge \|\nabla R_\alpha(\theta)\|\left(\rho_0 - \frac{(\rho_0 + 2r)J_d(\theta)}{\frac{\|\nabla R_{\alpha_0}(\theta)\|}{(\frac{1}{\alpha_0} - \frac{1}{\alpha})} - J_d(\theta)}\right), \quad (191)$$

where we used the inequality in (185). Thus, we finally obtain

$$\langle -\nabla R_\alpha(\theta), \theta_0 - \theta \rangle \ge \rho\|\nabla R_\alpha(\theta)\|, \quad (192)$$

where $\rho > 0$ is given by

$$\rho = \rho_0\left(1 - \frac{(1 + 2r\rho_0^{-1})J_d(\theta)}{\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_d(\theta)}\right), \quad (193)$$

as desired. Therefore by collecting all three parts, we have by Lemma 1 that R_α is $(\epsilon, \kappa, \theta_0)$-SLQC at θ. $\square$

3) Bootstrapping SLQC: Recall that the floor function, denoted $\lfloor \cdot \rfloor : \mathbb{R}^+ \to \mathbb{N}$, can alternatively be written as $\lfloor x \rfloor = x - q$, for some $q \in [0, 1)$.

Lemma 5: Fix $\theta \in \mathbb{B}_d(r)$. Suppose that $\rho_0 > 0$ and there exists $g_\theta > 0$ such that $\|\nabla R_{\alpha'}(\theta)\| > g_\theta$ for all $\alpha' \in [\alpha_0, \infty]$. Given $N \in \mathbb{N}$, for each $n \in [N]$ we define

$$\alpha_n = \alpha_{n-1} + \frac{1}{N}, \quad (194a)$$
$$\epsilon_n = \epsilon_{n-1} + 2L_d(\theta)\frac{1}{\alpha_n\alpha_{n-1}}\frac{1}{N}, \quad (194b)$$
$$\rho_n = \rho_{n-1} - \frac{(\rho_{n-1} + 2r)J_d(\theta)}{\alpha_n\alpha_{n-1}G_{n-1} - J_d(\theta)/N}\frac{1}{N}, \quad (194c)$$

where $G_{n-1} := \|\nabla R_{\alpha_{n-1}}(\theta)\|$. If $N > J_d(\theta)\left(\alpha_0^2 g_\theta\right)^{-1}$, then we have that $\{\alpha_n\}_{n=0}^N$, $\{\epsilon_n\}_{n=0}^N$, and $\{\rho_n\}_{n=0}^N$ are well-defined. Furthermore, we have that $\rho_n > 0$ for all $n \le \lfloor \alpha_0^2 g_\theta(1 + 2r\rho_0^{-1})^{-1}J_d(\theta)^{-1}N \rfloor$.

Proof: For ease of notation, let $J := J_d(\theta)$, $L := L_d(\theta)$, and $g := g_\theta$. Observe that $\{\alpha_n\}_{n=0}^N$ is well defined and so is $\{\epsilon_n\}_{n=0}^N$. It can be verified that if $N > J\left(\alpha_0^2 g\right)^{-1}$, then $\alpha_{n-1}\alpha_n G_{n-1} - J/N > 0$ and thus $\{\rho_n\}_{n=0}^N$ is well defined. Now we show by induction that $\rho_n > 0$ for

$$n < \left\lfloor \frac{\rho_0}{\rho_0 + 2r}\frac{\alpha_0^2 g}{J}N \right\rfloor. \quad (195)$$

By assumption, $\rho_0 > 0$. For the inductive hypothesis, assume that $\rho_0, \ldots, \rho_{n-1}$ are non-negative. Observe that, by definition, we have

$$\rho_k - \rho_{k+1} = \frac{(\rho_k + 2r)J}{\alpha_k\alpha_{k+1}G_k - J/N}\frac{1}{N}. \quad (196)$$

The previous equation and a telescoping sum lead to

$$\rho_0 - \rho_n = \sum_{k=0}^{n-1} \frac{(\rho_k + 2r)J}{\alpha_k\alpha_{k+1}G_k - J/N}\frac{1}{N}. \quad (197)$$

Since $\rho_k > 0$ for all $k \in [n-1]$, we have the following ordering $\rho_0 > \rho_1 > \cdots > \rho_n$ and, as a result,

$$\rho_0 - \rho_n < \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N}\frac{n}{N}. \quad (198)$$

It can be shown that our choice of n in (195) implies that

$$\rho_n > \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N} \frac{n}{N} > 0, \tag{199}$$

which implies that $\rho_n > 0$ as desired. $\qquad\square$

Theorem 4: Let $\alpha_0 \in [1, \infty)$, $\epsilon_0, \kappa_0 > 0$, and $\theta_0, \theta \in \mathbb{B}_d(r)$. Suppose that R_{α_0} is $(\epsilon_0, \kappa_0, \theta_0)$-SLQC at $\theta \in \mathbb{B}_d(r)$ and that there exists $g_\theta > 0$ such that $\|\nabla R_{\alpha'}(\theta)\| > g_\theta$ for every $\alpha' \in [\alpha_0, \infty]$. Then, for every $\lambda \in (0, 1)$, R_{α_λ} is $(\epsilon_\lambda, \kappa_\lambda, \theta_0)$-SLQC at θ where

$$\alpha_\lambda := \alpha_0 + \lambda \frac{\alpha_0^2 g_\theta}{J_d(\theta)\left(1 + 2r\frac{\kappa_0}{\epsilon_0}\right)}, \tag{200}$$

$$\epsilon_\lambda := \epsilon_0 + 2\lambda L_d(\theta)\left(\frac{\alpha_\lambda - \alpha_0}{\alpha_\lambda \alpha_0}\right)\frac{\alpha_0^2 g_\theta}{J_d(\theta)\left(1 + r\frac{\kappa_0}{\epsilon_0}\right)}, \tag{201}$$

$$\frac{\epsilon_\lambda}{\kappa_\lambda} > \frac{\epsilon_0}{\kappa_0}(1 - \lambda). \tag{202}$$

Proof: For ease of notation, let $J := J_d(\theta)$, $L := L_d(\theta)$, and $g := g_\theta$. Let $\lambda \in (0, 1)$ be given. For each

$$N > \frac{1 + 2r\rho_0^{-1}}{1 - \lambda}\frac{2J}{\alpha_0^2 g}, \tag{203}$$

we define

$$N_\lambda = \left\lfloor \lambda \frac{\rho_0}{\rho_0 + 2r}\frac{\alpha_0^2 g}{J}N \right\rfloor. \tag{204}$$

The bootstrapping proof strategy is as follows: 1) For fixed $N \in \mathbb{N}$ large enough (as given above), we show by induction that R_{α_n} is $(\epsilon_n, \kappa_n, \theta_0)$-SLQC at θ with $\rho_n = \epsilon_n/\kappa_n$ for $n \le N_\lambda$ using Lemma 5 and Theorem 3; 2) We take the limit as N approaches infinity in order to derive the largest range on α and the strongest SLQC parameters.

First, we show by induction that R_{α_n} is $(\epsilon_n, \kappa_n, \theta_0)$-SLQC at θ with $\rho_n = \epsilon_n/\kappa_n$ for $n \le N_\lambda$. By assumption, R_{α_0} is $(\epsilon_0, \kappa_0, \theta_0)$-SLQC at θ. For the inductive hypothesis, assume that R_{α_k} is $(\alpha_k, \epsilon_k, \kappa_k)$-SLQC at θ for all $k \in [n-1]$. In order to apply Lemma 5 to show that

$$\rho_0 > \rho_1 > \ldots > \rho_n > \cdots > \rho_{N_\lambda} > C_\lambda > 0, \tag{205}$$

for all $n \le N_\lambda$ and for some $C_\lambda > 0$, we first show that the assumptions of Lemma 5 are satisfied. Observe that, by our assumption on $N \in \mathbb{N}$, we have that

$$N > \frac{1 + 2r\rho_0^{-1}}{1 - \lambda}\frac{2J}{\alpha_0^2 g} > \frac{1 + r\rho_0^{-1}}{1 - \lambda}\frac{J}{\alpha_0^2 g} > \frac{J}{\alpha_0^2 g}, \tag{206}$$

which is the first requirement of Lemma 5. Next, we want to show that

$$n \le N_\lambda < \left\lfloor \frac{\rho_0}{\rho_0 + 2r}\frac{\alpha_0^2 g}{J}N \right\rfloor, \tag{207}$$

which is the last requirement of Lemma 5. This is achieved by observing that

$$N_\lambda = \left\lfloor \lambda\frac{\rho_0}{\rho_0 + 2r}\frac{\alpha_0^2 g}{J}N \right\rfloor = \lambda\frac{\rho_0}{\rho_0 + 2r}\frac{\alpha_0^2 g}{J}N - q, \tag{208}$$

for some $q \in [0, 1)$ and that

$$\left\lfloor \frac{\rho_0}{\rho_0 + 2r}\frac{\alpha_0^2 g}{J}N \right\rfloor = \frac{\rho_0}{\rho_0 + 2r}\frac{\alpha_0^2 g}{J}N - w, \tag{209}$$

also for some $w \in [0, 1)$. Note that (207) is equivalent to

$$(q - w)\frac{1 + r\rho_0^{-1}}{1 - \lambda}\frac{J}{\alpha_0^2 g} < N, \tag{210}$$

which holds by the fact that $N > \frac{1 + r\rho_0^{-1}}{1 - \lambda}\frac{J}{\alpha_0^2 g}$ in (206) and $q - w \le 1$. Thus by Lemma 5, we have that

$$\rho_n > \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N}\frac{n}{N} > 0, \tag{211}$$

for all $n \le N_\lambda$. In particular for $n = N_\lambda$, we have that

$$\rho_{N_\lambda} > \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N}\frac{N_\lambda}{N} \tag{212}$$

$$> \rho_0\left(1 - \lambda - \frac{\lambda J}{\alpha_0^2 g - J/N}\frac{1}{N}\right) \tag{213}$$

$$> \frac{\rho_0(1 - \lambda)}{2}, \tag{214}$$

where the second inequality follows by plugging in N_λ and adding and subtracting $\lambda J/N$ in the fraction and the last inequality follows from $N > \frac{1 + 2r\rho_0^{-1}}{1 - \lambda}\frac{2J}{\alpha_0^2 g} > \frac{1 + \lambda}{1 - \lambda}\frac{J}{\alpha_0^2 g}$ since $2r\rho_0^{-1} \ge \lambda$ for all $\lambda \in (0, 1)$. Therefore, we have that $C_\lambda = \frac{\rho_0(1 - \lambda)}{2}$; in other words,

$$\rho_0 > \rho_1 > \ldots > \rho_{n-1} > \rho_n > \cdots > \rho_{N_\lambda} > \frac{\rho_0(1 - \lambda)}{2} > 0. \tag{215}$$

Also, observe that

$$\alpha_n - \alpha_{n-1} = \frac{1}{N} < \frac{\alpha_0^2 g}{2J(1 + 2r\rho_0^{-1}(1 - \lambda)^{-1})}, \tag{216}$$

where the inequality follows from the fact that

$$N > \frac{1 + 2r\rho_0^{-1}}{1 - \lambda}\frac{2J}{\alpha_0^2 g} > \left(1 + \frac{2r\rho_0^{-1}}{1 - \lambda}\right)\frac{2J}{\alpha_0^2 g}. \tag{217}$$

In particular, (215) and (216) leads to

$$\alpha_n - \alpha_{n-1} < \frac{\alpha_0^2 g}{2J(1 + 2r\rho_0^{-1}(1 - \lambda)^{-1})} \tag{218}$$

$$< \frac{\alpha_{n-1}^2 G_{n-1}}{2J(1 + r\rho_{n-1}^{-1})}, \tag{219}$$

where we use the fact that $\alpha_n \ge \alpha_0$ and $G_{n-1} \ge g$. As a result, we can apply Theorem 3 to conclude that R_{α_n} is $(\epsilon_n, \rho_n, \theta_0)$-SLQC at θ with α_n, ϵ_n and ρ_n given as in (194a). In particular by unfolding the recursion, we have that $R_{\alpha_{N_\lambda}}$ is $(\epsilon_{N_\lambda}, \rho_{N_\lambda}, \theta_0)$-SLQC at θ with

$$\alpha_{N_\lambda} = \alpha_0 + \lambda(1 + 2r\rho_0^{-1})^{-1}\frac{\alpha_0^2 g}{J} - \frac{q}{N}, \tag{220}$$

$$\epsilon_{N_\lambda} = \epsilon_0 + 2L\sum_{n=0}^{N_\lambda - 1}\frac{1}{\alpha_n(\alpha_n + 1/N)}\frac{1}{N}, \tag{221}$$

$$\rho_{N_\lambda} = \rho_0\prod_{n=0}^{N_\lambda - 1}\left(1 - \frac{(1 + 2r\rho_n^{-1})J/N}{\alpha_{n+1}\alpha_n\|\nabla R_{\alpha_n}(\theta)\| - J/N}\right), \tag{222}$$

for some $q \in [0, 1)$.

Finally, we take the limit as N approaches infinity in order to derive the largest range on α and the strongest SLQC parameters. Recall that $N_\lambda = \left\lfloor \lambda \frac{\rho_0}{\rho_0 + 2r} \frac{\alpha_0^2 g}{J} N \right\rfloor = \lambda \frac{\rho_0}{\rho_0 + 2r} \frac{\alpha_0^2 g}{J} N - q$, for some $q \in [0, 1)$. Thus, we have the following relationship

$$\frac{1}{N} = \frac{\lambda \rho_0 \alpha_0^2 g}{(N_\lambda + q)(\rho_0 + r)J}. \tag{223}$$

Observe that taking the limit as N approaches infinity is equivalent to taking the limit as N_λ approaches infinity.

Examining (220) as N_λ approaches infinity, we have that

$$\alpha_\lambda := \lim_{N_\lambda \to \infty} \alpha_{N_\lambda} = \alpha_0 + \lambda(1 + 2r\rho_0^{-1})^{-1} \frac{\alpha_0^2 g}{J}. \tag{224}$$

Next considering (221), we rewrite to obtain

$$\epsilon_{N_\lambda} = \epsilon_0 + 2L \sum_{n=0}^{N_\lambda - 1} \frac{1}{\alpha_n(\alpha_n + 1/N)} \frac{1}{N} \tag{225}$$

$$= \epsilon_0 + \frac{2L}{N} \sum_{n=0}^{N_\lambda - 1} \left(\frac{1}{\alpha_n^2} + \frac{1}{N} \frac{1}{\alpha_n^3 - \alpha_n^2/N} \right), \tag{226}$$

where we used a partial fraction decomposition. Let μ_{N_λ} be the discrete measure given by

$$\mu_{N_\lambda} = \frac{1}{N_\lambda} \sum_{n=0}^{N_\lambda - 1} \delta_{\alpha_n}, \tag{227}$$

where δ_{α_n} is the point mass at α_n. In particular for large N, we can write (226) as

$$\epsilon_{N_\lambda} = \epsilon_0 + \frac{2L\lambda \alpha_0^2 g}{(1 + r\rho_0^{-1})J} \int \frac{1}{x^2} d\mu_{N_\lambda}(x) + O\left(\frac{1}{N_\lambda}\right). \tag{228}$$

Let μ_λ denote the uniform measure over $(\alpha_0, \alpha_\lambda]$, i.e., the Lebesgue measure on the interval $(\alpha_0, \alpha_\lambda]$. Note that μ_{N_λ} converges in distribution to μ_λ as N_λ goes to infinity. By taking limits, (228) becomes

$$\epsilon_\lambda = \lim_{N_\lambda \to \infty} \epsilon_{N_\lambda} \tag{229}$$

$$= \epsilon_0 + \frac{2L\lambda \alpha_0^2 g}{(1 + r\rho_0^{-1})J} \int_{\alpha_0}^{\alpha_\lambda} \frac{1}{x^2} dx \tag{230}$$

$$= \epsilon_0 + \frac{2L\lambda \alpha_0 g}{(1 + r\rho_0^{-1})J} \left(1 - \frac{\alpha_0}{\alpha_\lambda}\right). \tag{231}$$

Finally, we consider (222). Observe that from (199) we have

$$\rho_{N_\lambda} > \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N} \frac{N_\lambda}{N} \tag{232}$$

$$= \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N} \frac{\lambda \frac{N\alpha_0^2 g\rho_0}{J(\rho_0 + 2r)} - q}{N} \tag{233}$$

$$= \rho_0 - \left[\frac{N\lambda \rho_0 \alpha_0^2 g}{N\alpha_0^2 g - J} - \frac{q}{N}\left(\frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N}\right) \right], \tag{234}$$

for $q \in [0, 1)$, where we plugged in the definition of N_λ and simplified. Thus, taking the limit as N_λ approaches infinity

we have that

$$\rho_\lambda = \lim_{N_\lambda \to \infty} \rho_{N_\lambda} \tag{235}$$

$$> \lim_{N_\lambda \to \infty} \left(\rho_0 - \left[\frac{N\lambda \rho_0 \alpha_0^2 g}{N\alpha_0^2 g - J} - \frac{q}{N}\left(\frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N}\right) \right]\right) \tag{236}$$

$$= \rho_0(1 - \lambda). \tag{237}$$

Thus, we conclude that R_{α_λ} is $(\epsilon_\lambda, \kappa_\lambda, \theta_0)$-SLQC at θ with

$$\alpha_\lambda := \alpha_0 + \lambda(1 + 2r\rho_0^{-1})^{-1} \frac{\alpha_0^2 g}{J}, \tag{238}$$

$$\epsilon_\lambda := \epsilon_0 + \frac{2L\lambda \alpha_0 g}{(1 + r\rho_0^{-1})J} \left(1 - \frac{\alpha_0}{\alpha_\lambda}\right) \tag{239}$$

$$\rho_\lambda > \rho_0(1 - \lambda). \tag{240}$$

A change of variables leads to the desired result. $\square$

C. Rademacher Complexity Generalization and Asymptotic Optimality

Lemma 6: If $\alpha \in (0, \infty]$, then $\tilde{l}^\alpha(z)$ is $C_{r_0}(\alpha)$-Lipschitz in $z \in [-r_0, r_0]$ for every $r_0 > 0$, where for $\alpha \in (0, 1]$,

$$C_{r_0}(\alpha) := \sigma(r_0)\sigma(-r_0)^{1 - 1/\alpha}; \tag{241}$$

and, for $\alpha \in (1, \infty]$,

$$C_{r_0}(\alpha) := \begin{cases} \left(\frac{\alpha - 1}{2\alpha - 1}\right)^{1 - \frac{1}{\alpha}} \left(\frac{\alpha}{2\alpha - 1}\right) & e^{r_0} \geq \frac{\alpha - 1}{\alpha}, \\ \sigma(r_0)\sigma(-r_0)^{1 - \frac{1}{\alpha}} & e^{r_0} < \frac{\alpha - 1}{\alpha}. \end{cases} \tag{242}$$

Proof: The proof is analogous to the proof in Proposition 5. In order to show that $\tilde{l}^\alpha(z)$ is $C_{r_0}(\alpha)$-Lipschitz, we take the derivative of $\tilde{l}^\alpha(z)$ and seek to maximize it over $z \in [-r_0, r_0]$. Specifically, we have that for $\alpha \in (0, \infty]$,

$$\frac{d}{dz} \tilde{l}^\alpha(z) = \frac{d}{dz} \frac{\alpha}{\alpha - 1}\left(1 - \sigma(z)^{1 - 1/\alpha}\right) \tag{243}$$

$$= \sigma(z)^{2 - 1/\alpha} - \sigma(z)^{1 - 1/\alpha} \tag{244}$$

$$= (\sigma(z) - 1)\sigma(z)^{1 - 1/\alpha} \tag{245}$$

$$\leq |(\sigma(z) - 1)\sigma(z)^{1 - 1/\alpha}| \tag{246}$$

$$= \sigma(-z)\sigma(z)^{1 - 1/\alpha}, \tag{247}$$

where we used the fact that $\sigma(z) = 1 - \sigma(-z)$. If $\alpha \leq 1$, it can be shown that

$$\max_{z \in [-r_0, r_0]} \sigma(-z)\sigma(z)^{1 - 1/\alpha} = \sigma(r_0)\sigma(-r_0)^{1 - 1/\alpha}. \tag{248}$$

Similarly if $\alpha > 1$ and if $r_0 \geq \log(1 - 1/\alpha)$, then it can be shown that

$$\max_{z \in [-r_0, r_0]} \sigma(-z)\sigma(z)^{1 - 1/\alpha} = \left(\frac{\alpha - 1}{2\alpha - 1}\right)^{1 - 1/\alpha} \left(\frac{\alpha}{2\alpha - 1}\right), \tag{249}$$

where $z^* = \log(1 - 1/\alpha)$. Otherwise for $\alpha > 1$, if we have $r_0 < \log(1 - 1/\alpha)$, we obtain using local monotonicity,

$$\max_{z \in [-r_0, r_0]} \sigma(-z)\sigma(z)^{1 - 1/\alpha} = \sigma(r_0)\sigma(-r_0)^{1 - 1/\alpha}, \tag{250}$$

analogous to the case where $\alpha < 1$. Thus, combining the two regimes of α, we have the result. $\qquad\square$

Theorem 5: If $\alpha \in (0,\infty]$, then, with probability at least $1 - \delta$, for all $\theta \in \mathbb{B}_d(r)$,

$$\left| R_\alpha(\theta) - \hat{R}_\alpha(\theta) \right| \leq C_{r\sqrt{d}}(\alpha) \frac{r\sqrt{d}}{\sqrt{n}} + D_{r\sqrt{d}}(\alpha) \sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{n}}, \tag{251}$$

where $C_{r\sqrt{d}}(\alpha)$ is given in (52) and (53) and where $D_{r\sqrt{d}}(\alpha)$ is given by $D_{r\sqrt{d}}(\alpha) := 4\sqrt{2} \frac{\alpha}{\alpha-1}\left(1 - \sigma(-r\sqrt{d})^{1-1/\alpha}\right)$.

Proof: By Proposition 2, which gives a relation between α-loss and its margin-based form, we have

$$\mathcal{R}(l^\alpha \circ \mathcal{G} \circ S_n) = \mathbb{E}\left(\sup_{g_\theta \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i l^\alpha(y_i, g_\theta(x_i)) \right) \tag{252}$$

$$= \mathbb{E}\left(\sup_{\theta \in \mathbb{B}_d(r)} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{l}^\alpha(y_i\langle\theta, x_i\rangle) \right). \tag{253}$$

The right-hand-side of (252) can be rewritten as

$$\mathbb{E}\left(\sup_{\theta \in \mathbb{B}_d(r)} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{l}^\alpha(y_i\langle\theta, x_i\rangle) \right)$$
$$= \mathcal{R}(\{\tilde{l}^\alpha(y_1\langle\theta, x_1\rangle), \ldots, \tilde{l}^\alpha(y_n\langle\theta, x_n\rangle) : \theta \in \mathbb{B}_d(r)\}). \tag{254}$$

Observe that, for each $i \in [n]$, $y_i\langle\theta, x_i\rangle \leq r\sqrt{d}$ by the Cauchy-Schwarz inequality since $\theta \in \mathbb{B}_d(r)$ and for each $i \in [n]$, $x_i \in [0,1]^d$. Further, by Lemma 6, we know that $\tilde{l}^\alpha(z)$ is $C_{r_0}(\alpha)$-Lipschitz in $z \in [-r_0, r_0]$. Thus setting $r_0 = r\sqrt{d}$, we may apply Lemma 3 (Contraction Lemma) to obtain

$$\mathbb{E}\left(\sup_{\theta \in \mathbb{B}_d(r)} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \tilde{l}^\alpha(y_i\langle\theta, x_i\rangle) \right)$$
$$= \mathcal{R}\left(\{\tilde{l}^\alpha(y_1\langle\theta, x_1\rangle), \ldots, \tilde{l}^\alpha(y_n\langle\theta, x_n\rangle) : \theta \in \mathbb{B}_d(r)\} \right) \tag{255}$$

$$\leq C_{r\sqrt{d}}(\alpha)\, \mathcal{R}\left(\{(y_1\langle\theta, x_1\rangle, \ldots, y_n\langle\theta, x_n\rangle) : \theta \in \mathbb{B}_d(r)\} \right). \tag{256}$$

We absorb y_i into its corresponding x_i and apply Lemma 4 to obtain

$$C_{r\sqrt{d}}(\alpha)\, \mathcal{R}(\{(y_1\langle\theta, x_1\rangle, \ldots, y_n\langle\theta, x_n\rangle) : \theta \in \mathbb{B}_d(r)\})$$
$$\leq C_{r\sqrt{d}}(\alpha) \frac{r\sqrt{d}}{\sqrt{n}}, \tag{257}$$

which follows since we assume that $x_i \in [0,1]^d$ for each $i \in [n]$. In order to apply Proposition 6, it can readily be shown that for $\alpha \in (0,\infty]$

$$\max_{z \in [-r\sqrt{d}, r\sqrt{d}]} \tilde{l}^\alpha(z) \leq D_{r\sqrt{d}}(\alpha), \tag{258}$$

where $D_{r\sqrt{d}}(\alpha) = \frac{\alpha}{\alpha-1}\left(1 - \sigma(-r\sqrt{d})^{1-1/\alpha}\right)$. Thus, we apply Proposition 6 to achieve the desired result. $\qquad\square$

The following result attempts to quantify the uniform discrepancy between the empirical α-risk and the probability of error (true ∞-risk); the technique is a combination of Theorem 5 and Lemma 2. The result is most useful in the regime where $r\sqrt{d} \leq \alpha/\sqrt{n}$; this prohibits the second term in the right-hand-side of (259) from dominating the first, which is the most meaningful form of the bound.

Corollary 3: If $\alpha \in [1,\infty]$, then, with probability at least $1 - \delta$, for all $\theta \in \mathbb{B}_d(r)$,

$$\left| R_\infty(\theta) - \hat{R}_\alpha(\theta) \right| \leq \sigma\left(r\sqrt{d}\right)\left(\frac{2r\sqrt{d}}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}} \right)$$
$$+ \frac{\left(\log\sigma(-r\sqrt{d})\right)^2}{2\alpha}. \tag{259}$$

Proof: Consider the expression, $R_\infty(\theta) - \hat{R}_\alpha(\theta)$. Since $\hat{R}_\infty(\theta) \leq \hat{R}_\alpha(\theta)$ for all $\theta \in \mathbb{B}_d(r)$, the following holds

$$R_\infty(\theta) - \hat{R}_\alpha(\theta) \leq R_\infty(\theta) - \hat{R}_\infty(\theta) \tag{260}$$
$$\leq \sigma\left(r\sqrt{d}\right)\left(\frac{2r\sqrt{d}}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}} \right), \tag{261}$$

where we applied Theorem 5 for $\alpha = \infty$. Now, consider the reverse direction, $\hat{R}_\alpha(\theta) - R_\infty(\theta)$. For any $\theta \in \mathbb{B}_d(r)$, we add and subtract $\hat{R}_\infty(\theta)$ such that

$$\hat{R}_\alpha(\theta) - R_\infty(\theta)$$
$$= \hat{R}_\infty(\theta) - R_\infty(\theta) + \hat{R}_\alpha(\theta) - \hat{R}_\infty(\theta) \tag{262}$$
$$\leq \sigma\left(r\sqrt{d}\right)\left(\frac{2r\sqrt{d}}{\sqrt{n}} + 4\sqrt{\frac{2\log\left(\frac{4}{\delta}\right)}{n}} \right) + \frac{\left(\log\sigma(-r\sqrt{d})\right)^2}{2\alpha}, \tag{263}$$

where we apply Theorem 5 for the first term and Lemma 2 for the second term[3] on the maximum value of θ, i.e, $\|\theta\|_2 = r$. Thus, combining the two cases we have the desired statement for the corollary. $\qquad\square$

Theorem 6: Assume that the minimum α-risk is attained by the logistic model, i.e., (55) holds. Let S_n be a training dataset with $n \in \mathbb{N}$ samples as before. If for each $n \in \mathbb{N}$, $\hat{\theta}_n^\alpha$ is a global minimizer of the associated empirical α-risk $\theta \mapsto \hat{R}_\alpha(\theta)$, then the sequence $(\hat{\theta}_n^\alpha)_{n=1}^\infty$ is asymptotically optimal for the 0-1 risk, i.e., almost surely,

$$\lim_{n\to\infty} R(f_{\hat{\theta}_n^\alpha}) = R^*, \tag{264}$$

where $f_{\hat{\theta}_n^\alpha}(x) = \langle\hat{\theta}_n^\alpha, x\rangle$ for each $n \in \mathbb{N}$ and the Bayes risk R^* is given by $R^* := \min_{f:\mathcal{X}\to\mathbb{R}} \mathbb{P}[Y \neq \text{sign}(f(X))]$.

Proof: We begin by recalling the following proposition which establishes an important consequence of classification-calibration. In words, the following result assures that minimizing a classification-calibrated loss to optimality also minimizes the 0-1 loss to optimality.

---

[3] We apply Lemma 2 to the empirical distribution instead of the true distribution, leading to a bound for the empirical α-risk.

Proposition 9 (Thm. 3, [6]): Assume that ϕ 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 $\mathcal{X} \times \mathcal{Y}$,

$$\lim_{i \to \infty} R_\phi(f_i) = R_\phi^* \text{ implies that } \lim_{i \to \infty} R(f_i) = R^*, \quad (265)$$

where $R_\phi^* := \min_f R_\phi(f)$ and $R^* := \min_f R(f)$.

By the assumption that the minimum α-risk is obtained by the logistic model, we have that

$$\min_{\theta \in \mathbb{B}_d(r)} R_\alpha(\theta) = \min_{f : \mathcal{X} \to \mathbb{R}} R_\alpha(f), \quad (266)$$

where $R_\alpha(\theta)$ is given in (25) and $R_\alpha(f) = \mathbb{E}[\tilde{l}^\alpha(Yf(X))]$ for all measurable f. Thus, the proof strategy is to show that

$$\lim_{n \to \infty} R_\alpha(\hat{\theta}_n^\alpha) = \min_{\theta \in \mathbb{B}_d(r)} R_\alpha(\theta), \quad (267)$$

and then apply Proposition 9 to obtain the result.

Let θ_*^α be a minimizer of the α-risk, i.e.,

$$R_\alpha(\theta_*^\alpha) = \min_{\theta \in \mathbb{B}_d(r)} R_\alpha(\theta). \quad (268)$$

Observe that

$$0 \le R_\alpha(\hat{\theta}_n^\alpha) - R_\alpha(\theta_*^\alpha) = \mathrm{I}_n + \mathrm{II}_n, \quad (269)$$

where $\mathrm{I}_n := R_\alpha(\hat{\theta}_n^\alpha) - \hat{R}_\alpha(\hat{\theta}_n^\alpha)$ and $\mathrm{II}_n := \hat{R}_\alpha(\hat{\theta}_n^\alpha) - R_\alpha(\theta_*^\alpha)$. After some straightforward manipulations of Theorem 5, (54) implies that, for every $\epsilon > 0$,

$$\mathbb{P}\left(|R_\alpha(\hat{\theta}_n^\alpha) - \hat{R}_\alpha(\hat{\theta}_n^\alpha)| > \epsilon\right) \le 4\, e^{-n\left(\frac{\epsilon - C_{r\sqrt{d}}(\alpha)2r\sqrt{d}/n}{4\sqrt{2}D_{r\sqrt{d}}(\alpha)}\right)^2}, \quad (270)$$

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

$$\lim_{n \to \infty} \mathrm{I}_n = \lim_{n \to \infty} R_\alpha(\hat{\theta}_n^\alpha) - \hat{R}_\alpha(\hat{\theta}_n^\alpha) = 0. \quad (271)$$

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

$$\mathrm{II}_n = \hat{R}_\alpha(\hat{\theta}_n^\alpha) - R_\alpha(\theta_*^\alpha) \le \hat{R}_\alpha(\theta_*^\alpha) - R_\alpha(\theta_*^\alpha). \quad (272)$$

Again by Theorem 5, for every $\epsilon > 0$,

$$\mathbb{P}\left(|\hat{R}_\alpha(\theta_*^\alpha) - R_\alpha(\theta_*^\alpha)| > \epsilon\right) \le 4\, e^{-n\left(\frac{\epsilon - C_{r\sqrt{d}}(\alpha)2r\sqrt{d}/n}{4\sqrt{2}D_{r\sqrt{d}}(\alpha)}\right)^2}, \quad (273)$$

whenever n is large enough. Hence, the Borel-Cantelli lemma implies that, almost surely,

$$\lim_{n \to \infty} |\hat{R}_\alpha(\theta_*^\alpha) - R_\alpha(\theta_*^\alpha)| = 0. \quad (274)$$

In particular, we have that, almost surely,

$$\limsup_{n \to \infty} \mathrm{II}_n \le 0. \quad (275)$$

Plugging (271) and (275) in (269), we obtain, almost surely,

$$0 \le \limsup_{n \to \infty} \left[R_\alpha(\hat{\theta}_n^\alpha) - R_\alpha(\theta_*^\alpha)\right] \le 0, \quad (276)$$

from which (267) follows.

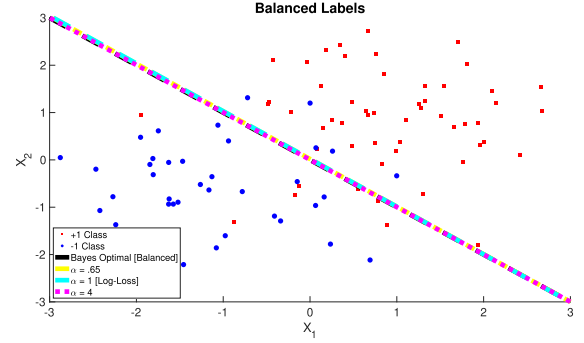Fig. 11. A synthetic experiment highlighting the collapse in trained linear predictors of α-loss for $\alpha \in \{0.65, 1, 4\}$ on clean, balanced data. Specifically, α-loss is trained until convergence under the logistic model for a 2D-GMM with mixing probability $\mathbb{P}[Y = -1] = \mathbb{P}[Y = +1]$, symmetric means $\mu_{X|Y=-1} = [-1, -1] = -\mu_{X|Y=1}$, and shared covariance matrix $\Sigma = \mathbb{I}_2$. Averaged linear predictors generated by training of α-loss averaged over 100 runs. Training data present in the figure is obtained from the last run.

For each $n \in \mathbb{N}$, let $f_{\hat{\theta}_n^\alpha} : \mathcal{X} \to \overline{\mathbb{R}}$ be $f_{\hat{\theta}_n^\alpha}(x) = \langle \hat{\theta}_n^\alpha, x \rangle$. Since we have

$$f_{\hat{\theta}_n^\alpha}(x) = \sigma^{-1}(\sigma(\hat{\theta}_n^\alpha \cdot x)) = \sigma^{-1}(g_{\hat{\theta}_n^\alpha}(x)), \quad (277)$$

Proposition 2, (266), and (267) imply that

$$\lim_{n \to \infty} R_\alpha(f_{\hat{\theta}_n^\alpha}) = \min_{\theta \in \mathbb{B}_d(r)} R_\alpha(f_\theta) = \min_{f : \mathcal{X} \to \mathbb{R}} R_\alpha(f) =: R_\alpha^*. \quad (278)$$

Since $\tilde{l}^\alpha$ is classification-calibrated as established in Theorem 1, Proposition 9 and (278) imply that

$$\lim_{n \to \infty} R(f_{\hat{\theta}_n^\alpha}) = \min_{f : \mathcal{X} \to \mathbb{R}} \mathbb{P}[Y \ne \mathrm{sign}(f(X))] =: R^*, \quad (279)$$

as required. $\quad\square$

D. Further Experimental Results and Details

1) Brief Review of the F1 Score: In binary classification, the F_1 score is a measure of a model's accuracy and is particularly useful when there is an imbalanced class, since it is known to give more precise performance information for an imbalanced class than simply using accuracy itself [80]. In words, the F_1 score is the harmonic mean of the precision and recall, where precision is defined as the number of true positives divided by the number of true positives plus false positives (all examples the model declares as positive) and where recall is defined as the number of true positives divided by the number of true positives plus false negatives (all the examples that the model should have declared as positive). Formally, the definition of the F_1 score is

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = \frac{\mathrm{TP}}{\mathrm{TP} + 0.5(\mathrm{FP} + \mathrm{FN})}, \quad (280)$$

where tp, fp, fn denote true positives, false positive, and false negatives, respectively. In practice, tp, fp, and fn are drawn from the confusion matrix of the model on test data. Note that the use of the term "positive", denoting the class name is arbitrarily chosen; in practice, one lets "positive" class denote the imbalanced class.

TABLE VII

FURTHER QUANTITATIVE RESULTS ASSOCIATED WITH FIG. 3(A) IN SECTION III-C WITH EXACTLY THE SAME EXPERIMENTAL SETUP. VALUES REPORTED IN THE TABLE ARE THE TEST ACCURACY (IN %) OF A LINEAR PREDICTIVE MODEL TESTED ON 1 MILLION EXAMPLES OF CLEAN, BALANCED SYNTHETIC TEST DATA. THE LINEAR MODEL WAS LEARNED BY AVERAGING MODELS FOR 100 TRAINING EXAMPLES OVER 100 RUNS. SUCH MODELS WERE LEARNED FOR DIFFERENT IMBALANCE LEVELS OF THE TRAINING DATA AS SHOWN IN THE TABLE. WE FOUND THAT THE BAYES ACCURACY OF THIS EXPERIMENT WAS 92.14%. IN GENERAL, WE FIND THAT $\alpha^* < 1$, WHICH ALIGNS WITH OUR THEORETICAL INTUITION. THIS CONTRASTS WITH THE NOTABLE EXCEPTION OF 1% IMBALANCE, WHERE $\alpha^* > 1$, WHICH POINTS TOWARDS THE USEFULNESS OF *class upweighting* IN ADDITION TO EMPLOYING α-LOSS FOR SUCH A HIGHLY IMBALANCED CLASS. ALSO OF NOTE, WE FIND THAT SMALLER α IS NOT ALWAYS BETTER (SEE <5% IMBALANCE), WHICH HINTS AT A TRADE-OFF BETWEEN EMPHASIZING THE IMBALANCED CLASS AND COMPUTATIONAL INFEASIBILITY (E.G., EXPLODING GRADIENTS) AS DISCUSSED AFTER PROPOSITION 5. LASTLY, WE NOTE THE CLOSENESS BETWEEN $\alpha = 8$ AND 10^{10} AND ∞; THIS FOLLOWS OUR THEORETICAL INTUITION DERIVED FROM THE *saturation effect* OF α-LOSS AS DEPICTED IN (40)

				$\leftarrow$	α's	$\rightarrow$				
		.4	.5	.65	.8	1	2.5	4	8	10^{10}	∞
	1	72.73	72.36	72.57	71.81	71.79	72.46	73.14	73.71	74.10	74.10
	2	79.54	79.55	78.51	77.81	76.87	74.13	74.59	75.32	75.71	75.71
	5	84.22	83.77	83.48	82.78	82.24	80.68	80.30	80.13	79.71	79.71
$\uparrow$	10	87.86	87.54	87.55	87.30	87.09	85.59	85.36	85.08	84.99	84.99
Imb %	15	89.01	88.98	88.74	88.66	88.63	88.32	88.09	88.14	87.97	87.97
$\downarrow$	20	90.09	90.11	89.96	89.88	89.79	89.61	89.59	89.73	89.60	89.60
	30	91.55	91.36	91.30	91.27	91.24	91.16	91.10	90.90	90.75	90.75
	40	92.00	91.97	91.98	91.97	91.98	92.05	92.07	92.08	92.08	92.08
	50	92.08	92.09	92.08	92.08	92.08	92.08	92.07	92.06	92.06	92.06

TABLE VIII

A TWIN TABLE OF TABLE VII, EXCEPT WITH F_1 SCORES REPORTED. FOR A BRIEF REVIEW OF THE F_1 SCORE, SEE APPENDIX D.1. GAINS OF $\alpha^* < 1$ OVER LOG-LOSS ($\alpha = 1$) ARE MORE EXAGGERATED BY THE F_1 SCORE, IN PARTICULAR SEE 2% AND 5% IMBALANCE

				$\leftarrow$	α's	$\rightarrow$				
		.4	.5	.65	.8	1	2.5	4	8	10^{10}	∞
	1	0.6261	0.6192	0.6231	0.6084	0.6081	0.6209	0.6338	0.6445	0.6517	0.6517
	2	0.7446	0.7448	0.7280	0.7165	0.7007	0.6524	0.6607	0.6739	0.6807	0.6807
	5	0.8146	0.8083	0.8040	0.7938	0.7857	0.7619	0.7560	0.7534	0.7467	0.7467
$\uparrow$	10	0.8648	0.8605	0.8606	0.8573	0.8545	0.8341	0.8309	0.8270	0.8257	0.8257
Imb %	15	0.8800	0.8797	0.8765	0.8755	0.8751	0.8710	0.8680	0.8687	0.8665	0.8665
$\downarrow$	20	0.8937	0.8940	0.8920	0.8910	0.8899	0.8876	0.8872	0.8892	0.8875	0.8875
	30	0.9124	0.9100	0.9092	0.9089	0.9084	0.9074	0.9066	0.9040	0.9021	0.9021
	40	0.9187	0.9183	0.9184	0.9183	0.9183	0.9195	0.9199	0.9200	0.9201	0.9201
	50	0.9207	0.9207	0.9207	0.9208	0.9208	0.9208	0.9207	0.9206	0.9205	0.9205

2) Experiments for Section III-C: In this section, we provide additional synthetic experiments, which follow the same experiment protocol as Fig. 3. They highlight some of the main themes of the paper, namely, $\alpha^* < 1$ in imbalanced experiments, $\alpha^* > 1$ in noisy experiments, trade-offs between computational feasibility and accuracy (for both regimes of α), and the saturation effect.

3) Commentary on Computational Feasibility of α-Loss: In this section, we provide further commentary regarding the computational feasibility of α-loss. In other words, we provide further reasoning for our choice of $\alpha \in [.8, 8]$ as a sufficient search space of α in the experiments in Section VI.

For $\alpha \rightarrow \infty$, we show through our theoretical landscape analysis (see Section IV, Theorem 4, and for a visual, Fig. 4) that the computational complexity increases because gradients tend to become "flatter"; another (perhaps simpler) way to see that the gradients become "flatter" is through Fig. 1(a), where the loss itself has smaller derivatives as α tends to ∞. Unfortunately, a standard gradient optimizer will get stuck in such flat regions of the landscape and learning ceases. Indeed in deep neural networks, the gradients are "back-propogated" through the network, and if the gradient values are small (as is

often the case for the very large α-losses), learning slows down or even stops. This motivates our choice of $\alpha = 8$ as the upper limiting point of our search space, and we argue that it is sufficient because of the saturation effect (see (40)).

For $\alpha \rightarrow 0$, we see the opposite effect, i.e., that the gradients explode as α decreases from 1 (see Proposition 5 with following commentary and Fig. 6 for a visual). Indeed, this motivated the choice of the lower limit of $\alpha = 0.65$ in Fig. 3(a). This issue was "pseudo-circumvented" in Tables VII, VIII, and IX because if there was a NaN, the code would disregard that run of the experiment for that small α and it wouldn't factor into that α's averaged linear predictor. To give a sense for how many NaNs occurred, for the 5% imbalance experiment, $\alpha = .4$ "NaN-ed" out 51 times out of the 100 runs. Thus, we argue that $\alpha = .8$ in general is sufficient as the lower limiting point of the α search space.

For another visual perspective of these considerations, see Fig. 10 which was obtained using [79] on a ResNet-18 learning the MNIST dataset. Interestingly, we see exploding gradients for $\alpha = .9$, loss of convexity (and increasing flatness) as α increases greater than 1, and saturation between $\alpha = 2$ and $\alpha = 8$. Thus, this visualization on a deep neural network

TABLE IX

FURTHER QUANTITATIVE RESULTS ASSOCIATED WITH FIG. 3(B) IN SECTION III-C WITH EXACTLY THE SAME EXPERIMENTAL SETUP (TRAINING DATA WITH LABEL NOISE). VALUES REPORTED IN THE TABLE ARE PERCENT ACCURACY OF AVERAGED LINEAR PREDICTORS, WHICH WERE TRAINED ON NOISY DATA, ON 1 MILLION EXAMPLES OF CLEAN, BALANCED SYNTHETIC TEST DATA. SIMILARLY AS IN TABLE VII, WE OBSERVE A SATURATION EFFECT. FURTHER, NOTE THAT $\alpha = \infty$ DOES NOT ALWAYS OUTPERFORM THE SMALLER α'S, IN PARTICULAR, SEE 20% NOISE WHERE $\alpha^* = 8$. THIS HINTS AT A TRADE-OFF BETWEEN α AND COMPUTATIONAL FEASIBILITY IN THE LARGE α REGIME ($\alpha > 1$), WHICH ALSO FOLLOWS FROM OUR THEORETICAL INTUITION AS STATED AT THE END OF SECTION IV

| | | $\leftarrow$ | | | α's | | $\rightarrow$ | | | |
	.4	.5	.65	.8	1	2.5	4	8	10^{10}	∞
1	92.18	92.17	92.16	92.17	92.17	92.18	92.16	92.13	92.12	92.12
2	92.06	92.07	92.08	92.09	92.11	92.14	92.14	92.14	92.15	92.15
5	91.34	91.41	91.61	91.68	91.85	92.11	92.12	92.13	92.13	92.13
10	90.41	90.34	90.53	90.89	91.29	92.01	92.04	92.05	92.06	92.06
15	88.45	88.72	89.03	89.53	90.14	91.95	92.02	92.02	92.03	92.03
20	87.84	86.21	86.52	87.38	88.85	91.17	91.53	91.91	91.46	91.54
30	80.43	80.34	81.48	82.36	83.55	90.15	90.68	90.86	90.98	90.98
40	75.02	75.20	75.11	75.38	75.89	83.00	84.51	85.59	85.82	85.82
50	67.66	67.45	67.26	67.22	67.08	70.61	73.33	75.67	76.89	76.89

(Row label at left: $\uparrow$ Noise % $\downarrow$)

TABLE X

MULTICLASS SYMMETRIC NOISY LABEL EXPERIMENT ON **MNIST**. SEE TABLE I FOR DESCRIPTIONS OF ACRONYMS

Data	Arch	LF %	LL Acc	α^* Acc	α^*	Gain %
MNIST	CNN 2+2	0	99.16	99.16	1	0.00
		10	94.15	99.00	8	5.15
		20	85.90	98.84	8	15.06
		30	73.54	98.52	8	33.97
		40	60.99	97.96	8	60.62

TABLE XI

MULTICLASS SYMMETRIC NOISY LABEL EXPERIMENT ON THE **FMNIST** DATASET

Data	Arch	LF %	LL Acc	α^* Acc	α^*	Gain %
FMNIST	CNN 2+2	0	90.45	90.45	1	0.00
		10	84.69	89.81	8	6.05
		20	77.51	89.27	7	15.18
		30	67.94	88.10	7	29.67
		40	68.28	88.20	8	28.91

hints at the generality of our theoretical results of the α-loss landscape in Section IV.

4) Multiclass Symmetric Label Flip Experiments: In this section, we present multiclass symmetric noisy label experiments for the MNIST and FMNIST datasets. Our goal is to evaluate the robustness of α-loss over log-loss ($\alpha = 1$) to symmetric noisy labels in the training data. We generate noise in the multiclass training data as follows:

1) For each run of an experiment, we randomly select 0-40% of the training data in increments of 10%.
2) For each training sample in the selected group, we remove the true underlying label number from a list of the ten classes, then we roll a fair nine-sided die over the nine remaining classes in the list; once we have a new label, we replace the true label with the new drawn label.

Note that the test data is clean, i.e., we do not flip the labels of the test dataset. Thus, we consider the canonical scenario where the labels of the training data have been flipped, but the test data is clean.

The results of the multiclass symmetric noisy label experiments are presented in Tables X and XI. Note that we use the same fixed learning rates as the binary symmetric noisy label experiments in Section VI-A. For the MNIST and FMNIST datasets with label flips, we find very strong gains in the test accuracy, which continue to improve as the percentage of label flips increases, through training α-loss for $\alpha > 1$ over log-loss ($\alpha = 1$). Once label flips are present in these two datasets, we note that $\alpha^* = 7$ or 8 for the CNN 2+2 architecture.

REFERENCES

[1] T. Sypherd, M. Diaz, L. Sankar, and P. Kairouz, "A tunable loss function for binary classification," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 2479–2483.

[2] T. Sypherd, M. Diaz, L. Sankar, and G. Dasarathy, "On the α-loss landscape in the logistic model," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2700–2705.

[3] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Springer Series in Statistics), vol. 1, no. 10. New York, NY, USA: Springer, 2001.

[4] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[5] S. Ben-David, N. Eiron, and P. M. Long, "On the difficulty of approximately maximizing agreements," *J. Comput. Syst. Sci.*, vol. 66, no. 3, pp. 496–514, May 2003.

[6] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Statist. Assoc.*, vol. 101, no. 473, pp. 138–156, Mar. 2006.

[7] H. Masnadi-Shirazi and N. Vasconcelos, "On the design of loss functions for classification: Theory, robustness to outliers, and SavageBoost," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1049–1056.

[8] Y. Lin, "A note on margin-based loss functions in classification," *Statist. Probab. Lett.*, vol. 68, no. 1, pp. 73–82, 2004.

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

[10] L. Rosasco, E. Vito, A. Caponnetto, M. Piana, and A. Verri, "Are loss functions all the same?" *Neural Comput.*, vol. 16, no. 5, pp. 1063–1076, May 2004.

[11] T. Nguyen and S. Sanner, "Algorithms for direct 0–1 loss optimization in binary classification," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1085–1093.

[12] A. Singh and J. C. Principe, "A loss function for classification based on a robust similarity metric," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2010, pp. 1–6.

[13] A. Tewari and P. L. Bartlett, "On the consistency of multiclass classification methods," *J. Mach. Learn. Res.*, vol. 8, pp. 1007–1025, May 2007.

[14] L. Zhao, M. Mammadov, and J. Yearwood, "From convex to nonconvex: A loss function analysis for binary classification," in *Proc. IEEE Int. Conf. Data Mining Workshops*, Dec. 2010, pp. 1281–1288.

[15] J. T. Barron, "A general and adaptive robust loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4331–4339.

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[17] J. Liao, O. Kosut, L. Sankar, and F. P. Calmon, "A tunable measure for information leakage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 701–705.

[18] J. Liao, L. Sankar, O. Kosut, and F. P. Calmon, "Robustness of maximal α-leakage to side information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 642–646.

[19] J. Liao, L. Sankar, O. Kosut, and F. P. Calmon, "Maximal α-leakage and its properties," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Jun. 2020, pp. 1–6.

[20] C. E. Shannon, "A mathematical theory of communication," *ACM SIG-MOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[21] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[22] S. Verdú, "α-mutual information," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2015, pp. 1–6.

[23] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1961, pp. 547–561. [Online]. Available: https://projecteuclid.org/euclid.bsmsp/1200512181

[24] S. Arimoto, "Information-theoretical considerations on estimation problems," *Inf. Control*, vol. 19, no. 3, pp. 181–194, Oct. 1971. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0019995871900659

[25] I. Sason and S. Verdú, "Arimoto–Rényi conditional entropy and Bayesian M-ary hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 4–25, Jan. 2018.

[26] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[27] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," in *Proc. Int. Conf. Learn. Represent.*, 2020. [Online]. Available: https://openreview.net/forum?id=ByexElSYDr

[28] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *Theor. Found. Mach. Learn.*, vol. 25, pp. 49–59, Feb. 2016.

[29] R. Nock and A. Menon, "Supervised learning: No loss no cry," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7370–7380.

[30] C. Walder and R. Nock, "All your loss are belong to Bayes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18505–18517.

[31] M. D. Reid and R. C. Williamson, "Composite binary losses," *J. Mach. Learn. Res.*, vol. 11, pp. 2387–2422, Sep. 2010.

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

[33] P. M. Long and R. A. Servedio, "Random classification noise defeats all convex potential boosters," *Mach. Learn.*, vol. 78, no. 3, pp. 287–304, 2010.

[34] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[35] E. Hazan, K. Levy, and S. Shalev-Shwartz, "Beyond convexity: Stochastic quasi-convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1594–1602.

[36] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6389–6399.

[37] Q. Nguyen and M. Hein, "The loss surface of deep and wide neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 2603–2612.

[38] H. Fu, Y. Chi, and Y. Liang, "Guaranteed recovery of one-hidden-layer neural networks via cross entropy," *IEEE Trans. Signal Process.*, vol. 68, pp. 3225–3235, 2020.

[39] S. Liang, R. Sun, Y. Li, and R. Srikant, "Understanding the loss surface of neural networks for binary classification," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2835–2843.

[40] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1802–1811.

[41] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier, "Deep relaxation: Partial differential equations for optimizing deep neural networks," *Res. Math. Sci.*, vol. 5, no. 3, p. 30, Sep. 2018.

[42] Y. E. Nesterov, "Minimization methods for nonsmooth convex and quasiconvex functions," *Matekon*, vol. 29, no. 3, pp. 519–531, 1984.

[43] L. Benigni and S. Péché, "Eigenvalue distribution of nonlinear models of random matrices," 2019, *arXiv:1904.03090*.

[44] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. Schoenholz, and J. Pennington, "Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5393–5402.

[45] J. Pennington and P. Worah, "Nonlinear random matrix theory for deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2637–2646.

[46] D. Russo and J. Zou, "How much does your data exploration overfit? Controlling bias via information usage," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 302–323, Jan. 2020.

[47] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2524–2533.

[48] A. T. Lopez and V. Jog, "Generalization error bounds using Wasserstein distances," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[49] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An information-theoretic view of generalization via Wasserstein distance," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 577–581.

[50] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 121–130, May 2020.

[51] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Proc. Conf. Learn. Theory*, 2020, pp. 3437–3452.

[52] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rényi-, f-divergences and maximal leakage," *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 4986–5004, Aug. 2021.

[53] B. R. Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via Wasserstein distance," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, 2021, pp. 1–21.

[54] G. Neu, G. K. Dziugaite, M. Haghifam, and D. M. Roy, "Information-theoretic generalization bounds for stochastic gradient descent," in *Proc. COLT*, 2021, pp. 1–21.

[55] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with Noisy labels," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 322–330.

[56] E. Amid, M. K. Warmuth, R. Anil, and T. Koren, "Robust bi-tempered logistic loss based on Bregman divergences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 15013–15022.

[57] T. Li, A. Beirami, M. Sanjabi, and V. Smith, "On tilted losses in machine learning: Theory and applications," 2021, *arXiv:2109.06141*.

[58] G. R. Kurri, T. Sypherd, and L. Sankar, "Realizing GANs via a tunable loss function," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Oct. 2021, pp. 1–6.

[59] T. Sypherd, R. Nock, and L. Sankar, "Being properly improper," 2021, *arXiv:2106.09920*.

[60] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," *Topics Inf. Theory*, to be published.

[61] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. Cambridge, MA, USA: MIT, 2021.

[62] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[63] T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 1999.

[64] Y. Wu and Y. Liu, "Robust truncated Hinge loss support vector machines," *J. Amer. Statist. Assoc.*, vol. 102, no. 479, pp. 974–983, 2007.

[65] O. Chapelle, C. B. Do, C. H. Teo, Q. V. Le, and A. J. Smola, "Tighter bounds for structured estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 281–288.

[66] J.-Y. Audibert and A. B. Tsybakov, "Fast learning rates for plug-in classifiers," *Ann. Statist.*, vol. 35, no. 2, pp. 608–633, 2007.

[67] Y. LeCun, C. Cortes, and C. J. C. Burges. (1998). *The MNIST Database of Handwritten Digits*. [Online]. Available: http://yann.lecun.com/exdb/mnist/index.html

[68] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2016, *arXiv:1708.07747*.

[69] A. Krizhevsky, V. Nair, and G. Hinton. (2014). *The CIFAR-10 Dataset*. [Online]. Available: http://www.cs.toronto.edu/kriz/cifar.html

[70] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[71] J. K. Cava. (2021). *Code for a Tunable Loss Function for Robust Classification*. [Online]. Available: https://github.com/SankarLab/AlphaLoss

[72] S. Fort, P. K. Nowak, S. Jastrzebski, and S. Narayanan, "Stiffness: A new perspective on generalization in neural networks," 2019, *arXiv:1901.09491*.

[73] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7472–7482.

[74] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–10.

[75] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5014–5026.

[76] P. Kairouz, J. Liao, C. Huang, and L. Sankar, "Censored and fair universal representations using generative adversarial models," 2019, *arXiv:1910.00411*.

[77] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[78] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York, NY, USA: McGraw-Hill, 2002.

[79] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 6391–6401.

[80] Y. Sasaki, *The Truth of the F-Measure*, 2007. [Online]. Available: https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/Fmeasure-YS-26Oct07.pdf

Tyler Sypherd (Graduate Student Member, IEEE) received the B.S.E. and M.S. degrees in electrical engineering from Arizona State University in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree in electrical engineering. His research interests include theoretical machine learning and information theory.

Mario Diaz (Member, IEEE) received the B.Eng. degree in electrical engineering from the Universidad de Guadalajara, Guadalajara, Mexico, in 2011, the M.Sc. degree in probability and statistics from the Centro de Investigación en Matemáticas, Guanajuato, Mexico, in 2013, and the Ph.D. degree in mathematics and statistics from Queen's University, Kingston, Canada, in 2017. He is currently a Research Associate with the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México, Mexico City, Mexico. Prior to this, he was a Post-Doctoral Scholar with Arizona State University, the Centro de Investigación en Matemáticas, and Harvard University. His research interests include the mathematical and statistical foundations of information privacy, theoretical machine learning, and random matrix theory.

John Kevin Cava (Graduate Student Member, IEEE) received the B.S. degrees in computer science, mathematics, and molecular biosciences and biotechnology from Arizona State University in 2018, where he is currently pursuing the Ph.D. degree in computer science. His research interests include computer vision, natural language processing, and their use in molecular biology applications.

Gautam Dasarathy (Senior Member, IEEE) received the bachelor's degree in electronics and communication engineering from VIT University in 2008 and the master's and doctorate degrees in electrical engineering from the University of Wisconsin–Madison in 2010 and 2014, respectively. He held post-doctoral positions at Rice University and Carnegie Mellon University. He is currently an Assistant Professor with the School of Electrical, Computer and Energy Engineering, Arizona State University. His research interests span topics in machine learning, statistics, signal processing and networked systems, and information theory. He was the recipient of the CAREER Award from the National Science Foundation (NSF) in 2021. He was also the winner of the 2022 Distinguished Alumnus Award from VIT University, India.

Peter Kairouz (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC). He is a Research Scientist at Google, where he coordinates research efforts on federated learning and privacy-preserving technologies. Before joining Google, he was a Post-Doctoral Research Fellow with Stanford University. He was the recipient of the 2012 Roberto Padovani Scholarship from Qualcomm's Research Center, the 2015 ACM SIGMETRICS Best Paper Award, the 2021 ACM Conference on Computer and Communications Security (CCS) Best Paper Award, the 2015 Qualcomm Innovation Fellowship Finalist Award, and the 2016 Harold L. Olesen Award for Excellence in Undergraduate Teaching from UIUC.

Lalitha Sankar (Senior Member, IEEE) received the B.Tech. degree from the Indian Institute of Technology, Mumbai, the M.S. degree from the University of Maryland, and the Ph.D. degree from Rutgers University. She is currently an Associate Professor with the School of Electrical, Computer, and Energy Engineering, Arizona State University. She currently leads an NSF HDR institute on data science for electric grid operations. Her research interests include applying information theory and data science to study reliable, responsible, and privacy-protected machine learning as well as cyber security and resilience in critical infrastructure networks. She received the National Science Foundation CAREER Award in 2014, the IEEE Globecom 2011 Best Paper Award for her work on privacy of side-information in multi-user data systems, and the Academic Excellence Award from Rutgers University in 2008.