On the α-loss Landscape in the Logistic Model

Tyler Sypherd*, Mario Diaz†, Lalitha Sankar*, and Gautam Dasarathy*
* Arizona State University, {tsypherd,lsankar,gautamd}@asu.edu
† Universidad Nacional Autónoma de México, mario.diaz@sigma.iimas.unam.mx

***Abstract*—We analyze the optimization landscape of a recently introduced tunable class of loss functions called α-loss, $\alpha \in (0, \infty]$, in the logistic model. This family encapsulates the exponential loss ($\alpha = 1/2$), the log-loss ($\alpha = 1$), and the 0-1 loss ($\alpha = \infty$) and contains compelling properties that enable the practitioner to discern among a host of operating conditions relevant to emerging learning methods. Specifically, we study the evolution of the optimization landscape of α-loss with respect to α using tools drawn from the study of strictly-locally-quasi-convex functions in addition to geometric techniques. We interpret these results in terms of optimization complexity via normalized gradient descent.**

I. Introduction

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. Consider a feature vector $X \in \mathcal{X}$, an unknown finite-valued label $Y \in \mathcal{Y}$, and a hypothesis $h : \mathcal{X} \to \mathcal{Y}$. The canonical 0-1 loss, given by $\mathbb{1}[h(X) \neq Y]$, is considered an ideal loss function that captures the probability of incorrectly guessing the true label Y using $h(X)$. However, since the 0-1 loss is neither continuous nor differentiable, its applicability in state-of-the-art learning algorithms is highly restricted.

Surrogate loss functions that approximate the 0-1 loss such as log-loss, exponential loss, sigmoid loss, etc. have generated much interest [1]–[14]. While early research was predominantly focused on convex losses [1], [3]–[5], more recent works propose the use of non-convex losses as a means to moderate the behavior of an algorithm [2], [6], [10], [12]. This is primarily due to the fact that modern learning models (e.g., deep learning) are inherently non-convex as they involve vast functional compositions [15]; further, non-convex losses are also believed to provide increased robustness over convex losses [2], [6], [10], [12].

There have been numerous theoretical attempts to capture the non-convex optimization landscape which is the loss surface induced by the learning model, underlying distribution, and the surrogate loss function itself [12], [16]–[22]. Notably, Hazan *et al.* [16] propose the notion of *Strict-Local-Quasi-Convexity* (SLQC) to parametrically quantify quasi-convex functions, and provide convergence guarantees for the efficiency of the Normalized Gradient Descent (NGD) algorithm (originally introduced in [23]) optimizing such functions.

In [13], Sypherd *et al.* introduce a tunable class of loss functions called α-loss, $\alpha \in [1, \infty]$, which includes log-loss ($\alpha = 1$) and the soft 0-1 loss ($\alpha = \infty$); they prove that it satisfies many desirable properties for surrogate losses including the notion of classification-calibration [1]. In the extended version of this paper [14], Sypherd *et al.* extend α-loss to the range $\alpha \in (0, \infty]$ which includes exponential loss ($\alpha = 1/2$); they prove that the extended range of α also induces classification-calibrated losses and has desirable convexity characteristics. Further, they show experimentally that, relative to log-loss ($\alpha = 1$), $\alpha > 1$ achieves increased robustness to noise while $\alpha < 1$ achieves better accuracy for imbalanced classes.

In this paper, we present three main contributions for the logistic model (the hypothesis class of sigmoid soft classifiers [24]): (i) we show that the expected risk under α-loss is strongly convex for $\alpha \in (0, 1]$ (under mild distribution assumptions); (ii) we provide, in a quantitative manner, bounds for the evolution of the SLQC parameters of the expected risk of α-loss as α increases, which is most useful in a neighborhood of $\alpha_0 = 1$ when combined with the first result; (iii) we study a saturation effect of α-loss in the logistic model, *i.e.*, how the distance between the expected risk for $\alpha \geq 1$ quickly resembles the expected risk of $\alpha = \infty$. As a byproduct of the analysis in the second point, we prove an equivalent form of the SLQC definition that can be of independent interest. Based on our theoretical analysis into the evolution of the optimization landscape with respect to α, we ultimately posit that there is a small range of α useful to the practitioner, thereby drastically reducing the search for the optimal value of α in practice. Further, via the saturation effect, we argue that this narrow search in α is sufficient for the logistic model.

II. Preliminaries

A. α-loss Definition and Interpretations

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

$$l^\alpha(y, P_Y) := \frac{\alpha}{\alpha - 1}\left[1 - P_Y(y)^{1-1/\alpha}\right], \tag{1}$$

and, by continuous extension, $l^1(y, P_Y) := -\log P_Y(y)$ and $l^\infty(y, P_Y) := 1 - P_Y(y)$.

Note that $l^{1/2}(y, P_Y) := P_Y^{-1}(y) - 1$. We refer to $l^{1/2}$ as the soft exponential loss and l^∞ as the soft 0-1 loss; observe that l^1 recovers log-loss. For (y, P_Y) fixed, note that $l^\alpha(y, P_Y)$ is continuous in α. The above definition of α-loss presents a tunable class of loss functions that value the probabilistic estimate of the label differently as a function of α.

Consider random variables $(X, Y) \sim P_{X,Y}$. Observing X, one can construct an estimate $\hat{Y}$ of Y such that $Y -$

$X - \hat{Y}$ form a Markov chain. One can use expected α-loss $\mathbb{E}_{X,Y}[l^\alpha(Y, P_{\hat{Y}|X})]$, hence called α-risk, to quantify the effectiveness of the estimated posterior $P_{\hat{Y}|X}$. In particular,

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

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

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

i.e., the expected α-loss for $\alpha = \infty$ equals the probability of error. Recall that the expectation of the 0-1 loss is also the probability of error [25]; thus, we say that α-loss for $\alpha = \infty$ is a soft version of the 0-1 loss. The following result by Liao *et al.* provides an explicit characterization of the risk-minimizing posterior under α-loss.

Proposition 1 ([26, Lemma 1]). *For each $\alpha \in [1, \infty]$, the minimal α-risk is*

$$\min_{P_{\hat{Y}|X}} \mathbb{E}_{X,Y}\left[l^\alpha(Y, P_{\hat{Y}|X})\right] = \frac{\alpha}{\alpha - 1}\left(1 - e^{\frac{1-\alpha}{\alpha} H_\alpha^A(Y|X)}\right), \quad (4)$$

where $H_\alpha^A(Y|X) = \dfrac{\alpha}{1-\alpha}\log\sum_y\left(\sum_x P_{X,Y}(x,y)^\alpha\right)^{1/\alpha}$ is the Arimoto conditional entropy of order α [27]. The resulting unique minimizer, $P^*_{\hat{Y}|X}(y|x)$, is the α-tilted true posterior

$$P^*_{\hat{Y}|X}(y|x) = \frac{P_{Y|X}(y|x)^\alpha}{\sum_y P_{Y|X}(y|x)^\alpha}. \quad (5)$$

The proof of Proposition 1 can be found in [26] and is easily extended to the case where $\alpha \in (0,1)$. For $\alpha = \infty$, minimizing the corresponding risk leads to making a single guess on the most likely label; on the other hand, for $\alpha = 1$, such a risk minimization involves minimizing the average log-loss, and therefore, obtaining the true posterior belief.

We note that α-loss exhibits different operating conditions through the choice of α; see [14] for experimental consideration of robustness and class imbalance trade-offs. With respect to (5), as α increases from 1 to ∞, α-loss increasingly limits the effect of low probability outcomes; on the other hand, as α decreases from 1 towards 0, α-loss places increasingly higher weights on low probability outcomes.

B. Strict-Local-Quasi-Convexity

We briefly review *Strict-Local-Quasi-Convexity* which was introduced by Hazan *et al.* in [16]. For $\theta_0 \in \mathbb{R}^d$ and $r > 0$, we let $\mathbb{B}_d(\theta_0, r) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\| \leq r\}$. For simplicity, we let $\mathbb{B}_d(r) = \mathbb{B}_d(\mathbf{0}, r)$; also note that all norms are Euclidean.

Definition 2. *Let $\theta, \theta_0 \in \mathbb{R}^d$, $\kappa, \epsilon > 0$. We say that $f : \mathbb{R}^d \to \mathbb{R}$ is $(\epsilon, \kappa, \theta_0)$-Strictly-Locally-Quasi-Convex (SLQC) in θ, if at least one of the following applies:*

1) $f(\theta) - f(\theta_0) \leq \epsilon$.
2) $\|\nabla f(\theta)\| > 0$, *and for every $\theta' \in \mathbb{B}_d(\theta_0, \epsilon/\kappa)$ it holds that $\langle \nabla f(\theta), \theta' - \theta\rangle \leq 0$.*

Intuitively, if θ_0 is fixed, then, for every θ, either $f(\theta)$ is ϵ-close to $f(\theta_0)$ or the constraint cone induced by the set of θ' about θ_0 requires quasi-convex functional descent behavior. This relaxed notion of quasi-convexity aligns with a natural adaptation of the Gradient Descent (GD) algorithm, namely, Normalized Gradient Descent (NGD) [16] as summarized in Algorithm 1 below. Similar to the convergence guarantees for GD for convex functions, the following result by Hazan *et al.* summarizes such guarantees of NGD for SLQC functions.

Algorithm 1 Normalized Gradient Descent (NGD)

1: **Input:** T Iterations, $\theta_1 \in \mathbb{R}^d$, learning rate $\eta > 0$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Update: $\theta_{t+1} = \theta_t - \eta\dfrac{\nabla f(\theta_t)}{\|\nabla f(\theta_t)\|}$
4: **Return** $\bar{\theta}_T = \arg\min_{\theta_1,\ldots,\theta_T} f(\theta_t)$

Proposition 2 ([16, Theorem 4.1]). *Fix $\epsilon > 0$, let $f : \mathbb{R}^d \to \mathbb{R}$, and $\theta^* = \arg\min_{\theta\in\mathbb{R}^d} f(\theta)$. If f is $(\epsilon, \kappa, \theta^*)$-SLQC in every $\theta \in \mathbb{R}^d$, then by running Algorithm 1 with $\eta = \epsilon/\kappa$ and $T \geq \kappa^2\|\theta_1 - \theta^*\|^2/\epsilon^2$, we have $f(\bar{\theta}_T) - f(\theta^*) \leq \epsilon$.*

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 including $\mathbb{B}_d(r)$.

III. Landscape Evolution in the Logistic Model

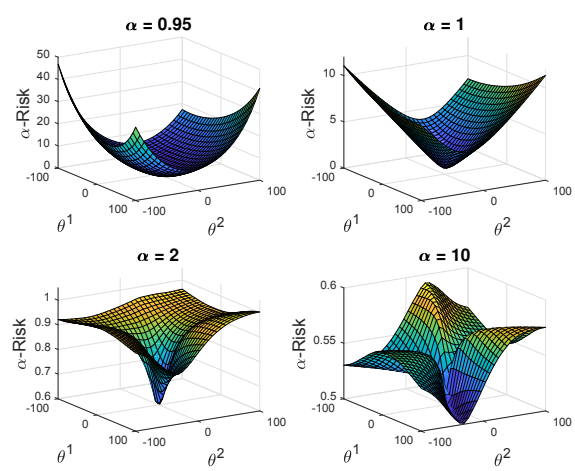

Figure 1. The landscape of α-loss (R_α for $\alpha = 0.95, 1, 2, 10$) in the logistic model, where features are normalized and $r = 100$, for a 2D-GMM with $\mathbb{P}[Y = -1] = 0.12 = 1 - \mathbb{P}[Y = 1]$, $\mu_{X|Y=-1} = [-0.18, 1.49]$, $\mu_{X|Y=1} = [-0.01, 0.16]$, $\Sigma_{-1} = [3.20, -2.02; -2.01, 2.71]$, and $\Sigma_1 = [4.19, 1.27; 1.27, 0.90]$.

In this section, we quantify the optimization complexity of NGD by characterizing the SLQC constants (ϵ and κ) of the α-risk within a neighborhood of $\alpha = 1$ (log-loss)

in the logistic model. For $\alpha \leq 1$, we find that the α-risk is strongly convex under mild distributional assumptions; for $\alpha > 1$ we reinterpret the SLQC definition to provide bounds on SLQC constants as α increases. Finally, we provide a result which characterizes a saturation effect of the α-risk in the logistic model, *i.e.*, the fact that the α-risk observes uniform convergence with respect to $\alpha = \infty$ as α increases.

Prior to stating our main results, we clarify the setting and provide necessary definitions. Let $X \in \mathbb{B}_d(1) := \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ be the normalized feature, $Y \in \{-1, +1\}$ the label and $S_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$ the training dataset where, for each $i \in \{1, \ldots, n\}$, the samples (X_i, Y_i) are independently drawn according to an unknown distribution $P_{X,Y}$. For a 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

$$g_\theta(x) = \sigma(\langle \theta, x \rangle), \tag{6}$$

with $\sigma : \mathbb{R} \to [0, 1]$ being the sigmoid function given by

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{7}$$

For convenience, we present the following short form of α-loss in the logistic model which is equivalent to the expanded expression in [13]. For $\alpha \in (0, \infty]$, α-loss is given by

$$l^\alpha(y, g_\theta(x)) = \frac{\alpha}{\alpha - 1}\left[1 - g_\theta(yx)^{1-1/\alpha}\right]. \tag{8}$$

For $\alpha = 1$, l^1 is the logistic loss and we recover logistic regression by optimizing this loss. Further, note that in this setting $y \cdot \langle x, \theta \rangle$ is the margin, and (8) is convex for $\alpha \in (0, 1]$ and quasi-convex for $\alpha > 1$ in $y \cdot \langle x, \theta \rangle$; see the extended version [14] for proofs of these facts.

For $\theta \in \mathbb{B}_d(r)$, we define the α-risk R_α as the risk of (8),

$$R_\alpha(\theta) := \mathbb{E}_{X,Y}[l^\alpha(Y, g_\theta(X))]. \tag{9}$$

The α-risk (9) is plotted for several values of α in a two-dimensional Gaussian Mixture Model (GMM) in Figure 1. Further, observe that, for all $\theta \in \mathbb{B}_d(r)$,

$$R_\infty(\theta) := \mathbb{E}_{X,Y}[l^\infty(Y, g_\theta(X))] = \mathbb{P}[Y \neq \hat{Y}_\theta], \tag{10}$$

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 (8), by employing the following useful properties of the sigmoid

$$\frac{d}{dz}\sigma(z) = \sigma(z)(1 - \sigma(z)); \qquad \sigma(-z) = 1 - \sigma(z). \tag{11}$$

Indeed, a straightforward computation shows that

$$\frac{\partial}{\partial \theta^j} l^\alpha(y, g_\theta(x)) = \left[-y g_\theta(yx)^{1-1/\alpha}(1 - g_\theta(yx))\right] x^j, \tag{12}$$

where θ^j, x^j denote the j-th components of θ and x, respectively. Thus, the gradient of α-loss in (8) is

$$\nabla_\theta l^\alpha(Y, g_\theta(X)) = F_1(\alpha, \theta, X, Y) X, \tag{13}$$

where $F_1(\alpha, \theta, x, y)$ is the expression within brackets in (12). Another straightforward computation yields

$$\nabla_\theta^2 l^\alpha(Y, g_\theta(X)) = F_2(\alpha, \theta, X, Y) X X^T, \tag{14}$$

where F_2 is given by

$$F_2(\alpha, \theta, x, y) = g_\theta^{1-\alpha^{-1}}(yx)\left(g_\theta'(yx) - \left(1 - \alpha^{-1}\right) g_\theta^2(-yx)\right).$$

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 Figure 1. Prior to stating the result, for two matrices $A, B \in \mathbb{R}^{d \times d}$, we let $\geq$ denote the Loewner (partial) order in the positive semi-definite cone. That is, we write $A \geq B$ 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 $m \geq 0$ [28].

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

$$\Lambda(\alpha, r) = \sigma^{1-1/\alpha}(r)\left(\sigma'(r) - \left(1 - \alpha^{-1}\right)\sigma^2(-r)\right). \tag{15}$$

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 $F_2(\alpha, \theta, x, y) \geq \Lambda(\alpha, r) \geq 0$. Now, consider $R_\alpha(\theta) = \mathbb{E}[l^\alpha(Y, g_\theta(X))]$. We have

$$\begin{aligned}
\nabla_\theta^2 R_\alpha(\theta) &= \mathbb{E}_{X,Y}[\nabla_\theta^2 l^\alpha(Y, g_\theta(X))] \\
&= \mathbb{E}_{X,Y}[F_2(\alpha, \theta, X, Y) X X^T] \\
&\geq \Lambda(\alpha, r)\mathbb{E}[XX^T] \qquad (16) \\
&= \Lambda(\alpha, r)\Sigma \geq 0, \qquad (17)
\end{aligned}$$

where we used an identity of positive semi-definite matrices for (16) (see, e.g., [29, Ch. 7]); for (17), we used the fact that $\Lambda(\alpha, r) \geq 0$ and we recognize that Σ is positive semi-definite as it is the autocorrelation of the random vector $X \in \mathbb{B}_d(1)$ (see, e.g., [30, Ch. 7]). We also note that $\min_{i \in [d]} \lambda_i(\Sigma) \geq 0$ (see, e.g., [29, Ch. 7]). Thus, $\nabla_\theta^2 R_\alpha(\theta)$ is positive semi-definite for every $\theta \in \mathbb{B}_d(r)$. Therefore, since $\lambda_{\min}(\nabla^2 R_\alpha(\theta)) \geq \Lambda(\alpha, r) \min_{i \in [d]} \lambda_i(\Sigma) \geq 0$ for every $\theta \in \mathbb{B}_d(r)$ [29, Corollary 4.3.12], we have that R_α is $\Lambda(\alpha, r) \min_{i \in [d]} \lambda_i(\Sigma)$-strongly convex for $\alpha \in (0, 1]$. □

Observe that for $r > 0$, $\Lambda(\alpha, r)$ is monotonically decreasing in α. Therefore, R_α becomes more strongly convex as α approaches zero. It can be shown that R_α is $C_{r,\alpha}$-Lipschitz in θ where $C_{r,\alpha} := \sigma(r)(1 - \sigma(r))^{1-1/\alpha}$. Thus, in conjunction with Theorem 1 and a result by Hazan *et al.* in [16] (after Definition 3) which holds by assuming $\Sigma > 0$, we provide the following corollary which explicitly characterizes SLQC constants of R_α for $\alpha \in (0, 1]$.

Corollary 1. *If $0 < \alpha \leq 1$, $\Sigma > 0$, and $\theta_0 \in \mathbb{B}_d(r)$, then, for every $\epsilon > 0$, the α-risk R_α is $(\epsilon, C_{r,\alpha}, \theta_0)$-SLQC in $\theta \in \mathbb{B}_d(r)$ where $C_{r,\alpha} = \sigma(r)(1 - \sigma(r))^{1-1/\alpha}$.*

As α tends to zero, $C_{r,\alpha}$ tends to infinity which implies that the learning rate of NGD, $\eta_\alpha = \epsilon/\kappa_\alpha = \epsilon/C_{r,\alpha}$ also tends

to zero. Thus, by Proposition 2, the number of iterations of NGD, T_α, tends to infinity as α tends to zero. Therefore, for $\alpha \in (0, 1]$, there is a trade-off in the desired strong-convexity of R_α and the computational complexity of NGD.

Next, we study the evolution of SLQC parameters of R_α in a neighborhood of $\alpha = 1$ as we increase α. Since R_α tends more towards the probability of error (expectation of 0-1 loss) as α approaches infinity, we find that SLQC constants deteriorate and the computational complexity of NGD increases as we increase α. Our next main result leverages the following novel lemma, which is a structural result for general differentiable functions that provides an alternative formulation of the second requirement of SLQC functions in Definition 2; proof details and illustrations can be found in the extended version [14].

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, Lemma 1 reformulates the SLQC requirement that the gradient points in the 'right' direction into an expression which is reminiscent of a Cauchy-Schwarz inequality.

We now present two Lipschitz inequalities which will be useful in the sequel. In the extended version [14], Sypherd *et al.* show that for $\alpha \in [1, \infty]$, R_α is L_r-Lipschitz in α where

$$L_r := \frac{(r + \log 2)^2}{2}. \tag{18}$$

It can similarly be shown that for $\alpha \in [1, \infty]$, ∇R_α is J_r-Lipschitz in $\alpha^{-1} \in [0, 1]$ where

$$J_r := (r + \log 2)\sigma(r). \tag{19}$$

Finally, for ease of notation, let

$$I_{\alpha_0,\epsilon_0,r}(\theta_0) = \inf_{\substack{\theta \in \mathbb{B}_d(r) \\ R_{\alpha_0}(\theta) - R_{\alpha_0}(\theta_0) > \epsilon_0}} \|\nabla R_{\alpha_0}(\theta)\|. \tag{20}$$

Using Lemma 1 and the Lipschitz relations (18) and (19), we provide the following result which gives precise bounds on the degradation of SLQC constants for any initial $\alpha_0 \in [1, \infty]$.

Theorem 2. *Let $\alpha_0 \in [1, \infty]$, $\epsilon_0, \kappa_0 > 0$, and $\theta_0 \in \mathbb{B}_d(r)$. If R_{α_0} is $(\epsilon_0, \kappa_0, \theta_0)$-SLQC in $\theta \in \mathbb{B}_d(r)$, and*

$$0 \leq \alpha - \alpha_0 < \frac{\alpha_0^2 I_{\alpha_0,\epsilon_0,r}(\theta_0)}{2J_r\left(1 + r\frac{\kappa_0}{\epsilon_0}\right)}, \tag{21}$$

then R_α is $(\epsilon, \kappa, \theta_0)$-SLQC in $\theta \in \mathbb{B}_d(r)$ with

$$\epsilon = \epsilon_0 + 2L_r(\alpha - \alpha_0), \tag{22}$$

and

$$\frac{\epsilon}{\kappa} = \frac{\epsilon_0}{\kappa_0}\left(1 - \frac{\left(1 + 2r\frac{\kappa_0}{\epsilon_0}\right) J_r(\alpha - \alpha_0)}{\alpha\alpha_0 I_{\alpha_0,\epsilon_0,r}(\theta_0) - J_r(\alpha - \alpha_0)}\right). \tag{23}$$

Proof. For ease of notation let $\rho_0 = \frac{\epsilon_0}{\kappa_0}$ and $\rho = \frac{\epsilon}{\kappa}$. Let $\theta \in \mathbb{B}_d(r)$ be arbitrary and consider the following cases.

Case 1: If $R_{\alpha_0}(\theta) - R_{\alpha_0}(\theta_0) \leq \epsilon_0$, then,

$$\begin{aligned} R_\alpha(\theta) - R_\alpha(\theta_0) &= R_\alpha(\theta) - R_{\alpha_0}(\theta) + R_{\alpha_0}(\theta) \\ &\quad - R_{\alpha_0}(\theta_0) + R_{\alpha_0}(\theta_0) - R_\alpha(\theta_0) \\ &\leq L_r|\alpha - \alpha_0| + \epsilon_0 + L_r(\alpha - \alpha_0). \end{aligned} \tag{24}$$

Since $\epsilon_0 + 2L_r(\alpha - \alpha_0) = \epsilon$, we have $R_\alpha(\theta) - R_\alpha(\theta_0) \leq \epsilon$.

Case 2: If $R_{\alpha_0}(\theta) - R_{\alpha_0}(\theta_0) > \epsilon_0$, then, since R_{α_0} is $(\epsilon_0, \kappa_0, \theta_0)$-SLQC in θ by assumption, we have that $\|\nabla R_{\alpha_0}(\theta)\| > 0$, and for every $\theta' \in \mathbb{B}(\theta_0, \rho_0)$ it holds that $\langle \nabla R_{\alpha_0}(\theta), \theta_0 - \theta \rangle \leq 0$. By Lemma 1, we want to prove that

$$\langle -\nabla R_\alpha(\theta), \theta_0 - \theta \rangle \geq \rho \|\nabla R_\alpha(\theta)\|, \tag{25}$$

for ρ given by (23). By the Cauchy-Schwarz inequality,

$$\begin{aligned} \langle -\nabla R_\alpha(\theta), \theta_0 - \theta \rangle &\geq \langle -\nabla R_{\alpha_0}(\theta), \theta_0 - \theta \rangle \\ &\quad - \|\nabla R_\alpha(\theta) - \nabla R_{\alpha_0}(\theta)\| \|\theta_0 - \theta\| \\ &\geq \rho_0 \|\nabla R_{\alpha_0}(\theta)\| - J_r(\alpha_0^{-1} - \alpha^{-1})2r, \end{aligned} \tag{26}$$

since ∇R_α is J_r-Lipschitz in α^{-1} and $\theta_0 - \theta \in \mathbb{B}_d(2r)$, and since R_{α_0} is SLQC, we apply Lemma 1. For ease of notation, we temporarily let $\Delta = J_r(\alpha_0^{-1} - \alpha^{-1})$. Continuing, we have

$$\begin{aligned} \rho_0 \|\nabla R_{\alpha_0}(\theta)\| - \Delta 2r &\geq \rho_0 \|\nabla R_\alpha(\theta)\| - \Delta 2r \\ &\quad - \rho_0 \|\nabla R_{\alpha_0}(\theta) - \nabla R_\alpha(\theta)\| \\ &\geq \rho_0 \|\nabla R_\alpha(\theta)\| - \Delta(\rho_0 + 2r), \end{aligned} \tag{27}$$

which follows by the reverse triangle inequality and since $\nabla R_\alpha(\theta)$ is J_r-Lipschitz in α^{-1}. Further, we have that

$$0 < I_{\alpha_0,\epsilon_0,r}(\theta_0) - J_r(\alpha_0^{-1} - \alpha^{-1}) \leq \|\nabla R_\alpha(\theta)\|, \tag{28}$$

which follows by the reverse triangle inequality, by the fact that $\nabla R_\alpha(\theta)$ is J_r-Lipschitz in α^{-1}, and the definition of α in (21) since $\alpha < \alpha_0^2 I_{\alpha_0,\epsilon_0,r}(\theta_0) J_r^{-1} + \alpha_0$ and $\alpha_0^2 \leq \alpha\alpha_0$. Thus, returning to (27), we let $\Gamma = \Delta(\rho_0 + 2r)$ and $I = I_{\alpha_0,\epsilon_0,r}(\theta_0)$ for ease of notation and we have

$$\begin{aligned} \rho_0 \|\nabla R_\alpha(\theta)\| - \Gamma &= \|\nabla R_\alpha(\theta)\| \left(\rho_0 - \frac{\Gamma}{\|\nabla R_\alpha(\theta)\|}\right) \\ &\geq \|\nabla R_\alpha(\theta)\| \left(\rho_0 - \frac{(\rho_0 + 2r)J_r}{I(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_r}\right) \end{aligned} \tag{29}$$

where we used the inequality in (28). Since we assume that

$$0 \leq \alpha - \alpha_0 < \frac{\alpha_0^2 I}{2J_r\left(1 + r\rho_0^{-1}\right)}, \tag{30}$$

returning to (29), it can be shown using $\alpha_0^2 \leq \alpha\alpha_0$ that

$$\frac{(1 + 2r\rho_0^{-1})J_r}{I(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_r} < 1. \tag{31}$$

Therefore, we finally obtain that

$$\langle -\nabla R_\alpha(\theta), \theta_0 - \theta \rangle \geq \rho \|\nabla R_\alpha(\theta)\|, \tag{32}$$

where $\rho > 0$ is given by

$$\rho = \rho_0 \left(1 - \frac{(1 + 2r\rho_0^{-1})J_r}{I(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_r}\right) \tag{33}$$

as desired. □

Combining Corollary 1 and Theorem 2, we provide the following corollary which quantifies the evolution of SLQC constants for $\alpha_0 = 1$ as α increases.

Corollary 2. *Let $\Sigma > 0$, $\alpha_0 = 1$, $\epsilon_0 > 0$, and $\theta_0 \in \mathbb{B}_d(r)$. If*

$$0 \leq \alpha - 1 < \frac{I_{\alpha_0,\epsilon_0,r}(\theta_0)}{2J_r\left(1 + r\frac{\sigma(r)}{\epsilon_0}\right)}, \tag{34}$$

then R_α is $(\epsilon, \kappa, \theta_0)$-SLQC in $\theta \in \mathbb{B}_d(r)$ with

$$\epsilon = \epsilon_0 + 2L_r(\alpha - 1), \tag{35}$$

and

$$\frac{\epsilon}{\kappa} = \frac{\epsilon_0}{\sigma(r)}\left(1 - \frac{\left(1 + 2r\frac{\sigma(r)}{\epsilon_0}\right)J_r(\alpha - 1)}{\alpha I_{\alpha_0,\epsilon_0,r}(\theta_0) - J_r(\alpha - 1)}\right). \tag{36}$$

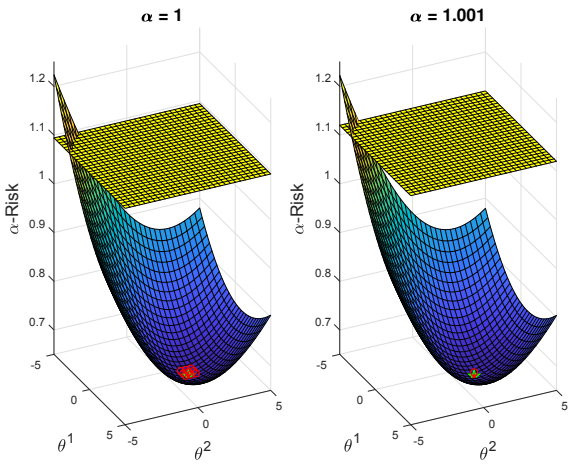

Figure 2. The landscape of α-loss, R_α ($\alpha = 1, 1.001$) in the logistic model, where the features are normalized and $r = 5$, for a 2D-GMM with $\mathbb{P}[Y = 1] = \mathbb{P}[Y = -1]$, $\mu_{X|Y=-1} = [0.4, 0.4]$, $\mu_{X|Y=1} = [1, 1]$, $\Sigma = [3, 0.2; 0.2, 1.5]$. For $\alpha = 1$, the red region depicts ϵ_0/κ_0 which is calculated using Corollary 1 about θ_0, where θ_0 is set to be the global minimum of R_1 and is depicted by the star; for illustrative purposes, we set $\epsilon_0 = 0.4$ and it is depicted by the yellow plane. For $\alpha = 1.001$, the red region depicts ϵ/κ about θ_0 (the star) and ϵ is also depicted by the yellow plane; both quantities approximate the bounds given by Corollary 2.

An illustration of the degradation of SLQC constants as specified by Corollary 2 for increasing α is presented in Figure 2 for a two-dimensional GMM. Intuitively, we find that for a fixed $\theta_0 \in \mathbb{B}_d(r)$, increasing α is equivalent to reducing the radius of the ϵ/κ ball about θ_0 and increasing the value of ϵ. Both of these effects hinder the optimization process and increase the required number of iterations of NGD as stated in Proposition 2.

While the learning practitioner would ultimately like to approximate the intractable 0-1 loss (approximated by $\alpha = \infty$), the bounds presented in Theorem 2 suggest that the computational complexity of NGD quickly worsens as α increases. Fortunately, in the logistic model, α-loss exhibits a saturation effect whereby smaller values of α resemble the landscape induced by $\alpha = \infty$. More concretely, the saturation effect of α-loss is the fact that the uniform distance between R_α and R_∞ decreases geometrically in α as summarized by the following lemma.

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

$$|R_\alpha(\theta) - R_{\alpha'}(\theta)| \leq L_r\left|\frac{1}{\alpha} - \frac{1}{\alpha'}\right|, \tag{37}$$

where L_r is given in (18).

See Figure 3 for an illustration which depicts how quickly the landscape for $\alpha > 1$ resembles the $\alpha = \infty$ landscape.

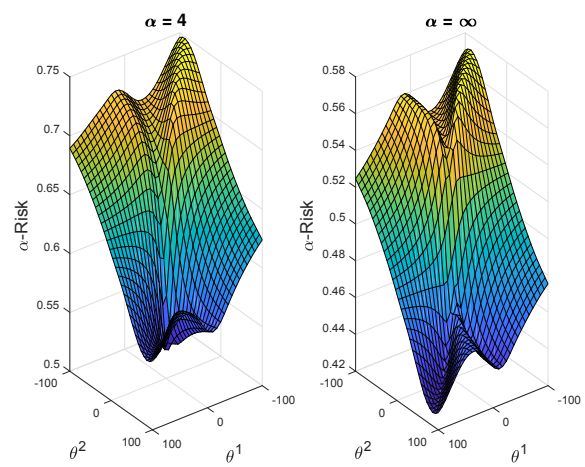

Figure 3. An illustration of the saturation effect of α-loss (R_α for $\alpha = 4, \infty$) in the logistic model, where features are normalized and $r = 100$, for a 2D-GMM with $\mathbb{P}[Y = -1] = 0.61 = 1 - \mathbb{P}[Y = 1]$, $\mu_{X|Y=-1} = [-0.14, 0.21]$, $\mu_{X|Y=1} = [0.06, 0.43]$, $\Sigma_{-1} = [0.38, 0.25; 0.25, 3.17]$, and $\Sigma_1 = [2.07, -1.62; -1.62, 1.97]$.

IV. Concluding Remarks

In this work, we analyze the evolution of the α-loss landscape in the logistic model by examining different regimes of α. As α approaches zero, R_α becomes more strongly convex (see Theorem 1), but the computational complexity of NGD increases since the Lipschitz constant of R_α grows. As α approaches infinity, R_α becomes more non-convex since SLQC parameters degrade (see Theorem 2), which also increases the computational complexity of NGD; however, accuracy increases since the landscape of α-loss tends towards that of the ∞-loss, *i.e.*, the 0-1 loss. Combining Corollary 1 and Theorem 2, we provide explicit bounds to quantify the evolution of SLQC parameters of R_α in the logistic model for α in a neighborhood of 1 (see Corollary 2). Using a moderately large α, α-loss leads to similar performance as the desired, and computationally harder to optimize, 0-1 loss (see Lemma 2); this is a saturation effect of α-loss in the logistic model. Therefore, for the logistic model, we ultimately posit that there is a narrow range of α useful to the practitioner.

REFERENCES

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

[2] H. Masnadi-Shirazi and N. Vasconcelos, "On the design of loss functions for classification: theory, robustness to outliers, and SavageBoost," in *Advances in Neural Information Processing Systems*, 2009, pp. 1049–1056.

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

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

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

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

[7] A. Singh and J. C. Principe, "A loss function for classification based on a robust similarity metric," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–6.

[8] A. Tewari and P. L. Bartlett, "On the consistency of multiclass classification methods," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 1007–1025, 2007.

[9] L. Zhao, M. Mammadov, and J. Yearwood, "From convex to nonconvex: a loss function analysis for binary classification," in *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 2010, pp. 1281–1288.

[10] J. T. Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4331–4339.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

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

[13] T. Sypherd, M. Diaz, L. Sankar, and P. Kairouz, "A tunable loss function for binary classification," in *2019 IEEE International Symposium on Information Theory (ISIT)*, July 2019, pp. 2479–2483.

[14] T. Sypherd, M. Diaz, H. Laddha, L. Sankar, P. Kairouz, and G. Dasarathy, "A class of parameterized loss functions for classification: Optimization tradeoffs and robustness characteristics," *CoRR*, vol. abs/1906.02314, 2019. [Online]. Available: http://arxiv.org/abs/1906.02314

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.

[16] E. Hazan, K. Levy, and S. Shalev-Shwartz, "Beyond convexity: Stochastic quasi-convex optimization," in *Advances in Neural Information Processing Systems*, 2015, pp. 1594–1602.

[17] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Advances in Neural Information Processing Systems*, 2018, pp. 6389–6399.

[18] Q. Nguyen and M. Hein, "The loss surface of deep and wide neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2603–2612.

[19] H. Fu, Y. Chi, and Y. Liang, "Guaranteed recovery of one-hidden-layer neural networks via cross entropy," *arXiv preprint arXiv:1802.06463*, 2018.

[20] S. Liang, R. Sun, Y. Li, and R. Srikant, "Understanding the loss surface of neural networks for binary classification," *arXiv preprint arXiv:1803.00909*, 2018.

[21] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry, "Exploring the landscape of spatial robustness," in *International Conference on Machine Learning*, 2019, pp. 1802–1811.

[22] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier, "Deep relaxation: Partial differential equations for optimizing deep neural networks," *Research in the Mathematical Sciences*, vol. 5, no. 3, p. 30, 2018.

[23] Y. E. Nesterov, "Minimization methods for nonsmooth convex and quasiconvex functions," *Matekon*, vol. 29, pp. 519–531, 1984.

[24] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[25] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

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

[27] S. Arimoto, "Information measures and capacity of order α for discrete memoryless channels," *Topics in Information Theory*, 1977.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[29] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge university press, 2012.

[30] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. Tata McGraw-Hill Education, 2002.