A Tunable Loss Function for Robust, Rigorous, and Reliable Machine Learning

by

Tyler Sypherd

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2022 by the
Graduate Supervisory Committee:

Lalitha Sankar, Chair
Visar Berisha
Gautam Dasarathy
Oliver Kosut

ARIZONA STATE UNIVERSITY

December 2022

ABSTRACT

In the era of big data, more and more decisions and recommendations are being made by machine learning (ML) systems and algorithms. Despite their many successes, there have been notable deficiencies in the robustness, rigor, and reliability of these ML systems, which have had detrimental societal impacts. In the next generation of ML, these significant challenges must be addressed through careful algorithmic design, and it is crucial that practitioners and meta-algorithms have the necessary tools to construct ML models that align with human values and interests.

In an effort to help address these problems, this dissertation studies a tunable loss function called α-loss for the ML setting of classification. The α-loss is a hyperparameterized loss function originating from information theory that continuously interpolates between the exponential ($\alpha = 1/2$), log ($\alpha = 1$), and 0-1 ($\alpha = \infty$) losses, hence providing a holistic perspective of several classical loss functions in ML. Furthermore, the α-loss exhibits unique operating characteristics depending on the value (and different regimes) of α; notably, for $\alpha > 1$, α-loss robustly trains models when noisy training data is present. Thus, the α-loss can provide robustness to ML systems for classification tasks, and this has bearing in many applications, e.g., social media, finance, academia, and medicine; indeed, results are presented where α-loss produces more robust logistic regression models for COVID-19 survey data with gains over state of the art algorithmic approaches.

DEDICATION

To my wife, K. Thank you for your constant love throughout this journey.

ACKNOWLEDGMENTS

I would like to express my sincere thanks and gratitude to my advisor, Professor Lalitha Sankar, for her guidance, inspiration, and support throughout the course of this research, and indeed my whole academic studies at ASU over the last eight years. I first met Professor Sankar in fall 2014 when she was my instructor for EEE 202 Circuits 1. She helped mentor me during my undergraduate studies, and she took a chance on me as one of her PhD students. This dissertation is the results of our research collaboration over the last six years.

I would like to sincerely thank Professor Mario Diaz (UNAM) for his mentorship and friendship, particularly during the first few years of my PhD studies, when he was a postdoc in Professor Sankar's lab. I would also like to thank Dr. Peter Kairouz (Google), Dr. Gowtham Kurri (ASU postdoc), and Dr. Richard Nock (Google) for their research collaborations; I learned a great deal from each of them. I am also grateful to my committee members, Professors Visar Berisha, Gautam Dasarathy, and Oliver Kosut for their insightful questions and comments, and I am also particularly thankful for the research collaborations with Professors Berisha and Dasarathy.

I would also like to thank and acknowledge all of the fellow students I met and collaborated with during my PhD studies: Dr. Chong Huang, Dr. Jiachun Liao, Corbin Ott, Harshit Laddha, John Kevin Cava, Monica Welfert, Nathan Stromberg, Gokulan Vikash Babu, and Kyle Otstot.

In addition, I would like to express my gratitude to the Graduate College at ASU, the ARCS organization, and the Wyandotte Native American Nation for their financial contributions to my graduate studies. Finally, I would like to give special thanks to my family and friends for their encouragements and support.

TABLE OF CONTENTS

Page

APPENDIX

LIST OF TABLES

LIST OF FIGURES

PREFACE

The work presented in this dissertation is aggregated from publications in leading venues: the “International Symposium on Information Theory”, the “Transactions on Information Theory”, the “International Conference on Machine Learning”, and the “Information Theory Workshop”.

Chapter 1

INTRODUCTION

1.1 Background

In the era of big data, more and more decisions and recommendations are being made by machine learning (ML) systems and algorithms. Despite their many successes, there have been notable deficiencies in the robustness, rigor, and reliability of these ML systems, which have had detrimental societal impacts. In the next generation of ML, these significant challenges must be addressed through careful algorithmic design, and it is crucial that practitioners and meta-algorithms have the necessary tools to construct ML models that align with our values and interests.

Within the supervised ML paradigm, the loss function plays a crucial role. As an algorithm learns on training data, the loss function gives feedback to the algorithm, allowing it to adjust and better fit the training data. For classification tasks, practitioners typically employ the log/logistic/cross-entropy loss, as it has good statistical characteristics and guarantees, and is easy to implement. However, as massive amounts of training data continue to be aggregated and labeled, it is inevitable that noise appears in the data; furthermore, the data accumulates human biases, which have implications for fairness in ML. Indeed, the usual log/logistic/cross-entropy loss choice is known to suffer in such situations, misleading the learning algorithm to perturbed and biased models.

Therefore, in an effort to help address these problems, in this dissertation we propose a tunable loss function called α-loss for the ML setting of classification. The α-loss is a hyperparameterized loss function originating from information theory that continuously interpolates between the exponential ($\alpha = 1/2$), log ($\alpha = 1$), and 0-1 ($\alpha = \infty$) losses, hence providing a holistic perspective of several classical loss functions in ML. Furthermore, the α-loss exhibits unique operating characteristics depending on the value (and different regimes) of α; notably, for $\alpha > 1$, α-loss is able to produce more robust models than state of the art for important tasks involving noisy data.

1.2 Outline of Thesis

The following Dissertation is organized in four main chapters, as follows:

- Chapter 2 presents the foundations of the α-loss.

  In Section 2.2, we articulate the information-theoretic motivations of α-loss; in Section 2.3 for the setting of binary classification, we show that α-loss is classification-calibrated for all $\alpha \in (0, \infty]$; in Section 2.4 for the logistic model, we provide results regarding the optimization landscape as a function of α, particularly noticing that convexity decreases as $\alpha > 1$ increases; in Section 2.5, we provide Rademacher complexity generalization bounds for all $\alpha \in (0, \infty]$ and asymptotic optimality results; finally in Section 2.6, we provide experiments for logistic regression and convolutional neural networks, observing sensitivity to class imbalances for $\alpha < 1$ and robustness to noisy labels for $\alpha > 1$.

- Chapter 3 presents a statistical theory of robustness for loss functions in the class probability estimation setting, proving that α-loss satisfies a notion of robustness called *twist-properness.*

  In Section 3.4 we present the notion of twist-properness for the setting of class probability estimation; we show that α-loss is twist-proper and a fixed $\alpha_0 > 1$ is statistically more robust than the log-loss ($\alpha = 1$) for symmetric label noise. In Section 3.5 we "properly" boost α-loss with PILBoost (pseudo-inverse-link), a convex boosting algorithm that has robustness characteristics that are shown experimentally in Section 4.5.

- Chapter 4 studies the robustness of α-loss for simple models, e.g., logistic regression and boosting with low maximum depth trees.

  In Section 4.3.1, we present a novel boosting algorithm, AdaBoost.α, that smoothly tunes through classical boosting algorithms (vanilla AdaBoost for $\alpha = 1/2$ and LogAdaBoost for $\alpha = 1$) to non-convex boosters ($\alpha > 1$). In Section 4.3.2, we show that AdaBoost.α is provably robust for $\alpha > 1$ on the hard Long-Servedio dataset, which pathologically defeats any convex booster; we support this theory with experiments in Section 4.5.1. In Section 4.4 we indicate robustness for α-loss in the logistic model, with upper and lower bounds showing that the noisy gradient is smaller for $\alpha > 1$ than when $\alpha \leq 1$; this theory is supported in Section 4.5.2 with an application on a COVID-19 dataset, indicating the real-world efficacy of α-loss for robustness tasks.

- Lastly, Chapter 5 presents a novel framework for generative adversarial networks (GAN) via α-loss, called α-GAN.

  In Section 5.2 we present the theoretical foundations for the α-GAN, showing how it recovers the vanilla GAN for $\alpha = 1$. In Section 5.3 we present the connections between the α-GAN and the classical Arimoto divergence, arguing for the efficacy of α-GAN for certain tasks.

Chapter 2

FOUNDATIONS: CALIBRATION, LANDSCAPE, AND GENERALIZATION

2.1 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 (Friedman *et al.*, 2001; Shalev-Shwartz and Ben-David, 2014). 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 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 (Ben-David *et al.*, 2003). 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 (Bartlett *et al.*, 2006b; Masnadi-Shirazi and Vasconcelos, 2009; Lin, 2004; Nguyen *et al.*, 2009a; Rosasco *et al.*, 2004; Nguyen and Sanner, 2013; Singh and Principe, 2010; Tewari and Bartlett, 2007; Zhao *et al.*, 2010; Barron, 2019; Lin *et al.*, 2017b).

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 (Liao *et al.*, 2018a, 2019, 2020). 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.* (2018a) derived a new privacy measure called α-*leakage* which continuously interpolates between Shannon's mutual information (Shannon, 2001) and maximal leakage introduced by Issa *et al.* (2019); indeed, Liao *et al.* showed that α-leakage is equivalent to the Arimoto mutual information (Verdú, 2015). 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.

2.1.1 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 (Rényi, 1961), and Arimoto extended the Rényi entropy to conditional distributions (Arimoto, 1971a). This led to the α-mutual information (Verdú, 2015; Sason and Verdú, 2017), which is directly related to a recently introduced privacy measure called α-leakage (Liao *et al.*, 2018a). More recently in networking, Mo and Walrand (2000) introduced α-fairness, 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 (Li *et al.*, 2019). Even more recently in machine learning, Barron (2019) 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 chapter. 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) (Janocha and Czarnecki, 2016). 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.* (2006b) proposed the notion of *classification-calibration* in a seminal paper. *Classification-calibration* is analogous to point-wise 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 (Nock and Menon, 2020; Walder and Nock, 2020; Reid and Williamson, 2010a). *Properness* of a loss function is a necessary condition for efficacy in the class probability estimation setting (see, e.g., (Reid and Williamson, 2010a)), but for the classification setting which is the focus of this chapter, 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 (Bartlett *et al.*, 2006b; Rosasco *et al.*, 2004; Nguyen *et al.*, 2009a; Lin, 2004), more recent works propose the use of non-convex losses as a means to moderate the behavior of an algorithm (Mei *et al.*, 2018; Nguyen and Sanner, 2013; Masnadi-Shirazi and Vasconcelos, 2009; Barron, 2019). This is due to the increased robustness non-convex losses offer over convex losses (Long and Servedio, 2010; Mei *et al.*, 2018; Barron, 2019) and the fact that modern learning models (e.g., deep learning) are inherently non-convex as they involve vast functional compositions (Goodfellow *et al.*, 2016). 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 (Mei *et al.*, 2018; Hazan *et al.*, 2015; Li *et al.*, 2018b; Nguyen and Hein, 2017; Fu *et al.*, 2020; Liang *et al.*, 2018; Engstrom *et al.*, 2019; Chaudhari *et al.*, 2018). To this end, Hazan *et al.* (2015) 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 (Nesterov, 1984)) 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 (Benigni and Péché, 2019; Xiao *et al.*, 2018; Pennington and Worah, 2017) 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 (Shalev-Shwartz and Ben-David, 2014, Ch. 26). A recent line of research, initiated by the works of Russo and Zou (2020) and Xu and Raginsky (2017), 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., (Lopez and Jog, 2018; Wang *et al.*, 2019a; Bu *et al.*, 2020; Steinke and Zakynthinou, 2020; Esposito *et al.*, 2021; Gálvez *et al.*, 2021; Neu *et al.*, 2021), these results are inherently algorithm dependent which makes them hard to instantiate and obfuscates the role of the loss function. Hence, in this chapter 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 (Wang *et al.*, 2019b; Amid *et al.*, 2019a; Nguyen and Sanner, 2013; Li *et al.*, 2021). Notably, the symmetric cross entropy loss introduced by Wang *et al.* (2019b) 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.* (2019a) 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.* (2021) introduced tilted empirical risk minimization, a framework which parametrically generalizes empirical risk minimization using a log-exponential transformation to induce fairness or robustness in the model. Contrasting with this chapter, we note that our study is exclusively focused on α-loss acting within empirical risk minimization. Summing up, the main distinctions that differentiate this chapter 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.

2.1.2 Contributions

The following are the main contributions presented in this chapter:

- 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.* (2018a) 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.

2.2 Information-Theoretic Motivations

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

$$\mathbb{E}_{X,Y} \left[\ell(Y, P_{\hat{Y}|X}) \right], \tag{2.1}$$

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. Liao *et al.* (2018a) proposed the definition of α-loss for $\alpha \in [1, \infty]$ in order to quantify adversarial

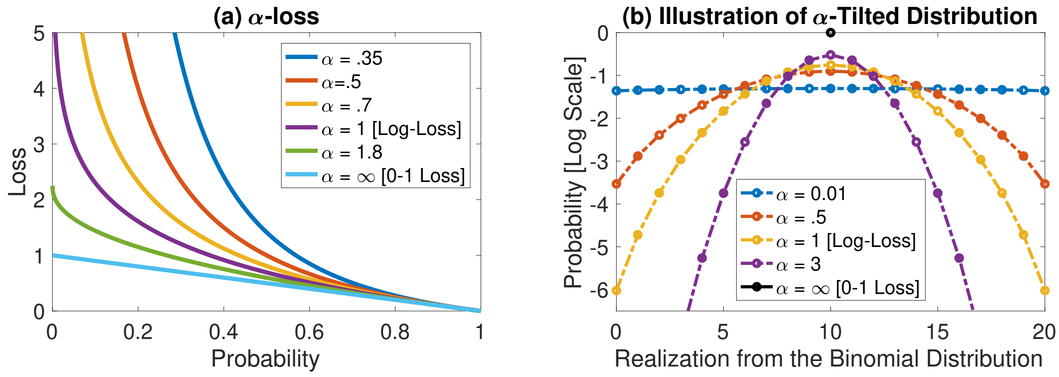


Figure 2.1: (a) α-loss (2.2) as a Function of the Probability for Several Values of α; (b) α-tilted Posterior (2.6) for Several Values of α Where the True Underlying Distribution Is the (20,.5)-binomial Distribution.

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*

$$l^\alpha(y, \hat{P}) := \frac{\alpha}{\alpha - 1}\left(1 - \hat{P}(y)^{1-1/\alpha}\right), \tag{2.2}$$

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[1] 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,

$$\mathbb{E}_{X,Y}\left[l^1(Y, P_{\hat{Y}|X})\right] = \mathbb{E}_X\left[H(P_{Y|X=x}, P_{\hat{Y}|X=x})\right], \tag{2.3}$$

[1] The continuity of α-loss in α has nontrivial importance in later robustness arguments.

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

$$\mathbb{E}_{X,Y}[l^{\infty}(Y, P_{\hat{Y}|X})] = \mathbb{P}[Y \neq \hat{Y}], \tag{2.4}$$

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 (Shalev-Shwartz and Ben-David, 2014). 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. 2.1(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.* (2018a) 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*

$$\min_{P_{\hat{Y}|X}} \mathbb{E}_{X,Y}\left[l^{\alpha}(Y, P_{\hat{Y}|X})\right] = \frac{\alpha}{\alpha-1}\left(1 - e^{\frac{1-\alpha}{\alpha}H_{\alpha}^{A}(Y|X)}\right), \tag{2.5}$$

where $H_{\alpha}^{A}(Y|X) := \dfrac{\alpha}{1-\alpha}\log\sum\limits_{x}\Big(\sum\limits_{y} P_{X,Y}(x,y)^{\alpha}\Big)^{1/\alpha}$ is the Arimoto conditional entropy of order α (Arimoto, 1977). The resulting unique minimizer, $\hat{P}_{\alpha}^{*}$, is the α-tilted true posterior

$$\hat{P}_{\alpha}^{*}(y|x) = \frac{P_{Y|X}(y|x)^{\alpha}}{\sum\limits_{y} P_{Y|X}(y|x)^{\alpha}}. \tag{2.6}$$

The proof of Proposition 1 for $\alpha \in [1, \infty]$ can be found in (Liao *et al.*, 2018a) 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 (2.5) given by the α-tilted distribution in (2.6) 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. 2.1(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 2.6 for experimental consideration of such class imbalance and noisy label trade-offs in logistic regression and convolutional neural networks.

Through Proposition 2.5, we observe that the minimization of α-loss recovers the Arimoto entropy (Arimoto, 1971b). As we will see in Chapter 5, utilizing α-loss in Generative Adversarial Networks also recovers Arimoto divergences (Vajda, 2009). Hence, α-loss is intimately related to Arimoto-type entropies, informations, and divergences. With the information-theoretic motivations of α-loss behind us, we now consider the setting of binary classification, where we study the statistical and robustness properties of α-loss.

2.3 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 $h : \mathcal{X} \rightarrow \{-1, +1\}$ that minimizes the probability of error, or the risk (expectation) of the 0-1 loss, given by

$$R(h) = \mathbb{P}[h(X) \neq Y], \tag{2.7}$$

where the true 0-1 loss given by $\mathbb{1}[h(X) \neq Y]$. Unfortunately, this optimization problem is NP-hard (Ben-David *et al.*, 2003). 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 $f : \mathcal{X} \rightarrow \overline{\mathbb{R}}$, $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, whose output captures the certainty of a model's prediction of the true underlying binary label $Y \in \{-1, 1\}$ associated with X (Bartlett *et al.*, 2006b; Lin, 2004; Nguyen *et al.*, 2009a; Masnadi-Shirazi and Vasconcelos, 2009; Sypherd *et al.*, 2019; Schapire and Freund, 2013; Shalev-Shwartz and Ben-David, 2014; Friedman *et al.*, 2001). 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 (Friedman *et al.*, 2001). 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.

2.3.1 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 $\tilde{l} : \overline{\mathbb{R}} \to \mathbb{R}_+$ (Bartlett *et al.*, 2006b; Lin, 2004; Masnadi-Shirazi and Vasconcelos, 2009; Nguyen *et al.*, 2009a; Janocha and Czarnecki, 2016). In this case, the loss of the pair $(y, f(x))$ only depends on the product $z := yf(x)$, the (unnormalized) margin (Schapire and Freund, 2013). 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. 2.2(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*

$$\tilde{l}^\alpha(z) := \frac{\alpha}{\alpha - 1}\left(1 - \left(1 + e^{-z}\right)^{1/\alpha - 1}\right), \tag{2.8}$$

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 α-loss; this is elaborated in the sequel. First, we discuss its relationship to the original

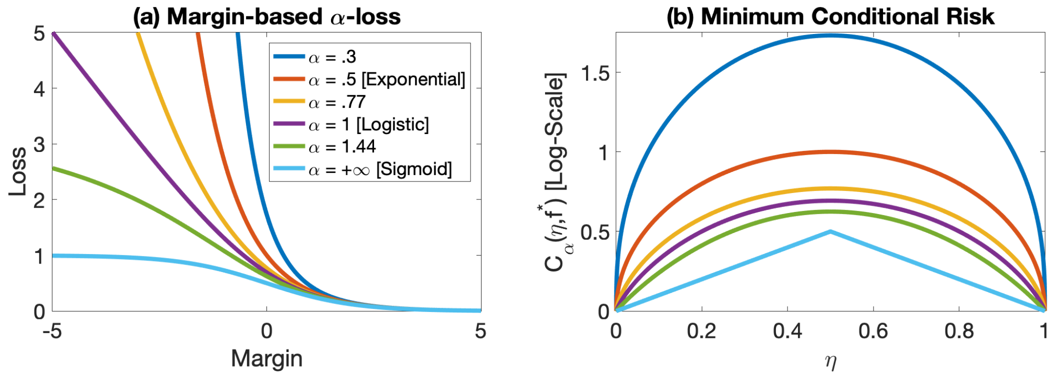


Figure 2.2: (a) Margin-based α-loss (4.3) as a Function of the Margin ($z := yf(x)$) for $\alpha \in \{.3, .5, .77, 1, 1.44, \infty\}$; (b) minimum Conditional Risk (2.14) for the Same Values of α.

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 $g : \mathcal{X} \to [0, 1]$ which encode the conditional distribution, namely, $g(x) := P_{\hat{Y}|X}(1|x)$. In essence, soft classifiers capture a model's *belief* of $Y|X$ (Shalev-Shwartz and Ben-David, 2014; Goodfellow *et al.*, 2016; Sypherd *et al.*, 2019). 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 $\sigma : \overline{\mathbb{R}} \to [0, 1]$ given by

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \tag{2.9}$$

which is generalized by the softmax function in the multiclass setting (Goodfellow *et al.*, 2016). Observe that σ is invertible and $\sigma^{-1} : [0, 1] \to \overline{\mathbb{R}}$ is given by

$$\sigma^{-1}(z) = \log\left(\frac{z}{1 - z}\right), \tag{2.10}$$

which is often referred to as the logistic link (Reid and Williamson, 2010a).

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 (Sypherd *et al.*, 2019).

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

$$l^{\alpha}(y, g(x)) = \tilde{l}^{\alpha}(yf(x)). \tag{2.11}$$

Conversely, if f is a classification function, then the soft classifier $g(x) := \sigma(f(x))$ *satisfies* (2.11)*. In particular, for every* $\alpha \in (0, \infty]$*,*

$$\min_{g} \mathbb{E}_{X,Y}(l^{\alpha}(Y, g(x))) = \min_{f} \mathbb{E}_{X,Y}(\tilde{l}^{\alpha}(Yf(X))). \tag{2.12}$$

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* (4.1)*, 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* (4.1) *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, $\tilde{l}^{\alpha} : \overline{\mathbb{R}} \to \mathbb{R}_+$ is convex for $\alpha \leq 1$ and quasi-convex for $\alpha > 1$.*

Recall that a real-valued function $f : \mathbb{R} \to \mathbb{R}$ 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., (Boyd and Vandenberghe, 2004a)). Intuitively through Fig. 2.2(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.

2.3.2 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.* (2006b) 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 $\phi : \overline{\mathbb{R}} \to \mathbb{R}_+$ 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 $f : \mathcal{X} \to \overline{\mathbb{R}}$ is a classification function. Thus, the conditional risk of the margin-based α-loss for $\alpha \in (0, \infty]$ is given by

$$C_\alpha(\eta(x), f(x)) = \mathbb{E}_Y[\tilde{l}^{\alpha}(Yf(X))|X = x]. \tag{2.13}$$

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

$$\inf_{f:\mathcal{X}\to\mathbb{R}} C_\phi(\eta(x), f(x)) = \inf_{f:\mathcal{X}\to\mathbb{R}} \eta(x)\phi(f(x)) + (1-\eta(x))\phi(-f(x)), \tag{2.14}$$

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

$$\operatorname{sign}(f^*(x)) = \operatorname{sign}(2\eta(x) - 1). \tag{2.15}$$

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 (Lin, 2004; Bartlett *et al.*, 2006b).

The expectation of the loss function ϕ, or the ϕ-risk, is denoted

$$R_\phi(f) = \mathbb{E}_X[C_\phi(\eta(X), f(X))], \tag{2.16}$$

and this notation will be useful in the sequel when we quantify the asymptotic behavior of α-loss. Finally, as is common in the literature (Masnadi-Shirazi and Vasconcelos, 2009; Bartlett *et al.*, 2006b), 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*

$$f_\alpha^*(\eta) = \alpha \cdot \sigma^{-1}(\eta) = \alpha \log\left(\frac{\eta}{1-\eta}\right). \tag{2.17}$$

See Appendix A.1 for full proof details. Examining the optimal classification function in (2.17) 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 (2.6), the optimal classification function in (2.17) 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 (2.17) into (2.13), 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*

$$\begin{cases} \frac{\alpha}{\alpha-1}\left(1 - (\eta^\alpha + (1-\eta)^\alpha)^{1/\alpha}\right) & \alpha \in (0,1) \cup (1, +\infty), \\ -\eta \log \eta - (1-\eta)\log(1-\eta) & \alpha = 1, \\ \min\{\eta, 1-\eta\} & \alpha \to +\infty. \end{cases} \tag{2.18}$$

Remark 2. *Observe that in* (2.18) *for $\alpha = 1$, the minimum conditional risk can be rewritten as*

$$C^*_1(\eta) = -\eta \log \eta - (1-\eta)\log(1-\eta) \tag{2.19}$$

$$= H(Y|X=x), \tag{2.20}$$

where $H(Y|X=x)$ is the Shannon conditional entropy for a Y given $X = x$ (Thomas and Joy, 2006). For $\alpha \in (0,1) \cup (1, +\infty)$, also note that in (2.18)*, the minimum conditional risk can be rewritten as*

$$C^*_\alpha(\eta) = \frac{\alpha}{\alpha-1}\left[1 - (\eta^\alpha + (1-\eta)^\alpha)^{1/\alpha}\right] \tag{2.21}$$

$$= \frac{\alpha}{\alpha-1}\left[1 - e^{\frac{1-\alpha}{\alpha} H^A_\alpha(Y|X=x)}\right], \tag{2.22}$$

where $H^A_\alpha(Y|X=x) = \frac{1}{1-\alpha}\log\left(\sum_y P_{Y|X}(y|x)^\alpha\right)$ is the Arimoto conditional entropy of order α (Arimoto, 1977). Finally, observe that $\mathbb{E}_X[C^*_\alpha(\eta(X))]$ recovers (2.5) *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. 2.2(b)); indeed, this is known to be a useful property for classification problems (Masnadi-Shirazi and Vasconcelos, 2009). 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.

2.3.3 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 (Mei *et al.*, 2018; Nguyen and Sanner, 2013; Barron, 2019; Sypherd *et al.*, 2019; Schapire and Freund, 2013; Wu and Liu, 2007; Chapelle *et al.*, 2009; Long and Servedio, 2010; Masnadi-Shirazi and Vasconcelos, 2009). In essence, non-convex loss functions tend to assign less weight to misclassified training examples[2] and therefore algorithms optimizing such losses are often less perturbed by outliers, i.e., examples which induce large negative margins. More concretely, consider Fig. 2.2(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) (Masnadi-Shirazi and Vasconcelos, 2009; Schapire and Freund, 2013).

We explore these ideas with the following synthetic experiment presented in Fig. 2.3. We assume the practitioner has access to modified training data which approximates the true underlying distribution given by a two-dimensional Gaussian Mixture Model

[2] Convex losses grow at least linearly with respect to the negative margin which results in an increased sensitivity to outliers. See Fig. 2.2(a) for $\alpha = 1$ as an example of this phenomenon.

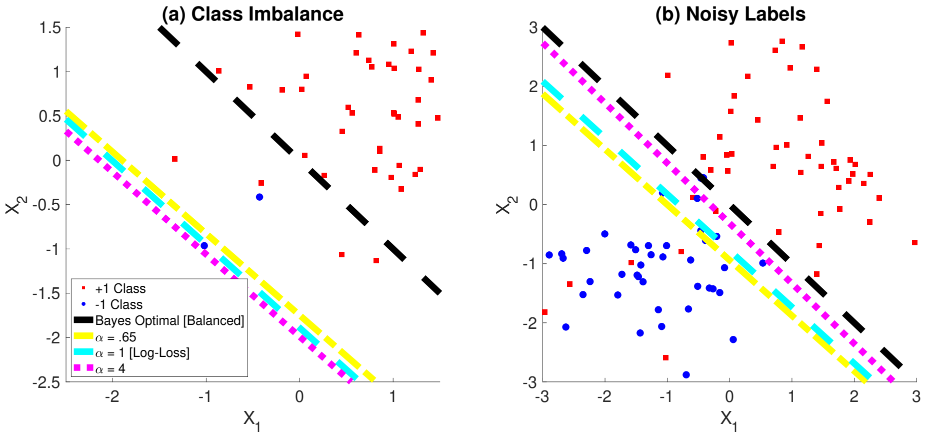


Figure 2.3: Two Synthetic Experiments Each Averaged over 100 Runs Highlighting the Differences in Trained Linear Predictors of α-loss for $\alpha \in \{.65, 1, 4\}$ on Imbalanced and Noisy Data, Which Are Compared with the Bayes Optimal Predictor for the Clean, Balanced Distribution. Training Data Present 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 $\alpha > 1$.

(2D-GMM) with equal mixing probability $\mathbb{P}[Y=-1]=\mathbb{P}[Y=+1]$, symmetric means

$$\mu_{X|Y=-1}=(-1,-1)^{\intercal}=-\mu_{X|Y=1}, \tag{2.23}$$

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. 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. 2.3 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. A.5 in Appendix A.4.2. However, training on noisy or imbalanced data affects the linear predictors of α-loss in different ways. In the class imbalance experiment in Fig. 2.3(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. 2.3(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 A.1, A.2, and A.3 in Appendix A.4.2.

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 2.6; theoretical investigations of the robustness of α-loss can be found in (Sypherd *et al.*, 2021). We now turn our attention to theoretically characterizing the optimization complexity of α-loss for the different regimes of α in the logistic model.

2.4 Optimization Landscape

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 A.2.

2.4.1 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

$$g_\theta(x) = \sigma(\langle \theta, x \rangle), \tag{2.24}$$

with $\sigma : \mathbb{R} \to [0, 1]$ being the sigmoid function given by (4.1). For convenience, we present the following short form of α-loss in the logistic model which is equivalent to the expanded expression in (Sypherd *et al.*, 2019). For $\alpha \in (0, \infty]$, α-loss is given by

$$l^\alpha(y, g_\theta(x)) = \frac{\alpha}{\alpha - 1} \left[1 - g_\theta(yx)^{1 - 1/\alpha} \right]. \tag{2.25}$$

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 7 that (2.25) 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 (2.25),

$$R_\alpha(\theta) := \mathbb{E}_{X,Y}[l^\alpha(Y, g_\theta(X))]. \tag{2.26}$$

The α-risk (2.26) is plotted for several values of α in a two-dimensional Gaussian Mixture Model (GMM) in Fig. 2.4. Further, observe that, for all $\theta \in \mathbb{B}_d(r)$,

$$R_\infty(\theta) := \mathbb{E}_{X,Y}[l^\infty(Y, g_\theta(X))] = \mathbb{P}[Y \neq \hat{Y}_\theta], \tag{2.27}$$

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

$$\sigma(-z) = 1 - \sigma(z) \quad \text{and} \quad \frac{d}{dz}\sigma(z) = \sigma(z)(1 - \sigma(z)). \tag{2.28}$$

Indeed, a straightforward computation shows that

$$\frac{\partial}{\partial \theta^j} l^\alpha(y, g_\theta(x)) = \left[-y g_\theta(yx)^{1-1/\alpha}(1 - g_\theta(yx))\right] x^j, \tag{2.29}$$

where θ^j, x^j denote the j-th components of θ and x, respectively. Thus, the gradient of α-loss in (2.25) is

$$\nabla_\theta l^\alpha(Y, g_\theta(X)) = F_1(\alpha, \theta, X, Y)X, \tag{2.30}$$

where $F_1(\alpha, \theta, x, y)$ is defined as the expression within brackets in (2.29). Another straightforward computation yields

$$\nabla_\theta^2 l^\alpha(Y, g_\theta(X)) = F_2(\alpha, \theta, X, Y)XX^\intercal, \tag{2.31}$$

where F_2 is defined as

$$F_2(\alpha, \theta, x, y) := g_\theta(yx)^{1-1/\alpha} g_\theta(-yx) \left(g_\theta(yx) - \left(1 - \frac{1}{\alpha}\right) g_\theta(-yx)\right). \tag{2.32}$$

2.4.2 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. 2.4 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

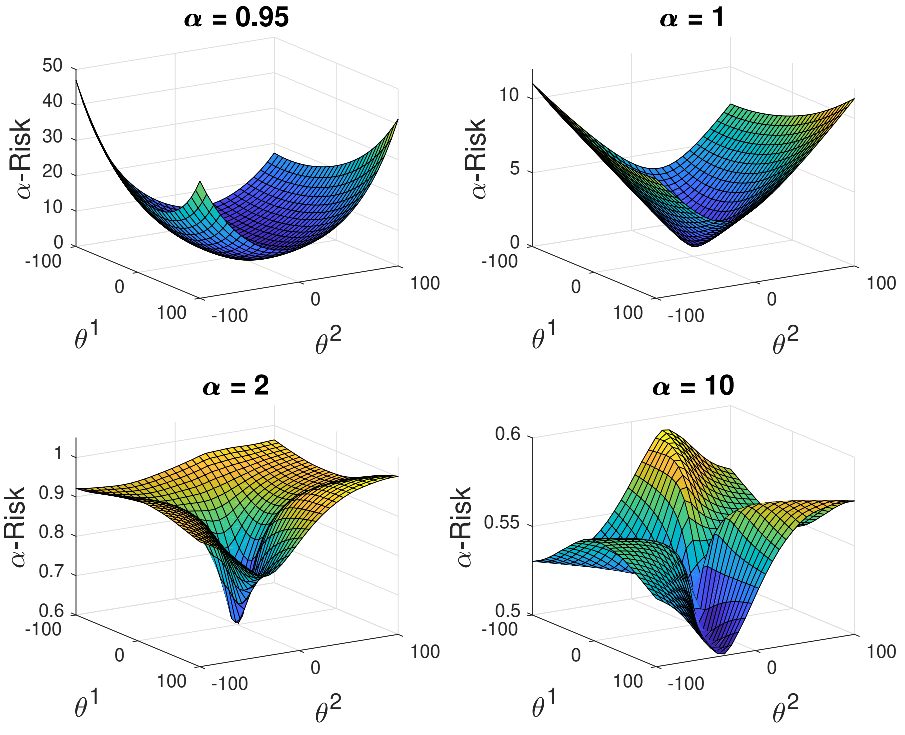


Figure 2.4: The Landscape of α-loss (R_α for $\alpha = .95, 1, 2, 10$) in the Logistic Model, Where Features Are Normalized, for a 2d-GMM with $\mathbb{P}[Y = -1] = .12$, $\mu_{X|y=-1} = (-0.18, 1.49)^\intercal$, $\mu_{X|y=1} = (-0.01, .16)^\intercal$, $\sigma_{-1} = [3.20, -2.02; -2.02, 2.71]$, and $\sigma_1 = [4.19, 1.27; 1.27, .90]$.

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 $m \geq 0$ (Boyd and Vandenberghe, 2004a).

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*

$$\Lambda(\alpha, r\sqrt{d}) := \sigma(r\sqrt{d})^{1-1/\alpha} \left(\sigma'(r\sqrt{d}) - \left(1 - \frac{1}{\alpha}\right) \sigma(-r\sqrt{d})^2 \right). \tag{2.33}$$

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 we have that* $R_\alpha(\theta)$ *is* $\tilde{\Lambda}(\alpha, r\sqrt{d}) \min_{i \in [d]} \lambda_i(\Sigma)$*-strongly convex in* $\theta \in \mathbb{B}_d(r)$ *for* $\alpha \in \left(0, (e^{2r\sqrt{d}} - e^{r\sqrt{d}})^{-1}\right]$, *where*

$$\tilde{\Lambda}(\alpha, r\sqrt{d}) := \sigma(-r\sqrt{d})^{2-1/\alpha}\sigma(r\sqrt{d})\left(1 - e^{r\sqrt{d}} + \frac{e^{-r\sqrt{d}}}{\alpha}\right). \tag{2.34}$$

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. 2.4 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. 2.4 for $\alpha = 2$ more closely, we see that it is reminiscent of a quasi-convex function. Recall that (e.g., Chapter 3.4 in (Boyd and Vandenberghe, 2004a)) 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

$$\langle -\nabla f(\theta), \theta_0 - \theta \rangle \geq 0. \tag{2.35}$$

In other words, the negative gradient of a quasi-convex function always points in the direction of descent. While α-loss (2.25) is quasi-convex for $\alpha > 1$, this does not imply that the α-*risk* (2.26) is quasi-convex for $\alpha > 1$ since the sum of quasi-convex functions is not guaranteed to be quasi-convex (Boyd and Vandenberghe, 2004a). 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.

2.4.3 Strictly Local Quasi-Convexity and its Extensions

We use a framework developed by Hazan *et al.* (2015) 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 (Hazan *et al.*, 2015).

Definition 3 (Definition 3.1, (Hazan *et al.*, 2015))**.** *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* $\theta' \in \mathbb{B}(\theta_0, \epsilon/\kappa)$,

$$\langle -\nabla f(\theta), \theta' - \theta \rangle \geq 0. \tag{2.36}$$

Briefly, in Hazan *et al.* (2015) 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 $\mathbb{B}(\theta_0, \epsilon/\kappa)$ as evidence through (2.35); see Fig. 2.5 for an illustration of the difference between classical quasi-convexity and SLQC in this regime.

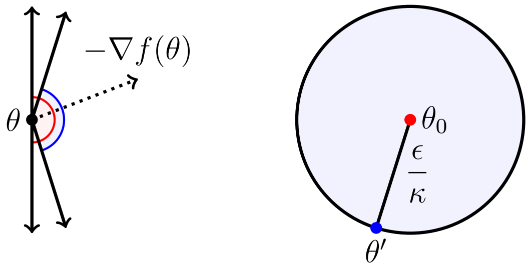


Figure 2.5: An Illustration Highlighting the Difference Between Quasi-convexity as given in (2.35) and the Second Slqc Condition of Definition 3. If f Is Quasi-convex, the Red Angle Describes the Possible Negative Gradients of f at θ with Respect to θ_0. If f Is Slqc, the Blue Angle Describes the Possible Negative Gradients of f at θ with Respect to θ_0 and the given ϵ/κ-radius Ball.

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 A.2.4.

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 $\theta' \in \mathbb{B}_d\left(\theta_0, \rho\right)$,*

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 Hazan *et al.* (2015), they 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 (Shalev-Shwartz and Ben-David, 2014)) except gradients are normalized such that the algorithm applies uniform-size

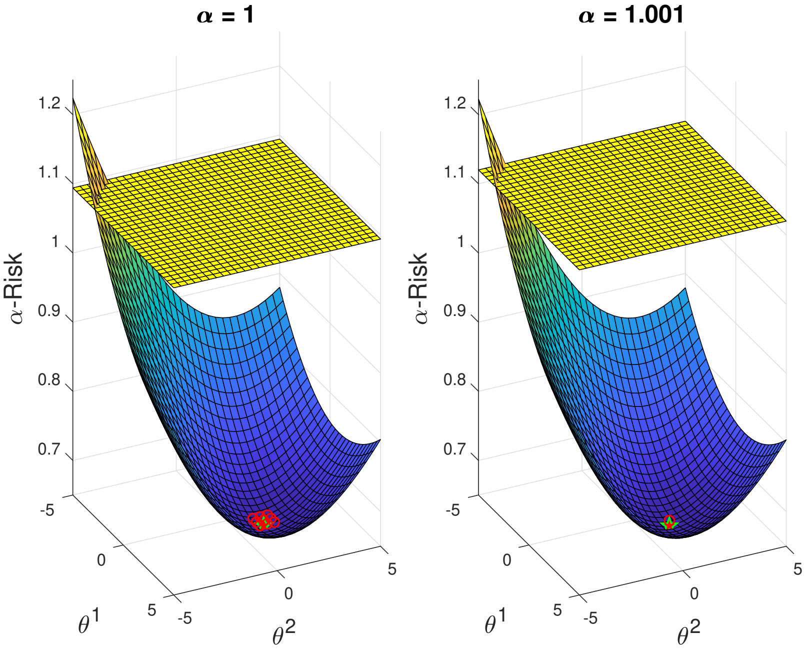


Figure 2.6: The Landscape of α-loss, R_α ($\alpha = 1, 1.001$) in the Logistic Model, Where the Features Are Normalized and $r = 5$, for a 2d-GMM with $\mathbb{P}[Y = 1] = \mathbb{P}[Y = -1]$, $\mu_{X|y=-1} = [.4, .4]$, $\mu_{X|y=1} = [1, 1]$, $\sigma = [3, .2; .2, 1.5]$. For $\alpha = 1$, the Red Region Depicts ϵ_0/κ_0 Which Is Calculated Using Theorem 2 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 $\epsilon_0 = .4$ 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 Theorem 3.

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 A.2.4. Indeed, the convergence guarantees of NGD for SLQC functions are similar to those of Gradient Descent for convex functions.

Proposition 4 (Thm. 4.1, (Hazan *et al.*, 2015))**.** *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 $\min\limits_{t=1,\ldots,T} f(\theta_t) - f(\theta^*) \leq \epsilon$.*

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 $r > 0$).

2.4.4 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 $\alpha \in (0, 1]$,

$$C_d(r, \alpha) := \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha}; \tag{2.37}$$

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

$$C_d(r, \alpha) := \begin{cases} \sqrt{d}\left(\frac{\alpha-1}{2\alpha-1}\right)^{1-1/\alpha}\left(\frac{\alpha}{2\alpha-1}\right) & e^{r\sqrt{d}} \geq \frac{\alpha-1}{\alpha}, \\ \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha} & e^{r\sqrt{d}} < \frac{\alpha-1}{\alpha}. \end{cases} \tag{2.38}$$

Thus, in conjunction with Theorem 2, Corollary 2, and a result by Hazan *et al.* (2015) (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 $r\sqrt{d} < \operatorname{arcsinh}(1/2)$, 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 (2.37) *and* (2.38)*;*

- *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 (2.37), 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. 2.4) 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 (Shalev-Shwartz and Ben-David, 2014). 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.

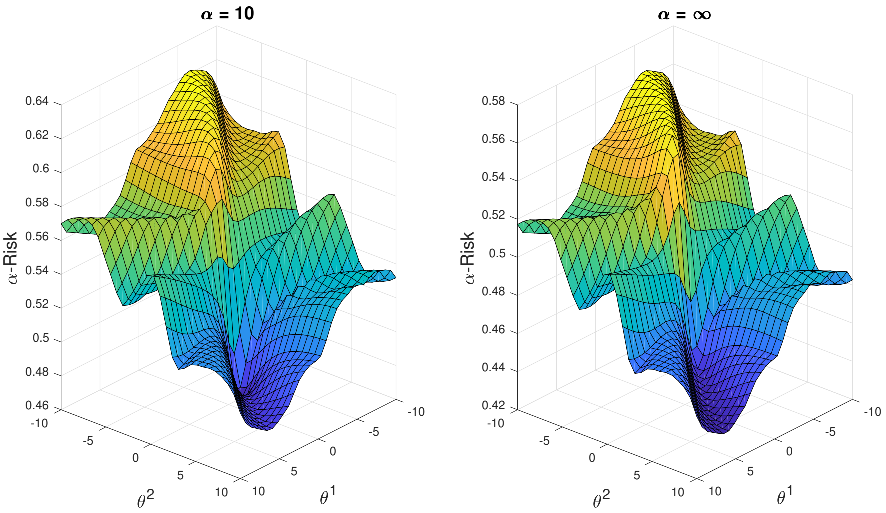


Figure 2.7: An Illustration of the Saturation Phenomenon of α-loss (R_α for $\alpha = 10, \infty$) in the Logistic Model for a 2d-GMM with $\mathbb{P}[Y = 1] = \mathbb{P}[Y = -1]$, $\mu_{X|y=-1} = (-.91, .50)^\intercal$, $\mu_{X|y=1} = (-.27, .20)^\intercal$, $\sigma = [1.38, .55; .55, 2.18]$. Note the Small Difference, Uniformly over the Parameter Space, Between R_{10} and R_∞.

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

$$|R_\alpha(\theta) - R_{\alpha'}(\theta)| \leq L_d(\theta) \left| \frac{\alpha - \alpha'}{\alpha\alpha'} \right|, \tag{2.39a}$$

$$\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\| \leq J_d(\theta) \left| \frac{\alpha - \alpha'}{\alpha\alpha'} \right|, \tag{2.39b}$$

where

$$L_d(\theta) := \frac{\left(\log\left(1 + e^{\|\theta\|\sqrt{d}}\right)\right)^2}{2}, \tag{2.40a}$$

$$J_d(\theta) := \sqrt{d} \log\left(1 + e^{\|\theta\|\sqrt{d}}\right) \sigma(\|\theta\|\sqrt{d}). \tag{2.40b}$$

This result is proved in Appendix A.2.6, 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

$$|R_{10}(\theta) - R_\infty(\theta)| \leq \frac{L_d(\theta)}{10}, \tag{2.41a}$$

$$|\nabla R_{10}(\theta) - \nabla R_\infty(\theta)| \leq \frac{J_d(\theta)}{10}, \tag{2.41b}$$

where $L_d(\theta)$ and $J_d(\theta)$ are both given in (2.40). In words, the pointwise distance between the $\alpha = 10$ landscape and the $\alpha = \infty$ landscape decreases geometrically; for a visual representation see Fig. 2.7.

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*

$$0 \leq \alpha - \alpha_0 < \frac{\alpha_0^2 \|\nabla R_{\alpha_0}(\theta)\|}{2J_d(\theta)\left(1 + r\frac{\kappa_0}{\epsilon_0}\right)}, \tag{2.42}$$

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

$$\epsilon = \epsilon_0 + 2L_d(\theta)\left(\frac{\alpha - \alpha_0}{\alpha\alpha_0}\right), \tag{2.43}$$

$$\frac{\epsilon}{\kappa} = \frac{\epsilon_0}{\kappa_0}\left(1 - \frac{\left(1 + 2r\frac{\kappa_0}{\epsilon_0}\right) J_d(\theta)(\alpha - \alpha_0)}{\alpha\alpha_0\|\nabla R_{\alpha_0}(\theta)\| - J_d(\theta)(\alpha - \alpha_0)}\right). \tag{2.44}$$

The proof of Theorem 3 can be found in Appendix A.2.6. 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 A.2.8.

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*

$$\alpha_\lambda := \alpha_0 + \lambda \frac{\alpha_0^2 g_\theta}{J_d(\theta)\left(1 + 2r\frac{\kappa_0}{\epsilon_0}\right)}, \tag{2.45}$$

$$\epsilon_\lambda := \epsilon_0 + 2\lambda L_d(\theta)\left(\frac{\alpha_\lambda - \alpha_0}{\alpha_\lambda \alpha_0}\right)\frac{\alpha_0^2 g_\theta}{J_d(\theta)\left(1 + r\frac{\kappa_0}{\epsilon_0}\right)}, \tag{2.46}$$

$$\frac{\epsilon_\lambda}{\kappa_\lambda} > \frac{\epsilon_0}{\kappa_0}(1 - \lambda). \tag{2.47}$$

We now provide three different interpretations and comments regarding the previous result. First regarding the SLQC parameters themselves, observe from (2.45) that the bound on α is improved over Theorem 3 as the factor of 2 in the denominator in (2.42) is moved into the parentheses; next, it can be observed (upon plugging in α_λ) that ϵ_λ in (2.46) is linear in λ, which is again an improvement over the first equation in (2.43); finally, note that the bound on $\epsilon_\lambda/\kappa_\lambda$ in (2.46) is vastly more tractable and informative than the second expression in (2.43). 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 (2.40). 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 (2.41) and Fig. 2.7 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 2.6.

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 $\alpha \in (0, \infty]$.

2.5 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.* (2006b). We invoke the same setting and definitions provided in Section 2.4.3. 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 $S_n = \{(X_i, Y_i) : i = 1, \dots, n\}$ be the training dataset where, for each $i \in \{1, \dots, n\}$, 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 (2.25), i.e., for each $\theta \in \mathbb{B}_d(r)$ we have

$$\hat{R}_\alpha(\theta) = \frac{1}{n} \sum_{i=1}^{n} l^\alpha(Y_i, g_\theta(X_i)). \tag{2.48}$$

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 (2.48). First, we recall classical results in Rademacher complexity generalization bounds.

2.5.1 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 (Shalev-Shwartz and Ben-David, 2014) 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*

$$\mathcal{R}(A) := \mathbb{E}\left(\sup_{a\in A} \frac{1}{n}\langle \sigma, a\rangle\right), \tag{2.49}$$

where $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_n)$ with $\sigma_1, \sigma_2, \dots, \sigma_n$ 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, (Shalev-Shwartz and Ben-David, 2014))**.** *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 $h \in \mathcal{H}$,*

$$\left|R_l(h) - \hat{R}_l(h)\right| \leq 2\mathcal{R}(l \circ \mathcal{H} \circ S_n) + 4D\sqrt{\frac{2\ln(4/\delta)}{n}}, \tag{2.50}$$

where $R_l(h)$ and $\hat{R}_l(h)$ denote the true risk and empirical risk of l, respectively, and[3] *$l \circ \mathcal{H} \circ S_n \subset \mathbb{R}^n$ which is equal to*

$$\{(l(h,(x_1,y_1)), \dots, l(h,(x_n,y_n))) : h \in \mathcal{H}\}. \tag{2.51}$$

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, (Shalev-Shwartz and Ben-David, 2014))**.** *Suppose $\tilde{l}_1, \dots, \tilde{l}_n : \mathbb{R} \to \mathbb{R}$ are r_0-Lipschitz functions with common constant $r_0 \geq 0$. If $\tilde{l} = (\tilde{l}_1, \dots, \tilde{l}_n)$ and $A \subset \mathbb{R}^n$, then $\mathcal{R}(\tilde{l}(A)) \leq r_0\mathcal{R}(A)$, where $\tilde{l}(A) := \{(\tilde{l}_1(a_1), \dots, \tilde{l}_n(a_n)) : a \in A\}$.*

[3] In (2.50) we present the two-sided version of Theorem 26.5 in (Shalev-Shwartz and Ben-David, 2014), which can be readily obtained via the symmetrization technique.

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, (Shalev-Shwartz and Ben-David, 2014))**.** *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,*

$$\mathcal{R}(\mathcal{H} \circ x_{1:n}) \leq \frac{r \max_{i \in [n]} \|x_i\|_2}{\sqrt{n}}. \tag{2.52}$$

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.

2.5.2 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 $\alpha \in (0, 1]$,

$$C_{r_0}(\alpha) := \sigma(r_0)\sigma(-r_0)^{1-1/\alpha}; \tag{2.53}$$

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

$$C_{r_0}(\alpha) := \begin{cases} \left(\frac{\alpha-1}{2\alpha-1}\right)^{1-\frac{1}{\alpha}} \left(\frac{\alpha}{2\alpha-1}\right) & e^{r_0} \geq \frac{\alpha-1}{\alpha}, \\ \sigma(r_0)\sigma(-r_0)^{1-\frac{1}{\alpha}} & e^{r_0} < \frac{\alpha-1}{\alpha}. \end{cases} \tag{2.54}$$

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 11 in Appendix A.3 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 $\theta \in \mathbb{B}_d(r)$,*

$$\left| R_\alpha(\theta) - \hat{R}_\alpha(\theta) \right| \leq C_{r\sqrt{d}}(\alpha)\, \frac{r\sqrt{d}}{\sqrt{n}} + D_{r\sqrt{d}}(\alpha) \sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{n}}, \tag{2.55}$$

where $C_{r\sqrt{d}}(\alpha)$ is given in (2.53) *and* (2.54) *and where $D_{r\sqrt{d}}(\alpha)$ is given by $D_{r\sqrt{d}}(\alpha) := 4\sqrt{2}\dfrac{\alpha}{\alpha-1}\left(1 - \sigma(-r\sqrt{d})^{1-1/\alpha}\right)$.*

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 A.3 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 Audibert *et al.* (2007), 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 chapter. 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

$$\min_{\theta \in \mathbb{B}_d(r)} R_\alpha(\theta) = \min_{f:\mathcal{X}\to\mathbb{R}} R_\alpha(f), \tag{2.56}$$

where $R_\alpha(\theta)$ is given in (2.26) 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.,* (2.56) *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,*

$$\lim_{n\to\infty} R(f_{\hat{\theta}_n^\alpha}) = R^*, \tag{2.57}$$

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 $R^* := \min\limits_{f:\mathcal{X}\to\mathbb{R}} \mathbb{P}[Y \neq \mathrm{sign}(f(X))]$.

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 (2.56) is a stipulation for the 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.

2.6 Logistic Regression and CNN Experiments

As was first introduced in Section 2.3.3, 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 (LeCun, 1998), Fashion MNIST (FMNIST) (Xiao *et al.*, 2017), and CIFAR-10 (Krizhevsky *et al.*, 2014). 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 A.4.4, 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 (Paszke *et al.*, 2019). 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

$$\text{rel acc gain \%} = \frac{|\alpha\text{-loss acc} - \text{log-loss acc}|}{\text{log-loss acc}} \times 100, \tag{2.58}$$

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 (Cava, 2021).

2.6.1 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 of classes *Cat* and *Dog*) are presented in Tables 2.1, 2.2, and 2.3, 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 2.1, 2.2, and 2.3, 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., 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 2.3.3, 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 2.1, 2.2, and 2.3 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 A.4.4. 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.

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 2.1: 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 (2.58). Also Note That Each Reported Accuracy Is Averaged over 10 Runs.

2.6.2 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.

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

Table 2.2: Symmetric Binary Noisy Label Experiment on Classes *T-shirt* and *Shirt* of the FMNIST Dataset.

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

Arch	LF %	LL Acc %	α^* Acc %	α^*	Gain %
CNN 2+1	0	80.59	80.68	0.99	0.11
	10	79.61	79.89	1.1	0.35
	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
CNN 3+2	0	85.80	85.80	1	0.00
	10	82.92	83.15	0.99	0.28
	20	77.61	80.88	3	4.21
	30	69.53	76.72	5	10.34
	40	59.44	67.19	6	13.04
CNN 4+2	0	87.49	87.59	0.9	0.12
	10	83.65	84.69	1.2	1.25
	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
CNN 6+2	0	87.31	87.93	1.2	0.70
	10	84.91	85.33	2	0.49
	20	78.92	81.80	6	3.64
	30	68.88	77.20	7	12.09
	40	58.54	65.16	7	11.32

Table 2.3: Symmetric Binary Noisy Label Experiment on CIFAR-10 Classes *Cat* and *Dog*.

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 2.4, 2.5, and 2.6. 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 2.3.3 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.* (Fort *et al.*, 2019). 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 2.5 when *Dog* is the minority class. Thus we conclude that the results in Tables 2.4, 2.5, and 2.6, 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.

		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 2.4: 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 A.4.1 for a Brief Review of the Definition of the F_1 Score. The Relative % Gain Is Defined as the Relative Percent Gain (2.58) 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	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 2.5: 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	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

Table 2.6: Binary Cifar-10 Cnn 6+2 Imbalance Experiments on *Cat* and *Dog* Classes.

2.6.3 Key Takeaways

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

Overall Performance Relative to Log-loss: The experimental results as evidenced through Tables 2.1 to 2.6 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 our earlier theoretical investigations (Section 2.4.3) 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 (Janocha and Czarnecki, 2016) regarding the lack of canonical alternatives to log-loss (cross-entropy loss) in modern machine learning.

In this chapter, 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 2.2), and then we studied the statistical calibration (Section 2.3), optimization landscape (Section 2.4.3), and generalization capabilities (Section 2.5) 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 2.3.3, we performed noisy label and class imbalance experiments on MNIST, FMNIST, and CIFAR-10 using logistic regression and convolutional neural networks (Section 2.6). 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 (Zhang *et al.*, 2019; Madry *et al.*, 2018a; Schmidt *et al.*, 2018) in addition to learning censored and fair representations that ensure statistical fairness for all downstream learning tasks (Kairouz *et al.*, 2019a).

Chapter 3

BEING PROPERLY IMPROPER: A STATISTICAL THEORY OF ROBUSTNESS FOR LOSS FUNCTIONS

3.1 Introduction

The loss function is a cornerstone of machine learning (ML). The founding theory of properness for supervised losses stipulates that the loss function shapes the learning algorithm towards the true posterior (Reid and Williamson, 2011). Consequently, a model trained with a proper loss function will try to closely approximate the Bayes rule of the data generating distribution. Historically, properness draws its roots from classical work in normative economics for *class probability estimation* (CPE) (Reid and Williamson, 2011; Savage, 1971; Shuford *et al.*, 1966) and Fisher consistency (Fisher, 1922); some of the most famous losses in supervised learning are proper, e.g., log, square, Matusita (Matusita, 1956), to name a few. Unfortunately, in many modern applications data can be corrupted or *twisted* in various ways (see Section 3.2); examples of twists include label noise, adversarial noise, and feature noise. Thus, optimizing a proper loss function on twisted data could perilously lead the learning algorithm towards the Bayes rule of the twisted posterior, rather than to the desired clean posterior. To ensure that a model trained with a proper loss function on twisted data properly generalizes to the clean distribution, a generalization of properness is clearly required.

To this end, we propose the notion of *twist-properness*. In words, a loss function is twist-proper if and only if (iff), for any twist, there exist hyperparameter(s) of the loss which allow its minimizer to "untwist" the twisted posterior into the clean posterior. Thus, twist-properness *certifies* loss functions that allow general posterior corrections, which is analogous to how PAC learning certifies computationally efficient and accurate learning algorithms (Valiant, 1984). This generalization of properness with twist-properness would be less impactful without a solid contender loss, and we show that a nontrivial extension of α-loss, which itself is an information-theoretic hyperparameterization of the log-loss (Arimoto, 1971c; Liao *et al.*, 2018b; Sypherd *et al.*, 2019), *is* twist-proper and exhibits desirable properties for local and global (namely, fixed hyperparameter) twist corrections. Furthermore, twist-properness is not vacuous as we provide a counterexample that another (popular) generalization of the log-loss, the focal loss (Lin *et al.*, 2017a), which was originally designed to solve specific twists, i.e., class imbalance, is *not* twist-proper. In addition, we provide a proof that a loss which acts as a general "wrapper" of a loss, the Super Loss (Castells *et al.*, 2020), is also *not* twist-proper. One of our key takeaways is that twist-properness necessitates a certain nontrivial *symmetry* of the loss, rather than merely a trivial extension of the hyperparameter(s).

Recently, α-loss was practically implemented in logistic regression and in deep neural networks (Sypherd *et al.*, 2022a). In both settings, it was shown to be more robust to symmetric label noise for fixed $\alpha > 1$ than the proper log-loss ($\alpha = 1$), thereby providing a hint at the twist-properness of α-loss. In order to complement our theory of twist-properness and these recent results regarding the robustness of α-loss, we also practically implement α-loss in boosting. Boosting is imbued with the computational constraint that strong learning happens from "weak updates" in polynomial time, thus inducing substantial convergence rates (Kearns and Vazirani, 1994). Furthermore, boosting algorithms are known to suffer under label noise, particularly for convex losses in low capacity models (Long and Servedio, 2010; Mansour *et al.*, 2022a). Thus, boosting presents as an ideal choice to further practically investigate the twist-properness of α-loss.

In order to implement α-loss in boosting, a popular route is to invert the canonical link of the loss which computes the weighting of the examples (Friedman, 2001; Nock and Nielsen, 2008; Nock and Williamson, 2019). While this is feasible for the log-loss (one gets the popular sigmoid function), it turns out to be nontrivial for α-loss. We address this issue by providing the first (to the best of our knowledge) general boosting scheme (called PilBoost) for any loss which requires only an approximation of the inverse canonical link, depending on a parameter $\zeta \in [0, 1]$ (the closer to 0, the better the approximation), and gives boosting-compliant convergence, further meeting the general optimum number of calls to the weak learner. The cost of this approximation is only a factor $O(1/(1-\zeta)^2)$ in number of iterations.

In Section 4.5, we implement PilBoost with the approximate inverse canonical link of α-loss on several tabular datasets, each suffering from various twists (label, feature, and adversarial noise), and compare against AdaBoost (Freund and Schapire, 1997a) and XGBoost (Chen and Guestrin, 2016). In general, we find improved algorithmic robustness to all twists through using simple (fixed) hyperparameter corrections via the α-loss, which aligns with our theoretical contributions (see Section 3.4).

3.2 Related Work

Studying data corruption in ML dates back to the 80s (Valiant, 1985). Remarkably, the first twist models assumed very strong corruption, possibly coming from an adversary with unbounded computational resources, *but* the data at hand was binary. Thus, because the feature space was as "complex" as the class space, the twist models lacked the unparalelled data complexity that we now face. Obtaining such twist models at scale with real world data has been a major problem in ML over the past decade for a number of reasons. Nevertheless, there have been several streams of recent research aimed at addressing specific twists.

Label noise is a twist which has recently drawn much attention and garnered many corrective attempts (Patrini *et al.*, 2017b; Zhang and Sabuncu, 2018b; Zhang *et al.*, 2021; Natarajan *et al.*, 2013; Long and Servedio, 2010; Sypherd *et al.*, 2022a; Liu and Guo, 2020; Ghosh *et al.*, 2017). Notably, Natarajan *et al.* (2013) theoretically study the presence of class conditional noise in binary classification. Their approach consists of augmenting proper loss functions with re-weighting coefficients, which is strictly dependent on the class conditional noise percentages, and hence requires knowledge of the noise proportions. As a byproduct of their analysis, they show that biased SVM and weighted logistic regression are provably noise-tolerant.

Setting label noise aside, there exists a zoo of other twists and corrective attempts. For instance, *data augmentation* techniques, with vicinal risk minimization standing as a pioneer (Chapelle *et al.*, 2000), seek to induce general robustness (Zhang *et al.*, 2018). In deep learning, *adversarial robustness* attempts to address the brittleness of neural networks to targeted adversarial noise (Szegedy *et al.*, 2013; Madry *et al.*, 2018b; Andriushchenko and Hein, 2019). *Data poisoning* twists in computer vision can be very sophisticated and require further investigation (Truong *et al.*, 2020). *Invariant risk minimization* aims at finding data representations yielding good classifiers but also invariant to "environment changes" (Arjovsky *et al.*, 2019); relatedly, *covariate shift* seeks to address changes between train and test, stemming from non-stationarity or bias in the data (Zhang *et al.*, 2020). A recent trend has also emerged with correcting losses due to model *confidence* issues (Guo *et al.*, 2017; Mukhoti *et al.*, 2020; Castells *et al.*, 2020).

Viewed more broadly, the abovementioned papers arguably study much different problems, but they tend to have a theme that goes substantially deeper than the superficial observation that they assume twisted data in some way: the core loss function is usually a *proper* loss. Therefore, they tend to start from the premise of a loss that inevitably fits the (unwanted) twist, and correct it mostly with a regularizer informed with some prior knowledge of the twist, on a "twist-by-twist" basis. There has been some positive work in this "loss + regularizer" direction (Amid *et al.*, 2019b; Ma *et al.*, 2020; Zhang *et al.*, 2021), but we note that this does not fully address the underlying issue of properness shaping the learning algorithm towards the twisted posterior.

Lastly, our generalization of properness with twist-properness is partly inspired by recent work by (Charoenphakdee *et al.*, 2021), where they theoretically investigate the focal loss. Notably, they show that the focal loss is classification-calibrated, but not strictly proper. From their work, we also gather the implicit notion that hyperparameterized losses that generalize proper losses (e.g., focal loss or α-loss generalizing log-loss), which may represent a next step for loss functions in ML, need to be carefully understood from the standpoint of what their hyperparameterization trades-off from properness.

3.3 Losses for Class-Probability Estimation

Our setting is that of losses for class probability estimation (CPE) and our notations follow (Reid and Williamson, 2010b, 2011). Given a domain of observations $\mathcal{X}$, we wish to learn a classifier $h : \mathrm{dom}(h) = \mathcal{X}$ that predicts the label $\mathsf{Y} \in \mathcal{Y} \doteq \{-1, 1\}$ (we assume two classes or labels) associated with every instance of data drawn from $\mathcal{X}$. Traditionally, there are two kinds of outputs sought: one requires $\mathrm{Im}(h) = [0, 1]$, in which case h provides an estimate of $\mathbb{P}[\mathsf{Y} = 1|\mathsf{X}]$, which is often called the Bayes posterior. This is the framework of class probability estimation. The other kind of output requires $\mathrm{Im}(h) = \mathbb{R}$, but is usually completed by a mapping to $[0, 1]$, e.g., via the softmax in deep learning. A loss for class probability estimation, $\ell : \mathcal{Y} \times [0, 1] \to \mathbb{R}$, has the general definition

$$\ell(y,u) \doteq [y=1]\cdot\ell_1(u) + [y=-1]\cdot\ell_{-1}(u), \tag{3.1}$$

where $[\cdot]$ is Iverson's bracket (Knuth, 1992). Functions ℓ_1, ℓ_{-1} are called *partial* losses, minimally assumed to satisfy $\mathrm{dom}(\ell_1) = \mathrm{dom}(\ell_{-1}) = [0,1]$ and $|\ell_1(u)| \ll \infty, |\ell_{-1}(u)| \ll \infty, \forall u \in (0,1)$ to be useful for ML. Key additional properties of partial losses are:

(M) Monotonicity: ℓ_1, ℓ_{-1} are non-increasing and non-decreasing, respectively;

(D) Differentiability: ℓ_1 and ℓ_{-1} are differentiable;

(S) Symmetry: $\ell_1(u) = \ell_{-1}(1-u), \forall u \in [0,1]$.

Commonly used proper losses such as log, square and Matusita all satisfy the above three assumptions. The pointwise conditional risk of the local guess $u \in [0,1]$ with respect to a *ground truth* $v \in [0,1]$ is

$$L(u,v) \doteq \mathbb{E}_{\mathsf{Y}\sim\mathrm{B}(v)}\left[\ell(\mathsf{Y},u)\right] = v\cdot\ell_1(u) + (1-v)\cdot\ell_{-1}(u), \tag{3.2}$$

where $\mathrm{B}(v)$ defines a Bernoulli distribution with v.

Properness $L(u,v)$ is the fundamental quantity that allows to distinguish proper losses: a loss is *proper* iff for any ground truth $v \in [0,1]$, $L(v,v) = \inf_u L(u,v)$, and strictly proper iff $u = v$ is the sole minimiser (Reid and Williamson, 2011). The (pointwise) *Bayes* risk is $\underline{L}(v) \doteq \inf_u L(u,v)$.

Surrogate loss Oftentimes, minimization occurs over the reals (e.g., boosting), hence it is useful to employ a surrogate to the 0-1 loss (Bartlett *et al.*, 2006a). (Nock and Nielsen, 2008) showed that the outputs in $[0,1]$ and $\mathbb{R}$ can be related via convex duality of the losses. Let $g^\star(z) \doteq \sup_t\{zt - g(t)\}$ denote the convex conjugate of g (Boyd and Vandenberghe, 2004b). The *surrogate* F of $\underline{L}$ is thus given by

$$F(z) \doteq (-\underline{L})^\star(-z), \forall z \in \mathbb{R}. \tag{3.3}$$

For example, picking the log-loss as ℓ gives the binary entropy for $\underline{L}$ and the logistic loss for F (see Appendix B.1.1 for a derivation). Convex duality implies that predictions in $[0, 1]$ and $\mathbb{R}$ are related via the (canonical) *link* of the loss, $(-\underline{L})'$ (Nock and Williamson, 2019) where we use the notation f' to denote the derivative of a function f with respect to its argument. In the sequel, we will see that boosting requires inverting the link of the loss, which we show is nontrivial for hyperparameterized losses, such as α-loss. Lastly, we provide summary properties of a CPE loss (not necessarily proper) and its surrogate, monotonicity being of primary importance. Some parts of the following Lemma are known in the literature (e.g., concavity in (Agarwal, 2014, Lemma 1)), or are folklore.

Lemma 5. $\forall \ell$ CPE *loss, $\underline{L}$ is concave and continuous; F is convex, continuous and non-increasing.*

3.4 Twist-Proper Losses

With the classical setting of properness in hand, we now provide fundamental definitions of *twists* and *twist-properness*, and study the twist-properness of several hyperparameterized loss functions. When it comes to correcting (or untwisting) twists, one needs a loss with the property that its minimizer in (3.2) is different from the now twisted value $\tilde{v}$ *and* recovers the "hidden" ground truth v.

Bayes tilted estimates We first characterize the minimizers of (3.2) when the CPE loss is not necessarily proper. We define the set-valued (pointwise) Bayes *tilted estimate* t_ℓ as

$$t_\ell(\tilde{v}) \doteq \arg \inf_{u \in [0,1]} L(u, \tilde{v}). \tag{3.4}$$

Ideally, we would like for $v \in t_\ell(\tilde{v})$, i.e., the Bayes tilted estimate $t_\ell(\tilde{v})$ untwists (with hyperparameter(s)) the twisted value $\tilde{v}$ and recovers the ground truth v. However, it follows that if the loss is proper, $\tilde{v} \in t_\ell(\tilde{v})$ and, if strictly proper, $t_\ell(\tilde{v}) = \{\tilde{v}\}$. This formally highlights the limitation of proper loss functions in twisted settings, namely, the inability of a proper loss to untwist the twisted value because the minimization of the loss is centered on what it "perceives" to be the ground truth. The following result stipulates the cardinality of the Bayes tilted estimates.

Lemma 6. *If the partial losses ℓ_1 and ℓ_{-1} of a given* CPE *loss $\ell : \mathcal{Y} \times [0,1] \to \mathbb{R}$ satisfy **(M)**, **(D)**, and **(S)** and are also strictly convex, then $|t_\ell(\tilde{v})| = 1$ for every $\tilde{v} \in [0,1]$.*

In Appendix B.1.3, we provide an extended version of Lemma 6, denoted Lemma 12, where we prove properties of Bayes tilted estimates for when t_ℓ is set-valued (e.g., set-valued monotonicity and symmetry, and analysis of extreme values). As a consequence, we show that strict monotonicity of the partial losses is not sufficient to guarantee that t_ℓ is a singleton; in fact, strict convexity is required as in Lemma 6.

An important class of twists We now adopt more conventional ML notations and instead of a hidden ground truth v and twisted ground truth $\tilde{v}$, we use η_{c} and η_{t} to denote the "clean" and "twisted" posterior probabilities that $Y = 1$ given $X = x$, respectively. Further, a "**twist**" refers to a general mapping $\eta_{\mathrm{c}} \mapsto \eta_{\mathrm{t}}$, which could be a consequence of label/feature/adversarial noise. The following delineates a fundamental class of twists, important in the sequel.

Definition 5. *A twist $\eta_c \mapsto \eta_t$ is said to be Bayes blunting if and only if $(\eta_c \leq \eta_t \leq 1/2) \vee (\eta_c \geq \eta_t \geq 1/2)$.*

The term "blunting" is inspired by adversarial training (Cranko *et al.*, 2019). Intuitively, a Bayes blunting twist does not change the *maximum a posteriori* guess for the label given the observation, but it does reduce algorithmic confidence in learned posterior estimates, which is particularly damaging in practice where the learning algorithm only has a finite number of (twisted) training examples. Furthermore, Bayes blunting twists capture a very important twist (see Section 3.2): *symmetric label noise* (SLN). Under symmetric label noise with flip probability $p \in [0, 1]$, the twisted posterior η_{t} is given by $\eta_{\mathrm{t}} = \eta_{\mathrm{c}}(1-p) + (1-\eta_{\mathrm{c}})p$ (Reid and Williamson, 2010b). The following result readily follows via Definition 5 from consideration of p for fixed η_{c}.

Lemma 7. *SLN is Bayes blunting for $p < 1/2$.*

Historically, Reid and Williamson (2010b) showed that proper loss functions are not robust to this twist which further motivates our consideration of twist-proper losses.

Twist-proper losses To overcome these limitations of properness, we propose a generalized notion, called twist-properness, which utilizes hyperparameterization of the loss to untwist twisted posteriors into clean posteriors.

Definition 6. *A loss ℓ is twist-proper (respectively, strictly twist-proper) iff for any twist, there exists hyperparameter(s) such that $\eta_{\mathrm{c}} \in t_\ell(\eta_t)$ (respectively, $\{\eta_{\mathrm{c}}\} = t_\ell(\eta_t)$).*

Where a proper loss could perilously lead the learning algorithm to estimate η_{t}, a *twist-proper* loss employs hyperparameters so that its Bayes tilted estimate recovers η_{c}, hence guiding the algorithm to untwist the twisted posterior. We emphasize the need for hyperparameters as otherwise, twist-properness would trivially enforce $t_\ell(\cdot) = [0, 1]$. Recently, hyperparameterized loss functions have garnered much interest in ML, to name a few (Barron, 2019; Lin *et al.*, 2017a; Amid *et al.*, 2019b; Li *et al.*, 2021; Sypherd *et al.*, 2022a), possibly because such losses allow practitioners to induce variegated models. Indeed, hyperparameterized loss functions could be efficiently implemented via meta-algorithms, such as AutoML (He *et al.*, 2021), *or* practically utilized in the burgeoning field of federated learning (Kairouz *et al.*, 2019b), where the hyperparameter(s) might yield more fine-grained ML model customization for edge devices.

Ostensibly, “optimal” hyperparameters requires explicit knowledge of the distribution and twist, *and* each example in the training set requires a different hyperparameter to untwist its twisted posterior. However, in the sequel, we show that a twist-proper loss, namely α-loss, with a *fixed* hyperparameter (α) can untwist a large class of twists, i.e. Bayes blunting twists (such as SLN), better than log-loss. Thus, we posit through our experimental results in Section 4.5 that the practitioner only needs peripheral, rather than explicit, knowledge of a Bayes blunting twist in the data.

Twist-(im)proper losses Lin *et al.* (2017a) introduced the focal loss to improve class imbalance issues associated with dense object detection. It generalizes the log-loss and has become popular due to its success in such domains. Recently, the focal loss has received increased scrutiny (Charoenphakdee *et al.*, 2021), where it was shown to be classification-calibrated but not strictly proper. Here, we determine the *twist-properness* of the focal loss.

Lemma 8. *Define the* ***focal loss*** *via the following partial losses:* $\ell_1^{FL}(u) \doteq -(1-u)^\gamma \log u$ *and* $\ell_{-1}^{FL}(u) \doteq \ell_1^{FL}(1-u)$*, with* $\gamma \geq 0$*. Then the focal loss is* ***not*** *twist proper.*

In the proof (see Appendix B.1.4), we also provide a proof that a loss which acts as a general “wrapper” of a loss, the Super Loss (Castells *et al.*, 2020), is *not* twist proper. Concerning the focal loss, Lemma 8 is not necessarily an impediment for this loss function, which was designed to deal with a specific twist, class imbalance, and it does not prevent *generalizations* of the focal loss that would be twist proper. However, our proof suggests that the Bayes tilted estimate (3.4) of such generalizations risks not being in a simple analytical form. Intuitively, twist-properness requires more than a trivial extension of the hyperparameter of the loss; it also seems to require a certain *symmetry*, which we observe with the following twist-proper loss, α-loss.

A twist-proper loss The α-loss was first introduced in information theory in the early 70s (Arimoto, 1971c) for $\alpha \in \mathbb{R}_+$ and recently received increased scrutiny in privacy and ML (Liao *et al.*, 2018b; Sypherd *et al.*, 2019) for $\alpha \geq 1$. Most recently, Sypherd *et al.* (2022a) studied the calibration, optimization, and generalization characteristics of α-loss in ML for $\alpha \in \mathbb{R}_+$. In particular, they experimentally found that α-loss is robust to noisy labels under logistic regression and convolutional neural-networks for $\alpha > 1$. We now provide our (extended) definition of the α-loss in CPE.

Definition 7. *For $\alpha \geq 0$, the α-loss has the following partial losses: $\forall u \in [0,1]$, $\ell_1^\alpha(u) \doteq \ell_{-1}^\alpha(1-u)$ where*

$$\ell_1^\alpha(u) \doteq \frac{\alpha}{\alpha-1} \cdot \left(1 - u^{\frac{\alpha-1}{\alpha}}\right), \tag{3.5}$$

and by continuity we have $\ell_1^0(u) \doteq \infty$, $\ell_1^1(u) \doteq -\log u$, and $\ell_1^\infty(u) \doteq 1-u$. For $\alpha < 0$, we let $\forall u \in [0,1]$,

$$\ell_1^\alpha(u) \doteq \ell_{-1}^{-\alpha}(u) = \ell_1^{-\alpha}(1-u). \tag{3.6}$$

For a plot of (3.5), see Figure B.1 in Appendix B. Note that the α-loss is **(S)**ymmetric by construction, and that it continuously interpolates the log-loss ($\alpha = 1$) which is proper (Reid and Williamson, 2010b). Our definition extends the previous definitions with (3.6), which induces a fundamental symmetry that is required for twist-properness and is utilized in the following result. For any $u \in [0,1]$, we let $\iota(u) \doteq \log(u/(1-u))$ denote the logit of u.

Lemma 9. *The following four properties, labeled **(a)-(d)**, all hold for α-loss: **(a)** **(M)**, **(D)**, **(S)** all hold, $\forall \alpha \in \mathbb{R} \setminus \{0\}$; **(b)** if $(\alpha = 0) \vee (\alpha = \pm\infty \wedge \eta_t = 1/2)$, then*

$t_{\ell^\alpha}(\eta_t) = [0, 1]$, *if* $\alpha \in \mathbb{R} \setminus \{0, \pm\infty\}$, *then* $t_{\ell^\alpha}(\eta_t) = \left\{\frac{\eta_t^\alpha}{\eta_t^\alpha + (1-\eta_t)^\alpha}\right\}$, *and if* $\alpha \to \pm\infty$, *then* $t_{\ell^{\pm\infty}}(\eta_t) = \pm 1$ *or* ∓ 1, *depending on the sign of* $\eta_{\mathrm{t}} - 1/2$*;* ***(c)*** *hence,* α*-loss is twist-proper with* $\alpha^* = \iota(\eta_{\mathrm{c}})/\iota(\eta_{\mathrm{t}})$*;* ***(d)*** *for any Bayes blunting twist,* $\alpha^* \geq 1$*.*

The proof of Lemma 9 can be found in Appendix B.1.5. Note that **(a)** readily follows from Definition 7. The Bayes tilted estimate in **(b)**, i.e. t_{ℓ^α} for $\alpha \in \mathbb{R} \setminus \{0, \pm\infty\}$, is known in the literature as the α-tilted distribution (Arimoto, 1971c; Liao *et al.*, 2018b; Sypherd *et al.*, 2022a). We observe that the α-tilted distribution is symmetric upon permuting $(\eta_{\mathrm{t}}, \alpha)$ and $(1 - \eta_{\mathrm{t}}, -\alpha)$. Hence, our nontrivial extension of the α-loss induces a *symmetry*, particularly useful for untwisting malevolent twists, which thereby yields twist-properness, **(c)**. Lastly, **(d)** indicates that $\alpha^* \geq 1$ for any Bayes blunting twist (e.g., SLN with $p < 1/2$); however, note that this holds merely for a given x, not over the whole domain $\mathcal{X}$.

Untwisting over the whole domain $\mathcal{X}$ Just as classification-calibration is a pointwise form of consistency (Bartlett *et al.*, 2006a), twist-properness is a *pointwise form of correction.* Extending twist-properness to the entire domain $\mathcal{X}$ seems to require learning a *mapping* $\alpha : \mathcal{X} \to [-\infty, \infty]$, which is infeasible under standard ML assumptions, since one would need explicit knowledge of the distribution and twist. Nevertheless, we show here that for a large class of twists, namely Bayes blunting twists, a fixed $\alpha_0 > 1$ obtained *non-constructively*, is strictly "better" than the proper choice, log-loss ($\alpha = 1$). We also provide a general *constructive* formula for a fixed $\alpha^{**} \in \mathbb{R}$, calculated from distributional and twist information.

In order to represent population quantities, we assume a marginal distribution M over $\mathcal{X}$ (following notation by (Reid and Williamson, 2011)), from which the expected value of a *loss* ℓ quantifies its true risk of a given classifier h. With a slight abuse of notation, we also let $\eta_{\mathrm{c}}, \eta_{\mathrm{t}} : \mathcal{X} \to [0, 1]$ denote the clean and twisted posterior *mappings*, respectively. To evaluate the efficacy of the Bayes tilted estimate of α-

loss at untwisting the twisted posterior mapping and recovering the clean posterior mapping, we define the following averaged cross-entropy, given by

$$\mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha) \doteq \mathbb{E}_{\mathsf{X}\sim\mathrm{M}}[\eta_{\mathrm{c}}(\mathsf{X}) \cdot - \log t_{\ell^\alpha}(\eta_{\mathrm{t}}(\mathsf{X})) + (1 - \eta_{\mathrm{c}}(\mathsf{X})) \cdot - \log t_{\ell^\alpha}(1 - \eta_{\mathrm{t}}(\mathsf{X}))], \tag{3.7}$$

where for convenience we used the symmetry property of t_ℓ from Appendix B.1.3, i.e., $t_{\ell^\alpha}(1-\eta_{\mathrm{t}}(\mathsf{X})) = 1-t_{\ell^\alpha}(\eta_{\mathrm{t}}(\mathsf{X}))$. Following (Schapire and Freund, 2012), we denote the binary entropy as $H_{\mathrm{b}}(u) \doteq -u \cdot \log(u) - (1 - u) \cdot \log(1 - u)$, for $u \in [0, 1]$. We let $H(\eta_{\mathrm{c}})$ represent an averaged binary entropy of the η_{c}-mapping, given by

$$H(\eta_{\mathrm{c}}) \doteq \mathbb{E}_{\mathsf{X}\sim\mathrm{M}}[H_{\mathrm{b}}(\eta_{\mathrm{c}}(X))]. \tag{3.8}$$

With (3.7) and (3.8), we obtain (cf. (Thomas and Joy, 2006)),

$$D_{\mathrm{KL}}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha) \doteq \mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha) - H(\eta_{\mathrm{c}}), \tag{3.9}$$

that is, a KL-divergence between the α-Bayes tilted estimate of the twisted posterior and the clean posterior mappings. Intuitively, $D_{\mathrm{KL}}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha)$ aggregates a series of information-trajectories, strictly dependent on α (either fixed or a mapping), each tracing a path on the probability *simplex* between the two posterior mappings for every $x \in \mathcal{X}$. Slightly more restrictive than Definition 5, we define a *strictly* Bayes blunting twist as a Bayes blunting twist where there exists $\epsilon > 0$ such that $(\eta_{\mathrm{c}} + \epsilon \leq \eta_{\mathrm{t}} \leq 1/2) \vee (\eta_{\mathrm{c}} - \epsilon \geq \eta_{\mathrm{t}} \geq 1/2)$; we state one of our main results whose proof is in Appendix B.1.6.

Theorem 7. *For any strictly Bayes blunting twist $\eta_c \mapsto \eta_t$, there exists a fixed $\alpha_0 > 1$ and an optimal $\alpha^\star$-mapping, $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$, which induces the following ordering*

$$D_{\mathrm{KL}}(\eta_c, \eta_t; 1) > D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha_0) \geq D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha^\star). \tag{3.10}$$

This result answers in the affirmative that untwisting $\mathcal{X}$ for a large class of twists with a fixed hyperparameter $\alpha_0 > 1$ is *strictly* better than simply using the proper choice, i.e., $\alpha = 1$ (log-loss). Specifically, Theorem 7 holds for SLN, which is a strictly Bayes blunting twist for flip probability $0 < p < 1/2$ (Lemma 7). The result also states that there exists a mapping $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$ which optimally untwists the strictly Bayes blunting twist; indeed, $\alpha^\star$ can be recovered from Lemma 9**(c)**, i.e., $\alpha^\star(x) := \iota(\eta_{\mathrm{c}}(x))/\iota(\eta_{\mathrm{t}}(x))$, for every $x \in \mathcal{X}$. Thus, by the twist-properness of α-loss, $D_{\mathrm{KL}}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha^\star) = 0$ (more details in Appendix B.1.6). Regarding the search for a fixed $\alpha_0 > 1$ in practice, Sypherd *et al.* (2022a) showed via optimization landscape analysis and experiments on SLN for logistic regression and neural-networks that the search space for α_0 is bounded (due to saturation), typically $\alpha_0 \in [1.1, 8]$. In Section 4.5, we report experimental results for several α; we also incorporate a method inspired by (Menon *et al.*, 2015) to estimate the amount of SLN in training data and thus estimate α_0 using Lemma 9**(c)** as motivated by Theorem 7.

Theorem 7 gave a *nonconstructive* indication for the optimal regime of α for strictly Bayes blunting twists. Our next result gives a *constructive* formula for a fixed α for any twist. Given $B > 0$, let $\mathrm{M}(B)$ denote the distribution restricted to the support over $\mathcal{X}$ for which we have almost surely

$$(1 + \exp(B))^{-1} \leq \eta_{\mathrm{t}}(x) \leq (1 + \exp(-B))^{-1}, \tag{3.11}$$

and let $p(B) \in [0, 1]$ be the weight of this support in M. We let $\mathrm{D}(B)$ denote the product distribution on examples ($\mathcal{X} \times \mathcal{Y}$) induced by marginal $\mathrm{M}(B)$ and posterior η_{c} (see (Reid and Williamson, 2011, Section 4)). We define the logit-*edge* as

$$\mathsf{e} \doteq (1/B) \cdot \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim \mathrm{D}(B)} \left[\mathsf{Y} \cdot \iota(\eta_{\mathrm{t}}(\mathsf{X})) \right], \tag{3.12}$$

where we note that $\mathsf{e} \in [-1, 1]$ due to the assumption in (3.11). Finally, we let $q \doteq (1 + \mathsf{e})/2 \in [0, 1]$.

Theorem 8. *Let $B > 0$. If $p(B) = 1$ and we fix $\alpha = \alpha^{**}$ with $\alpha^{**} \doteq \iota(q)/B$, then the following bound holds*

$$D_{\mathrm{KL}}(\eta_c, \eta_\iota; \alpha^{**}) \leq H_b(q) - H(\eta_c). \tag{3.13}$$

The proof of Theorem 8 is in Appendix B.1.7, where we also prove an extended version of the result when $p(B) < 1$. In addition, we provide a simple example where $D_{\mathrm{KL}}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha^{**})$ can vanish with respect to $D_{\mathrm{KL}}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; 1)$ (the "proper" choice). Intuitively, the difference on the right-hand-size of (3.13) in Theorem 8 is reminiscent of a Jensen's gap. Also in the proof of Theorem 8, we find that if $|\alpha^{**}|$ is large, there is more "flatness" in the bounded terms near α^{**}. Hence, this suggests that a choice of α_0 "close-enough" to α^{**} could yield similar performance.

3.5 Sideways Boosting a Surrogate Loss

With the theory of twist-properness and the twist-proper α-loss in hand, we now turn towards the algorithmic contribution presented in this chapter. As stated in the introduction, α-loss was recently implemented in logistic regression and in deep neural networks (Sypherd *et al.*, 2022a), and was found to be more robust to symmetric label noise for fixed $\alpha > 1$ than the proper log-loss ($\alpha = 1$). Thus, in order to complement our theory of twist-properness and these recent results of α-loss, we also practically implement α-loss in *boosting.*

Formally, we have a training sample $\mathcal{S} \doteq \{(\boldsymbol{x}_i, y_i), i \in [m]\} \subset \mathcal{X} \times \mathcal{Y}$ of m examples, where $[m] \doteq \{1, 2, ..., m\}$ and note that $\mathcal{Y} = \{-1, +1\}$. We write $i \sim \mathcal{S}$ to indicate sampling example $(\boldsymbol{x}_i, y_i)$ according to $\mathcal{S}$. Following (Schapire and Singer, 1999; Collins *et al.*, 2000; Nock and Nielsen, 2008), the boosting algorithm minimizes an expected surrogate loss with respect to $\mathcal{S}$ in order to learn a real-valued classifier

$H : \mathcal{X} \to \mathbb{R}$ given by

$$H_{\boldsymbol{\beta}} \doteq \sum_j \beta_j h_j, \tag{3.14}$$

where $\{h_{\cdot} : \mathcal{X} \to \mathbb{R}\}$ are WL (weak learning) classifiers with slightly better than random classification accuracy. The oracle WL returns an index $j \in \mathbb{N}$, and the task for the boosting algorithm is to learn the coordinates of $\boldsymbol{\beta}$, initialized to the null vector. In our general framework, the losses we consider are the surrogates F in Lemma 5, essentially convex and non-increasing functions, adding the condition that they are twice differentiable. We compute weights using the blueprint of (Friedman, 2001), which uses the full $H_{\boldsymbol{\beta}}$,

$$w_i \doteq -F'(y_i H_{\boldsymbol{\beta}}(\boldsymbol{x}_i)), \forall i \in [m]. \tag{3.15}$$

Via Lemma 5, weights w_i are non-negative and tend to increase for an example given the wrong class by the current weak classifier h_j, thus, weighting puts emphasis on "hard" examples. For an underlying CPE loss ℓ, we have that (see Appendix B.1.8 for a derivation)

$$-F'(z) = (\ell_{-1} \circ t_\ell - \ell_1 \circ t_\ell)^{-1}(-z). \tag{3.16}$$

We thus need to invert the *difference* of the partial losses to recover $-F'$. The inversion is easy for the log-loss because of properties of the log function and for the square loss because its partial losses are quadratic functions. However, for hyperparameterized losses, such as the α-loss, the inversion in (3.16) is nontrivial. We circumvent this difficulty by proposing a novel boosting algorithm, PilBoost, given in Algorithm 1. Rather than using $-F'$ as in (3.15) for the weight update in Step 2.1, PilBoost uses an approximation function $\widetilde{f}$, which is non-negative and increasing, that we dub *pseudo-inverse link* (Pil), which is studied in general in Appendix B.1.8.

Algorithm 1 PilBoost

```
Input sample $\mathcal{S}$, number of iterations $T$, $a_f > 0$, Pil $\widetilde{f}$;
Step 1 : let $\boldsymbol{\beta} \leftarrow \mathbf{0}$; // first classifier, $H_{\mathbf{0}} = 0$
Step 2 : for $t = 1, 2, ..., T$
        Step 2.1 : let $w_i \leftarrow \widetilde{f}(-y_i H_{\boldsymbol{\beta}}(x_i)), \forall i \in [m]$;
        Step 2.2 : let $j \leftarrow \textsc{wl}(\mathcal{S}, \boldsymbol{w})$;
        Step 2.3 : let $\mathsf{e}_j \leftarrow (1/m) \cdot \sum_i w_i y_i h_j(\boldsymbol{x}_i)$;
        Step 2.4 : let $\beta_j \leftarrow \beta_j + a_f \mathsf{e}_j$;
Output $H_{\boldsymbol{\beta}}$.
```

Specifically, in Lemma 17, we provide $\widetilde{f}_\ell$ for α-loss, given in (B.145). Furthermore in Lemma 18, we show that there exists $K > 0$ such that, for almost all $z \in \mathbb{R}$, $|(\widetilde{f}_\ell - (-\underline{L}')^{-1})(z)| \lesssim K/\alpha$. We now theoretically analyze PilBoost, and we make two classical assumptions on wl (Schapire and Singer, 1999; Nock and Williamson, 2019).

Assumption 1. ***(R)*** *The weak classifiers have bounded range:* $\exists M > 0$ *such that* $|h_j(x_i)| \leq M, \forall j$.

Let $\tilde{\mathsf{e}}_j \doteq m \cdot \mathsf{e}_j/(\mathbf{1}^\top \boldsymbol{w}_j) \in [-M, M]$ be the normalized edge of the j-th weak classifier, where with a slight abuse of notation of (3.12), e_j is the (unnormalized) edge (Step 2.3).

Assumption 2. ***(WLA)*** *The weak classifiers are not random:* $\exists \gamma > 0$ *such that* $|\tilde{\mathsf{e}}_j| \geq \gamma \cdot M, \forall j$.

Note that "WLA" denotes the Weak Learning Assumption, which is a pillar of boosting theory (cf. (Freund *et al.*, 1999)). Since we employ $\widetilde{f}$ instead of F' in PilBoost, we need two more functional assumptions on the first- and second-order derivatives of F. The *edge discrepancy* of a function F on weak classifier h_j at iteration t is given by

$$\Delta_j(F) \doteq \left|\mathbb{E}_{i\sim\mathcal{S}}\left[y_i h_j(\boldsymbol{x}_i) F'(y_i H_{\boldsymbol{\beta}}(\boldsymbol{x}_i))\right] - \mathsf{e}_j\right|, \tag{3.17}$$

which is the absolute difference of the edge using (the derivative of) F vs. using PilBoost's $\widetilde{f}$ (implicit in e_j).

Assumption 3. ***(E, C)*** *$\exists \zeta, \pi \in [0,1)$ such that:*

> ***(E)*** *the edge discrepancy is bounded $\forall t$: $\Delta_j(F) \leq \zeta \cdot \mathsf{e}_j$, where j is returned by* WL *at iteration t;*
>
> ***(C)*** *the curvature of F is bounded: $F^* \doteq \sup_z F''(z) \leq (1-\zeta)(1+\pi)/(a_f M^2)$.*

Note that **(C)** is quite mild for specific sets of functions, e.g., proper canonical losses are Lipschitz (Reid and Williamson, 2010b), so **(C)** can in general be ensured by a simple renormalization of the loss. On the other hand, **(E)** can become progressively harder to ensure as the number of iterations increases because the choices of the WL will become restricted; nevertheless, it is not prohibitory in practice as our experiments in Section 4.5 suggest (also see the remark in Appendix B.1.10 for further commentary on this assumption). Let $\tilde{w}_t \doteq \mathbf{1}^\top \boldsymbol{w}_t$, the total weight at iteration t in PilBoost.

Theorem 9. *Suppose **(R, WLA)** hold on* WL *and **(E, C)** hold on F, for each iteration of* PilBoost. *Denote $Q(F) \doteq 2F^*/(\gamma^2(1-\zeta)^2(1-\pi^2))$. The following holds:*

- *on the risk defined by F: $\forall z^* \in \mathbb{R}, \forall T > 0$, if we observe $\sum_{t=0}^{T} \tilde{w}_t^2 \geq Q(F) \cdot (F(0) - F(z^*))$, then*

$$\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_{\boldsymbol{\beta}}(x_i))\right] \leq F(z^*). \tag{3.18}$$

- *on edge distribution:* $\forall\theta \geq 0, \forall\varepsilon \in [0,1], \forall T > 0$, *letting* $F_{\varepsilon,\theta} \doteq (1-\varepsilon)\inf F + \varepsilon F(\theta)$, *if the number of iterations satisfiees* $T \geq \frac{1}{\varepsilon^2} \cdot \frac{Q(F)\cdot(F(0)-F_{\varepsilon,\theta})}{\widetilde{f}^2(-\theta)}$, *then*

$$\mathbb{P}_{i\sim\mathcal{S}}\left[y_i H_{\boldsymbol{\beta}}(x_i) \leq \theta\right] \leq \varepsilon. \tag{3.19}$$

Thus, Theorem 9 gives boosting compliant convergence on training, and the synthesis of (3.18) and (3.19) provides a very strong convergence guarantee. When classical assumptions about the loss of interest are satisfied, such as it being Lipschitz (ensured for proper canonical losses (Reid and Williamson, 2010b)), there is a natural extension to generalization following standard approaches (Bartlett and Mendelson, 2002; Schapire *et al.*, 1998). See Appendix B.1.10 for the proof of Theorem 9, and for additional remarks regarding its *optimality* and further application to addressing discrepancies due to *machine type approximations*.

3.6 Experiments

We provide experimental results on PilBoost (for $\alpha \in \{1.1, 2, 4\}$) and compare with AdaBoost (Freund and Schapire, 1997a) and XGBoost (Chen and Guestrin, 2016) on four binary classification datasets, namely, cancer (Wolberg *et al.*, 1995), xd6 (Buntine and Niblett, 1992), diabetes (Smith *et al.*, 1988), and online shoppers intention (Sakar *et al.*, 2019). We performed 10 runs per algorithm with randomization over the train/test split and the twisters. All experiments use regression decision trees (of varying depths 1-3) in order to align with XGBoost. Hyperparameters of XGBoost were kept to default to maintain the fairest comparison between the three algorithms; for more of these experimental details, please refer to Appendix B.2.5. In order to demonstrate the *twist-properness* of α-loss as implemented in PilBoost, we corrupt the training examples of these datasets according to three different (malicious) twisters.

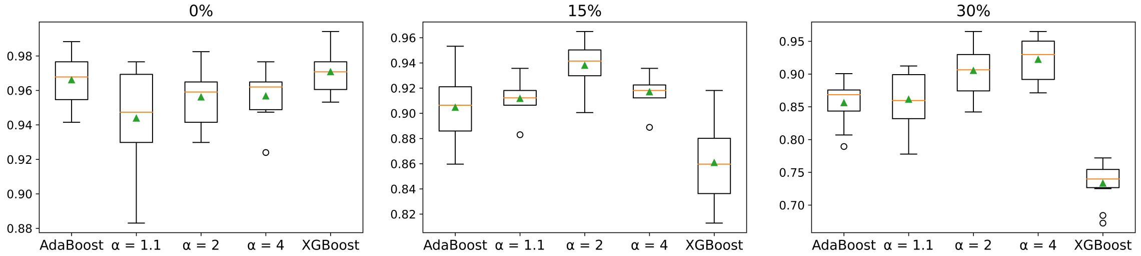


Figure 3.1: Box and Whisker Plots Reporting the *Classification Accuracy* of Adaboost, PilBoost (for $\alpha \in \{1.1, 2, 4\}$), and Xgboost on the Cancer Dataset Affected by the Class Noise Twister with 0%, 15%, and 30% Twist. Note That the Orange Line Is the Median, the Green Triangle Is the Mean, the Box Is the Interquartile Range, and the Circles Outside of the Whiskers Are Outliers. All Three Algorithms Were Trained with Decision Stumps (Depth 1 Regression Trees). For $\alpha = 1.1, 2$, and 4, We Set $a_F = 7, 2$, and 4, Respectively. Numeric Values Corresponding to the Box and Whisker Plots Are Provided in Table B.1 in Section B.2.3. We Find That PilBoost has Gains over Adaboost and Xgboost When There Is Twist Present, and α^* (of Our Set) Increases as the Amount of Twist Increases, Which Follows Theoretical Intuition (Lemma 9).

Class Noise Twister (all datasets): This twister is equivalent to SLN in the training sample. Results on this twister for the cancer dataset are presented in Figure 3.1 and see Appendix B.2.3 for further results. In general, we find that PilBoost is more robust to the Class Noise Twister than AdaBoost and XGBoost, and we find that α^* increases as the amount of twist increases, which complies with our theory (Lemma 9 and Theorem 7). We also present an *adaptive* α *experiment* in Figure 3.2. We denote the adaptive method Menon PilBoost, since we take inspiration from (Menon *et al.*, 2015), where they show that one can estimate the level of label noise (see their Appendix D.1) from the minimum and maximum posterior values.

Dataset	Algorithm	Feature Noise Twister			
		$p = 0$	0.15	0.25	0.5
xd6	AdaBoost	1.00 ± 0.00	0.99 ± 0.01	0.97 ± 0.01	0.88 ± 0.02
	us ($\alpha = 1.1$)	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01	0.91 ± 0.02
	us ($\alpha = 2.0$)	1.00 ± 0.00	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	0.91 ± 0.03
	us ($\alpha = 4.0$)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	$\mathbf{0.96 \pm 0.02}$
	XGBoost	1.00 ± 0.00	0.97 ± 0.02	0.96 ± 0.01	0.83 ± 0.03

Table 3.1: Accuracies of AdaBoost, PilBoost (for $\alpha \in \{1.1, 2, 4\}$), and XGBoost on the xd6 dataset affected by the feature noise twister with the flipping probability $p = \{0, 0.15, 0.25, 0.5\}$. All three algorithms were trained with depth 3 regression trees. For each value of α, we set $a_f = 8$. Note that the xd6 dataset is perfectly classified (when there is no twist) by a Boolean formula on the features, given in (Buntine and Niblett, 1992), which explains the performance when $p = 0$.

Using a single decision tree classifier with $O(\log(m))$ leaves and $O(\sqrt{m})$ samples per leaf ($m \approx 681$ examples for xd6 dataset with 70/30 train/test-split), and information gain as the splitting criterion, we estimate the minimum and maximum posterior values directly from the training data with local counts of number of samples classified such that $Y = 1$ at each leaf. Once we obtain η_{min} and η_{max} in this way, we estimate the symmetric noise value $p \in [0, 1]$ with the geometric mean $p = \sqrt{\eta_{\text{min}}(1 - \eta_{\text{max}})}$. Finally, to estimate α_0 for each noise level, we apply the formula in Lemma 9**(c)** and the SLN formula given just before Lemma 7 where we estimate η_{c} with the average posterior from the decision tree classifier. Further experimental consideration is given in Appendix B.2.6.

Feature Noise Twister (xd6 dataset): This twister perturbs the training sample by randomly flipping features. More precisely, for each training example, the example is selected if Ber(p_1) returns 1. Then, for each selected training example, and for each feature independently, the feature is flipped (the features of xd6 are Booleans) to the other symbol if Ber(p_2) also returns 1. Results on this twister are presented in Table 3.1 where $p_1 = p_2 = p$. In general, we find that PilBoost is more robust to the Feature Noise Twister than AdaBoost and XGBoost, and we find that α^* increases as the amount of twist increases.

Insider Twister (online shoppers intention dataset): This twister assumes more knowledge about the model than the previous two twisters. In essence, the insider twister adds noise to a few of the most informative features for predicting the class. Specifically for the online shoppers intention dataset, the insider twister adds noise to feature 8 (*page values* - numeric type with range in $[-250, 435]$), feature 10 (*month*), and feature 15 (*visitor type* - ternary alphabet). For *page values*, the insider twister adds i.i.d. $\mathcal{N}(0, 60)$ to the entries; for both *month* and *visitor type*, the insider twister independently increments (with probability $1/2$) the symbol according to their respective alphabets such that about 50% of each of these features are perturbed. Results on this twister are presented in Figure 3.3 and further discussion in Appendix B.2.4 (Figure B.10); post-twister, the feature importance profile of XGBoost is almost uniform, displaying damages to the algorithm's discriminative abilities (Figure 3.3, right), while the feature importance profile of PilBoost is much less perturbed overall.

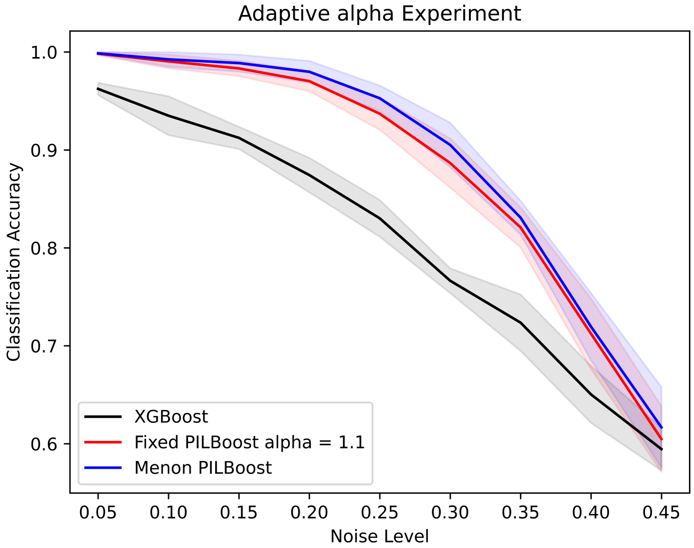


Figure 3.2: Adaptive α Experiment on the Xd6 Dataset with Depth 3 Regression Trees. Solid Curves Correspond to Mean Classification Accuracy and Shaded Areas Are the Associated 95% Confidence Intervals Obtained from a t-test. For Each Label Noise Value, We Train Three Algorithms: 1) Vanilla Xgboost; 2) PilBoost with Fixed $\alpha = 1.1$; 3) and, an Adaptive α PilBoost (We Refer to as Menon PilBoost). For Details Regarding Menon PilBoost, Refer to Class Noise Twister in the Main Body. The Result Suggests That a Fixed Value of $\alpha = 1.1$ in PilBoost is Good, but Approximating α_0 Does Induce Slightly Better Model Performance. For General Twists, We Suggest This Heuristic (or Some Variant) as Inspired By (Menon *et al.*, 2015) Could Be Used to Learn α_0. Further Experimental Consideration Is given in Appendix B.2.6.

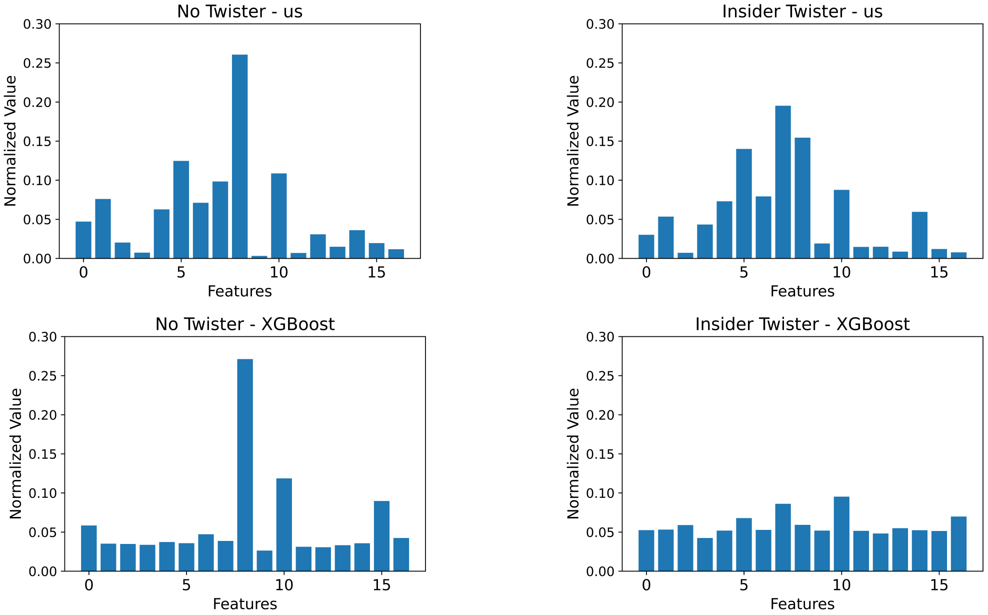


Figure 3.3: Normalized Feature Importance Profiles for PilBoost with $\alpha = 1.1$ and $a_f = 7$ (*Top*) and for Xgboost (*Bottom*) on the Online Shoppers Intention Dataset (Both for Depth 3 Trees) with and Without the Insider Twister. We Find That the Insider Twister Significantly Perturbs the Feature Importance of Xgboost as Evidenced in the Plot (*Far Right*), and Hence Significantly Reduces the Inferential Capacity of the Learned Model. More Details Can Be Found in Insider Twister (Main Body) and Appendix B.2.4.

Chapter 4

SMOOTHLY GIVING UP: ROBUSTNESS FOR SIMPLE MODELS

4.1 Introduction

In several critical infrastructure applications, simple models are favored over complex models. In health care analytics, simple models are typically preferred for their interpretability so that practitioners can audit the correlations the model uses for decision making (Rudin, 2019; Caruana *et al.*, 2015; Nori *et al.*, 2021; Chen *et al.*, 2021). In federated learning, simple models can be preferred for computational and energy efficiency, since edge devices are heterogeneous (Kairouz *et al.*, 2019b; Viola and Jones, 2001). Examples of learning algorithms that train simple models include logistic regression and boosting, particularly when the weak learner of the boosting algorithm is *weaker* (e.g., decision/regression trees with low maximum depth).

While simple models may offer more interpretability or energy efficiency, they are known to suffer, provably, from label noise (Ben-David *et al.*, 2012; Ji *et al.*, 2022; Rolnick *et al.*, 2017). Indeed, Long and Servedio (2010) showed that boosting algorithms that minimize convex losses over linear weak learners can achieve fair coin test accuracy after being trained with an arbitrarily small amount of (symmetric) label noise. In essence, Long and Servedio (2010) construct a pathological dataset which exploits the sensitivity of linear classifiers and the *inability* of convex losses to "give up" on noisy training examples, even if the convex boosting algorithm is regularized or stopped early.

Recent work argues that the negative result of Long and Servedio (2010) could perhaps be circumvented by increasing the complexity of the weak learner (Mansour *et al.*, 2022b), however, there are certain benefits for utilizing simple models. Thus, one remaining degree of freedom to robustly train a simple model is by tuning the loss function itself. We use the recently introduced margin-based α-loss, which smoothly tunes through the exponential ($\alpha = 1/2$), logistic ($\alpha = 1$), and sigmoid ($\alpha = \infty$) losses (Sypherd *et al.*, 2022b). The α hyperparameter controls the convexity of the loss, since for $0 < \alpha \leq 1$ the loss is convex, and for $\alpha > 1$ the loss is quasi-convex. We show that tuning $\alpha > 1$ allows the loss to "give up", which refers to how it evaluates large negative margins (preview Figure 4.1 and see the exponential vs. sigmoid losses). Hence, "giving up" on noisy training examples reduces the sensitivity of a simple hypothesis class to adverse perturbations.

Our contributions are as follows:

1. In Theorem 10, we show that there exist robust solutions of the margin-based α-loss for $\alpha > 1$ to the problem of Long and Servedio (2010); we verify this result with simulation (Figure 4.2) and experimental results (Section 4.5.1), where we show increased gains when the maximum depth of the (decision/regression) tree weak learner is restricted, i.e., for simpler models.

2. Building on the results in 1, we present a novel boosting algorithm (Algorithm 2 in Section 4.3.1), called AdaBoost.α, that may be of independent interest. The novelty of AdaBoost.α is that it smoothly tunes through vanilla AdaBoost (minimizing the exponential loss, $\alpha = 1/2$), LogAdaBoost (minimizing the logistic loss, $\alpha = 1$) (Schapire and Freund, 2013), to non-convex "AdaBoost-type" algorithms for $\alpha > 1$, all with the single α hyperparameter.

3. Noticing that the boosting setup of Long and Servedio (2010) ultimately reduces to a two-dimensional linear problem, we theoretically demonstrate robustness of the margin-based α-loss for $\alpha > 1$ under linear models of *arbitrary* dimensions

with an upperbound (Theorem 11) and dominating terms also appearing in a lowerbound (Theorem 12). In essence, we provide guarantees on the quality of optima, showing with upper and lower bounds on the noisy gradient that $\alpha > 1$ is smaller for "good solutions" than $\alpha \leq 1$.

4. Finally, in Section 4.5.2, we report experimental results on the logistic model for a synthetic Gaussian Mixture Model (GMM) and a COVID-19 survey dataset (Salomon *et al.*, 2021). In particular, we show that $\alpha > 1$ is able to preserve the interpretability of the linear model for the COVID-19 data, while also providing robustness to label noise. In addition, we provide straightforward heuristics for tuning α.

4.1.1 Related Work

Convex and Non-Convex Losses Label noise is an important problem (Frénay and Verleysen, 2013; Rauscher *et al.*, 2008; Gorber *et al.*, 2009), and many works propose reweighting/augmenting/regularizing/tuning *convex* loss functions to train robust models (Natarajan *et al.*, 2013; Ma *et al.*, 2020; Liu and Guo, 2020; Ghosh *et al.*, 2017; Patrini *et al.*, 2017a; Lee *et al.*, 2006; Lin *et al.*, 2017b; Leng *et al.*, 2022). Other approaches include abstention (Thulasidasan *et al.*, 2019; Ziyin *et al.*, 2020) and early stopping (Bai *et al.*, 2021), however, both techniques also typically revolve around a convex loss. Despite the fact that providing strong optimization guarantees for *non-convex* losses is nontrivial (Mei *et al.*, 2018), non-convex loss functions (satisfying certain basic conditions, e.g., differentiability, classification-calibration (Lin, 2004; Bartlett *et al.*, 2006b)) have been observed to provide superior robustness over convex losses (Beigman and Klebanov, 2009; Manwani and Sastry, 2013; Nguyen and Sanner, 2013; Barron, 2019; Zhang and Sabuncu, 2018a; Zhao *et al.*, 2010; Sypherd *et al.*, 2019; Chapelle *et al.*, 2008; Wu and Liu, 2007; Cheamanunkul *et al.*, 2014; Masnadi-Shirazi and Vasconcelos, 2009). Intuitively, non-convex loss functions seem to have a sophisticated regularization ability where they tend to assign less weight to misclassified training examples, and thus algorithms optimizing such losses are often less perturbed by outliers during training.

α-loss The α-loss, where $\alpha \in (0, \infty]$, arose in information theory (Liao *et al.*, 2018a; Arimoto, 1971b), and was recently introduced to ML (Sypherd *et al.*, 2019). It smoothly tunes through several important losses, and has statistical, optimization, and generalization tradeoffs dependent on α (Sypherd *et al.*, 2022b). Indeed, for shallow CNNs the α-loss is more robust for $\alpha > 1$, however, the loss becomes increasingly more non-convex as α increases greater than 1, hence an optimization/robustness tradeoff (Sypherd *et al.*, 2020). The α-loss is equivalent (under hyperparameter restriction) to the Generalized Cross Entropy loss (Zhang and Sabuncu, 2018a), which was motivated by the Box-Cox transformation in statistics (Box and Cox, 1964). Also, the α-loss was shown to satisfy a statistical notion of robustness for loss functions in the class probability estimation setting (Sypherd *et al.*, 2022c).

Convex and Non-Convex Boosting AdaBoost (Freund and Schapire, 1997b) (which minimizes the exponential loss (Schapire and Freund, 2013)) is the groundbreaking convex boosting algorithm. Later, the LogAdaBoost (which minimizes the logistic loss) was proposed as a more robust convex variant (Collins *et al.*, 2002; McDonald *et al.*, 2003). Indeed, a SOTA boosting algorithm, XGBoost, minimizes (an approximated) logistic loss, rather than the exponential loss (Chen and Guestrin, 2016). Sypherd *et al.* (2022c) recently introduced a novel boosting algorithm called PILBoost, which minimizes a *convex* (proper) surrogate approximation of the α-loss (Nock and Williamson, 2019; Reid and Williamson, 2010a), and presented experimental results on the robustness of PILBoost.

However, the seminal work of Long and Servedio (2010) showed that convex boosters provably suffer from label noise, particularly for simple weak learners (Mansour *et al.*, 2022b). Van Rooyen *et al.* (2015) proposed relaxing the nonnegativity condition of the convex loss in order to yield robustness, but it seems that this is unable to completely fix the problem (Long and Servedio, 2022). For this reason, non-convex boosting algorithms have been considered before (Masnadi-Shirazi and Vasconcelos, 2009; Cheamanunkul *et al.*, 2014), but there remains a large gap between the convex and non-convex realms. Therefore, we propose using the margin-based α-loss, which continuously tunes through several canonical convex and quasi-convex losses, in order to smoothly perform non-convex boosting.

4.2 Preliminaries

4.2.1 Margin-Based α-loss

We consider the setting of binary classification. The learner ideally wants to output a classifier $\overline{H} : \mathcal{X} \to \{-1, +1\}$ that minimizes the probability of error, the expectation of the 0-1 loss, however, this is NP-hard (Ben-David *et al.*, 2003). Thus, the problem is relaxed by optimizing a surrogate to the 0-1 loss over functions $H : \mathcal{X} \to \mathbb{R}$, whose output captures the certainty of prediction of the binary label $Y \in \{-1, 1\}$ associated with the feature vector $X \in \mathcal{X}$ (Bartlett *et al.*, 2006b). The classifier is obtained by making a hard decision, i.e., $\overline{H}(X) = \text{sign}(H(X))$. A surrogate loss is said to be margin-based if, the loss associated to a pair $(y, H(x))$ is given by $\tilde{l}(yH(x))$ for $\tilde{l} : \mathbb{R} \to \mathbb{R}_{+}$ (Lin, 2004). The loss of the pair $(y, H(x))$ only depends on the product $z := yH(x)$, i.e., the (unnormalized) margin (Schapire and Freund, 2013). A negative margin corresponds to a mismatch between the signs of $H(x)$ and y, i.e., a classification error by H; a positive margin corresponds to a correct classification by

H.

Since probabilities are typically the inputs to loss functions (e.g., log-loss, Matusita's loss (Matusita, 1956), α-loss (Sypherd *et al.*, 2019)), an important function we use is the sigmoid function $\sigma : \mathbb{R} \to [0,1]$, given by

$$\sigma(z) := \frac{1}{1+e^{-z}}, \tag{4.1}$$

where $z := yH(x)$ is the margin. The sigmoid smoothly maps real-valued predictions $H : \mathcal{X} \to \mathbb{R}$ to probabilities, and in the multiclass setting, the sigmoid is generalized by the softmax function (Goodfellow *et al.*, 2016). Note that the inverse of σ is the logit link (Reid and Williamson, 2010a). Noticing that $\sigma(-z) = 1 - \sigma(z)$, we have that

$$\sigma'(z) := \frac{d}{dz}\sigma(z) = \sigma(z)\sigma(-z) = \frac{e^z}{(1+e^z)^2}, \tag{4.2}$$

and note that σ' is an even function.

We now provide the definition of the margin-based α-loss, which was first presented in (Sypherd *et al.*, 2019) for $\alpha \in [1,\infty]$ and extended to $\alpha \in (0,\infty]$ (Sypherd *et al.*, 2022b).

Definition 8. *The margin-based α-loss, $\tilde{l}^\alpha : \mathbb{R} \to \mathbb{R}_+$, $\alpha \in (0,\infty]$, is given by, for $\alpha \in (0,1) \cup (1,\infty)$,*

$$\tilde{l}^\alpha(z) := \frac{\alpha}{\alpha-1}\left(1 - \sigma(z)^{1-1/\alpha}\right), \tag{4.3}$$

with $\tilde{l}^1(z) := -\log \sigma(z)$ and $\tilde{l}^\infty(z) := 1 - \sigma(z)$ by continuous extension, and note that $\tilde{l}^{1/2}(z) := e^{-z}$.

Indeed, $\tilde{l}^{1/2}$, $\tilde{l}^1$, and $\tilde{l}^\infty$ recover the exponential (AdaBoost), logistic (logistic regression), and sigmoid (smooth 0-1) losses, respectively (Shalev-Shwartz and Ben-David, 2014); see Figure 4.1(a) for a plot of $\tilde{l}^\alpha$ for several values of α versus the margin. Note that for fixed $z \in \mathbb{R}$, $\tilde{l}^\alpha(z)$ is continuous in α. Sypherd *et al.* (2022b) showed

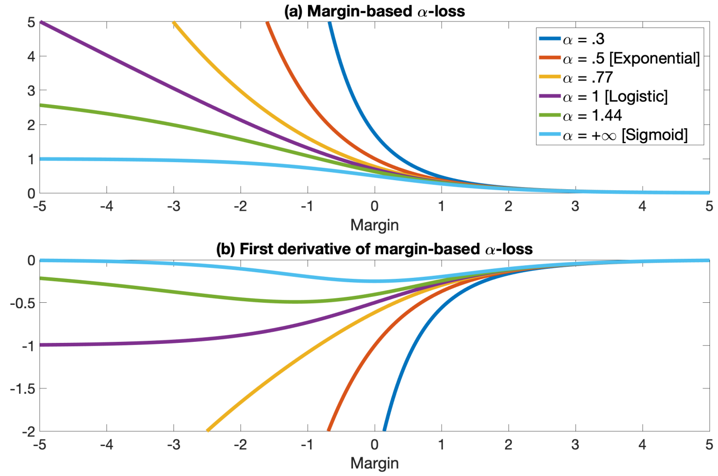


Figure 4.1: (a) Margin-based α-loss (4.3) as a Function of the Margin ($z := yH(x)$) for $\alpha \in \{.3, .5, .77, 1, 1.44, \infty\}$; (b) Its First Derivative (See Lemma 20 in Appendix C.1) with Respect to the Margin for the Same Set of α. The "Giving Up" Ability of the Margin-based α-loss for $\alpha > 1$ Can Be Seen from Its First Derivative, Where It Is More Constrained (than $\alpha \leq 1$) for Large Negative Values of the Margin.

that the margin-based α-loss is classification-calibrated for all $\alpha \in (0, \infty]$ (Bartlett *et al.*, 2006b). Thus, tuning the single α hyperparameter allows continuous interpolation through calibrated, important loss functions, however, different regimes of α have differing robustness properties. To this end, Sypherd *et al.* (2022b) presented the following result regarding the convexity characteristics of $\tilde{l}^{\alpha}$.

Proposition 7. $\tilde{l}^{\alpha} : \mathbb{R} \to \mathbb{R}_{+}$ *is convex for* $0 < \alpha \leq 1$ *and quasi-convex for* $\alpha > 1$.

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is quasi-convex if, for all $x, y \in \mathbb{R}$ and $\lambda \in [0, 1]$, $f(\lambda x + (1 - \lambda)y) \leq \max\{f(x), f(y)\}$, and also that any monotonic function is quasi-convex (cf. (Boyd and Vandenberghe, 2004a)).

In light of Proposition 7, consider Figure 4.1(a) for $\alpha = 1/2$ (convex) and $\alpha = 1.44$ (quasi-convex), and suppose for concreteness that $z_1 = -1$ and $z_2 = -5$. The difference in loss evaluations for these two negative values of the margin, which are representative of misclassified training examples, is approximately exponential vs. sublinear; this is similarly observed in Figure 4.1(b) with the first derivative of $\tilde{l}^{\alpha}$ (see Lemma 20 in Appendix C.1). Intuitively, if a training example is not fit well by the currently learned parameter values, then its margin will be (large and) negative and it will incur more derivative update; if such a training example is noisy, convex losses (e.g., $\alpha \leq 1$) encourage the algorithm to continue fitting the bad example, whereas non-convex losses (e.g., $\alpha > 1$) would instead allow the algorithm to "give up". This tendency of convex losses could be exacerbated for simpler models because they can suffer significant perturbation by label noise (preview Figure 4.2) vs. more nuanced function classes (Rolnick *et al.*, 2017).

4.2.2 Boosting Setup

For the boosting context, we assume access to a training sample $\mathcal{S} := \{(x^i, y^i), i \in [m]\} \subset \mathcal{X} \times \{-1, +1\}$ of m examples, where $[m] := \{1, 2, ..., m\}$. Following the functional gradient perspective of boosting (i.e., the blueprint of (Friedman, 2001)), the boosting algorithm minimizes a margin-based loss $\tilde{l}$ with respect to $\mathcal{S}$ over $t \in [T]$ iterations in order to learn a function $H_T : \mathcal{X} \to \mathbb{R}$, given by

$$H_T(\cdot) := \sum_{t \in [T]} \theta_t h_t(\cdot), \tag{4.4}$$

where θ_t are the learned parameters and the $h_t : \mathcal{X} \to \mathbb{R}$ are weak learners with slightly better than random classification accuracy. On each iteration $t \in [T]$, we compute weights for each training example using the full H_{t-1} via

$$D_t(i) := -\tilde{l}'(y^i H_{t-1}(x^i)), \forall i \in [m]. \tag{4.5}$$

The weights $D_t(i)$ are non-negative, normalized to form a distribution over the training examples, and tend to increase for an example that is incorrectly predicted (negative margin) by the previously learned H_{t-1}. Thus, weighting puts emphasis on "hard" examples using the first derivative of the loss function, which is a kind of functional gradient descent (cf. (Schapire and Freund, 2013)). Then, the distribution over training examples D_t is passed to the weak learning oracle (see Algorithm 2 for the general procedure).

In the next section, we show that using the derivative of the margin-based α-loss in (4.5) recovers a novel robust boosting algorithm, which may be of independent interest. We also show that this algorithm has provable robustness guarantees on the negative result of Long and Servedio (2010).

4.3 Robustness for Boosting

4.3.1 AdaBoost.α: Boosting with a Give Up Option

Using the smooth tuning of the margin-based α-loss, we present a novel robust boosting algorithm, AdaBoost.α in Algorithm 2, which is obtained by noticing (from the functional gradient perspective (Schapire and Freund, 2013)) that the exponential weighting of vanilla AdaBoost is really the negative first derivative of the exponential loss (i.e., $\alpha = 1/2$). Generalizing this observation for all $\alpha \in (0, \infty]$ (via Lemma 20 in Appendix C.1) in (4.6), we obtain a hyperparameterized family of "AdaBoost-type" algorithms.

Indeed, AdaBoost.α also recovers LogAdaBoost (see Section 4.1.1) for $\alpha = 1$. For $\alpha > 1$, AdaBoost.α becomes a non-convex boosting algorithm minimizing the quasi-convex margin-based α-losses (Proposition 7). As argued in Section 4.2.1, non-convex losses enable the boosting algorithm to give up on noisy examples, and hence yield a more robust model H_T. Indeed, for these same robustness reasons, non-convex boosting algorithms have been considered before (see Section 4.1.1). However, the novelty

Algorithm 2 AdaBoost.α

```
Given: $(x^1, y^1), \ldots, (x^m, y^m)$ where $x^i \in \mathcal{X}$, $y^i \in \{-1, +1\}$, and $\alpha \in (0, \infty]$
Initialize: $H_0 = 0$.
for $t = 1, 2, \ldots, T$:
    Update, for $i = 1, \ldots, m$:
        $$D_t(i) = \frac{\sigma'(y^i H_{t-1}(x^i))\sigma(y^i H_{t-1}(x^i))^{-\frac{1}{\alpha}}}{\mathcal{Z}_t}, \tag{4.6}$$
    where $\mathcal{Z}_t$ is a normalization factor.
    Return $h_t$, weakly learned on $D_t$.
    Compute error of weak hypothesis $h_t$:
        $$\epsilon_t = \sum_{i: h_t(x^i) \neq y^i} D_t(i). \tag{4.7}$$
    Let $\theta_t = \frac{1}{2} \log \left( \frac{1-\epsilon_t}{\epsilon_t} \right)$.
    Update: $H_t = H_{t-1} + \theta_t h_t$
Return $\overline{H}(\cdot) = \mathrm{sign}\left(H_T(\cdot)\right)$
```

of AdaBoost.α is that it continuously interpolates through convex AdaBoost variants ($\alpha \leq 1$) to non-convex "AdaBoost-type" algorithms ($\alpha > 1$). Thus, AdaBoost.α allows the practitioner or meta-algorithm (He *et al.*, 2021) to tune how much one would like the algorithm to give up on hard, possible noisy, training examples, which may be useful in a distributed context (Cooper and Reyzin, 2017).

4.3.2 Robustness on the Long-Servedio Dataset

In Long and Servedio (2010), the training sample $\mathcal{S}$ is a multiset consisting of three distinct examples, one of which is repeated twice, where the data margin $0 < \gamma < 1/6$:

- $\mathcal{S}$ contains one copy of the example $x = (1, 0)$ with label $y = +1$. (Called the

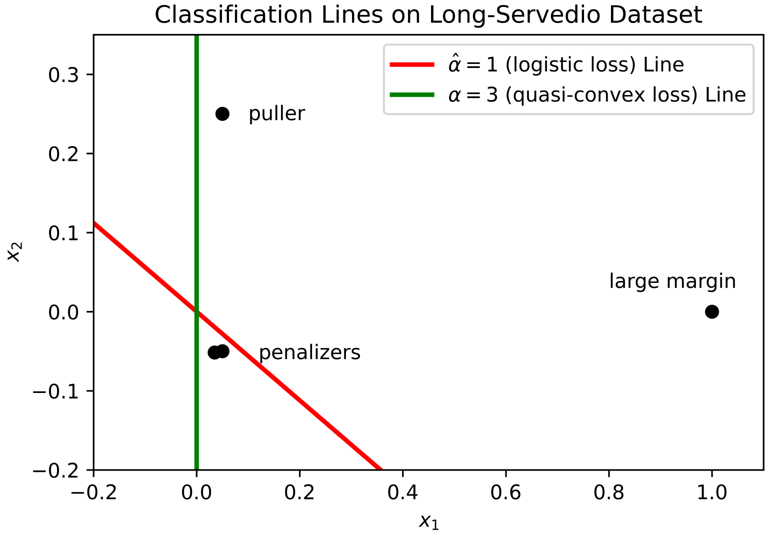


Figure 4.2: A Plot Depicting Optimal Classification Lines of $\hat{\alpha} = 1$ and $\alpha = 3$ for the *Clean* Long-servedio Dataset $\mathcal{S}$, Where the Penalizer Examples Are Slightly Separated for Display. The $\hat{\alpha}, \alpha$ Optima Are Obtained by Grid-search on the Noisy Long-servedio Dataset $\hat{\mathcal{S}}$, Where the Noise Level Is Chosen as $p = 1/3$, and $\gamma_{\hat{\alpha}} = 1/20$ Is Subsequently Chosen for the Negative Result Of Long and Servedio (2010) to "kick-in". The $\hat{\alpha} = 1$ (Logistic Loss) Line (Red) Is given by $(\theta_1^{\hat{\alpha}}, \theta_2^{\hat{\alpha}}) = (.79, 1.41)$ For (4.8), and Has Fair Coin Accuracy on $\mathcal{S}$, Misclassifying Both Penalizers. The $\alpha = 3$ (Quasi-convex Loss) Line (Green) Is given by $(\theta_1^{\alpha}, \theta_2^{\alpha}) = (41.59, -1.19 \times 10^{-11})$, and Has Perfect Accuracy on $\mathcal{S}$. This Simulation Aligns with Theorem 10 in That the Quasi-convex $\alpha = 3$ Loss Is Able to "give Up" on the Noisy Copies of the Training Examples and Recover Perfect Classification Parameters. More α's Are Presented in Appendix C.1.1.

"large margin" example.)

- $\mathcal{S}$ contains two copies of the example $x = (\gamma, -\gamma)$ with label $y = +1$. (Called the "penalizers" since these are the points that the booster will misclassify.)

- $\mathcal{S}$ contains one copy of the example $x = (\gamma, 5\gamma)$ with label $y = +1$. (Called the "puller".)

Thus, all four examples in $\mathcal{S}$ have positive label and lie in the unit disc $\{x : \|x\| \leq 1\}$; see Figure 4.2 for a plot of the dataset. Notice that $\overline{H}(x) = \text{sign}(x_1)$ (sign of first coordinate of x) corrrectly classifies all four examples in $\mathcal{S}$ with margin $\gamma > 0$, so the weak learner hypothesis class $\mathcal{H} = \{h_1(x) = x_1, h_2(x) = x_2\}$ is sufficient for perfect classification of the dataset. The task for the boosting algorithm is to learn parameters (θ_1, θ_2) such that, from (4.4),

$$H_{\tilde{l},\gamma}(x_1, x_2) := \theta_1 x_1 + \theta_2 x_2, \tag{4.8}$$

achieves perfect classification accuracy on $\mathcal{S}$, where the dependency on the loss $\tilde{l}$ and data margin γ is clear. Note that (4.8) (we abbreviate $H_{\tilde{l},\gamma} = (\theta_1, \theta_2)$) is a 2D linear model, so this setup parallels with logistic regression, which we consider in the sequel. Following (Mansour *et al.*, 2022b), we obtain a noisy sample $\hat{\mathcal{S}}$ with label flip probability $0 < p < 1/2$ by including $p^{-1} - 1$ copies of $\mathcal{S}$ and 1 copy of $\mathcal{S}$ with the labels flipped. Long and Servedio (2010) showed that for any calibrated, *convex* loss $\tilde{l}$:

- When $p = 0$, i.e., the training sample is $\mathcal{S}$, the optimal $H_{\tilde{l}} = (\theta_1^{\tilde{l}}, \theta_2^{\tilde{l}})$ of $\tilde{l}$ has *perfect* accuracy on $\mathcal{S}$.

- For any $0 < p < 1/2$ generating training sample $\hat{\mathcal{S}}$, there exists $0 < \gamma_{\tilde{l}} < 1/6$ such that the optimal $H_{\tilde{l},\gamma_{\tilde{l}}} = (\theta_1^{\tilde{l}}, \theta_2^{\tilde{l}})$ of $\tilde{l}$ has *fair coin* accuracy on $\mathcal{S}$.

Intuitively, the interplay between the "large margin" and "puller" examples forces a convex booster, boosting $\mathcal{H}$, to try to fit the noisy examples in $\hat{\mathcal{S}}$; this holds even if the booster is regularized or stopped early, ultimately outputing a model that misclassifies both "penalizers" of $\mathcal{S}$ (Long and Servedio, 2010). Taking stock with $\tilde{l}^{\alpha}$, we see that this pathology holds for $\alpha \leq 1$, since these are convex losses. However, tuning $\alpha > 1$ to quasi-convex losses is able to induce the existence of optima which can fix the problem.

Theorem 10. *Let* $0 < p < 1/2$ *for* $\hat{\mathcal{S}}$, *and* $\hat{\alpha} \leq 1$ *for* $\tilde{l}^{\alpha}$. *By Long and Servedio (2010), there exists* $0 < \gamma_{\hat{\alpha}} < 1/6$ *such that the optimal* $H_{\hat{\alpha},\gamma_{\hat{\alpha}}} = (\theta_1^{\hat{\alpha}}, \theta_2^{\hat{\alpha}})$ *is a fair coin on* $\mathcal{S}$. *On the other hand, for* $\alpha \in (1, \infty)$, $\tilde{l}^{\alpha}$ *has optimum* $H_{\alpha,\gamma_{\hat{\alpha}}} = (\theta_1^{\alpha}, \theta_2^{\alpha})$, *where* $\theta_1^{\alpha} = \mathcal{O}\left(\alpha\gamma_{\hat{\alpha}}^{-1} \log\left(p^{-1} - 1\right)\right)$ *and* $\theta_2^{\alpha} = 0$, *with perfect classification accuracy on* $\mathcal{S}$.

The proof of Theorem 10 (in Appendix C.1.1) is nontrivial since $\alpha > 1$ has a non-convex optimization landscape. In Figure 4.2 where $p = 1/3$ and $\gamma_{\hat{\alpha}} = 1/20$, the grid search returns $(\theta_1^{\alpha}, \theta_2^{\alpha}) = (41.59, -1.19 \times 10^{-11})$, which aligns with Theorem 10, namely that $\theta_1^{\alpha} \approx 3 \times 20 \times \log(2) \approx 41.59$ and $\theta_2^{\alpha} \approx 0$. Intuitively, increasing $\alpha \in (1, \infty)$ increases θ_1^{α}, which may have practical utility (see Section 4.5.1), but the rate for θ_1^{α} hints at why $\alpha = \infty$ is not included, since $\alpha = \infty$ "pushes" θ_1^{α} to ∞, an impossibility; this is an example of the robustness/optimization complexity tradeoff inherent in the margin-based α-loss (Sypherd *et al.*, 2020).

4.4 Robustness for Linear Models

Taking inspiration from the boosting setup in Section 4.3.2, where the weak learner recovered a 2D linear model in (4.8), we now consider a generalization of that hypothesis class to $d \in \mathbb{N}$ dimensions, which is equivalent to the logistic model (Sypherd *et al.*, 2022b). Similar to Theorem 10, we provide guarantees on the quality of optima, showing with upper and lower bounds that the noisy gradient for $\alpha > 1$ is smaller for "good solutions" than when $\alpha \leq 1$.

We let $X \in [0,1]^d$ be the normalized feature vector, $Y \in \{-1,+1\}$ the label, and we assume that the pair is drawn according to an unknown distribution $P_{X,Y}$. We assume that the parameter vector $\theta \in \mathbb{B}_d(r)$ where $r > 0$ and $\mathbb{B}_d(r) := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq r\}$. Thus, in this setting $\langle yx, \theta\rangle$ (inner product) is the margin, and note by the Cauchy-Schwarz inequality that $\langle yx, \theta\rangle \leq r\sqrt{d}$. Also, note that for $\alpha = 1$, we recover logistic regression.

For $\alpha \in (0,\infty]$, the expected margin-based α-loss, abbreviated the α-risk, evaluated at $\theta \in \mathbb{B}_d(r)$ is given by

$$R_\alpha(\theta) := \mathbb{E}_{X,Y}\left[\tilde{l}^\alpha(\langle YX, \theta\rangle)\right], \tag{4.9}$$

and for symmetric label noise rate $0 < p < 1/2$,

$$R_\alpha^p(\theta) := \mathbb{E}_{X,Y}\left[\mathbb{E}_{\tau\sim\mathrm{Rad}(p)}\left(\tilde{l}^\alpha(\langle -\tau YX, \theta\rangle)\right)\right], \tag{4.10}$$

is called the noisy α-risk, where τ is a Rademacher random variable with parameter p. In order to assess the efficacy of a given parameter vector $\theta \in \mathbb{B}_d(r)$, we are interested in the gradient of the loss function, due to the use of gradient methods for optimization (Boyd and Vandenberghe, 2004a). Thus, the gradient of the α-risk in (4.9) is

$$\nabla_\theta R_\alpha(\theta) := \mathbb{E}_{X,Y}\left[\nabla_\theta \tilde{l}^\alpha(\langle YX, \theta\rangle)\right], \tag{4.11}$$

$\nabla_\theta \tilde{l}^\alpha(\langle YX, \theta\rangle) := -\sigma'(\langle YX, \theta\rangle)\sigma(\langle YX, \theta\rangle)^{-\frac{1}{\alpha}}YX$ for $\alpha \in (0,\infty]$ from Lemma 20 in Appendix C.1. Hence, the gradient of the noisy α-risk (4.10) is given by

$$\nabla_\theta R_\alpha^p(\theta) := \mathbb{E}_{X,Y}\left[\mathbb{E}_{\tau\sim\mathrm{Rad}(p)}\left(\nabla_\theta \tilde{l}^\alpha(\langle -\tau YX, \theta\rangle)\right)\right]. \tag{4.12}$$

We now present a result in the realizable setting, indicating (4.12) is smaller for $\alpha = \infty$ (soft 0-1 loss) at the data generating vector $\theta^* \in \mathbb{B}_d(r)$ than for $\alpha = 1$ (logistic loss).

Theorem 11. *Let $0 < p < 1/2$ and let $\hat{\theta}^1, \hat{\theta}^\infty \in \mathbb{B}_d(r)$ be such that $\nabla_\theta R_1^p(\hat{\theta}^1) = \mathbf{0}$ and $\nabla_\theta R_\infty^p(\hat{\theta}^\infty) = \mathbf{0}$. We assume that there exists $\theta^* \in \mathbb{B}_d(r)$, such that the following ordering holds for all $(x, y) \in \mathcal{X} \times \{-1, +1\}$,*

$$\langle yx, \theta^* \rangle \geq \langle yx, \hat{\theta}^\infty \rangle \geq \langle yx, \hat{\theta}^1 \rangle > \ln(2 + \sqrt{3}). \tag{4.13}$$

Then, we have that for $\alpha = 1$ or ∞,

$$\frac{\|\nabla_\theta R_\alpha^p(\theta^*)\|_\infty}{C_\alpha} \leq d^{\frac{1}{2}} r \left|\tilde{l}^{\alpha''}(z_\alpha^*)\right| + dr^2 \left|\tilde{l}^{\alpha'''}(z_\alpha^*)\right|, \tag{4.14}$$

where $C_\alpha = 2$ for $\alpha = 1$ and $C_\alpha = 2-4p$ for $\alpha = \infty$, and $z_\alpha^* := \arg\max_{z \in \{\langle yx, \hat{\theta}^\alpha \rangle\}} \left|\tilde{l}^{\alpha''}(z)\right|$. Furthermore,

$$1 - 2p < \frac{d^{\frac{1}{2}} r \left|\tilde{l}_1^{1''}(z_1^*)\right| + dr^2 \left|\tilde{l}^{1'''}(z_1^*)\right|}{d^{\frac{1}{2}} r \left|\tilde{l}^{\infty''}(z_\infty^*)\right| + dr^2 \left|\tilde{l}^{\infty'''}(z_\infty^*)\right|}. \tag{4.15}$$

Theorem 11 uses symmetries of the first derivative of $\tilde{l}^\alpha$ for $\alpha = 1$ and ∞; see Appendix C.1.2 for proof details. Intuitively, (4.15) indicates that there is a significant discrepancy between the two upper bounds as the noise rate $p \to 1/2$, and the assumption in (4.13) is mild because all three vectors are assumed to have perfect accuracy on the clean data.

In support of the upper bounds in Theorem 11, we now present a uniform lower bound on the norm of (4.12) for the skew-symmetric family of distributions (e.g., a GMM).

Theorem 12. *Let $0 < p < 1/2$, and for each $y \in \{-1, 1\}$, let $X^{[y]}$ have the distribution of X conditioned on $Y = y$. We assume a skew-symmetric distribution, namely, that $X^{[1]} \overset{\mathrm{d}}{=} -X^{[-1]}$, and $\mathbb{E}[X^{[1]}] \neq \mathbf{0}$. We also assume that $r > 0$ is small enough such that both of the following hold:*

$$(1 - 2p)(1 - \sigma'(r\sqrt{d})) < \frac{\|\mathbb{E}(X^{[1]})\|_2}{\mathbb{E}(\|X^{[1]}\|_2)}, \tag{4.16}$$

and, for all $\alpha \in [1, \infty]$*,*

$$e^{\frac{r\sqrt{d}}{\alpha}} \log\left(e^{r\sqrt{d}} + 1\right) < \left(p^{-1} - 1\right) \log\left(e^{-r\sqrt{d}} + 1\right). \tag{4.17}$$

Then, we have that for every $\theta \in \mathbb{B}_d(r)$*,*

$$\|\nabla_\theta R_\alpha^p(\theta)\|_2 \geq \|\mathbb{E}[X^{[1]}]\|_2 - \gamma \mathbb{E}[\|X^{[1]}\|_2] > 0, \tag{4.18}$$

where (letting $\tilde{\gamma} := \sigma(r\sqrt{d})^{1-\frac{1}{\alpha}} \sigma(-r\sqrt{d}) - 1$*)*

$$\gamma := \begin{cases} \sigma(r\sqrt{d}) - p & \alpha = 1 \\ p\tilde{\gamma} - (1-p)\tilde{\gamma} & \alpha \in (1, \infty) \\ (1-2p)(1 - \sigma'(r\sqrt{d})) & \alpha = \infty, \end{cases} \tag{4.19}$$

and γ *is monotonically increasing in* $\alpha \in [1, \infty]$*.*

The proof of Theorem 12 (in Appendix C.1.3) is inspired by the Morse landscape analysis in (Sypherd *et al.*, 2019). Intuitively, (4.19) implies that the RHS in (4.18) is monotonically *decreasing* in $\alpha \in [1, \infty]$, which aligns with the ordering given by the upper bounds in Theorem 11. Regarding the assumptions in (4.16) and (4.17), they are both more easily satisfied for smaller $r > 0$, indicating alignment with the underlying optimization landscape phenomena. Taken together, Theorems 11 and 12 suggest that larger $\alpha > 1$ are more robust than $\alpha = 1$ (logistic regression); also, note the $1 - 2p$ coefficient for $\alpha = \infty$ in both bounds.

4.5 Experiments

We now provide empirical results in support of the previous sections, namely the efficacy of AdaBoost.α (Algorithm 2) on the Long-Servedio dataset and the robustness of the margin-based α-loss (Definition 8) in linear models, both for $\alpha > 1$. Further details and results are in Appendix C.2.

4.5.1 Boosting

For the boosting experiments, we utilize the *experiment version* of the Long-Servedio dataset (Long and Servedio, 2010; Cheamanunkul *et al.*, 2014), where the feature vectors are 21D, which differs from the theory version presented in Section 4.3.2, where the feature vectors are 2D. A full description of the dataset is presented in Appendix C.2.1. We introduce symmetric label noise in the training data with flip probability $0 < p < 1/2$.

Robustness for simple models In Figure 4.3, we report results of AdaBoost.α with $\alpha > 1$ (quasi-convex) vs. SOTA convex boosters: vanilla AdaBoost (AdaBoost.α with $\alpha = 1/2$), LogAdaBoost (AdaBoost.α with $\alpha = 1$), XGBoost, and PILBoost (see Section 4.1.1). For lower maximum tree depth of the weak learner (i.e., simpler models), $\alpha > 1$ boosters are better able to "give up" on the noisy labels during training and the learned model yields better accuracy on the *clean* test set, aligning with Theorem 10. When the maximum depth is increased, all of the algorithms perform roughly the same (Mansour *et al.*, 2022b).

Giving up In Figure 4.4, we plot the clean test accuracy of AdaBoost.α boosting decision stumps for several values of α versus iterations (i.e., number of weak learners). We see that for $\alpha \leq 1$, increasing iterations does not increase accuracy; however, the $\alpha > 1$ (non-convex) boosters continue "giving up" on the noisy training examples, resulting in a $\approx 25\%$ gain. For the large $\alpha > 1$, i.e. $\alpha = 8$ or 20, the confidence intervals widen, which is an example of the robustness/non-convexity tradeoff inherent in the α hyperparameter (Sypherd *et al.*, 2020).

Smooth tuning It is not difficult to tune α for AdaBoost.α, see Figure C.8 in Appendix C.2.1. Sypherd *et al.* (2022b) indicated that the effective range of α is typically bounded, e.g., $\alpha^* \in [.8, 8]$ for shallow CNNs; AdaBoost.α appears to be no

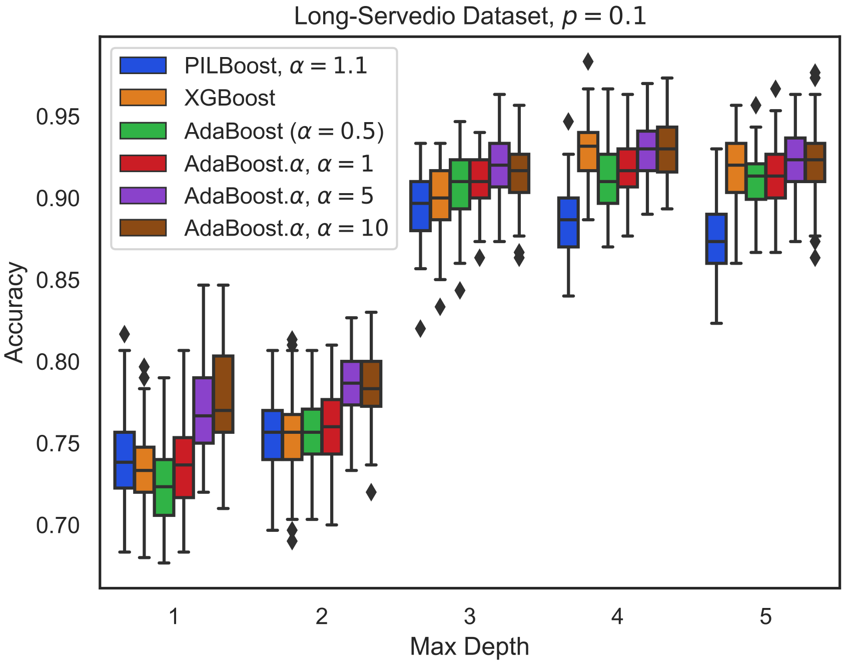


Figure 4.3: Box and Whisker Plots of the Clean Test Accuracies of Several Boosters with 100 Decision Trees of Varying Maximum Depth on the Long-servedio Dataset for $p = .1$ Symmetric Label Noise. The Boxes Are the Interquartile Ranges, the Lines in the Boxes Are the Medians, and the Diamonds Are the Outliers. Note That Adaboost.α with $\alpha > 1$ (Quasi-convex), Outperforms the Convex Boosters When the Maximum Depth Is 1 or 2. Further Commentary Is in Section 4.5.1, and More Noise Levels Are in Appendix C.2.1.

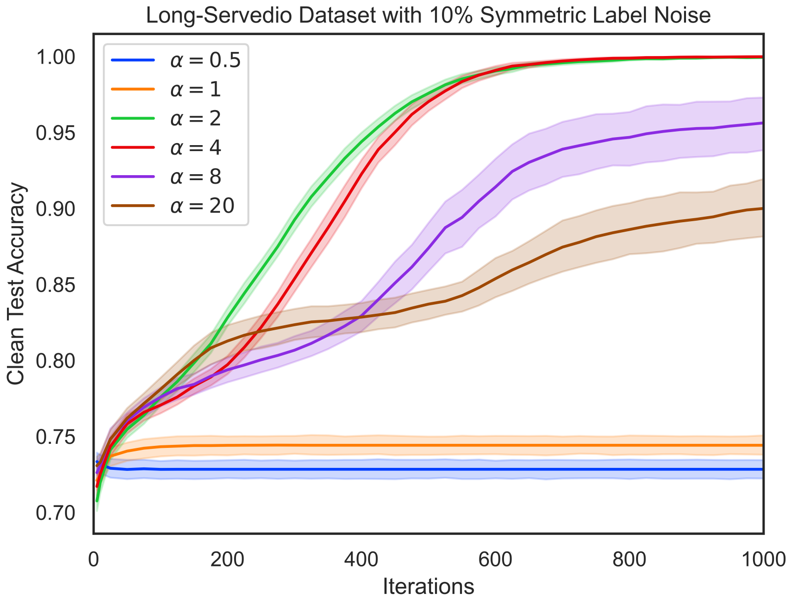


Figure 4.4: We Plot Clean Test Accuracies Vs. the Number of Iterations of Adaboost.α Boosting Decision Stumps for Several Values of α on the Long-servedio Dataset with $p = .1$ Symmetric Label Noise. Note That the Solid Curves Correspond to Mean Accuracy and Shaded Areas Are the Associated 95% Confidence Intervals (from 80 Runs of the Experiment). This Result Reflects the Tendency of the Convex $\alpha \leq 1$ Boosters to Continue Overfitting on the Noisy Training Examples, and the Ability of the Non-convex $\alpha > 1$ Boosters to Continue Judiciously "giving Up" on the Noisy Training Examples. Further Commentary Is in Section 4.5.1, and More Noise Levels Are in Appendix C.2.1.

different. In part, this is due to a *saturation effect*, where $\alpha > 1$ quickly "resembles" the ∞-loss (Sypherd *et al.*, 2020). Hence, tuning $\alpha > 1$, but not too large, trades a reasonable amount of non-convexity for robustness.

In Appendix C.2.1, we also present results of AdaBoost.α on the breast cancer dataset (Wolberg *et al.*, 1995), similarly observing gains for smaller maximum tree depths.

4.5.2 Linear Model

For the linear model experiments, we consider two datasets: a 2D GMM, and a real-world COVID-19 survey dataset (Salomon *et al.*, 2021). We introduce symmetric label noise into the training data for both.

For the effectiveness metric of using the margin-based α-loss, we consider the model parameters themselves, as they have clear interpretations in the form of odds ratios for the linear setting. Specifically, we examine a linear classifier trained with α-loss on noisy data and calculate the mean squared error (MSE) of its learned parameters and those of some baseline (further described for each dataset). By ensuring that the model parameters are close to those of a clean model, we preserve interpretability and accuracy.

2D GMM We first consider a 2D GMM with $\mu_1 = (1, 1) = -\mu_{-1}$, identity covariance, and $\mathbb{P}[Y = 1] = 0.14$ (aligning with the next experiment). Thus, the Bayes-optimal classifier is linear, and we compare with the separator learned by training α-loss on noisy data. In Figure 4.5, we see that tuning $\alpha > 1$ results in a decreased MSE for every non-zero noise level, and implies that the model learned by $\alpha > 1$ is closer to the Bayes optimal line than the model learned by $\alpha \leq 1$, aligning with Theorems 11 and 12. Tuning on this simple dataset is quite easy as the MSE is fairly flat for $\alpha > 1$, see Appendix C.2.2 for more details.

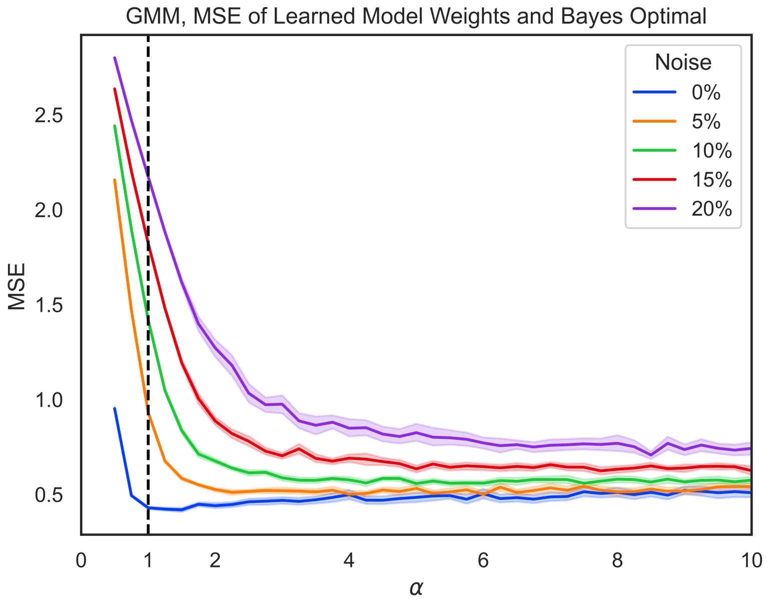


Figure 4.5: Mse of Bayes Optimal Line and the Parameters Learned by α-loss, on a 2d Gmm with 86 : 14 Class Imbalance and Varying Label Noise Levels. We See That $\alpha > 1$ Is Able to More Closely Approximate the Clean Parameters than $\alpha \leq 1$, and the Mse Is Fairly Flat in the Large α Regime, Indicating That It Is Not Difficult to Tune α. Note That the 95% Confidence Intervals Grow Wider for Larger α, Indicative of the Optimization/Robustness Tradeoff (Sypherd *et al.*, 2020).

COVID-19 survey data We now consider the US COVID-19 Trends and Impact Survey (US CTIS) dataset (Salomon *et al.*, 2021), which consists of self-reported survey data. We compress the dataset from 71 features to 42 categorical and real-valued features including symptom data, behaviors, and comorbidities. For simplicity and interpretability, 8 features, listed in Table 4.1, were chosen using cross validation which contributed the most to the final prediction (largest odds ratios). Each example is labeled either as RT-PCR-confirmed COVID positive (1) or negative (-1), based on self-reported diagnoses by study participants. Examples with clearly spurious responses (e.g., a negative number of people in a household) or responses with missing features were removed. This pre-processing resulted in a dataset of $864,154$ training examples with a class imbalance of $14 : 86$ of positive to negative COVID cases.

Feature	Type
`Age`	Categorical
`Gender`	Categorical
`LossOfSmellTaste`	Binary
`ShortBreath`	Binary
`Aches`	Binary
`Tired`	Binary
`Cough`	Binary
`Fever`	Binary

Table 4.1: Top 8 Features of the Us Covid-19 Survey Dataset (Salomon *et al.*, 2021), Selected via the Largest Odds Ratios on the Validation Set.

In Figure 4.6, we compare the model parameters learned by the margin-based α-loss on noisy data with those of the $\alpha = 1$ (logistic regression) trained on *clean* data,

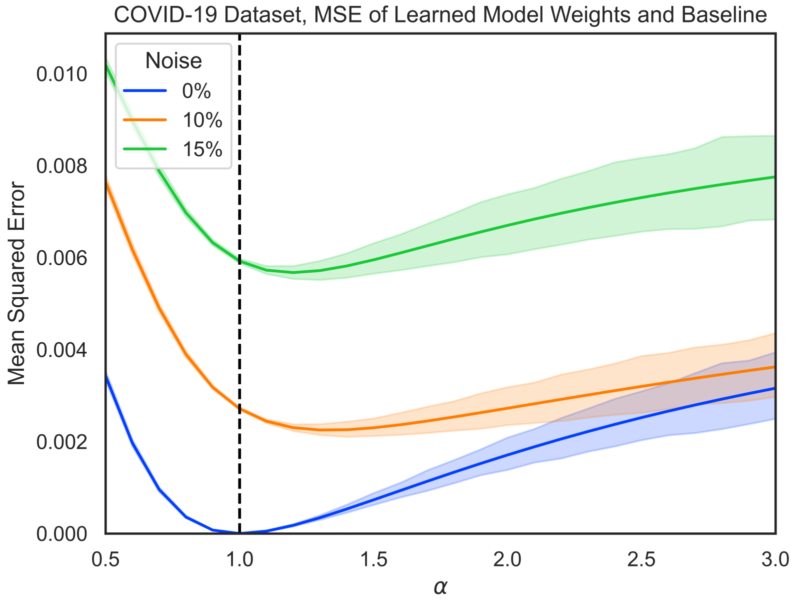


Figure 4.6: A US Covid-19 Survey Dataset (Salomon *et al.*, 2021), Plotting Mse of Logistic Regression ($\alpha = 1$) Baseline Parameters on Clean Data and Model Parameters Learned Using α-loss on Noisy Data Vs. α. For Non-zero Noise the Mse Is Minimized for $\alpha > 1$, but Some Care Is Required in Increasing $\alpha \gg 1$ as the Confidence Intervals Widen, Due to This Being Non-realizable and Highly Imbalanced Data.

which is a calibrated model (Tu, 1996); we are interested in the utility of $\alpha > 1$ to "give up" on the noisy training data and recover the clean model parameters. We see that tuning $\alpha > 1$ gives gains for both non-zero noise levels, but there is a clear tradeoff with optimization complexity; this is indicated by the widening confidence intervals as α increases (Sypherd *et al.*, 2020), which could be due to the COVID-19 survey data being non-realizable and highly imbalanced. However, we note that reduced MSE for $\alpha > 1$ directly translates to gains on test-time accuracy; in Figure C.23 in Appendix C.2 we show that the sensitivity of the model increases with increasing α.

Chapter 5

α-GAN

We introduce a tunable GAN, called α-GAN, parameterized by $\alpha \in (0, \infty]$, which interpolates between various f-GANs and Integral Probability Metric based GANs (under constrained discriminator set). We construct α-GAN using a supervised loss function, namely, α-loss, which is a tunable loss function capturing several canonical losses. We show that α-GAN is intimately related to the Arimoto divergence, which was first proposed by Österriecher (1996), and later studied by Liese and Vajda (2006). We posit that the holistic understanding that α-GAN introduces will have practical benefits of addressing both the issues of vanishing gradients and mode collapse.

Goodfellow *et al.* (2014) introduced *generative adversarial networks* (GANs), a novel technique for training *generative models* to produce samples from an unknown (true) distribution using a finite number of real samples. A GAN involves two learning models (both represented by deep neural networks in practice): a generator model G that takes a random seed in a low-dimensional (relative to the data) *latent* space to generate synthetic samples (by implicitly learning the true distribution without explicit probability models), and a discriminator model D which classifies inputs (from either the true distribution or the generator) as real or fake. The generator wants to fool the discriminator while the discriminator wants to maximize the discrimination power between the true and generated samples. The opposing goals of G and D lead to a zero-sum min-max game in which a chosen value function is minimized and maximized over the model parameters of G and D, respectively.

For the value function considered in *vanilla* GAN (we refer to the GAN introduced by (Goodfellow *et al.*, 2014) as *vanilla* GAN, as done in the literature (Lim and Ye, 2017; Cai *et al.*, 2020) to distinguish it from others introduced later) (Goodfellow *et al.*, 2014), when G and D are given enough training time and capacity, the min-max game is shown to have a Nash equilibrium leading to the generator minimizing the Jensen-Shannon divergence (JSD) between the true and the generated distributions. Subsequently, Nowozin *et al.* (2016) showed that the GAN framework can minimize several f-divergences, including JSD, leading to f-GANs. Arguing that vanishing gradients are due to the sensitivity of f-divergences to mismatch in distribution supports, Arjovsky *et al.* (2017) proposed Wasserstein GAN (WGAN) using a "weaker" Euclidean distance between distributions. This has led to a broader class of GANs based on integral probability metric (IPM) distances (Liang, 2018). Yet neither the vanilla GAN nor the IPM GANs perform consistently well in practice due to a variety of issues that arise during training (e.g., *mode collapse, vanishing gradients, oscillatory convergence, to name a few*) (Huszár, 2015; Metz *et al.*, 2016; Salimans *et al.*, 2016; Arjovsky and Bottou, 2017; Gulrajani *et al.*, 2017), thus providing even less clarity on how to choose the value function.

In this chapter, we first formalize a supervised loss function perspective of GANs and propose a tunable α-GAN based on α-loss, a class of tunable loss functions (Sypherd *et al.*, 2019; Sypherd *et al.*, 2020) parameterized by $\alpha \in (0, \infty]$ that captures the well-known exponential loss ($\alpha = 1/2$) (Freund and Schapire, 1997b), the log-loss ($\alpha = 1$) (Merhav and Feder, 1998; Courtade and Wesel, 2011), and the 0-1 loss ($\alpha = \infty$) (Nguyen *et al.*, 2009b; Bartlett *et al.*, 2006c). Ultimately, we find that α-GAN reveals a holistic structure in relating several canonical GANs, thereby unifying convergence and performance analyses. Our main contributions are as follows:

- We present a unique global Nash equilibrium to the min-max optimization problem induced by the α-GAN, provided G and D have sufficiently large capacity and the models can be trained sufficiently long (Theorem 13). When the discriminator is trained to optimality (where its strategy under α-loss is a tilted distribution), the generator seeks to minimize the *Arimoto divergence* (which has wide applications in statistics and information theory (Liese and Vajda, 2006; Österreicher and Vajda, 2003)) between the true and the generated distributions, thereby providing an operational interpretation to the divergence. We note that our approach differs from Nowozin *et al.* f-GAN approach, please see Remark 3 for clarification.

- We show that α-GAN interpolates between various f-GANs including vanilla GAN ($\alpha = 1$), Hellinger GAN (Nowozin *et al.*, 2016) ($\alpha = 1/2$), Total Variation GAN (Nowozin *et al.*, 2016) ($\alpha = \infty$), and IPM-based GANs including WGANs (when the discriminator set is appropriately constrained) by smoothly tuning the hyperparameter α (see Theorem 14 and (5.13)). Thus, α-GAN allows a practitioner to determine how much they want to resemble vanilla GAN, for instance, since certain datasets/distributions may favor certain GANs (or even interpolation between certain GANs). Analogous to results on α-loss in classification (Sypherd *et al.*, 2020, 2019), where the model performance saturates quickly for $\alpha \to \infty$, we expect a similar saturation for α-GAN (see Figure 5.1). Thus, we posit that smooth tuning from JSD to IPM that results from increasing α from 1 to ∞ can address issues like mode collapse, vanishing gradients, etc.

- Finally in Theorem 15, we reconstruct the Arimoto divergence using the margin-based form of α-loss (Sypherd *et al.*, 2019) and the variational formulation of (Nguyen *et al.*, 2009b), which sheds more light on the convexity of the generator function of the divergence first proposed by (Österreicher, 1996), and later studied by (Liese and Vajda, 2006).

5.1 α-loss and GANs

5.1.1 Background on GANs

Let P_r be a probability distribution over $\mathcal{X} \subset \mathbb{R}^d$, which the generator wants to learn *implicitly* by producing samples by playing a competitive game with a discriminator in an adversarial manner. We parameterize the generator G and the discriminator D by vectors $\theta \in \Theta \subset \mathbb{R}^{n_g}$ and $\omega \in \Omega \subset \mathbb{R}^{n_d}$, respectively, and write G_θ and D_ω (θ and ω are typically the weights of neural network models for the generator and the discriminator, respectively). The generator G_θ takes as input a $d'(\ll d)$-dimensional latent noise $Z \sim P_Z$ and maps it to a data point in $\mathcal{X}$ via the mapping $z \mapsto G_\theta(z)$. For an input $x \in \mathcal{X}$, the discriminator outputs $D_\omega(x) \in [0,1]$, the probability that x comes from P_r (real) as opposed to P_{G_θ} (synthetic). The generator and the discriminator play a two-player min-max game with a value function $V(\theta, \omega)$, resulting in a saddle-point optimization problem given by

$$\inf_{\theta \in \Theta} \sup_{\omega \in \Omega} V(\theta, \omega). \tag{5.1}$$

Goodfellow *et al.* (2014) introduced a value function

$$\begin{aligned} V_{\mathrm{VG}}(\theta, \omega) &= \mathbb{E}_{X \sim P_r}[\log D_\omega(X)] + \mathbb{E}_{Z \sim P_Z}[\log(1 - D_\omega(G_\theta(Z)))] && (5.2) \\ &= \mathbb{E}_{X \sim P_r}[\log D_\omega(X)] + \mathbb{E}_{X \sim P_{G_\theta}}[\log(1 - D_\omega(X))] && (5.3) \end{aligned}$$

and showed that when the discriminator class $\{D_\omega\}$, parametrized by ω, is rich enough, (5.1) simplifies to finding the $\inf_{\theta \in \Theta} 2D_{\mathrm{JS}}(P_r || P_{G_\theta}) - \log 4$, where $D_{\mathrm{JS}}(P_r || P_{G_\theta})$ is the Jensen-Shannon divergence (Lin, 1991) between P_r and P_{G_θ}. This simplification is achieved, for any G_θ, by choosing the optimal discriminator

$$D_{\omega^*}(x) = \frac{p_r(x)}{p_r(x) + p_{G_\theta}(x)}, \tag{5.4}$$

where p_r and p_{G_θ} are the corresponding densities of the distributions P_r and P_{G_θ}, respectively, with respect to a base measure dx (e.g., Lebesgue measure).

Generalizing this, Nowozin *et al.* (2016) derived value function

$$V_f(\theta, \omega) = \mathbb{E}_{X \sim P_r}[D_\omega(X)] + \mathbb{E}_{X \sim P_{G_\theta}}[f^*(D_\omega(X))], \tag{5.5}$$

where[1] $D_\omega : \mathcal{X} \to \mathbb{R}$ and $f^*(t) \triangleq \sup_u \{ut - f(u)\}$ is the Fenchel conjugate of a convex lower semincontinuous function f, for any f-divergence $D_f(P_r||P_{G_\theta}) := \int_{\mathcal{X}} p_{G_\theta}(x) f\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right) dx$ (Rényi, 1961; Csiszár, 1967; Ali and Silvey, 1966) (not just the Jensen-Shannon divergence) leveraging its variational characterization (Nguyen *et al.*, 2010). In particular, $\sup_{\omega \in \Omega} V_f(\theta, \omega) = D_f(P_r||P_{G_\theta})$ when there exists $\omega^* \in \Omega$ such that $T_{\omega^*}(x) = f'\left(\frac{p_r(x)}{p_{G_\theta}(x)}\right)$. Rényi divergence measures are also studied in the context of GANs (Pantazis *et al.*, 2020; Bhatia *et al.*, 2021; Sarraf and Nie, 2021).

Highlighting the problems with the continuity of various f-divergences (e.g., Jensen-Shannon, KL, reverse KL, total variation) over the parameter space Θ (Arjovsky and Bottou, 2017), Arjovsky *et al.* (2017) proposed Wasserstein-GAN (WGAN) using the following Earth Mover's (also called Wasserstein-1) distance:

$$W(P_r, P_{G_\theta}) = \inf_{\Gamma_{X_1 X_2} \in \Pi(P_r, P_{G_\theta})} \mathbb{E}_{(X_1, X_2) \sim \Gamma_{X_1 X_2}} \|X_1 - X_2\|_2, \tag{5.6}$$

where $\Pi(P_r, P_{G_\theta})$ is the set of all joint distributions $\Gamma_{X_1 X_2}$ with marginals P_r and P_{G_θ}. WGAN employs the Kantorovich-Rubinstein duality (Villani, 2008) using the value function

$$V_{\text{WGAN}}(\theta, \omega) = \mathbb{E}_{X \sim P_r}[D_\omega(X)] - \mathbb{E}_{X \sim P_{G_\theta}}[D_\omega(X)], \tag{5.7}$$

where the functions $D_\omega : \mathcal{X} \to \mathbb{R}$ are all 1-Lipschitz, to simplify $\sup_{\omega \in \Omega} V_{\text{WGAN}}(\theta, \omega)$ to $W(P_r, P_{G_\theta})$ when the class Ω is rich enough. Although, various GANs have been

[1] This is a slight abuse of notation in that D_ω is not a probability here. However, we chose this for consistency in notation of discriminator across various GANs.

proposed in the literature, each of them exhibits their own strengths and weaknesses in terms of convergence, vanishing gradients, mode collapse, computational complexity, etc. leaving the problem of instability unsolved (Wiatrak *et al.*, 2019).

5.2 Tunable α-GAN

Noting that a GAN involves a classifier (i.e., discriminator), it is well known that the value function $V_{\mathrm{VG}}(\theta, \omega)$ in (5.3) considered by Goodfellow *et al.* (2014) is related to cross entropy loss. While perhaps it has not been explicitly articulated heretofore in the literature, we first formalize this loss function perspective of GANs and propose a tunable GAN based on α-loss generalizing vanilla GAN and various other GANs. Arora *et al.* (2017) observed that the log function in (5.3) can be replaced by any concave function $\phi(x)$ (e.g., $\phi(x) = x$ for WGANs). More generally, we show that one can write $V(\theta, \omega)$ in terms of a classification loss $\ell(y, \hat{y})$ with inputs $y \in \{0, 1\}$ (the true label) and $\hat{y} \in [0, 1]$ (soft prediction of y). For a GAN, we have $(X|y = 1) \sim P_r$, $(X|y = 0) \sim P_{G_\theta}$, and $\hat{y} = D_\omega(x)$. With this, we observe that the value function V_{VG} in (5.3) for the vanilla GAN can be expressed in terms of cross-entropy loss $\ell_{\mathrm{CE}}(y, \hat{y}) \triangleq -y \log \hat{y} - (1 - y) \log (1 - \hat{y})$ as

$$\begin{aligned} V_{\mathrm{VG}}(\theta, \omega) &= \mathbb{E}_{X|y=1}[-\ell_{\mathrm{CE}}(y, D_\omega(X))] + \mathbb{E}_{X|y=0}[-\ell_{\mathrm{CE}}(y, D_\omega(X))] && (5.8) \\ &= \mathbb{E}_{X \sim P_r}[-\ell_{\mathrm{CE}}(1, D_\omega(X))] + \mathbb{E}_{X \sim P_{G_\theta}}[-\ell_{\mathrm{CE}}(0, D_\omega(X))]. && (5.9) \end{aligned}$$

Now we write α-loss in (2.2) analogous to ℓ_{CE} to obtain

$$\ell_\alpha(y, \hat{y}) := \frac{\alpha}{\alpha - 1} \left(1 - y\hat{y}^{\frac{\alpha-1}{\alpha}} - (1 - y)(1 - \hat{y})^{\frac{\alpha-1}{\alpha}}\right), \tag{5.10}$$

for $\alpha \in (0,1) \cup (1,\infty)$. Note that (5.10) recovers ℓ_{CE} as $\alpha \to 1$. Now consider a *tunable* α*-GAN* with a value function

$$V_\alpha(\theta,\omega) = \mathbb{E}_{X\sim P_r}[-\ell_\alpha(1, D_\omega(X))] + \mathbb{E}_{X\sim P_{G_\theta}}[-\ell_\alpha(0, D_\omega(X))] \tag{5.11}$$

$$= \frac{\alpha}{\alpha-1}\left(\mathbb{E}_{X\sim P_r}\left[D_\omega(X)^{\frac{\alpha-1}{\alpha}}\right] + \mathbb{E}_{X\sim P_{G_\theta}}\left[(1-D_\omega(X))^{\frac{\alpha-1}{\alpha}}\right] - 2\right). \tag{5.12}$$

We can verify that $\lim_{\alpha\to 1} V_\alpha(\theta,\omega) = V_{\mathrm{VG}}(\theta,\omega)$ recovering the value function of the vanilla GAN. Also, notice that

$$\lim_{\alpha\to\infty} V_\alpha(\theta,\omega) = \mathbb{E}_{X\sim P_r}\left[D_\omega(x)\right] - \mathbb{E}_{X\sim P_{G_\theta}}\left[D_\omega(x)\right] - 1 \tag{5.13}$$

is the value function (modulo a constant) used in Intergral Probability Metric (IPM) based GANs[2], e.g., WGAN, McGan (Mroueh *et al.*, 2017b), Fisher GAN (Mroueh and Sercu, 2017), and Sobolev GAN (Mroueh *et al.*, 2017a). The resulting min-max game in α-GAN is given by

$$\inf_{\theta\in\Theta}\sup_{\omega\in\Omega} V_\alpha(\theta,\omega). \tag{5.14}$$

The following theorem provides the min-max solution, i.e., Nash equilibrium, to the two-player game in (5.14) for the non-parametric setting, i.e., when the discriminator set Ω is large enough.

Theorem 13 (min-max solution)**.** *For a fixed generator* G_θ*, the discriminator* $D_{\omega^*}(x)$ *optimizing the* sup *in* (5.14) *is given by*

$$D_{\omega^*}(x) = \frac{p_r(x)^\alpha}{p_r(x)^\alpha + p_{G_\theta}(x)^\alpha}. \tag{5.15}$$

For this $D_{\omega^*}(x)$*,* (5.14) *simplifies to minimizing a non-negative symmetric* f_α*-divergence* $D_{f_\alpha}(\cdot||\cdot)$ *as*

$$\inf_{\theta\in\Theta} D_{f_\alpha}(P_r||P_{G_\theta}) + \frac{\alpha}{\alpha-1}\left(2^{\frac{1}{\alpha}} - 2\right), \tag{5.16}$$

[2] Note that IPMs do not restrict the function D_ω to be a probability.

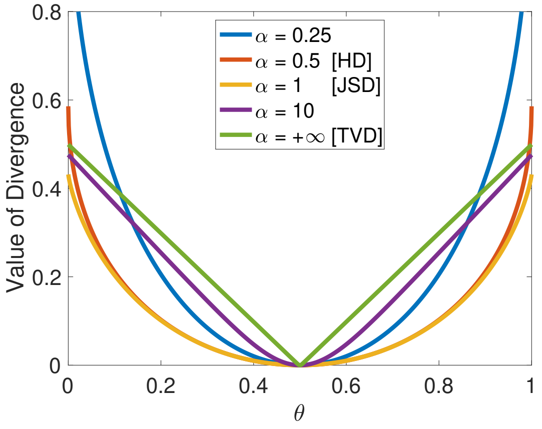


Figure 5.1: A Plot of D_{f_α} In (5.18) for Several Values of α Where $p \sim \text{Ber}(1/2)$ and $q \sim \text{Ber}(\theta)$. Note That HD, JSD, and TVD, Are Abbreviations for Hellinger, Jensen-shannon, and Total Variation Divergences, Respectively. As $\alpha \to 0$, the Curvature of the Divergence Increases, Placing Increasingly More Weight on $\theta \neq 1/2$. Conversely, for $\alpha \to \infty$, D_{f_α} Quickly Resembles D_{f_∞}, Hence a Saturation Effect of D_{f_α}.

where

$$f_\alpha(u) = \frac{\alpha}{\alpha - 1}\left((1 + u^\alpha)^{\frac{1}{\alpha}} - (1 + u) - 2^{\frac{1}{\alpha}} + 2\right), \tag{5.17}$$

for $u \geq 0$ and[3]

$$D_{f_\alpha}(P||Q) = \frac{\alpha}{\alpha - 1}\left(\int_{\mathcal{X}} (p(x)^\alpha + q(x)^\alpha)^{\frac{1}{\alpha}}\, dx - 2^{\frac{1}{\alpha}}\right), \tag{5.18}$$

which is minimized iff $P_{G_\theta} = P_r$.

For intuition on the construction of (5.17), see Theorem 15.

Remark 3. *It can be inferred from* (5.16) *that when the discriminator is trained to optimality, the generator has to minimize the f_α-divergence hinting at an application*

[3] We note that the divergence D_{f_α} has been referred to as *Arimoto divergence* in the literature (Österreicher, 1996; Österreicher and Vajda, 2003; Liese and Vajda, 2006). We refer the reader to Section 5.3 for more details.

of f-GAN instead. Implementing f_α-GAN directly via value function in (5.5) *(for f_α) involves finding convex conjugate of f_α, which is challenging in terms of computational complexity making it inconvenient for optimization in the training phase of GANs. In contrast, our approach of using supervised losses circumvents this tedious effort* and *also provides an operational interpretation of f_α-divergence via losses. A related work where an f-divergence (in particular, α-divergence (Amari, 1985)) shows up in the context of GANs, even when the problem formulation is not via f-GAN, is by (Cai* et al.*, 2020). However, our work presented in this chapter differs from (Cai* et al.*, 2020) in that the value function we use is well motivated via supervised loss functions of binary classification and also recovers the basic GAN (Goodfellow* et al.*, 2014) (among others).*

Remark 4. *As $\alpha \to 0$, note that* (5.15) *implies a more cautious discriminator, i.e., if $p_{G_\theta}(x) \geq p_r(x)$, then $D_{w^*}(x)$ decays more slowly from* 1/2*, and if $p_{G_\theta}(x) \leq p_r(x)$, $D_{w^*}(x)$ increases more slowly from* 1/2*. Conversely, as $\alpha \to \infty$,* (5.15) *simplifies to $D_{\omega^*}(x) = \mathbb{1}\{p_r(x) > p_{G_\theta}(x)\} + \frac{1}{2}\mathbb{1}\{p_r(x) = p_{G_\theta}(x)\}$, where the discriminator implements the Maximum Likelihood (ML) decision rule, i.e., a hard decision whenever $p_r(x) \neq p_{G_\theta}(x)$. In other words,* (5.15) *for $\alpha \to \infty$ induces a very confident discriminator. Regarding the generator's perspective,* (5.16) *(and Figure 5.1) implies that the generator seeks to minimize the discrepancy between P_r and P_{G_θ} according to the geometry induced by D_{f_α}. Thus, the optimization trajectory traversed by the generator during training is strongly dependent on the practitioner's choice of $\alpha \in (0, \infty]$. Please refer to Figure 5.2 for an illustration of this observation.*

A detailed proof of Theorem 13 is in Appendix D.1. Next we show that α-GAN recovers various well known f-GANs.

Theorem 14 (f-GANs)**.** *α-GAN recovers vanilla GAN, Hellinger GAN (H-GAN) (Nowozin*

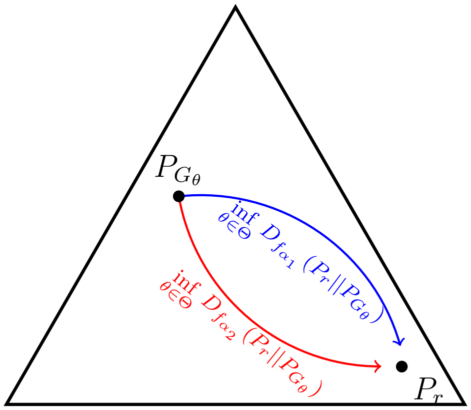

P_{G_θ}
$\inf_{\theta\in\Theta} D_{f_{\alpha_1}}(P_r||P_{G_\theta})$
$\inf_{\theta\in\Theta} D_{f_{\alpha_2}}(P_r||P_{G_\theta})$
P_r

Figure 5.2: An Idealized Illustration on the Probability Simplex of the Infimum over θ In (5.16) for $\alpha_1, \alpha_2 \in (0, \infty]$ Such That $\alpha_1 \neq \alpha_2$. The Choice of α in the Min-max Game for the α-GAN In (5.14) Defines the Optimization Trajectory Taken by the Generator (Versus an Optimal Discriminator as Specified In (5.15)) by Distorting the Underlying Geometry According to D_{f_α}.

et al.*, 2016), and Total Variation GAN (TV-GAN) (Nowozin* et al.*, 2016) as $\alpha \to 1$, $\alpha = \frac{1}{2}$, and $\alpha \to \infty$, respectively.*

A detailed proof is in Appendix.

5.3 Reconstructing Arimoto Divergence

It is interesting to note that the divergence $D_{f_\alpha}(\cdot||\cdot)$ (in (5.18)) that naturally emerges from the analysis of α-GAN was first proposed by (Österreicher, 1996) in the context of statistics and was later referred to as the *Arimoto divergence* by (Liese and Vajda, 2006). It was shown to have several desirable properties with applications in statistics and information theory (Cerone *et al.*, 2004; Vajda, 2009). For example:

- A geometric interpretation of the divergence D_{f_α} in the context of hypothesis testing (Österreicher, 1996).

- $D_{f_\alpha}(P||Q)^{\min\{\alpha,\frac{1}{2}\}}$ defines a distance metric (satisfying the triangle inequality) on the set of probability distributions (Österreicher and Vajda, 2003).

When the Arimoto divergence D_{f_α} was proposed, the convexity of the generating function f_α was proved via the traditional second derivative test (Österreicher, 1996, Lemma 1). We present an alternative approach to arriving at the Arimoto divergence by utilizing the margin-based[4] form of α-loss (see (Sypherd *et al.*, 2019)) where the convexity of f_α (and also the symmetric property of $D_{f_\alpha}(\cdot||\cdot)$) arises in a rather natural manner, thereby reconstructing the Arimoto divergence through a distinct conceptual perspective.

We do this by noticing that the Arimoto divergence falls into the category of a broad class of f-divergences that can be obtained from margin-based loss functions. Such a connection between margin-based losses in classification and the corresponding f-divergences was introduced by (Nguyen *et al.*, 2009b, Theorem 1). They observed that, for a given margin-based loss function $\tilde{\ell}$, there is a corresponding f-divergence with the convex function f defined as $f(u) := -\inf_t \left(u\tilde{\ell}(t) + \tilde{\ell}(-t)\right)$. The convexity of f follows simply because the infimum of affine functions is concave, and this argument does not require $\tilde{\ell}$ to be convex[5]. Additionally, the f-divergence obtained is always symmetric because f satisfies $f(u) = uf(\frac{1}{u})$ since $\inf_t u\tilde{\ell}(t) + \tilde{\ell}(-t) = \inf_t \tilde{\ell}(t) + u\tilde{\ell}(-t)$.

The margin-based α-loss (Sypherd *et al.*, 2019) for $\alpha \in (0,1)\cup(1,\infty)$, $\tilde{\ell}_\alpha : \bar{\mathbb{R}} \to \mathbb{R}_+$ is defined as

$$\tilde{\ell}_\alpha(t) \triangleq \frac{\alpha}{\alpha - 1}\left(1 - \sigma(t)^{\frac{\alpha-1}{\alpha}}\right), \tag{5.19}$$

[4] In the binary classification context, the margin is represented by $t := yf(x)$, where $x \in \mathcal{X}$ is the feature vector, $y \in \{-1, +1\}$ is the label, and $f : \mathcal{X} \to \mathbb{R}$ is the prediction function produced by a learning algorithm.

[5] in fact α-loss in its margin-based form is only quasi-convex for $\alpha > 1$

where $\sigma : \bar{\mathbb{R}} \to \mathbb{R}_+$ is the sigmoid function given by $\sigma(t) = (1 + \mathrm{e}^{-t})^{-1}$. With these preliminaries in hand, we have the following result.

Theorem 15. *For the function f_α in* (5.17)*, it holds that*

$$f_\alpha(u) = -\inf_t \left(u\tilde{\ell}_\alpha(t) + \tilde{\ell}_\alpha(-t)\right) - \frac{\alpha}{\alpha - 1}\left(2^{\frac{1}{\alpha}} - 2\right), \quad \textit{for } u \geq 0. \tag{5.20}$$

A detailed proof is in Appendix D.2.

REFERENCES

Agarwal, S., “Surrogate regret bounds for bipartite ranking via strongly proper losses”, JMLR **15**, 1, 1653–1674 (2014).

Ali, S. M. and S. D. Silvey, “A general class of coefficients of divergence of one distribution from another”, Journal of the Royal Statistical Society. Series B (Methodological) **28**, 1, 131–142 (1966).

Alon, N., A. Gonen, E. Hazan and S. Moran, “Boosting simple learners”, in “STOC’21”, (2021).

Amari, S.-i., *α-Divergence and α-Projection in Statistical Manifold*, pp. 66–103 (Springer New York, New York, NY, 1985).

Amid, E., M. K. Warmuth, R. Anil and T. Koren, “Robust bi-tempered logistic loss based on bregman divergences”, in “Advances in Neural Information Processing Systems”, pp. 15013–15022 (2019a).

Amid, E., M.-K. Warmuth, R. Anil and T. Koren, “Robust bi-tempered logistic loss based on bregman divergences”, in “NeurIPS*32”, pp. 14987–14996 (2019b).

Andriushchenko, M. and M. Hein, “Provably robust boosted decision stumps and trees against adversarial attacks”, in “NeurIPS*32”, (2019).

Arimoto, S., “Information-theoretical considerations on estimation problems”, Information and Control **19**, 3, 181 – 194, URL `http://www.sciencedirect.com/science/article/pii/S0019995871900659` (1971a).

Arimoto, S., “Information-theoretical considerations on estimation problems”, Information and control **19**, 3, 181–194 (1971b).

Arimoto, S., “Information-theoretical considerations on estimation problems”, Information and control **19**, 181–194 (1971c).

Arimoto, S., “Information measures and capacity of order α for discrete memoryless channels”, Topics in information theory (1977).

Arjovsky, M. and L. Bottou, “Towards principled methods for training generative adversarial networks”, arXiv preprint arXiv:1701.04862 (2017).

Arjovsky, M., L. Bottou, I. Gulrajani and D. Lopez-Paz, “Invariant risk minimization”, CoRR **abs/1907.02893** (2019).

Arjovsky, M., S. Chintala and L. Bottou, “Wasserstein generative adversarial networks”, in “Proceedings of the 34th International Conference on Machine Learning”, vol. 70, pp. 214–223 (2017).

Arora, S., R. Ge, Y. Liang, T. Ma and Y. Zhang, “Generalization and equilibrium in generative adversarial nets (GANs)”, in “Proceedings of the 34th International Conference on Machine Learning”, vol. 70, pp. 224–232 (2017).

Audibert, J.-Y., A. B. Tsybakov *et al.*, “Fast learning rates for plug-in classifiers”, The Annals of statistics **35**, 2, 608–633 (2007).

Bai, Y., E. Yang, B. Han, Y. Yang, J. Li, Y. Mao, G. Niu and T. Liu, “Understanding and improving early stopping for learning with noisy labels”, Advances in Neural Information Processing Systems **34**, 24392–24403 (2021).

Barron, J. T., “A general and adaptive robust loss function”, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, pp. 4331–4339 (2019).

Bartlett, P., M. Jordan and J. D. McAuliffe, “Convexity, classification, and risk bounds”, J. of the Am. Stat. Assoc. **101**, 138–156 (2006a).

Bartlett, P. L., M. I. Jordan and J. D. McAuliffe, “Convexity, classification, and risk bounds”, Journal of the American Statistical Association **101**, 473, 138–156 (2006b).

Bartlett, P. L., M. I. Jordan and J. D. Mcauliffe, “Convexity, classification, and risk bounds”, Journal of the American Statistical Association **101**, 473, 138–156 (2006c).

Bartlett, P.-L. and S. Mendelson, “Rademacher and gaussian complexities: Risk bounds and structural results”, JMLR **3**, 463–482 (2002).

Beigman, E. and B. B. Klebanov, “Learning with annotation noise”, in “Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP”, pp. 280–287 (2009).

Ben-David, S., N. Eiron and P. M. Long, “On the difficulty of approximately maximizing agreements”, Journal of Computer and System Sciences **66**, 3, 496–514 (2003).

Ben-David, S., D. Loker, N. Srebro and K. Sridharan, “Minimizing the misclassification error rate using a surrogate convex loss”, arXiv preprint arXiv:1206.6442 (2012).

Benigni, L. and S. Péché, “Eigenvalue distribution of nonlinear models of random matrices”, arXiv preprint arXiv:1904.03090 (2019).

Bhatia, H., W. Paul, F. Alajaji, B. Gharesifard and P. Burlina, “Least k th-order and rényi generative adversarial networks”, Neural Computation **33**, 9, 2473–2510 (2021).

Box, G. E. and D. R. Cox, “An analysis of transformations”, Journal of the Royal Statistical Society: Series B (Methodological) **26**, 2, 211–243 (1964).

Boyd, S. and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, 2004a).

Boyd, S. and L. Vandenberghe, *Convex optimization* (Cambridge University Press, 2004b).

Bu, Y., S. Zou and V. V. Veeravalli, “Tightening mutual information-based bounds on generalization error”, IEEE Journal on Selected Areas in Information Theory **1**, 1, 121–130 (2020).

Bun, M., M.-L. Carmosino and J. Sorrell, “Efficient, noise-tolerant, and private learning via boosting”, in “33 th COLT”, Proceedings of Machine Learning Research, pp. 1031–1077 (PMLR, 2020).

Buntine, W. and T. Niblett, “A further comparison of splitting rules for decision-tree induction”, Machine Learning **8**, 1, 75–85, URL `https://www.openml.org/d/40693` (1992).

Cai, L., Y. Chen, N. Cai, W. Cheng and H. Wang, “Utilizing amari-alpha divergence to stabilize the training of generative adversarial networks”, Entropy **22**, 4, 410 (2020).

Caruana, R., Y. Lou, J. Gehrke, P. Koch, M. Sturm and N. Elhadad, “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission”, in “Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining”, KDD ’15, p. 1721–1730 (Association for Computing Machinery, New York, NY, USA, 2015).

Castells, T., P. Weinzaepfel and J. Revaud, “SuperLoss: A generic loss for robust curriculum learning”, in “NeurIPS*33”, (2020).

Cava, J. K., “Code for a tunable loss function for robust classification”, URL `https://github.com/SankarLab/AlphaLoss` (2021).

Cerone, P., S. S. Dragomir and F. Österreicher, “Bound on extended f-divergences for a variety of classes”, Kybernetika **40**, 6, 745–756 (2004).

Chapelle, O., C. B. Do, C. H. Teo, Q. V. Le and A. J. Smola, “Tighter bounds for structured estimation”, in “Advances in neural information processing systems”, pp. 281–288 (2009).

Chapelle, O., C. Teo, Q. Le, A. Smola *et al.*, “Tighter bounds for structured estimation”, Advances in neural information processing systems **21** (2008).

Chapelle, O., J. Weston, L. Bottou and V. Vapnik, “Vicinal risk minimization”, in “Advances in Neural Information Processing Systems*13”, (2000).

Charoenphakdee, N., J. Vongkulbhisal, N. Chairatanakul and M. Sugiyama, “On focal loss for class-posterior probability estimation: A theoretical perspective”, in “34th IEEE CVPR”, pp. 5202–5211 (2021).

Chaudhari, P., A. Oberman, S. Osher, S. Soatto and G. Carlier, “Deep relaxation: Partial differential equations for optimizing deep neural networks”, Research in the Mathematical Sciences **5**, 3, 30 (2018).

Cheamanunkul, S., E. Ettinger and Y. Freund, "Non-convex boosting overcomes random label noise", arXiv preprint arXiv:1409.2905 (2014).

Chen, T. and C. Guestrin, "Xgboost: A scalable tree boosting system", in "Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining", pp. 785–794 (ACM, 2016).

Chen, Z., S. Tan, H. Nori, K. Inkpen, Y. Lou and R. Caruana, "Using explainable boosting machines (ebms) to detect common flaws in data", in "Machine Learning and Principles and Practice of Knowledge Discovery in Databases", pp. 534–551 (Springer International Publishing, Cham, 2021).

Collins, M., R. Schapire and Y. Singer, "Logistic regression, adaboost and Bregman distances", in "Proc. of the 13 th International Conference on Computational Learning Theory", pp. 158–169 (2000).

Collins, M., R. E. Schapire and Y. Singer, "Logistic regression, adaboost and bregman distances", Machine Learning **48**, 1-3, 253–285 (2002).

Cooper, J. and L. Reyzin, "Improved algorithms for distributed boosting", in "2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)", pp. 806–813 (IEEE, 2017).

Courtade, T. A. and R. D. Wesel, "Multiterminal source coding with an entropy-based distortion measure", in "IEEE International Symposium on Information Theory", pp. 2040–2044 (2011).

Cranko, Z., A.-K. Menon, R. Nock, C. S. Ong, Z. Shi and C.-J. Walder, "Monge blunts Bayes: Hardness results for adversarial training", in "36th ICML", pp. 1406–1415 (2019).

Csiszár, I., "Information-type measures of difference of probability distributions and indirect observation", Studia Scientiarum Mathematicarum Hungarica **2**, 229–318 (1967).

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

Esposito, A. R., M. Gastpar and I. Issa, "Generalization error bounds via rényi-, f-divergences and maximal leakage", IEEE Transactions on Information Theory (2021).

Fisher, R. A., "On the mathematical foundations of theoretical statistics", Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character **222**, 594-604, 309–368 (1922).

Fort, S., P. K. Nowak, S. Jastrzebski and S. Narayanan, "Stiffness: A new perspective on generalization in neural networks", arXiv preprint arXiv:1901.09491 (2019).

Frénay, B. and M. Verleysen, “Classification in the presence of label noise: a survey”, IEEE transactions on neural networks and learning systems **25**, 5, 845–869 (2013).

Freund, Y., R. Schapire and N. Abe, “A short introduction to boosting”, Journal-Japanese Society For Artificial Intelligence **14**, 771-780, 1612 (1999).

Freund, Y. and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting”, Journal of computer and system sciences **55**, 1, 119–139 (1997a).

Freund, Y. and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting”, Journal of Computer and System Sciences **55**, 1, 119 – 139 (1997b).

Friedman, J., T. Hastie and R. Tibshirani, “Additive Logistic Regression : a Statistical View of Boosting”, Ann. of Stat. **28**, 337–374 (2000).

Friedman, J., T. Hastie and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1 (Springer series in statistics New York, 2001).

Friedman, J. H., “Greedy function approximation: a gradient boosting machine”, Ann. of Stat. **29**, 1189–1232 (2001).

Fu, H., Y. Chi and Y. Liang, “Guaranteed recovery of one-hidden-layer neural networks via cross entropy”, IEEE transactions on signal processing **68**, 3225–3235 (2020).

Gálvez, B. R., G. Bassi, R. Thobaben and M. Skoglund, “Tighter expected generalization error bounds via wasserstein distance”, in “Thirty-Fifth Conference on Neural Information Processing Systems”, (2021).

Ghosh, A., H. Kumar and P. S. Sastry, “Robust loss functions under label noise for deep neural networks”, in “Proceedings of the AAAI conference on artificial intelligence”, vol. 31 (2017).

Goodfellow, I., Y. Bengio and A. Courville, *Deep Learning* (MIT press, 2016).

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative adversarial nets”, in “Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2”, p. 2672–2680 (2014).

Gorber, S. C., S. Schofield-Hurwitz, J. Hardt, G. Levasseur and M. Tremblay, “The accuracy of self-reported smoking: A systematic review of the relationship between self-reported and cotinine-assessed smoking status”, Nicotine & Tobacco Research **11**, 1, 12–24 (2009).

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, “Improved training of Wasserstein GANs”, in “Advances in Neural Information Processing Systems”, vol. 30 (2017).

Guo, C., G. Pleiss, Y. Sun and K.-Q. Weinberger, "On calibration of modern neural networks", in "34^{th} ICML", pp. 1321–1330 (2017).

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

He, X., K. Zhao and X. Chu, "Automl: A survey of the state-of-the-art", Knowledge-Based Systems **212**, 106622 (2021).

Horn, R. A. and C. R. Johnson, *Matrix Analysis* (Cambridge university press, 2012).

Huszár, F., "How (not) to train your generative model: Scheduled sampling, likelihood, adversary?", arXiv preprint arXiv:1511.05101 (2015).

Issa, I., A. B. Wagner and S. Kamath, "An operational approach to information leakage", IEEE Transactions on Information Theory **66**, 3, 1625–1657 (2019).

Janocha, K. and W. M. Czarnecki, "On loss functions for deep neural networks in classification", Schedae Informaticae **25**, 49–59 (2016).

Ji, Z., K. Ahn, P. Awasthi, S. Kale and S. Karp, "Agnostic learnability of halfspaces via logistic loss", in "Proceedings of the 39th International Conference on Machine Learning", edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato, vol. 162 of *Proceedings of Machine Learning Research*, pp. 10068–10103 (PMLR, 2022).

Kairouz, P., J. Liao, C. Huang and L. Sankar, "Censored and fair universal representations using generative adversarial models", arXiv pp. arXiv–1910 (2019a).

Kairouz, P., H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning", arXiv preprint arXiv:1912.04977 (2019b).

Kearns, M. J. and U. V. Vazirani, *An Introduction to Computational Learning Theory* (M.I.T. Press, 1994).

Kline, M., *Calculus: an intuitive and physical approach* (Courier Corporation, 1998).

Knuth, D.-E., "Two notes on notation", The American Mathematical Monthly **99**, 5, 403–422 (1992).

Krizhevsky, A., V. Nair and G. Hinton, "The cifar-10 dataset", online: http://www. cs. toronto. edu/kriz/cifar. html **55** (2014).

LeCun, Y., "The mnist database of handwritten digits", http://yann. lecun. com/exdb/mnist/ (1998).

Lee, S.-I., H. Lee, P. Abbeel and A. Y. Ng, "Efficient l˜ 1 regularized logistic regression", in "Aaai", vol. 6, pp. 401–408 (2006).

Leng, Z., M. Tan, C. Liu, E. D. Cubuk, X. Shi, S. Cheng and D. Anguelov, “Polyloss: A polynomial expansion perspective of classification loss functions”, arXiv preprint arXiv:2204.12511 (2022).

Li, H., Z. Xu, G. Taylor, C. Studer and T. Goldstein, “Visualizing the loss landscape of neural nets”, Advances in neural information processing systems **31** (2018a).

Li, H., Z. Xu, G. Taylor, C. Studer and T. Goldstein, “Visualizing the loss landscape of neural nets”, in “Advances in Neural Information Processing Systems”, pp. 6389–6399 (2018b).

Li, T., A. Beirami, M. Sanjabi and V. Smith, “On tilted losses in machine learning: Theory and applications”, arXiv preprint arXiv:2109.06141 (2021).

Li, T., M. Sanjabi, A. Beirami and V. Smith, “Fair resource allocation in federated learning”, in “International Conference on Learning Representations”, (2019).

Liang, S., R. Sun, Y. Li and R. Srikant, “Understanding the loss surface of neural networks for binary classification”, in “International Conference on Machine Learning”, pp. 2835–2843 (PMLR, 2018).

Liang, T., “How well generative adversarial networks learn distributions”, arXiv preprint arXiv:1811.03179 (2018).

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

Liao, J., O. Kosut, L. Sankar and F. du Pin Calmon, “A tunable measure for information leakage”, in “2018 IEEE International Symposium on Information Theory, ISIT 2018”, pp. 701–705 (IEEE, 2018b).

Liao, J., L. Sankar, O. Kosut and F. P. Calmon, “Robustness of maximal α-leakage to side information”, in “2019 IEEE International Symposium on Information Theory (ISIT)”, pp. 642–646 (IEEE, 2019).

Liao, J., L. Sankar, O. Kosut and F. P. Calmon, “Maximal α-leakage and its properties”, in “2020 IEEE Conference on Communications and Network Security (CNS)”, pp. 1–6 (IEEE, 2020).

Liese, F. and I. Vajda, “On divergences and informations in statistics and information theory”, IEEE Transactions on Information Theory **52**, 10, 4394–4412 (2006).

Lim, J. H. and J. C. Ye, “Geometric GAN”, arXiv preprint arXiv:1705.02894 (2017).

Lin, J., “Divergence measures based on the shannon entropy”, IEEE Transactions on Information Theory **37**, 1, 145–151 (1991).

Lin, T., P. Goyal, R.-B. Girshick, K. He and P. Dollár, “Focal loss for dense object detection”, in “Proc. of the 23 rd IEEE ICCV”, pp. 2999–3007 (2017a).

Lin, T.-Y., 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", pp. 2980–2988 (2017b).

Lin, Y., "A note on margin-based loss functions in classification", Statistical & Probability Letters **68**, 1, 73–82 (2004).

Liu, Y. and H. Guo, "Peer loss functions: Learning from noisy labels without knowing noise rates", in "International Conference on Machine Learning", pp. 6226–6236 (PMLR, 2020).

Long, P. M. and R. A. Servedio, "Random classification noise defeats all convex potential boosters", Machine learning **78**, 3, 287–304 (2010).

Long, P. M. and R. A. Servedio, "The perils of being unhinged: On the accuracy of classifiers minimizing a noise-robust convex loss", Neural Computation **34**, 6, 1488–1499 (2022).

Lopez, A. T. and V. Jog, "Generalization error bounds using wasserstein distances", in "2018 IEEE Information Theory Workshop (ITW)", pp. 1–5 (IEEE, 2018).

Ma, X., H. Huang, Y. Wang, S. Romano, S. Erfani and J. Bailey, "Normalized loss functions for deep learning with noisy labels", in "International Conference on Machine Learning", pp. 6543–6553 (PMLR, 2020).

Madry, A., A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards deep learning models resistant to adversarial attacks", in "International Conference on Learning Representations", (2018a).

Madry, A., A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards deep learning models resistant to adversarial attacks", in "6^{th} ICLR", (2018b).

Mansour, Y., R. Nock and R. C. Williamson, "What killed the convex booster ?", URL `https://arxiv.org/abs/2205.09628` (2022a).

Mansour, Y., R. Nock and R. C. Williamson, "What killed the convex booster ?", URL `https://arxiv.org/abs/2205.09628` (2022b).

Manwani, N. and P. Sastry, "Noise tolerance under risk minimization", IEEE transactions on cybernetics **43**, 3, 1146–1151 (2013).

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

Matusita, K., "Decision rule, based on the distance, for the classification problem", Annals of the institute of statistical mathematics **8**, 2, 67–77 (1956).

McDonald, R. A., D. J. Hand and I. A. Eckley, "An empirical comparison of three boosting algorithms on real data sets with artificial class noise", in "International Workshop on Multiple Classifier Systems", pp. 35–44 (Springer, 2003).

Mei, S., Y. Bai and A. Montanari, "The landscape of empirical risk for nonconvex losses", The Annals of Statistics **46**, 6A, 2747–2774 (2018).

Menon, A. K., B. van Rooyen, C.-S. Ong and R.-C. Williamson, "Learning from corrupted labels via class probability estimation", in "32^{nd} ICML", pp. 125–134 (2015).

Merhav, N. and M. Feder, "Universal prediction", IEEE Transactions on Information Theory **44**, 6, 2124–2147 (1998).

Metz, L., B. Poole, D. Pfau and J. Sohl-Dickstein, "Unrolled generative adversarial networks", arXiv preprint arXiv:1611.02163 (2016).

Mo, J. and J. Walrand, "Fair end to end window-based congestion control", IEEE/ACM Transactions on networking **8**, 5, 556–567 (2000).

Mroueh, Y., C.-L. Li, T. Sercu, A. Raj and Y. Cheng, "Sobolev GAN", arXiv preprint arXiv:1711.04894 (2017a).

Mroueh, Y. and T. Sercu, "Fisher GAN", in "Advances in Neural Information Processing Systems", vol. 30 (2017).

Mroueh, Y., T. Sercu and V. Goel, "McGan: Mean and covariance feature matching GAN", in "Proceedings of the 34th International Conference on Machine Learning", vol. 70, pp. 2527–2535 (2017b).

Mukhoti, J., V. Kulharia, A. Sanyal, S. Golodetz, P.-H.-S. Torr and P.-K. Dokania, "Calibrating deep neural networks using focal loss", in "NeurIPS*33", (2020).

Natarajan, N., I. S. Dhillon, P. K. Ravikumar and A. Tewari, "Learning with noisy labels", Advances in neural information processing systems **26**, 1196–1204 (2013).

Nesterov, Y. E., "Minimization methods for nonsmooth convex and quasiconvex functions", Matekon **29**, 519–531 (1984).

Neu, G., G. K. Dziugaite, M. Haghifam and D. M. Roy, "Information-theoretic generalization bounds for stochastic gradient descent", in "COLT", (2021).

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

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

Nguyen, X., M. J. Wainwright and M. I. Jordan, "On surrogate loss functions and f-divergences", The Annals of Statistics **37**, 2, 876–904 (2009a).

Nguyen, X., M. J. Wainwright and M. I. Jordan, "On surrogate loss functions and f-divergences", The Annals of Statistics **37**, 2, 876–904 (2009b).

Nguyen, X., M. J. Wainwright and M. I. Jordan, “Estimating divergence functionals and the likelihood ratio by convex risk minimization”, IEEE Transactions on Information Theory **56**, 11, 5847–5861 (2010).

Nock, R. and A. Menon, “Supervised learning: No loss no cry”, in “International Conference on Machine Learning”, pp. 7370–7380 (PMLR, 2020).

Nock, R. and F. Nielsen, “On the efficient minimization of classification-calibrated surrogates”, in “NIPS*21”, pp. 1201–1208 (2008).

Nock, R. and R.-C. Williamson, “Lossless or quantized boosting with integer arithmetic”, in “36^{th} ICML”, pp. 4829–4838 (2019).

Nori, H., R. Caruana, Z. Bu, J. H. Shen and J. Kulkarni, “Accuracy, interpretability, and differential privacy via explainable boosting”, in “Proceedings of the 38th International Conference on Machine Learning”, edited by M. Meila and T. Zhang, vol. 139 of *Proceedings of Machine Learning Research*, pp. 8227–8237 (PMLR, 2021).

Nowozin, S., B. Cseke and R. Tomioka, “f gan: Training generative neural samplers using variational divergence minimization”, in “Proceedings of the 30th International Conference on Neural Information Processing Systems”, p. 271–279 (2016).

Österreicher, F., “On a class of perimeter-type distances of probability distributions”, Kybernetika **32**, 4, 389–393 (1996).

Österreicher, F. and I. Vajda, “A new class of metric divergences on probability spaces and its applicability in statistics”, Annals of the Institute of Statistical Mathematics **55**, 3, 639–653 (2003).

Pantazis, Y., D. Paul, M. Fasoulakis, Y. Stylianou and M. Katsoulakis, “Cumulant gan”, arXiv preprint arXiv:2006.06625 (2020).

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

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “Pytorch: An imperative style, high-performance deep learning library”, in “Advances in neural information processing systems”, pp. 8026–8037 (2019).

Patrini, G., A. Rozza, A. Krishna Menon, R. Nock and L. Qu, “Making deep neural networks robust to label noise: A loss correction approach”, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, pp. 1944–1952 (2017a).

Patrini, G., A. Rozza, A.-K. Menon, R. Nock and L. Qu, “Making deep neural networks robust to label noise: A loss correction approach”, in “CVPR17”, pp. 2233–2241 (2017b).

Pennington, J. and P. Worah, “Nonlinear random matrix theory for deep learning”, in “Advances in Neural Information Processing Systems”, pp. 2637–2646 (2017).

Rauscher, G. H., T. P. Johnson, Y. I. Cho and J. A. Walk, “Accuracy of Self-Reported Cancer-Screening Histories: A Meta-analysis”, Cancer Epidemiology, Biomarkers & Prevention **17**, 4, 748–757 (2008).

Reid, M. D. and R. C. Williamson, “Composite binary losses”, The Journal of Machine Learning Research **11**, 2387–2422 (2010a).

Reid, M.-D. and R.-C. Williamson, “Composite binary losses”, JMLR **11**, 2387–2422 (2010b).

Reid, M.-D. and R.-C. Williamson, “Information, divergence and risk for binary experiments”, JMLR **12**, 731–817 (2011).

Rényi, A., “On measures of entropy and information”, in “Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability”, pp. 547–561 (1961).

Rolnick, D., A. Veit, S. Belongie and N. Shavit, “Deep learning is robust to massive label noise”, arXiv preprint arXiv:1705.10694 (2017).

Rosasco, L., E. D. Vito, A. Caponnetto, M. Piana and A. Verri, “Are loss functions all the same?”, Neural Computation **16**, 5, 1063–1076 (2004).

Rudin, C., “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”, Nature Machine Intelligence (2019).

Russo, D. and J. Zou, “How much does your data exploration overfit? controlling bias via information usage”, IEEE Transactions on Information Theory **66**, 1, 302–323 (2020).

Rényi, A., “On measures of entropy and information”, in “Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics”, pp. 547–561 (University of California Press, Berkeley, Calif., 1961), URL `https://projecteuclid.org/euclid.bsmsp/1200512181`.

Sakar, C. O., S. O. Polat, M. Katircioglu and Y. Kastro, “Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and lstm recurrent neural networks”, Neural Computing and Applications **31**, 10, 6893–6908 (2019).

Salimans, T., I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, “Improved techniques for training GANs”, arXiv preprint arXiv:1606.03498 (2016).

Salomon, J. A., A. Reinhart, A. Bilinski, E. J. Chua, W. La Motte-Kerr, M. M. Rönn, M. B. Reitsma, K. A. Morris, S. LaRocca, T. H. Farag, F. Kreuter, R. Rosenfeld and R. J. Tibshirani, “The U.S. COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination”, Proceedings of the National Academy of Sciences **118**, 51 (2021).

Sarraf, A. and Y. Nie, “Rgan: Rényi generative adversarial network”, SN Computer Science **2**, 1, 1–8 (2021).

Sasaki, Y., “The truth of the f-measure”, Teach Tutor Mater (2007).

Sason, I. and S. Verdú, “Arimoto–rényi conditional entropy and bayesian m-ary hypothesis testing”, IEEE Transactions on Information theory **64**, 1, 4–25 (2017).

Savage, L.-J., “Elicitation of personal probabilities and expectations”, J. of the Am. Stat. Assoc. pp. 783–801 (1971).

Schapire, R.-E. and Y. Freund, *Boosting, Foundations and Algorithms* (MIT Press, 2012).

Schapire, R. E. and Y. Freund, “Boosting: Foundations and algorithms”, Kybernetes (2013).

Schapire, R. E., Y. Freund, P. Bartlett and W. S. Lee, “Boosting the margin : a new explanation for the effectiveness of voting methods”, Annals of statistics **26**, 1651–1686 (1998).

Schapire, R. E. and Y. Singer, “Improved boosting algorithms using confidence-rated predictions”, MLJ **37**, 297–336 (1999).

Schmidt, L., S. Santurkar, D. Tsipras, K. Talwar and A. Madry, “Adversarially robust generalization requires more data”, in “Advances in Neural Information Processing Systems”, pp. 5014–5026 (2018).

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

Shannon, C. E., “A mathematical theory of communication”, ACM SIGMOBILE mobile computing and communications review **5**, 1, 3–55 (2001).

Shuford, E., A. Albert and H.-E. Massengil, “Admissible probability measurement procedures”, Psychometrika pp. 125–145 (1966).

Singh, A. 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)”, pp. 1–6 (IEEE, 2010).

Smith, J. W., J. E. Everhart, W. Dickson, W. C. Knowler and R. S. Johannes, “Using the adap learning algorithm to forecast the onset of diabetes mellitus”, in “Proceedings of the annual symposium on computer application in medical care”, p. 261 (American Medical Informatics Association, 1988), URL `https://www.kaggle.com/uciml/pima-indians-diabetes-database`.

Steinke, T. and L. Zakynthinou, “Reasoning about generalization via conditional mutual information”, in “Conference on Learning Theory”, pp. 3437–3452 (PMLR, 2020).

Sypherd, T., M. Diaz, J. K. Cava, G. Dasarathy, P. Kairouz and L. Sankar, “A tunable loss function for robust classification: Calibration, landscape, and generalization”, arXiv preprint arXiv:1906.02314 (2019).

Sypherd, T., M. Diaz, J. K. Cava, G. Dasarathy, P. Kairouz and L. Sankar, “A tunable loss function for robust classification: Calibration, landscape, and generalization”, IEEE Transactions on Information Theory pp. 1–1 (2022a).

Sypherd, T., M. Diaz, J. K. Cava, G. Dasarathy, P. Kairouz and L. Sankar, “A tunable loss function for robust classification: Calibration, landscape, and generalization”, IEEE Transactions on Information Theory **68**, 9, 6021–6051 (2022b).

Sypherd, T., M. Diaz, L. Sankar and G. Dasarathy, “On the α-loss landscape in the logistic model”, in “2020 IEEE International Symposium on Information Theory (ISIT)”, pp. 2700–2705 (2020).

Sypherd, T., M. Diaz, L. Sankar and G. Dasarathy, “On the α-loss landscape in the logistic model”, in “IEEE International Symposium on Information Theory”, pp. 2700–2705 (2020).

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

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

Sypherd, T., R. Nock and L. Sankar, “Being Properly Improper”, arXiv e-prints p. arXiv:2106.09920 (2021).

Sypherd, T., R. Nock and L. Sankar, “Being properly improper”, in “Proceedings of the 39th International Conference on Machine Learning”, edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu and S. Sabato, vol. 162 of *Proceedings of Machine Learning Research*, pp. 20891–20932 (PMLR, 2022c).

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, “Intriguing properties of neural networks”, CoRR **abs/1312.6199** (2013).

Telgarsky, M., “Boosting with the logistic loss is consistent”, arXiv preprint arXiv:1305.2648 (2013).

Tewari, A. and P. L. Bartlett, “On the consistency of multiclass classification methods”, Journal of Machine Learning Research **8**, May, 1007–1025 (2007).

Thekumparampil, K.-K., P. Jain, P. Netrapalli and S. Oh, “Projection efficient subgradient method and optimal nonsmooth Frank-Wolfe method”, in “NeurIPS*33”, (2020).

Thomas, M. and A. T. Joy, *Elements of information theory* (Wiley-Interscience, 2006).

Thulasidasan, S., T. Bhattacharya, J. Bilmes, G. Chennupati and J. Mohd-Yusof, “Combating label noise in deep learning using abstention”, arXiv preprint arXiv:1905.10964 (2019).

Truong, L., C. Jones, B. Hutchinson, A. August, B. Praggastis, R. Jasper, N. Nichols and A. Tuor, “Systematic evaluation of backdoor data poisoning attacks on image classifiers”, in “CVPR’20”, pp. 3422–3431 (2020).

Tu, J. V., “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes”, Journal of clinical epidemiology **49**, 11, 1225–1231 (1996).

Vajda, I., “On metric divergences of probability measures”, Kybernetika **45**, 6, 885–900 (2009).

Valiant, L. G., “A theory of the learnable”, Communications of the ACM **27**, 1134–1142 (1984).

Valiant, L. G., “Learning disjunctions of conjunctions”, in “Proc. of the 9 th International Joint Conference on Artificial Intelligence”, pp. 560–566 (1985).

Van Rooyen, B., A. Menon and R. C. Williamson, “Learning with symmetric label noise: The importance of being unhinged”, Advances in neural information processing systems **28** (2015).

Verdú, S., “α-mutual information”, in “2015 Information Theory and Applications Workshop (ITA)”, pp. 1–6 (2015).

Villani, C., *Optimal transport: old and new*, vol. 338 (Springer Science & Business Media, 2008).

Viola, P. and M. Jones, “Rapid object detection using a boosted cascade of simple features”, in “Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001”, vol. 1, pp. I–I (Ieee, 2001).

Walder, C. and R. Nock, “All your loss are belong to bayes”, Advances in Neural Information Processing Systems **33**, 18505–18517 (2020).

Wang, H., M. Diaz, J. C. S. Santos Filho and F. P. Calmon, “An information-theoretic view of generalization via wasserstein distance”, in “2019 IEEE International Symposium on Information Theory (ISIT)”, pp. 577–581 (IEEE, 2019a).

Wang, Y., X. Ma, Z. Chen, Y. Luo, J. Yi and J. Bailey, “Symmetric cross entropy for robust learning with noisy labels”, in “Proceedings of the IEEE International Conference on Computer Vision”, pp. 322–330 (2019b).

Wiatrak, M., S. V. Albrecht and A. Nystrom, “Stabilizing generative adversarial networks: A survey”, arXiv preprint arXiv:1910.00927 (2019).

Wolberg, D. W., N. Street and O. Mangasarian, “Breast cancer wisconsin (diagnostic) data set”, URL `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)` (1995).

Wu, Y. and Y. Liu, “Robust truncated hinge loss support vector machines”, Journal of the American Statistical Association **102**, 479, 974–983 (2007).

Xiao, H., K. Rasul and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms", (2017).

Xiao, L., 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 "International Conference on Machine Learning", pp. 5393–5402 (PMLR, 2018).

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

Zhang, H., M. Cisse, Y.-D. Dauphin and D. Lopez-Paz, "*mixup*: beyond empirical risk minimization", in "6^{th} ICLR", (2018).

Zhang, H., Y. Yu, J. Jiao, E. Xing, L. El Ghaoui and M. Jordan, "Theoretically principled trade-off between robustness and accuracy", in "International conference on machine learning", pp. 7472–7482 (PMLR, 2019).

Zhang, T., I. Yamane, N. Lu and M. Sugiyama, "A one-step approach to covariate shift adaptation", CoRR **abs/2007.04043** (2020).

Zhang, Y., G. Niu and M. Sugiyama, "Learning noise transition matrix from only noisy labels via total variation regularization", in "38^{th} ICML", pp. 12501–12512 (2021).

Zhang, Z. and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels", Advances in neural information processing systems **31** (2018a).

Zhang, Z. and M.-R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels", in "NeurIPS*31", (2018b).

Zhao, L., 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", pp. 1281–1288 (IEEE, 2010).

Ziyin, L., B. Chen, R. Wang, P. P. Liang, R. Salakhutdinov, L.-P. Morency and M. Ueda, "Learning not to learn in the presence of noisy labels", arXiv preprint arXiv:2002.06541 (2020).

APPENDIX A

APPENDIX TO CHAPTER 2

A.1 α-loss in Binary Classification

A.1.1 Proof of Proposition 2

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

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

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

$$\begin{align}
l^{\alpha}(1, P_{\hat{Y}|X}(1|x)) &= l^{\alpha}(1, \sigma(f(x))) \tag{A.2}\\
&= \frac{\alpha}{\alpha - 1}\left[1 - \sigma(f(x))^{1-1/\alpha}\right] \tag{A.3}\\
&= \tilde{l}^{\alpha}(f(x)). \tag{A.4}
\end{align}$$

If $y = -1$, then

$$\begin{align}
l^{\alpha}(-1, P_{\hat{Y}|X}(-1|x)) &= l^{\alpha}(-1, 1 - P_{\hat{Y}|X}(1|x)) \tag{A.5}\\
&= l^{\alpha}(-1, 1 - \sigma(f(x))) \tag{A.6}\\
&= l^{\alpha}(-1, \sigma(-f(x))) \tag{A.7}\\
&= \frac{\alpha}{\alpha - 1}[1 - \sigma(-f(x))^{1-1/\alpha}] \tag{A.8}\\
&= \tilde{l}^{\alpha}(-f(x)), \tag{A.9}
\end{align}$$

where (A.7) follows from

$$\sigma(x) + \sigma(-x) = 1, \tag{A.10}$$

which can be observed by (4.1). To show the reverse direction of (A.1) we substitute

$$f(x) = \sigma^{-1}(g(x)) = \sigma^{-1}(P_{\hat{Y}|X}(1|x)), \tag{A.11}$$

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

$$\begin{align}
\tilde{l}^{\alpha}(f(x)) &= \tilde{l}^{\alpha}(\sigma^{-1}(P_{\hat{Y}|X}(1|x))) \tag{A.12}\\
&= \frac{\alpha}{\alpha - 1}[1 - (\sigma(\sigma^{-1}(P_{\hat{Y}|X}(1|x))))^{1-1/\alpha}] \tag{A.13}\\
&= \frac{\alpha}{\alpha - 1}[1 - P_{\hat{Y}|X}(1|x)^{1-1/\alpha}] \tag{A.14}\\
&= l^{\alpha}(1, P_{\hat{Y}|X}(1|x)). \tag{A.15}
\end{align}$$

For $y = -1$,

$$\tilde{l}^{\alpha}(-f(x)) = \tilde{l}^{\alpha}(-\sigma^{-1}(P_{\hat{Y}|X}(1|x))) \tag{A.16}$$
$$= \frac{\alpha}{\alpha-1}[1 - \sigma(-\sigma^{-1}(P_{\hat{Y}|X}(1|x)))^{1-1/\alpha}] \tag{A.17}$$
$$= \frac{\alpha}{\alpha-1}[1 - (1-\sigma(\sigma^{-1}(P_{\hat{Y}|X}(1|x))))^{1-1/\alpha}] \tag{A.18}$$
$$= \frac{\alpha}{\alpha-1}[1 - P_{\hat{Y}|X}(-1|x)^{1-1/\alpha}] \tag{A.19}$$
$$= l^{\alpha}(-1, P_{\hat{Y}|X}(-1|x)), \tag{A.20}$$

where (A.18) follows from (A.10).

The equality in the results of the minimization procedures follows from the equality between l^{α} and $\tilde{l}^{\alpha}$. As was shown in Liao *et al.* (2018a), the minimizer of the left-hand-side is

$$P^{*}_{\hat{Y}|X}(y|x) = \frac{P_{Y|X}(y|x)^{\alpha}}{\sum\limits_{y} P_{Y|X}(y|x)^{\alpha}}. \tag{A.21}$$

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

A.1.2 Proof of Proposition 7

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

$$\frac{d^2}{dz^2}\tilde{l}^{\alpha}(z) = \frac{(e^{-z}+1)^{1/\alpha} e^{z}(\alpha e^{z} - \alpha + 1)}{\alpha(e^{z}+1)^3}. \tag{A.22}$$

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 Boyd and Vandenberghe (2004a). 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 $z < \log(1 - \alpha^{-1})$. Thus, the margin-based α-loss, $\tilde{l}^{\alpha}$, is not convex for $\alpha \in (1, \infty]$. However, observe that

$$\frac{d}{dz}\tilde{l}^{\alpha}(z) = \frac{-(e^{-z}+1)^{1/\alpha} e^{z}}{(1+e^{z})^2}. \tag{A.23}$$

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 Boyd and Vandenberghe (2004a), we have that $\tilde{l}^{\alpha}$ is quasi-convex for $\alpha > 1$.

A.1.3 Proof of Theorem 1

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.*, 2006b).

Proposition 8 (Thm. 6, Bartlett *et al.* (2006b))**.** *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 7. Thus, $\tilde{l}^{\alpha}$ satisfies Proposition 8, 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 (2.14) 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

$$\inf_{f \in \mathbb{R}} C_{\tilde{l}^{\alpha}}(\eta, f) = \inf_{f \in \mathbb{R}} \eta \tilde{l}^{\alpha}(f) + (1 - \eta)\tilde{l}^{\alpha}(-f) \tag{A.24}$$

$$= \frac{\alpha}{\alpha - 1}\left(1 - \sup_{f \in \mathbb{R}} \left[\eta \sigma(f)^{1-1/\alpha} + (1 - \eta)\sigma(-f)^{1-1/\alpha}\right]\right), \tag{A.25}$$

where we substituted $\tilde{l}^{\alpha}$ into (A.24) and pulled the infimum through. We take the derivative of the expression inside the supremum, which we denote $g(\eta, \alpha, f)$, and obtain

$$\frac{d}{df} g(\eta, \alpha, f) = \left(1 - \frac{1}{\alpha}\right)\left(\frac{1}{e^{f} + 2 + e^{-f}}\right)\left[\eta \left(1 + e^{-f}\right)^{\frac{1}{\alpha}} - (1 - \eta)\left(1 + e^{f}\right)^{\frac{1}{\alpha}}\right]. \tag{A.26}$$

One can then obtain the f^* minimizing (A.24) by setting $\frac{d}{df} g(\eta, \alpha, f) = 0$, i.e.,

$$\eta \left(1 + e^{-f^*}\right)^{1/\alpha} = (1 - \eta)\left(1 + e^{f^*}\right)^{1/\alpha}, \tag{A.27}$$

and solving for f^* we have

$$f^*_{\alpha}(\eta) = \alpha \log\left(\frac{\eta}{1 - \eta}\right) = \alpha \cdot \sigma^{-1}(\eta). \tag{A.28}$$

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

$$\tilde{l}^{\infty}(z) = 1 - \sigma(z) = \frac{e^{z}}{1 + e^{z}}, \tag{A.29}$$

which is sigmoid loss. Similarly, sigmoid loss can be shown to be classification-calibrated as is given in Bartlett *et al.* (2006b). 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 (A.28) 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 (2.6) for a binary label $\mathcal{Y} = \{-1, +1\}$ into (2.10) as stated by Proposition 2. Indeed, we have that

$$f^*(x) = \sigma^{-1}(P^*_{\hat{Y}|X}(1|x)) \tag{A.30}$$

$$= \log\left(\frac{P_{Y|X}(1|x)^\alpha}{P_{Y|X}(-1|x)^\alpha}\right) \tag{A.31}$$

$$= \alpha \log\left(\frac{\eta(x)}{1-\eta(x)}\right), \tag{A.32}$$

which aligns with (2.17).

A.1.4 Proof of Corollary 1

For $\alpha = 1$, we recover logistic loss and we know from Masnadi-Shirazi and Vasconcelos (2009) and Sypherd *et al.* (2019) that the minimum conditional risk is given by

$$C_1^*(\eta) = -\eta \log \eta - (1-\eta)\log(1-\eta). \tag{A.33}$$

Similarly, for $\alpha = \infty$, we recover the sigmoid loss and we know from Bartlett *et al.* (2006b) and Sypherd *et al.* (2019) that the minimum conditional risk is given by

$$C_\infty^*(\eta) = \min\{\eta, 1-\eta\}. \tag{A.34}$$

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

$$C_\alpha(\eta, f) = \eta \tilde{l}^\alpha(f) + (1-\eta)\tilde{l}^\alpha(-f) \tag{A.35}$$

$$= \frac{\alpha}{\alpha-1}\left[1 - \eta\sigma(f)^{1-1/\alpha} - (1-\eta)\sigma(-f)^{1-1/\alpha}\right], \tag{A.36}$$

where we substituted (4.3) into (A.35). We can obtain the minimum conditional risk upon substituting (2.17) into (A.36) which yields

$$C_\alpha^*(\eta) = \frac{\alpha}{\alpha-1} - \frac{\alpha}{\alpha-1}(1-\eta)\left(\frac{(1-\eta)^\alpha}{\eta^\alpha + (1-\eta)^\alpha}\right)^{1-1/\alpha} - \frac{\alpha}{\alpha-1}\eta\left(\frac{\eta^\alpha}{\eta^\alpha + (1-\eta)^\alpha}\right)^{1-1/\alpha} \tag{A.37}$$

$$= \frac{\alpha}{\alpha-1}\left[1 - (\eta^\alpha + (1-\eta)^\alpha)^{1/\alpha}\right], \tag{A.38}$$

where the last equation is obtained after some algebra. Finally, observe that $C_{1/2}^*(\eta) = 2\sqrt{\eta(1-\eta)}$, which aligns with Masnadi-Shirazi and Vasconcelos (2009).

A.2 Optimization Guarantees for α-loss in the Logistic Model

A.2.1 Proof of Theorem 2

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

$$\begin{aligned}\nabla^2_\theta R_\alpha(\theta) &= \mathbb{E}_{X,Y}[\nabla^2_\theta l^\alpha(Y, g_\theta(X))] && \text{(A.39)}\\ &= \mathbb{E}_{X,Y}[F_2(\alpha, \theta, X, Y)XX^\intercal] && \text{(A.40)}\\ &\succeq \Lambda(\alpha, r\sqrt{d})\mathbb{E}[XX^\intercal] && \text{(A.41)}\\ &= \Lambda(\alpha, r\sqrt{d})\Sigma \succeq 0, && \text{(A.42)}\end{aligned}$$

where we used an identity of positive semi-definite matrices for (A.41) (see, e.g., (Horn and Johnson, 2012, Ch. 7)); for (A.42), 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., (Papoulis and Pillai, 2002, Ch. 7)). We also note that $\min_{i\in[d]} \lambda_i(\Sigma) \geq 0$ (see, e.g., (Horn and Johnson, 2012, Ch. 7)). Thus, $\nabla^2_\theta R_\alpha(\theta)$ 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 (Horn and Johnson, 2012, 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 $\alpha \in (0, 1]$.

A.2.2 Proof of Corollary 2

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

$$\begin{aligned}\nabla^2_\theta R_\alpha(\theta) &= \mathbb{E}_{X,Y}[\nabla^2_\theta l^\alpha(Y, g_\theta(X))] && \text{(A.43)}\\ &= \mathbb{E}_{X,Y}[g_\theta(YX)^{1-1/\alpha}(g'_\theta(YX) - \left(1 - \frac{1}{\alpha}\right) g_\theta(-YX)^2)XX^\intercal] && \text{(A.44)}\\ &= \mathbb{E}_{X,Y}[g_\theta(YX)^{1-1/\alpha} g_\theta(-YX)(g_\theta(YX) - \left(1 - \frac{1}{\alpha}\right) g_\theta(-YX))XX^\intercal], && \text{(A.45)}\end{aligned}$$

where we recall (A.40) and factored out $g_\theta(-YX)$. Considering the expression in parentheses in (A.45), we note that this is the only part of the Hessian which can become negative. Examining this term more closely, we find that

$$\begin{aligned}g_\theta(YX) - \left(1 - \frac{1}{\alpha}\right) g_\theta(-YX) &= \frac{1}{1 + e^{-\langle\theta, YX\rangle}} - \left(1 - \frac{1}{\alpha}\right) \frac{1}{1 + e^{\langle\theta, YX\rangle}} && \text{(A.46)}\\ &= g_\theta(YX)\left[1 - \left(1 - \frac{1}{\alpha}\right) \frac{1 + e^{-\langle\theta, YX\rangle}}{1 + e^{\langle\theta, YX\rangle}}\right] && \text{(A.47)}\\ &= g_\theta(YX)\left[1 - \left(1 - \frac{1}{\alpha}\right) e^{-\langle\theta, YX\rangle}\right]. && \text{(A.48)}\end{aligned}$$

Continuing, observe that

$$\begin{aligned}1 - \left(1 - \frac{1}{\alpha}\right) e^{-\langle\theta, YX\rangle} &= 1 - e^{-\langle\theta, YX\rangle} + \frac{e^{-\langle\theta, YX\rangle}}{\alpha} && \text{(A.49)}\\ &\geq 1 - e^{r\sqrt{d}} + \frac{e^{-r\sqrt{d}}}{\alpha} \geq 0, && \text{(A.50)}\end{aligned}$$

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 (A.50) holds if $\alpha \leq e^{-r\sqrt{d}}(e^{r\sqrt{d}}-1)^{-1}$. Thus, returning to (A.45), we have that

$$\nabla^2_\theta R_\alpha(\theta) = \mathbb{E}_{X,Y}[g_\theta(YX)^{1-\frac{1}{\alpha}} g'_\theta(YX)(1-(1-\frac{1}{\alpha})e^{-\langle \theta, YX \rangle})XX^\intercal] \tag{A.51}$$

$$\succeq \sigma(-r\sqrt{d})^{2-\frac{1}{\alpha}} \sigma(r\sqrt{d}) \left(1 - e^{r\sqrt{d}} + \frac{e^{-r\sqrt{d}}}{\alpha}\right) \mathbb{E}\left[XX^\intercal\right] \tag{A.52}$$

$$= \sigma(-r\sqrt{d})^{2-\frac{1}{\alpha}} \sigma(r\sqrt{d}) \left(1 - e^{r\sqrt{d}} + \frac{e^{-r\sqrt{d}}}{\alpha}\right) \Sigma \succeq 0, \tag{A.53}$$

where in (A.51) we used (A.48) and the fact as given in (2.28) that $\sigma'(z) = \sigma(z)\sigma(-z)$, and in (A.52) and (A.53) we use the upper-bound derived above and the same arguments as Theorem 2, *mutatis mudandis*. 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 $\theta \in \mathbb{B}_d(r)$,

$$\tilde{\Lambda}(\alpha, r\sqrt{d}) := \sigma(-r\sqrt{d})^{2-1/\alpha} \sigma(r\sqrt{d}) \left(1 - e^{r\sqrt{d}} + \alpha^{-1} e^{-r\sqrt{d}}\right). \tag{A.54}$$

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 $\operatorname{arcsinh}(1/2) \approx 0.48$.

A.2.3 Proof of Proposition 5

In order to prove the result, we apply a result by Hazan, *et al.* Hazan *et al.* (2015) 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 $\alpha \in (0,1]$,

$$C_d(r,\alpha) := \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha}; \tag{A.55}$$

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

$$C_d(r,\alpha) := \begin{cases} \sqrt{d}\left(\frac{\alpha-1}{2\alpha-1}\right)^{1-1/\alpha}\left(\frac{\alpha}{2\alpha-1}\right) & e^{r\sqrt{d}} \geq \frac{\alpha-1}{\alpha}, \\ \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha} & e^{r\sqrt{d}} < \frac{\alpha-1}{\alpha}. \end{cases} \tag{A.56}$$

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

$$|R_\alpha(\theta) - R_\alpha(\theta')| \leq C\|\theta - \theta'\|, \tag{A.57}$$

where $C := \sup_{\theta \in \mathbb{B}_d(r)} \|\nabla R_\alpha(\theta)\|$. Recall from (2.30) that

$$\begin{aligned}\nabla_\theta R_\alpha(\theta) &= \mathbb{E}[\nabla_\theta l^\alpha(Y, g_\theta(X)] &\text{(A.58)}\\ &= \mathbb{E}[F_1(\alpha, \theta, X, Y)X], &\text{(A.59)}\end{aligned}$$

where from (2.29) we have

$$F_1(\alpha, \theta, x, y) = -yg_\theta(yx)^{1-1/\alpha}(1 - g_\theta(yx)). \tag{A.60}$$

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

$$|F_1(\alpha, \theta, x, y)| = g_\theta(yx)^{1-1/\alpha}(1 - g_\theta(yx)), \tag{A.61}$$

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

$$C = \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha}. \tag{A.62}$$

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 $\alpha > 1$,

$$C = \sqrt{d}\left(\frac{\alpha - 1}{2\alpha - 1}\right)^{1-1/\alpha}\left(\frac{\alpha}{2\alpha - 1}\right). \tag{A.63}$$

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

$$C = \sqrt{d}\sigma(r\sqrt{d})\sigma(-r\sqrt{d})^{1-1/\alpha}, \tag{A.64}$$

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 (2.37) and (2.38).

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 $r\sqrt{d} < \text{arcsinh}\,(1/2)$. Similarly, we apply Theorem 2 to obtain strong-convexity of R_α for $\alpha \in (0, 1]$, otherwise.

A.2.4 Fundamentals of SLQC and Reformulation

In this subsection, we briefly review *strictly locally quasi-convexity* (SLQC) which was introduced by Hazan *et al.* in Hazan *et al.* (2015). Recall that in Hazan *et al.* (2015) 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 $\mathbb{B}(\theta_0, r) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\| \leq r\}$. SLQC definition from Hazan *et al.* (2015). 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 $\theta' \in \mathbb{B}(\theta_0, \epsilon/\kappa)$,

$$\langle -\nabla f(\theta), \theta' - \theta \rangle \geq 0. \tag{A.65}$$

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. 2.5 for an illustration of the difference between classical quasi-convexity and SLQC in this regime. In Hazan *et al.* (2015), 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 Hazan *et al.* (2015), the convergence guarantees of Normalized Gradient Descent (NGD, given in Algorithm 3) for SLQC functions are similar to those of Gradient Descent for convex functions.

Algorithm 3 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: $t = 0, 1, \ldots, T-1$
3:         Update: $\theta_{t+1} = \theta_t - \eta \frac{\nabla f(\theta_t)}{\|\nabla f(\theta_t)\|}$
4: Return $\bar{\theta}_T = \underset{\theta_1,\ldots,\theta_T}{\arg\min} f(\theta_t)$
```

From Hazan *et al.* (2015), we have the following result.

Proposition 9. *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. Hazan, *et al.* Hazan *et al.* (2015) 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 10. *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

$$\theta' := \theta + \delta \nabla f(\theta) \in \mathbb{B}(\theta_0, \epsilon/\kappa). \tag{A.66}$$

Thus, we have that

$$\begin{aligned} 0 &\leq \langle -\nabla f(\theta), \theta' - \theta \rangle && \text{(A.67)} \\ &= \langle -\nabla f(\theta), \theta + \delta \nabla f(\theta) - \theta \rangle && \text{(A.68)} \\ &= -\delta \langle \nabla f(\theta), \nabla f(\theta) \rangle && \text{(A.69)} \\ &= -\delta \|\nabla f(\theta)\|^2, && \text{(A.70)} \end{aligned}$$

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 $\theta \in \overline{\mathbb{B}_d(\theta_0, \epsilon/\kappa) \cap \Theta}$. □

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. A.1 for a picture of the relevant quantities.

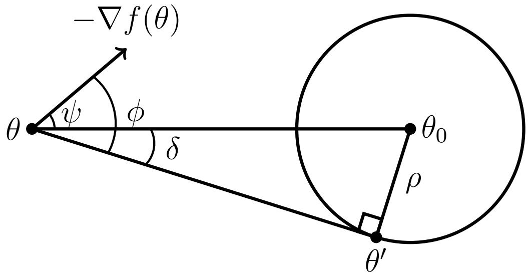


Figure A.1: A Companion Illustration for Lemma 1 Which Depicts the Relevant Quantities Involved. Note That There Are Three Different Configurations of the Angles δ, ϕ and ψ. Refer to Fig. A.2 for This Illustration.

A.2.5 Proof of 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) $\langle -\nabla f(\theta), \theta_0 - \theta \rangle \geq \rho \|\nabla f(\theta)\|$.

Clearly (1) ⇒ (2). (2) ⇒ (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. A.1. 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

$$\langle u, v\rangle = \|u\|\|v\| \cos(\varphi_{u,v}), \tag{A.71}$$

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

$$\|\nabla f(\theta)\|\|\theta' - \theta\| \cos(\phi) = \langle -\nabla f(\theta), \theta' - \theta\rangle \geq 0, \tag{A.72}$$

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

$$\cos(\psi) \geq \cos\left(\frac{\pi}{2} - \delta\right) = \sin(\delta). \tag{A.73}$$

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

$$\cos(\psi) \geq \frac{\rho}{\|\theta_0 - \theta\|}. \tag{A.74}$$

Therefore, we conclude that

$$\begin{aligned} \langle -\nabla f(\theta), \theta_0 - \theta\rangle &= \|\nabla f(\theta)\|\|\theta_0 - \theta\| \cos(\psi) && \text{(A.75)}\\ &\geq \rho\|\nabla f(\theta)\|, && \text{(A.76)} \end{aligned}$$

as we wanted to prove.

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

$$\begin{aligned} \|\nabla f(\theta)\|\|\theta_0 - \theta\| \cos(\psi) &= \langle -\nabla f(\theta), \theta_0 - \theta\rangle && \text{(A.77)}\\ &\geq \rho\|\nabla f(\theta)\| \geq 0. && \text{(A.78)} \end{aligned}$$

Since $\cos^{-1}(\cdot)$ is decreasing over $[-1, 1]$, (A.77) implies that

$$\psi \leq \cos^{-1}\left(\frac{\rho}{\|\theta_0 - \theta\|}\right). \tag{A.79}$$

Also, an immediate application of the law of cosines yields

$$\delta = \cos^{-1}\left(\frac{\|\theta_0 - \theta\|^2 + \|\theta' - \theta\|^2 - \|\theta' - \theta_0\|^2}{2\|\theta_0 - \theta\|\|\theta' - \theta\|}\right). \tag{A.80}$$

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

$$\delta < \cos^{-1}\left(\frac{\|\theta_0 - \theta\|^2 + \|\theta' - \theta\|^2 - \rho^2}{2\|\theta_0 - \theta\|\|\theta' - \theta\|}\right). \tag{A.81}$$

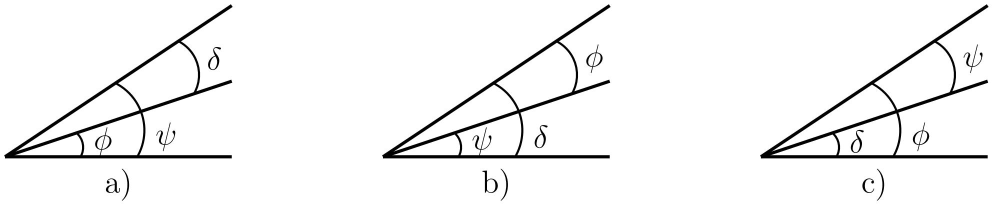


Figure A.2: Three Different Configurations of the Angles δ, ϕ and ψ.

A routine minimization argument further implies that

$$\delta < \cos^{-1}\left(\sqrt{1-\left(\frac{\rho}{\|\theta_0-\theta\|}\right)^2}\right) = \sin^{-1}\left(\frac{\rho}{\|\theta_0-\theta\|}\right), \tag{A.82}$$

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

$$\langle -\nabla f(\theta), \theta' - \theta\rangle = \|\nabla f(\theta)\|\|\theta'-\theta\|\cos(\phi) > 0, \tag{A.83}$$

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. A.2.

a) Since $\frac{\rho}{\|\theta_0-\theta\|} > 0$, (A.79) implies that $\psi < \frac{\pi}{2}$. Therefore, $\phi < \frac{\pi}{2}$ as $\phi \leq \psi$.

b) Since $\frac{\rho}{\|\theta_0-\theta\|} < 1$, (A.82) implies that $\delta < \frac{\pi}{2}$. Therefore, $\phi < \frac{\pi}{2}$ as $\phi \leq \delta$.

c) Since $\sin^{-1}(x) + \cos^{-1}(x) = \frac{\pi}{2}$, (A.79) and (A.82) imply that $\phi = \psi + \delta < \frac{\pi}{2}$.

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

A.2.6 Lipschitz Inequalities in α^{-1} and Main SLQC Result for the α-risk

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

$$|R_\alpha(\theta) - R_{\alpha'}(\theta)| \leq L_d(\theta)\left|\frac{\alpha-\alpha'}{\alpha\alpha'}\right|, \tag{A.84a}$$

$$\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\| \leq J_d(\theta)\left|\frac{\alpha-\alpha'}{\alpha\alpha'}\right|, \tag{A.84b}$$

where,

$$L_d(\theta) := \frac{\left(\log\left(1+e^{\|\theta\|\sqrt{d}}\right)\right)^2}{2}, \tag{A.85a}$$

$$J_d(\theta) := \sqrt{d}\log\left(1+e^{\|\theta\|\sqrt{d}}\right)\sigma(\|\theta\|\sqrt{d}). \tag{A.85b}$$

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

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

$$
\begin{aligned}
R_\alpha(\theta) &= \mathbb{E}[l^\alpha(Y, g_\theta(X))] && \text{(A.86)}\\
&= \mathbb{E}\left[\frac{1}{1-\beta}\left(1 - g_\theta(yx)^{1-\beta}\right)\right] && \text{(A.87)}\\
&= R_\beta(\theta). && \text{(A.88)}
\end{aligned}
$$

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

$$
\frac{d}{d\beta} R_\beta(\theta) = \mathbb{E}\left[\frac{d}{d\beta}\frac{1}{1-\beta}\left(1 - g_\theta(yx)^{1-\beta}\right)\right], \tag{A.89}
$$

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

$$
f(\beta, \theta, yx) = \frac{d}{d\beta}\frac{1}{1-\beta}\left(1 - g_\theta(yx)^{1-\beta}\right). \tag{A.90}
$$

It can be shown that

$$
f(\beta, \theta, yx) = \frac{g_\theta(yx)^{1-\beta}\log\left(g_\theta(yx)\right)}{1-\beta} + \frac{1 - g_\theta(yx)^{1-\beta}}{(1-\beta)^2}, \tag{A.91}
$$

and

$$
f(1, \theta, yx) = \frac{\left(\log g_\theta(yx)\right)^2}{2}. \tag{A.92}
$$

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 $\theta \in \mathbb{B}_d(r)$,

$$
\begin{aligned}
f(\beta, \theta, yx) &\leq f(1, \theta, yx) && \text{(A.93)}\\
&= \frac{\left(\log g_\theta(yx)\right)^2}{2} && \text{(A.94)}\\
&\leq \frac{\left(\log \sigma(-\|\theta\|\sqrt{d})\right)^2}{2}. && \text{(A.95)}
\end{aligned}
$$

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., $\beta \in [0, 1]$,

$$
\begin{aligned}
\nabla R_\alpha(\theta) &= \mathbb{E}[F_1(\alpha, \theta, X, Y)X] && \text{(A.96)}\\
&= \mathbb{E}[-Y g_\theta(YX)^{1-\beta}(1 - g_\theta(YX))X], && \text{(A.97)}
\end{aligned}
$$

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

$$\begin{aligned}
&\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\| \\
&= \|\mathbb{E}[(\tilde{F}_1(\beta, \theta, X, Y) - \tilde{F}_1(\beta', \theta, X, Y))X]\| && \text{(A.98)} \\
&\leq \mathbb{E}[|(\tilde{F}_1(\beta, \theta, X, Y) - \tilde{F}_1(\beta', \theta, X, Y))|\|X\|] && \text{(A.99)} \\
&\leq \sqrt{d}\mathbb{E}[|(\tilde{F}_1(\beta, \theta, X, Y) - \tilde{F}_1(\beta', \theta, X, Y))|], && \text{(A.100)}
\end{aligned}$$

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

$$\begin{aligned}
\frac{d}{d\beta} F_1(\beta, \theta, x, y) &= \frac{d}{d\beta} - yg_\theta(yx)^{1-\beta}(1-g_\theta(yx)) && \text{(A.101)} \\
&= y(1-g_\theta(yx))g_\theta(yx)^{1-\beta} \log g_\theta(yx), && \text{(A.102)}
\end{aligned}$$

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

$$\begin{aligned}
&y(1-g_\theta(yx))g_\theta(yx)^{1-\beta}\log g_\theta(yx) \\
&\leq \log\left(1+e^{\|\theta\|\sqrt{d}}\right)\sigma(\|\theta\|\sqrt{d})\sigma(\|\theta\|\sqrt{d})^{1-\beta} && \text{(A.103)} \\
&= \log\left(1+e^{\|\theta\|\sqrt{d}}\right)\sigma(\|\theta\|\sqrt{d})^{2-\beta} && \text{(A.104)} \\
&\leq \log\left(1+e^{\|\theta\|\sqrt{d}}\right)\sigma(\|\theta\|\sqrt{d}). && \text{(A.105)}
\end{aligned}$$

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

$$\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\| \leq J_d(\theta)|\beta - \beta'|, \tag{A.106}$$

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

$$\|\nabla R_\alpha(\theta) - \nabla R_{\alpha'}(\theta)\| \leq J_d(\theta)\left|\frac{1}{\alpha} - \frac{1}{\alpha'}\right|, \tag{A.107}$$

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

A.2.7 Proof of Theorem 3

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

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

$$\begin{aligned}
&R_\alpha(\theta) - R_\alpha(\theta_0) \\
&= R_\alpha(\theta) - R_{\alpha_0}(\theta) + R_{\alpha_0}(\theta) - R_{\alpha_0}(\theta_0) + R_{\alpha_0}(\theta_0) - R_\alpha(\theta_0) && \text{(A.108)} \\
&\leq L_d(\theta)\left(\frac{\alpha - \alpha_0}{\alpha\alpha_0}\right) + \epsilon_0 + L_d(\theta)\left(\frac{\alpha - \alpha_0}{\alpha\alpha_0}\right). && \text{(A.109)}
\end{aligned}$$

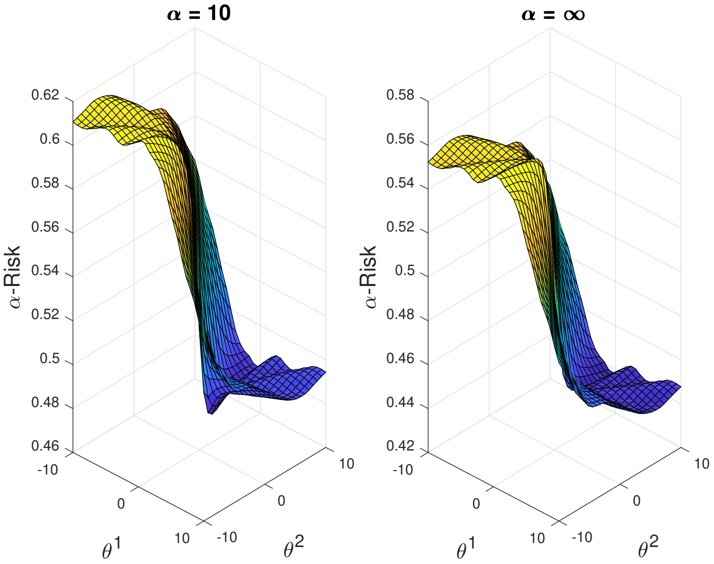


Figure A.3: Another Illustration Highlighting the Saturation of α-loss (r_α for $\alpha = 10, \infty$) in the Logistic Model for a 2D-GMM with $\mathbb{P}[Y = 1] = .5$, $\mu_{X|y=-1} = [.5, .5]$, $\mu_{X|y=1} = [1, 1]$, and Shared Covariance Matrix $\sigma = [1, .5; .5, 3]$.

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 (2.43). If $\|\theta - \theta_0\| > \rho$, $\|\nabla R_\alpha(\theta)\| > 0$ and

$$\langle -\nabla R_\alpha(\theta), \theta_0 - \theta\rangle \geq \rho\|\nabla R_\alpha(\theta)\|, \tag{A.110}$$

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.,

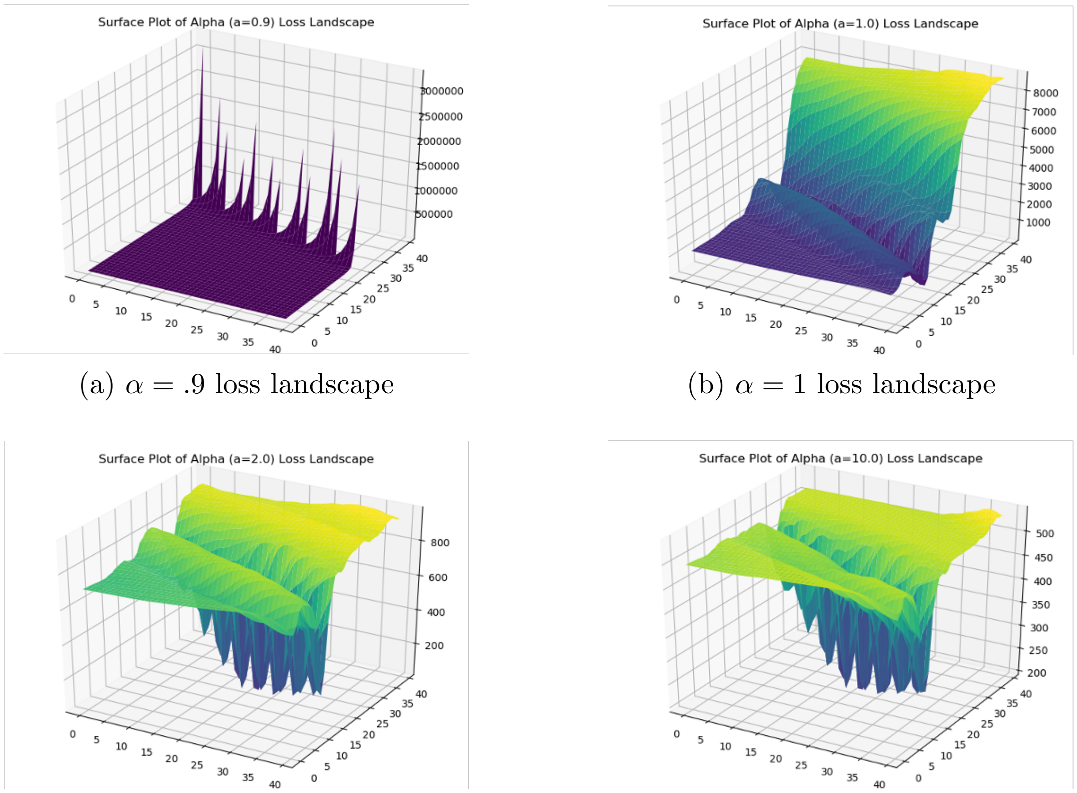
$$\|\nabla R_{\alpha_0}(\theta) - \nabla R_\alpha(\theta)\| \geq |\|\nabla R_\alpha(\theta)\| - \|\nabla R_{\alpha_0}(\theta)\||. \tag{A.111}$$

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

$$\left|\frac{1}{\alpha_0} - \frac{1}{\alpha}\right| J_d(\theta) \geq \|\nabla R_{\alpha_0}(\theta) - \nabla R_\alpha(\theta)\|. \tag{A.112}$$

The third follows from a manipulation of (2.42), i.e.,

$$\begin{aligned}\|\nabla R_{\alpha_0}(\theta)\| &> 2J_d(\theta)\left(1 + r\rho_0^{-1}\right)\left(\alpha_0^{-1} - \alpha^{-1}\right) &\text{(A.113)}\\ &> J_d(\theta)(\alpha_0^{-1} - \alpha^{-1}), &\text{(A.114)}\end{aligned}$$



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

(b) $\alpha = 1$ loss landscape

(c) $\alpha = 2$ loss landscape

(d) $\alpha = 10$ loss landscape

Figure A.4: Loss Landscape Visualizations Obtained Using Li *et al.* (2018a) 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 2.4.3 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 2.4.3.

using the fact that $\alpha_0^2 \leq \alpha\alpha_0$ and since $r\rho_0^{-1} \geq 1$. 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 10 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 (2.43). If we can show that $\rho_0 > \rho$, then we have the desired conclusion. Rearranging the left-hand-side of (A.113), we have that

$$\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} > 2J_d(\theta)(1 + r\rho_0^{-1}), \tag{A.115}$$

which can be rewritten to obtain

$$\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_d(\theta) > J_d(\theta)(1 + 2r\rho_0^{-1}). \tag{A.116}$$

Since by the right-hand-side of (A.113) we have that

$$\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_d(\theta) > 0, \tag{A.117}$$

it follows by (A.116) that

$$1 > \frac{J_d(\theta)(1 + 2r\rho_0^{-1})}{\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_d(\theta)}. \tag{A.118}$$

Thus examining (2.43) in light of (A.118), 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 (A.111) to (A.112) we obtain

$$\|\nabla R_\alpha(\theta)\| \geq \|\nabla R_{\alpha_0}(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1}) > 0, \tag{A.119}$$

where the right-hand-side inequality again follows by (A.113). Thus, we have that $\|\nabla R_\alpha(\theta)\| > 0$, as desired.

Finally, we show the expression in (A.110), i.e., $\langle -\nabla R_\alpha(\theta), \theta_0 - \theta\rangle \geq \rho\|\nabla R_\alpha(\theta)\|$. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\langle -\nabla R_\alpha(\theta), \theta_0 - \theta\rangle \\ &\geq \langle -\nabla R_{\alpha_0}(\theta), \theta_0 - \theta\rangle - \|\nabla R_\alpha(\theta) - \nabla R_{\alpha_0}(\theta)\|\|\theta_0 - \theta\| \qquad (A.120) \\ &\geq \rho_0\|\nabla R_{\alpha_0}(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})2r, \qquad (A.121) \end{aligned}$$

where in (A.120) 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 (A.112) and the fact that $\theta_0 - \theta \in \mathbb{B}_d(2r)$. Continuing from (A.121), we have that

$$\begin{aligned} &\langle -\nabla R_\alpha(\theta), \theta_0 - \theta\rangle \\ &\geq \rho_0\|\nabla R_\alpha(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})2r - \rho_0\|\nabla R_{\alpha_0}(\theta) - \nabla R_\alpha(\theta)\| \qquad (A.122) \\ &\geq \rho_0\|\nabla R_\alpha(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})(\rho_0 + 2r), \qquad (A.123) \end{aligned}$$

where we first apply the reverse triangle inequality in (A.111) and then we use the fact that $\nabla R_\alpha(\theta)$ is $J_d(\theta)$-Lipschitz in α^{-1}, i.e., the expression in (A.112). Rearranging the expression in (A.123), we obtain

$$\begin{aligned} &\rho_0\|\nabla R_\alpha(\theta)\| - J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})(\rho_0 + 2r) \\ &= \|\nabla R_\alpha(\theta)\|\left(\rho_0 - \frac{J_d(\theta)(\alpha_0^{-1} - \alpha^{-1})(\rho_0 + 2r)}{\|\nabla R_\alpha(\theta)\|}\right) \qquad (A.124) \\ &\geq \|\nabla R_\alpha(\theta)\|\left(\rho_0 - \frac{(\rho_0 + 2r)J_d(\theta)}{\frac{\|\nabla R_{\alpha_0}(\theta)\|}{(\frac{1}{\alpha_0} - \frac{1}{\alpha})} - J_d(\theta)}\right), \qquad (A.125) \end{aligned}$$

where we used the inequality in (A.119). Thus, we finally obtain

$$\langle -\nabla R_\alpha(\theta), \theta_0 - \theta\rangle \geq \rho \|\nabla R_\alpha(\theta)\|, \tag{A.126}$$

where $\rho > 0$ is given by

$$\rho = \rho_0 \left(1 - \frac{(1 + 2r\rho_0^{-1})J_d(\theta)}{\|\nabla R_{\alpha_0}(\theta)\|(\alpha_0^{-1} - \alpha^{-1})^{-1} - J_d(\theta)}\right), \tag{A.127}$$

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

A.2.8 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 10. *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*

$$\alpha_n = \alpha_{n-1} + \frac{1}{N}, \tag{A.128a}$$

$$\epsilon_n = \epsilon_{n-1} + 2L_d(\theta)\frac{1}{\alpha_n \alpha_{n-1}}\frac{1}{N}, \tag{A.128b}$$

$$\rho_n = \rho_{n-1} - \frac{(\rho_{n-1} + 2r)J_d(\theta)}{\alpha_n\alpha_{n-1}G_{n-1} - J_d(\theta)/N}\frac{1}{N}, \tag{A.128c}$$

where $G_{n-1} \coloneqq \|\nabla R_{\alpha_{n-1}}(\theta)\|$. 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 $n \leq \left\lfloor \alpha_0^2 g_\theta (1 + 2r\rho_0^{-1})^{-1} J_d(\theta)^{-1} N \right\rfloor$.

Proof. For ease of notation, let $J \coloneqq J_d(\theta)$, $L \coloneqq L_d(\theta)$, and $g \coloneqq g_\theta$. 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

$$n < \left\lfloor \frac{\rho_0}{\rho_0 + 2r}\frac{\alpha_0^2 g}{J} N \right\rfloor. \tag{A.129}$$

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

$$\rho_k - \rho_{k+1} = \frac{(\rho_k + 2r)J}{\alpha_k \alpha_{k+1} G_k - J/N}\frac{1}{N}. \tag{A.130}$$

The previous equation and a telescoping sum lead to

$$\rho_0 - \rho_n = \sum_{k=0}^{n-1} \frac{(\rho_k + 2r)J}{\alpha_k \alpha_{k+1} G_k - J/N}\frac{1}{N}. \tag{A.131}$$

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,

$$\rho_0 - \rho_n < \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N} \frac{n}{N}. \tag{A.132}$$

It can be shown that our choice of n in (A.129) implies that

$$\rho_n > \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N} \frac{n}{N} > 0, \tag{A.133}$$

which implies that $\rho_n > 0$ as desired. □

A.2.9 Proof of Theorem 4

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

$$N > \frac{1 + 2r\rho_0^{-1}}{1-\lambda} \frac{2J}{\alpha_0^2 g}, \tag{A.134}$$

we define

$$N_\lambda = \left\lfloor \lambda \frac{\rho_0}{\rho_0 + 2r} \frac{\alpha_0^2 g}{J} N \right\rfloor. \tag{A.135}$$

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 $n \leq N_\lambda$ using Lemma 10 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 $n \leq N_\lambda$. 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 10 to show that

$$\rho_0 > \rho_1 > \ldots > \rho_n > \cdots > \rho_{N_\lambda} > C_\lambda > 0, \tag{A.136}$$

for all $n \leq N_\lambda$ and for some $C_\lambda > 0$, we first show that the assumptions of Lemma 10 are satisfied. Observe that, by our assumption on $N \in \mathbb{N}$, we have that

$$N > \frac{1 + 2r\rho_0^{-1}}{1-\lambda} \frac{2J}{\alpha_0^2 g} > \frac{1 + r\rho_0^{-1}}{1-\lambda} \frac{J}{\alpha_0^2 g} > \frac{J}{\alpha_0^2 g}, \tag{A.137}$$

which is the first requirement of Lemma 10. Next, we want to show that

$$n \leq N_\lambda < \left\lfloor \frac{\rho_0}{\rho_0 + 2r} \frac{\alpha_0^2 g}{J} N \right\rfloor, \tag{A.138}$$

which is the last requirement of Lemma 10. This is achieved by observing that

$$N_\lambda = \left\lfloor \lambda \frac{\rho_0}{\rho_0 + 2r} \frac{\alpha_0^2 g}{J} N \right\rfloor = \lambda \frac{\rho_0}{\rho_0 + 2r} \frac{\alpha_0^2 g}{J} N - q, \tag{A.139}$$

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

$$\left\lfloor \frac{\rho_0}{\rho_0 + 2r} \frac{\alpha_0^2 g}{J} N \right\rfloor = \frac{\rho_0}{\rho_0 + 2r} \frac{\alpha_0^2 g}{J} N - w, \tag{A.140}$$

also for some $w \in [0, 1)$. Note that (A.138) is equivalent to

$$(q - w) \frac{1 + r\rho_0^{-1}}{1 - \lambda} \frac{J}{\alpha_0^2 g} < N, \tag{A.141}$$

which holds by the fact that $N > \frac{1 + r\rho_0^{-1}}{1 - \lambda} \frac{J}{\alpha_0^2 g}$ in (A.137) and $q - w \leq 1$. Thus by Lemma 10, we have that

$$\rho_n > \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N} \frac{n}{N} > 0, \tag{A.142}$$

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

$$\begin{aligned}
\rho_{N_\lambda} &> \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N} \frac{N_\lambda}{N} && \text{(A.143)} \\
&> \rho_0 \left(1 - \lambda - \frac{\lambda J}{\alpha_0^2 g - J/N} \frac{1}{N} \right) && \text{(A.144)} \\
&> \frac{\rho_0(1 - \lambda)}{2}, && \text{(A.145)}
\end{aligned}$$

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 $2r\rho_0^{-1} \geq \lambda$ for all $\lambda \in (0, 1)$. Therefore, we have that $C_\lambda = \frac{\rho_0(1 - \lambda)}{2}$; in other words,

$$\rho_0 > \rho_1 > \ldots > \rho_{n-1} > \rho_n > \cdots > \rho_{N_\lambda} > \frac{\rho_0(1 - \lambda)}{2} > 0. \tag{A.146}$$

Also, observe that

$$\alpha_n - \alpha_{n-1} = \frac{1}{N} < \frac{\alpha_0^2 g}{2J(1 + 2r\rho_0^{-1}(1 - \lambda)^{-1})}, \tag{A.147}$$

where the inequality follows from the fact that

$$N > \frac{1 + 2r\rho_0^{-1}}{1 - \lambda} \frac{2J}{\alpha_0^2 g} > \left(1 + \frac{2r\rho_0^{-1}}{1 - \lambda} \right) \frac{2J}{\alpha_0^2 g}. \tag{A.148}$$

In particular, (A.146) and (A.147) leads to

$$\alpha_n - \alpha_{n-1} < \frac{\alpha_0^2 g}{2J(1 + 2r\rho_0^{-1}(1-\lambda)^{-1})} < \frac{\alpha_{n-1}^2 G_{n-1}}{2J(1 + r\rho_{n-1}^{-1})}, \tag{A.149}$$

where we use the fact that $\alpha_n \geq \alpha_0$ and $G_{n-1} \geq g$. 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 (A.128a). 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

$$\alpha_{N_\lambda} = \alpha_0 + \lambda(1 + 2r\rho_0^{-1})^{-1}\frac{\alpha_0^2 g}{J} - \frac{q}{N}, \tag{A.150}$$

$$\epsilon_{N_\lambda} = \epsilon_0 + 2L\sum_{n=0}^{N_\lambda - 1} \frac{1}{\alpha_n(\alpha_n + 1/N)}\frac{1}{N}, \tag{A.151}$$

$$\rho_{N_\lambda} = \rho_0 \prod_{n=0}^{N_\lambda - 1}\left(1 - \frac{(1 + 2r\rho_n^{-1})J/N}{\alpha_{n+1}\alpha_n\|\nabla R_{\alpha_n}(\theta)\| - J/N}\right), \tag{A.152}$$

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

$$\frac{1}{N} = \frac{\lambda\rho_0\alpha_0^2 g}{(N_\lambda + q)(\rho_0 + r)J}. \tag{A.153}$$

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

Examining (A.150) as N_λ approaches infinity, we have that

$$\alpha_\lambda := \lim_{N_\lambda \to \infty} \alpha_{N_\lambda} = \alpha_0 + \lambda(1 + 2r\rho_0^{-1})^{-1}\frac{\alpha_0^2 g}{J}. \tag{A.154}$$

Next considering (A.151), we rewrite to obtain

$$\epsilon_{N_\lambda} = \epsilon_0 + 2L\sum_{n=0}^{N_\lambda - 1}\frac{1}{\alpha_n(\alpha_n + 1/N)}\frac{1}{N} \tag{A.155}$$

$$= \epsilon_0 + \frac{2L}{N}\sum_{n=0}^{N_\lambda - 1}\left(\frac{1}{\alpha_n^2} + \frac{1}{N}\frac{1}{\alpha_n^3 - \alpha_n^2/N}\right), \tag{A.156}$$

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

$$\mu_{N_\lambda} = \frac{1}{N_\lambda}\sum_{n=0}^{N_\lambda - 1}\delta_{\alpha_n}, \tag{A.157}$$

where δ_{α_n} is the point mass at α_n. In particular for large N, we can write (A.156) as

$$\epsilon_{N_\lambda} = \epsilon_0 + \frac{2L\lambda\alpha_0^2 g}{(1+r\rho_0^{-1})J}\int \frac{1}{x^2} d\mu_{N_\lambda}(x) + O\left(\frac{1}{N_\lambda}\right). \tag{A.158}$$

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, (A.158) becomes

$$\begin{aligned}
\epsilon_\lambda &= \lim_{N_\lambda \to \infty} \epsilon_{N_\lambda} && \text{(A.159)}\\
&= \epsilon_0 + \frac{2L\lambda\alpha_0^2 g}{(1+r\rho_0^{-1})J}\int\limits_{\alpha_0}^{\alpha_\lambda} \frac{1}{x^2} dx && \text{(A.160)}\\
&= \epsilon_0 + \frac{2L\lambda\alpha_0 g}{(1+r\rho_0^{-1})J}\left(1 - \frac{\alpha_0}{\alpha_\lambda}\right). && \text{(A.161)}
\end{aligned}$$

Finally, we consider (A.152). Observe that from (A.133) we have

$$\begin{aligned}
\rho_{N_\lambda} &> \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N}\frac{N_\lambda}{N} && \text{(A.162)}\\
&= \rho_0 - \frac{(\rho_0 + 2r)J}{\alpha_0^2 g - J/N}\frac{\lambda\frac{N\alpha_0^2 g\rho_0}{J(\rho_0+2r)} - q}{N} && \text{(A.163)}\\
&= \rho_0 - \left[\frac{N\lambda\rho_0\alpha_0^2 g}{N\alpha_0^2 g - J} - \frac{q}{N}\left(\frac{(\rho_0+2r)J}{\alpha_0^2 g - J/N}\right)\right], && \text{(A.164)}
\end{aligned}$$

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

$$\begin{aligned}
\rho_\lambda &= \lim_{N_\lambda \to \infty} \rho_{N_\lambda} && \text{(A.165)}\\
&> \lim_{N_\lambda \to \infty}\left(\rho_0 - \left[\frac{N\lambda\rho_0\alpha_0^2 g}{N\alpha_0^2 g - J} - \frac{q}{N}\left(\frac{(\rho_0+2r)J}{\alpha_0^2 g - J/N}\right)\right]\right) && \text{(A.166)}\\
&= \rho_0(1-\lambda). && \text{(A.167)}
\end{aligned}$$

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

$$\begin{aligned}
\alpha_\lambda &\coloneqq \alpha_0 + \lambda(1 + 2r\rho_0^{-1})^{-1}\frac{\alpha_0^2 g}{J}, && \text{(A.168)}\\
\epsilon_\lambda &\coloneqq \epsilon_0 + \frac{2L\lambda\alpha_0 g}{(1+r\rho_0^{-1})J}\left(1 - \frac{\alpha_0}{\alpha_\lambda}\right) && \text{(A.169)}\\
\rho_\lambda &> \rho_0(1-\lambda). && \text{(A.170)}
\end{aligned}$$

A change of variables leads to the desired result.

Lemma 11. *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 $\alpha \in (0, 1]$,*

$$C_{r_0}(\alpha) := \sigma(r_0)\sigma(-r_0)^{1-1/\alpha}; \tag{A.171}$$

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

$$C_{r_0}(\alpha) := \begin{cases} \left(\frac{\alpha-1}{2\alpha-1}\right)^{1-\frac{1}{\alpha}} \left(\frac{\alpha}{2\alpha-1}\right) & e^{r_0} \geq \frac{\alpha-1}{\alpha}, \\ \sigma(r_0)\sigma(-r_0)^{1-\frac{1}{\alpha}} & e^{r_0} < \frac{\alpha-1}{\alpha}. \end{cases} \tag{A.172}$$

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 $\alpha \in (0, \infty]$,

$$\begin{aligned} \frac{d}{dz}\tilde{l}^\alpha(z) &= \frac{d}{dz}\frac{\alpha}{\alpha-1}\left(1 - \sigma(z)^{1-1/\alpha}\right) && \text{(A.173)} \\ &= \sigma(z)^{2-1/\alpha} - \sigma(z)^{1-1/\alpha} && \text{(A.174)} \\ &= (\sigma(z) - 1)\sigma(z)^{1-1/\alpha} && \text{(A.175)} \\ &\leq |(\sigma(z) - 1)\sigma(z)^{1-1/\alpha}| && \text{(A.176)} \\ &= \sigma(-z)\sigma(z)^{1-1/\alpha}, && \text{(A.177)} \end{aligned}$$

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

$$\max_{z \in [-r_0, r_0]} \sigma(-z)\sigma(z)^{1-1/\alpha} = \sigma(r_0)\sigma(-r_0)^{1-1/\alpha}. \tag{A.178}$$

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

$$\max_{z \in [-r_0, r_0]} \sigma(-z)\sigma(z)^{1-1/\alpha} = \left(\frac{\alpha-1}{2\alpha-1}\right)^{1-1/\alpha} \left(\frac{\alpha}{2\alpha-1}\right), \tag{A.179}$$

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

$$\max_{z \in [-r_0, r_0]} \sigma(-z)\sigma(z)^{1-1/\alpha} = \sigma(r_0)\sigma(-r_0)^{1-1/\alpha}, \tag{A.180}$$

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

A.3.1 Proof of Theorem 5

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

$$\left|R_\alpha(\theta) - \hat{R}_\alpha(\theta)\right| \leq C_{r\sqrt{d}}(\alpha)\frac{r\sqrt{d}}{\sqrt{n}} + D_{r\sqrt{d}}(\alpha)\sqrt{\frac{\log\left(\frac{4}{\delta}\right)}{n}}, \tag{A.181}$$

where $C_{r\sqrt{d}}(\alpha)$ is given in (2.53) and (2.54) 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

$$\mathcal{R}(l^\alpha \circ \mathcal{G} \circ S_n) = \mathbb{E}\left(\sup_{g_\theta \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i l^\alpha(y_i, g_\theta(x_i))\right) = \mathbb{E}\left(\sup_{\theta \in \mathbb{B}_d(r)} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \tilde{l}^\alpha(y_i\langle\theta, x_i\rangle)\right). \tag{A.182}$$

The right-hand-side of (A.182) can be rewritten as

$$\mathbb{E}\left(\sup_{\theta \in \mathbb{B}_d(r)} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \tilde{l}^\alpha(y_i\langle\theta, x_i\rangle)\right) = \mathcal{R}(\{\tilde{l}^\alpha(y_1\langle\theta, x_1\rangle), \ldots, \tilde{l}^\alpha(y_n\langle\theta, x_n\rangle) : \theta \in \mathbb{B}_d(r)\}). \tag{A.183}$$

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 11, 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

$$\mathbb{E}\left(\sup_{\theta \in \mathbb{B}_d(r)} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \tilde{l}^\alpha(y_i\langle\theta, x_i\rangle)\right) = \mathcal{R}\left(\{\tilde{l}^\alpha(y_1\langle\theta, x_1\rangle), \ldots, \tilde{l}^\alpha(y_n\langle\theta, x_n\rangle) : \theta \in \mathbb{B}_d(r)\}\right) \tag{A.184}$$

$$\leq C_{r\sqrt{d}}(\alpha)\, \mathcal{R}\left(\{(y_1\langle\theta, x_1\rangle, \ldots, y_n\langle\theta, x_n\rangle) : \theta \in \mathbb{B}_d(r)\}\right). \tag{A.185}$$

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

$$C_{r\sqrt{d}}(\alpha)\, \mathcal{R}(\{(y_1\langle\theta, x_1\rangle, \ldots, y_n\langle\theta, x_n\rangle) : \theta \in \mathbb{B}_d(r)\}) \leq C_{r\sqrt{d}}(\alpha)\, \frac{r\sqrt{d}}{\sqrt{n}}, \tag{A.186}$$

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 $\alpha \in (0, \infty]$

$$\max_{z \in \left[-r\sqrt{d}, r\sqrt{d}\right]} \tilde{l}^\alpha(z) \leq D_{r\sqrt{d}}(\alpha), \tag{A.187}$$

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. □

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 (A.188) 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 $\theta \in \mathbb{B}_d(r)$,*

$$\left| R_\infty(\theta) - \hat{R}_\alpha(\theta) \right| \leq \sigma\left(r\sqrt{d}\right)\left(\frac{2r\sqrt{d}}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}}\right) + \frac{\left(\log\sigma(-r\sqrt{d})\right)^2}{2\alpha}. \tag{A.188}$$

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

$$R_\infty(\theta) - \hat{R}_\alpha(\theta) \leq R_\infty(\theta) - \hat{R}_\infty(\theta) \leq \sigma\left(r\sqrt{d}\right)\left(\frac{2r\sqrt{d}}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}}\right), \tag{A.189}$$

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

$$\hat{R}_\alpha(\theta) - R_\infty(\theta) = \hat{R}_\infty(\theta) - R_\infty(\theta) + \hat{R}_\alpha(\theta) - \hat{R}_\infty(\theta) \tag{A.190}$$

$$\leq \sigma\left(r\sqrt{d}\right)\left(\frac{2r\sqrt{d}}{\sqrt{n}} + 4\sqrt{\frac{2\log\left(\frac{4}{\delta}\right)}{n}}\right) + \frac{\left(\log\sigma(-r\sqrt{d})\right)^2}{2\alpha}, \tag{A.191}$$

where we apply Theorem 5 for the first term and Lemma 2 for the second term[1] on the maximum value of θ, i.e, $\|\theta\|_2 = r$. Thus, combining the two cases we have the desired statement for the corollary. □

A.3.2 Proof of Theorem 6

Assume that the minimum α-risk is attained by the logistic model, i.e., (2.56) 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,

$$\lim_{n\to\infty} R(f_{\hat{\theta}_n^\alpha}) = R^*, \tag{A.192}$$

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 $R^* := \min\limits_{f:\mathcal{X}\to\mathbb{R}} \mathbb{P}[Y \neq \text{sign}(f(X))]$.

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.

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

Proposition 11 (Thm. 3, Bartlett *et al.* (2006b))**.** *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 $\mathcal{X} \times \mathcal{Y}$,*

$$\lim_{i\to\infty} R_\phi(f_i) = R_\phi^* \text{ implies that } \lim_{i\to\infty} R(f_i) = R^*, \tag{A.193}$$

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

$$\min_{\theta\in\mathbb{B}_d(r)} R_\alpha(\theta) = \min_{f:\mathcal{X}\to\mathbb{R}} R_\alpha(f), \tag{A.194}$$

where $R_\alpha(\theta)$ is given in (2.26) and $R_\alpha(f) = \mathbb{E}[\tilde{l}^\alpha(Yf(X))]$ for all measurable f. Thus, the proof strategy is to show that

$$\lim_{n\to\infty} R_\alpha(\hat{\theta}_n^\alpha) = \min_{\theta\in\mathbb{B}_d(r)} R_\alpha(\theta), \tag{A.195}$$

and then apply Proposition 11 to obtain the result.

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

$$R_\alpha(\theta_*^\alpha) = \min_{\theta\in\mathbb{B}_d(r)} R_\alpha(\theta). \tag{A.196}$$

Observe that

$$0 \le R_\alpha(\hat{\theta}_n^\alpha) - R_\alpha(\theta_*^\alpha) = \mathrm{I}_n + \mathrm{II}_n, \tag{A.197}$$

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, (2.55) implies that, for every $\epsilon > 0$,

$$\mathbb{P}\left(|R_\alpha(\hat{\theta}_n^\alpha) - \hat{R}_\alpha(\hat{\theta}_n^\alpha)| > \epsilon\right) \le 4e^{-n\left(\frac{\epsilon - C_{r\sqrt{d}}(\alpha)2r\sqrt{d}/n}{4\sqrt{2}D_{r\sqrt{d}}(\alpha)}\right)^2}, \tag{A.198}$$

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

$$\lim_{n\to\infty} \mathrm{I}_n = \lim_{n\to\infty} R_\alpha(\hat{\theta}_n^\alpha) - \hat{R}_\alpha(\hat{\theta}_n^\alpha) = 0. \tag{A.199}$$

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

$$\mathrm{II}_n = \hat{R}_\alpha(\hat{\theta}_n^\alpha) - R_\alpha(\theta_*^\alpha) \le \hat{R}_\alpha(\theta_*^\alpha) - R_\alpha(\theta_*^\alpha). \tag{A.200}$$

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

$$\mathbb{P}\left(|\hat{R}_\alpha(\theta_*^\alpha) - R_\alpha(\theta_*^\alpha)| > \epsilon\right) \le 4e^{-n\left(\frac{\epsilon - C_{r\sqrt{d}}(\alpha)2r\sqrt{d}/n}{4\sqrt{2}D_{r\sqrt{d}}(\alpha)}\right)^2}, \tag{A.201}$$

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

$$\lim_{n\to\infty} |\hat{R}_\alpha(\theta_*^\alpha) - R_\alpha(\theta_*^\alpha)| = 0. \tag{A.202}$$

In particular, we have that, almost surely,

$$\limsup_{n\to\infty} \mathrm{II}_n \leq 0. \tag{A.203}$$

Plugging (A.199) and (A.203) in (A.197), we obtain, almost surely,

$$0 \leq \limsup_{n\to\infty} \left[R_\alpha(\hat{\theta}_n^\alpha) - R_\alpha(\theta_*^\alpha)\right] \leq 0, \tag{A.204}$$

from which (A.195) follows.

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

$$f_{\hat{\theta}_n^\alpha}(x) = \sigma^{-1}(\sigma(\hat{\theta}_n^\alpha \cdot x)) = \sigma^{-1}(g_{\hat{\theta}_n^\alpha}(x)), \tag{A.205}$$

Proposition 2, (A.194), and (A.195) imply that

$$\lim_{n\to\infty} R_\alpha(f_{\hat{\theta}_n^\alpha}) = \min_{\theta\in\mathbb{B}_d(r)} R_\alpha(f_\theta) = \min_{f:\mathcal{X}\to\mathbb{R}} R_\alpha(f) =: R_\alpha^*. \tag{A.206}$$

Since $\tilde{l}^\alpha$ is classification-calibrated as established in Theorem 1, Proposition 11 and (A.206) imply that

$$\lim_{n\to\infty} R(f_{\hat{\theta}_n^\alpha}) = \min_{f:\mathcal{X}\to\mathbb{R}} \mathbb{P}[Y \neq \mathrm{sign}(f(X))] =: R^*, \tag{A.207}$$

as required.

A.4 Further Experimental Results and Details

A.4.1 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 Sasaki (2007). 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

$$\mathrm{F}_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = \frac{\mathrm{TP}}{\mathrm{TP} + 0.5(\mathrm{FP} + \mathrm{FN})}, \tag{A.208}$$

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.

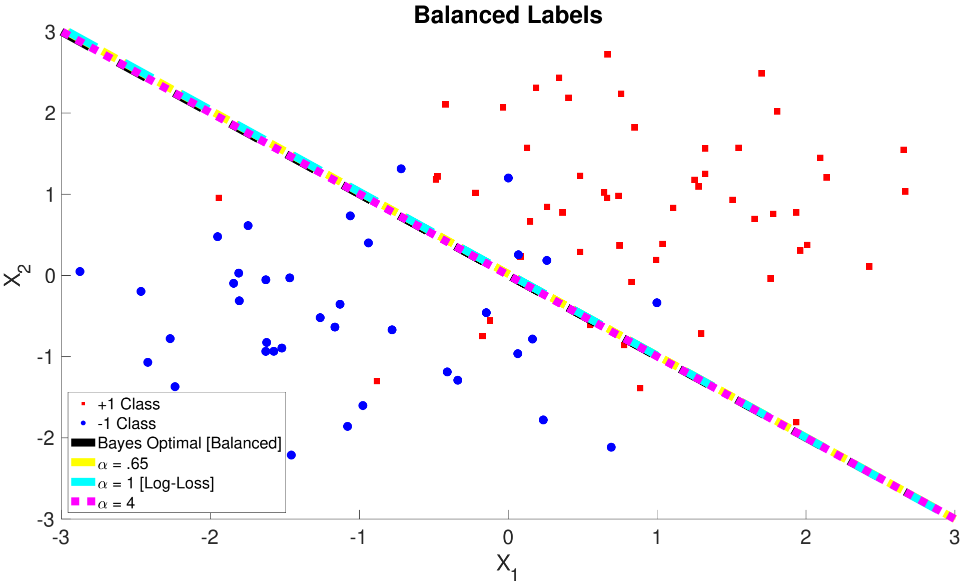


Figure A.5: A Synthetic Experiment Highlighting the Collapse in Trained Linear Predictors of α-loss for $\alpha \in \{.65, 1, 4\}$ on Clean, Balanced Data. Specifically, α-loss Is Trained Until Convergence Under the Logistic Model for a 2D-GMM With Mixing Probability $\mathbb{P}[Y = -1] = \mathbb{P}[Y = +1]$, 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.

A.4.2 Experiments for Section 2.3.3

In this section, we provide additional synthetic experiments, which follow the same experiment protocol as Fig. 2.3. 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.

A.4.3 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 2.6.

For $\alpha \to \infty$, we show through our theoretical landscape analysis (see Section 2.4.3, Theorem 4, and for a visual, Fig. 2.4) 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. 2.1(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 (2.41)).

For $\alpha \to 0$, we see the opposite effect, i.e., that the gradients explode as α decreases from 1 (see Proposition 5 with following commentary and Fig. 2.7 for a visual). Indeed, this motivated the choice of the lower limit of $\alpha = 0.65$ in Fig. 2.3(a). This issue was "pseudo-circumvented" in Tables A.1, A.2, and A.3 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. A.4 which was obtained using Li *et al.* (2018a) 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 hints at the generality of our theoretical results of the α-loss landscape in Section 2.4.3.

					←	α's	→				
		.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
↑	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
↓	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 A.1: Further Quantitative Results Associated with Fig. 2.3(a) in Section 2.3.3 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 (2.41).

					←	α's	→				
		.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
↑	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
↓	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

Table A.2: A Twin Table of Table A.1, Except with F_1 Scores Reported. For a Brief Review of the F_1 Score, See Appendix A.4.1. Gains of $\alpha^* < 1$ over Log-loss ($\alpha = 1$) Are More Exaggerated by the F_1 Score, in Particular See 2% and 5% Imbalance.

					←	α's	→				
		.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
Noise %	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

Table A.3: Further Quantitative Results Associated with Fig. 2.3(B) in Section 2.3.3 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 A.1, 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 2.4.3.

A.4.4 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 A.4 and A.5. Note that we use the same fixed learning rates as the binary symmetric noisy label experiments in Section 2.6.1. 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.

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 A.4: Multiclass Symmetric Noisy Label Experiment on Mnist. See Table 2.1 for Descriptions of Acronyms.

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

Table A.5: Multiclass Symmetric Noisy Label Experiment on the FMNIST Dataset.

APPENDIX B

APPENDIX TO CHAPTER 3

B.1 Proofs, Further Theoretical Results, and Additional Commentary

B.1.1 Illustration of Proper Loss to Surrogate through the Convex Conjugate

In this subsection, we provide a worked-out example for how picking the log-loss as ℓ gives the binary entropy for $\underline{L}$ and the logistic loss for F.

From Reid and Williamson (2010b), we have that the log-loss has partial losses $\ell_1(u) = -\log u$, $\ell_{-1}(u) = -\log(1-u)$ and is a proper loss. In order to compute the (pointwise) *Bayes* risk $\underline{L}$ for the log-loss, we first obtain from (3.2),

$$L(u,v) = v \cdot \ell_1(u) + (1-v) \cdot \ell_{-1}(u) = v \cdot -\log u + (1-v) \cdot -\log(1-u). \tag{B.1}$$

Recall that $\underline{L}(v) \doteq \inf_u L(u,v)$. In (B.1), taking the derivative with respect to u and setting the expression equal to zero, i.e., $\frac{d}{du}L(u,v) = 0$, and solving for u, obtains that $u = v$, in other words, the log-loss is indeed proper. Plugging $u = v$ back into (B.1), we find that the pointwise Bayes risk of the log-loss is

$$\underline{L}(v) = -v\log v - (1-v) \cdot \log(1-v), \tag{B.2}$$

which is indeed the binary (Shannon) entropy (Thomas and Joy, 2006). Finally, to obtain the logistic loss as the surrogate, we compute the convex conjugate of (B.2). Formally, we have from (3.3) that $\forall z \in \mathbb{R}$,

$$F(z) = (-\underline{L})^\star(-z) = (v\log v + (1-v) \cdot \log(1-v))^\star(-z). \tag{B.3}$$

Indeed, we have that $\forall z \in \mathbb{R}$,

$$(v\log v + (1-v) \cdot \log(1-v))^\star = \sup_v\{z \cdot v - v\log v - (1-v) \cdot \log(1-v)\}, \tag{B.4}$$

which is similarly obtained by setting the derivative equal to zero and solving, i.e.,

$$\frac{d}{dv}\left[z \cdot v - v\log v - (1-v) \cdot \log(1-v)\right] = 0 \tag{B.5}$$

$$z + \log(1-v) - \log v = 0 \tag{B.6}$$

$$z = \log\left(\frac{v}{1-v}\right) \tag{B.7}$$

$$v = \frac{1}{1+e^{-z}}, \tag{B.8}$$

which is obtained after a few steps of algebra. Plugging (B.8) back into (B.4), we obtain that

$$\sup_v\{z \cdot v - v\log v - (1-v) \cdot \log(1-v)\} = \frac{z}{1+e^{-z}} + \frac{\log(1+e^{-z})}{1+e^{-z}} + \frac{\log(1+e^{z})}{1+e^{z}}. \tag{B.9}$$

Noticing that $\log(1+e^{-z}) = \log((e^{-z})\cdot(1+e^{z}))$, we have from (B.9) that

$$\sup_{v}\{z\cdot v - v\log v - (1-v)\cdot\log(1-v)\} = \log(1+e^{z}). \tag{B.10}$$

Plugging this back into (B.3), we have that for the log-loss, the surrogate is given by $\forall z \in \mathbb{R}$,

$$F(z) = (-\underline{L})^{\star}(-z) = \log(1+e^{-z}), \tag{B.11}$$

which is indeed the logistic loss (cf. Bartlett *et al.* (2006a)), useful in the margin setting.

B.1.2 Proof of Lemma 5

We study $U \doteq (-\underline{L})^{\star}$, which is convex by definition, and show that it is non-decreasing. Monotonicity follows from the non-negativity of the argument of the partial losses and the definition of the convex conjugate: suppose $z' \geq z$ and let $u^{*} \in \arg\sup_{u} zu + \underline{L}(u)$. We have

$$\begin{aligned} U(z') &\doteq \sup_{u\in[0,1]} z'u + \underline{L}(u) && \text{(B.12)}\\ &= \sup_{u\in[0,1]} (z'-z)u + zu + \underline{L}(u) && \text{(B.13)}\\ &\geq (z'-z)u^{*} + zu^{*} + \underline{L}(u^{*}) && \text{(B.14)}\\ &= (z'-z)u^{*} + U(z) && \text{(B.15)}\\ &\geq U(z), && \text{(B.16)} \end{aligned}$$

which completes the proof that U is non-decreasing and therefore $F(z) \doteq U(-z)$ non-increasing.

Concavity of $\underline{L}$ follows from definition. We show continuity of $\underline{L}$, the continuity of F then following from the definition of the convex conjugate F (Boyd and Vandenberghe, 2004b). Let $a, u \in (0,1)$, let $u^{*} \in t_{\ell}(u), a^{*} \in t_{\ell}(a)$. We get:

$$\begin{aligned} \underline{L}(u) &\doteq u\ell_{1}(u^{*}) + (1-u)\ell_{-1}(u^{*}) && \text{(B.17)}\\ &\leq u\ell_{1}(a^{*}) + (1-u)\ell_{-1}(a^{*}) && \text{(B.18)}\\ &= \underline{L}(a) + (u-a)(\ell_{1}(a^{*}) - \ell_{-1}(a^{*})), && \text{(B.19)} \end{aligned}$$

(the inequality holds since otherwise $u^{*} \not\in t_{\ell}(u)$) Permuting the roles of u and a, we also get

$$\underline{L}(a) \;\leq\; \underline{L}(u) + (a-u)(\ell_{1}(u^{*}) - \ell_{-1}(u^{*})), \tag{B.20}$$

from which we get

$$|\underline{L}(a) - \underline{L}(u)| \;\leq\; Z\cdot|a-u|, \tag{B.21}$$

with $Z \doteq \max_{v\in\{a,u\}} \sup|\ell_{1}(t_{\ell}(v)) - \ell_{-1}(t_{\ell}(v))|$ (where we use set differences if t_{ℓ}s are not singletons). Since $Z \ll \infty$, (B.21) is enough to show the continuity of $\underline{L}$ (we have by assumption $\mathrm{dom}(\underline{L}) = [0,1]$).

B.1.3 Bayes Tilted Estimates

The proof of Lemma 6 readily follows from Definition 3.4 and standard properties of convex functions, see e.g., Boyd and Vandenberghe (2004b).

Below, we also provide analysis of the properties of Bayes tilted estimates for more general losses which induce set-valued functions. Following convention, we denote the set valued inequality $A \leq B$, such that, $\forall a \in A, \exists b \in B, a \leq b$ and the set-valued (Minkowski) difference $A - B \doteq \{a - b : a \in A, b \in B\}$.

Lemma 12. *The following properties of t_ℓ follow from assumptions M, D or S on partial losses:*
(M) *implies set-valued monotonicity: $\forall u_1 < u_3 \in [0,1]$, we have $t_\ell(u_1) \leq t_\ell(u_3)$ and $t_\ell(u_1) \cap t_\ell(u_3) \subseteq t_\ell(u_2), \forall u_2 \in (u_1, u_3)$;*
(D) *and t_ℓ differentiable imply monotonicity: $\forall u \in [0,1]$, $\ell'_1(t_\ell(u)) \leq \ell'_{-1}(t_\ell(u)) \Leftrightarrow t'_\ell(u) \geq 0$;*
(S) *implies set-valued symmetry: $t_\ell(1-u) = \{1\} - t_\ell(u), \forall u \in [0,1]$;*
(E) *Extreme values: $\ell_1(1) = \ell_{-1}(0) = 0$, $\ell_1([0,1]) \subseteq \mathbb{R}_+, \ell_{-1}([0,1]) \subseteq \mathbb{R}_+$. Further, this implies properness on extreme values, as $0 \in t_\ell(0), 1 \in t_\ell(1)$.*

Case (M) – Suppose $t_\ell(a) \cap t_\ell(a') \neq \emptyset$ for some $a \neq a'$ and let $v^* \in t_\ell(a) \cap t_\ell(a')$. It means $\forall v \in [0,1]$,

$$\begin{aligned} a\ell_1(v^*) + (1-a)\ell_{-1}(v^*) &\leq a\ell_1(v) + (1-a)\ell_{-1}(v), && \text{(B.22)} \\ a'\ell_1(v^*) + (1-a')\ell_{-1}(v^*) &\leq a'\ell_1(v) + (1-a')\ell_{-1}(v), && \text{(B.23)} \end{aligned}$$

and so $\forall \delta \in [0,1]$, if we let $a_\delta \doteq a + \delta(a' - a)$, a $1-\delta, \delta$ convex combination of both inequalities yields $\forall v \in [0,1]$,

$$a_\delta\ell_1(v^*) + (1-a_\delta)\ell_{-1}(v^*) \leq a_\delta\ell_1(v) + (1-a_\delta)\ell_{-1}(v), \forall v \in [0,1], \quad \text{(B.24)}$$

which implies $v^* \in t_\ell(a_\delta)$ and shows the right part of **Case (M)**.
To show show the left part of **Case (M)**; we add to (B.22) and (B.23) we now add the inequality:

$$a\ell_1(v^\circ) + (1-a)\ell_{-1}(v^\circ) \leq a\ell_1(v) + (1-a)\ell_{-1}(v), \quad \text{(B.25)}$$

with therefore $v^\circ \in t_\ell(a)$, implying $a\ell_1(v^\circ) + (1-a)\ell_{-1}(v^\circ) = a\ell_1(v^*) + (1-a)\ell_{-1}(v^*)$ as otherwise one of v°, v^* would not be in $t_\ell(a)$. We then get

$$\begin{aligned} a'\ell_1(v^\circ) + (1-a')\ell_{-1}(v^\circ) &= a\ell_1(v^\circ) + (1-a)\ell_{-1}(v^\circ) + (a'-a) \cdot (\ell_1(v^\circ) - \ell_{-1}(v^\circ)) \\ &= a\ell_1(v^*) + (1-a)\ell_{-1}(v^*) + (a'-a) \cdot (\ell_1(v^\circ) - \ell_{-1}(v^\circ)) \\ &= a'\ell_1(v^*) + (1-a')\ell_{-1}(v^*) + (a'-a) \cdot \Delta, && \text{(B.26)} \end{aligned}$$

with $\Delta \doteq \ell_1(v^\circ) - \ell_{-1}(v^\circ) - (\ell_1(v^*) - \ell_{-1}(v^*))$. Considering (B.26), we deduce from (B.23) that to have $v^\circ \in t_\ell(a')$, we equivalently need $(a'-a) \cdot \Delta \leq 0$. We also know by assumption that ℓ_1 is non-increasing and ℓ_{-1} is non-decreasing, so $g(u) \doteq \ell_1(u) - \ell_{-1}(u)$ is non-increasing. We thus have $(a'-a) \cdot \Delta \leq 0$ iff one of the two possibilities hold:

- $a' \geq a$ and $v^\circ \geq v^*$, or
- $a' \leq a$ and $v^\circ \leq v^*$,

which shows the right part of **Case (M)**.

Case (E) – we have $\underline{L}(0) = \inf_{v\in[0,1]} \ell_{-1}(v) = 0$ for $v = 0$, hence $0 \in t_\ell(0)$. Similarly, $\underline{L}(1) = \inf_{v\in[0,1]} \ell_1(v) = 0$ for $v = 1$, hence $1 \in t_\ell(1)$.

Case (D) – we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u}\underline{L}(u) &= \ell_1(t_\ell(u)) + u\ell_1'(t_\ell(u))t_\ell'(u) - \ell_{-1}(t_\ell(u)) + (1-u)\ell_{-1}'(t_\ell(u))t_\ell'(u)\\ &= \ell_1(t_\ell(u)) - \ell_{-1}(t_\ell(u)) + t_\ell'(u)\cdot(u\ell_1'(t_\ell(u)) + (1-u)\ell_{-1}'(t_\ell(u))), \quad \text{(B.27)}\end{aligned}$$

but since $v = t_\ell(u)$ is the solution to (B.22) it satisfies $u\ell_1'(t_\ell(u))+(1-u)\ell_{-1}'(t_\ell(u)) = 0$, so that (B.27) simplifies to

$$\frac{\mathrm{d}}{\mathrm{d}u}\underline{L}(u) \quad = \quad \ell_1(t_\ell(u)) - \ell_{-1}(t_\ell(u)), \quad \text{(B.28)}$$

and since $\underline{L}$ is concave and the partial losses are differentiable,

$$\frac{\mathrm{d}^2}{\mathrm{d}u^2}\underline{L}(u) \quad = \quad t_\ell'(u)\cdot(\ell_1'(t_\ell(u)) - \ell_{-1}'(t_\ell(u))) \leq 0, \forall u, \quad \text{(B.29)}$$

which proves the statement of the Lemma.

Case (S) – Suppose $v^* \in t_\ell(a)$, which implies

$$a\ell_1(v^*) + (1-a)\ell_{-1}(v^*) \quad \leq \quad a\ell_1(v) + (1-a)\ell_{-1}(v), \forall v \in [0,1]. \quad \text{(B.30)}$$

We also note that since symmetry holds, $a\ell_1(v^*)+(1-a)\ell_{-1}(v^*) = (1-a)\ell_1(1-v^*) + a\ell_{-1}(1-v^*)$, which implies because of (B.30) $1 - v^* \in t_\ell(1-a)$.

Remark: even if we assume the partial losses to be strictly monotonic, the tilted estimate can still be set valued. To see this, craft the partial losses such that $v \in t_\ell(u)$ and then for some $w > v$, replace the partial losses in the interval $[v, w]$ by affine parts w/ slope $-a < 0$ for ℓ_1, $b > 0$ for ℓ_{-1} and such that $b/a = u/(1-u)$ which guarantees $L(u, v) = L(u, w)$ and thus $w \in t_\ell(u)$;

B.1.4 Proof of Lemma 8

We recall the focal loss' corresponding pointwise conditional risk in lieu of (3.2):

$$L_\gamma(u, v) \quad \doteq \quad -v\cdot(1-u)^\gamma \log u - (1-v)\cdot u^\gamma \log(1-u), \quad \text{(B.31)}$$

and if it is twist proper, then for any $\eta_{\mathrm{t}}, \eta_{\mathrm{c}} \in [0,1]$, there exists $\gamma \geq 0$ such that

$$\left.\frac{\partial}{\partial u}L_\gamma(u, \eta_{\mathrm{t}})\right|_{u=\eta_{\mathrm{c}}} \quad = \quad 0. \quad \text{(B.32)}$$

Equivalently, we must find $\gamma \geq 0$ such that (keeping notations $u \doteq \eta_{\mathrm{c}}, v \doteq \eta_{\mathrm{t}}$ for clarity):

$$(1-v)u^{\gamma} \cdot (\gamma(1-u)\log(1-u) - u) \quad = \quad v(1-u)^{\gamma} \cdot (\gamma u \log u + u - 1)\,, \quad \text{(B.33)}$$

and we see that twist properness implies the statement that for any $K \geq 0$ (note that $K = \frac{v}{1-v}$) and any $u \in [0,1)$, there exists $\gamma \geq 0$ such that

$$\frac{f(u,\gamma)}{f(1-u,\gamma)} \quad = \quad K, \quad \text{(B.34)}$$

$$f(u,\gamma) \quad \doteq \quad u^{\gamma} \cdot (\gamma(1-u)\log(1-u) - u)\,. \quad \text{(B.35)}$$

We study the ratio for $u \in [0,1/2]$. We have $f(u,\gamma) \leq 0, \forall u \in [0,1], \forall \gamma \geq 0$ and

$$\frac{\partial}{\partial \gamma} f(u,\gamma) \quad = \quad u^{\gamma}(\gamma \cdot a(u) - b(u)), \quad \text{(B.36)}$$

with $a(u) \doteq (1-u)\log(1-u)\log(u) \geq 0$ and $b(u) \doteq u \log u - (1-u)\log(1-u)$, satisfying $b(1-u) = -b(u)$ and $ua(1-u) = (1-u)a(u)$. Hence, we arrive after some derivations to

$$\frac{\partial}{\partial \gamma}\frac{f(u,\gamma)}{f(1-u,\gamma)} = \frac{(u(1-u))^{\gamma}}{f^2(1-u,\gamma)} \cdot \left(A(u)\gamma^2 + B(u)\gamma + C(u)\right), \quad \text{(B.37)}$$

$$A(u) \doteq u(1-u)\log(u)\log(1-u)\log(u(1-u)), \quad \text{(B.38)}$$

$$B(u) \doteq -(u^2 \log^2 u + (1-u)^2 \log^2(1-u) + (1-2u)^2 \log u \log(1-u)), \quad \text{(B.39)}$$

$$C(u) \doteq (1-2u)b(u). \quad \text{(B.40)}$$

All functions A, B, C are non positive for any fixed $u \in [0,1/2]$, so the ratio in (B.34) is non-increasing over $\gamma \geq 0$ and as a consequence, for any fixed $u \in [0,1/2]$,

$$\frac{f(u,\gamma)}{f(1-u,\gamma)} \quad \leq \quad \frac{f(u,0)}{f(1-u,0)} \quad \text{(B.41)}$$

$$= \frac{u}{1-u}, \forall \gamma \geq 0, \quad \text{(B.42)}$$

so we see that (B.34) cannot be satisfied when $K > u/(1-u)$ and as a consequence, the focal loss is not twist-proper.

Twist-improperness of the Super Loss The Super Loss (Castells *et al.*, 2020) works as a "wrapper" of a loss, its partial losses being defined as

$$L_{b,\lambda}(\ell, \sigma_i) = (\ell_b - \tau)\sigma_i + \lambda(\log \sigma_i)^2, \quad \text{(B.43)}$$

where $b \in \{-1, 1\}$ indicates the partial loss of a loss of interest, $\tau \in \mathrm{Im}\ell_b, \lambda > 0$ are user-defined parameters. σ_i is a functional computed to minimize the partial losses, and we get the optimal expression:

$$\sigma^*(\ell_b) = \exp\left(-W(1/2 \max\left(-2/e, \beta\right))\right), \quad \text{(B.44)}$$

with $\beta = \frac{\ell_b - \tau}{\lambda}$ (notice this is also a function via the partial loss). W is called Lambert's function. It does not have an analytical form.

Lemma 13. *Suppose loss ℓ in the Super Loss is such that its partial loss ℓ_1 is strictly decreasing and ℓ_{-1} is strictly increasing. Then the corresponding Super Loss with partial losses $L_{b,\lambda}(\ell, \sigma^*(\ell_b))$ ($b \in \{-1, 1\}$) is not twist proper.*

Remark: the assumptions about the partial losses are very weak and would be satisfied by all popular losses (e.g. log, square, Matusita, etc.).

Proof. The notable facts about W, useful for our proof are:

$$e^{W(z)} = \frac{z}{W(z)} \quad \text{and} \quad \frac{d}{dz}W(z) = \frac{1}{z + e^{W(z)}} \quad \text{and} \quad \sup \exp(-W(z)) = e. \tag{B.45}$$

Simplifying notations above, we end up studing a loss with partial losses defined as

$$L^*_\lambda(\ell_b) = (\ell_b - \tau)\sigma_i + \lambda(\log \sigma^*(\ell_b))^2. \tag{B.46}$$

Recall that a loss ℓ is twist-proper iff for any twist, there exists hyperparameters such that $\eta_{\text{c}} \in t_\ell(\eta_{\text{t}})$. Examining this for the Super Loss, we obtain

$$\begin{aligned} t_L(v) &= \text{arginf}_{u\in[0,1]} L(u, v) && \text{(B.47)}\\ &= \text{arginf}_{u\in[0,1]} v \cdot L^*_\lambda(\ell_1(u)) + (1 - v) \cdot L^*_\lambda(\ell_{-1}(u)) && \text{(B.48)}\\ &= \text{arginf}_{u\in[0,1]} \left\{ \begin{array}{l} v \cdot [(\ell_1(u) - \tau)\sigma^*(\ell_1) + \lambda(\log \sigma^*(\ell_1))^2] \\ +(1 - v) \cdot [(\ell_{-1}(u) - \tau)\sigma^*(\ell_{-1}) + \lambda(\log \sigma^*(\ell_{-1}))^2] \end{array} \right. && \text{(B.49)} \end{aligned}$$

We note that if ℓ is proper, then $v \in t_\ell(v)$. Computing the minimum in (B.49), we obtain

$$\begin{aligned} 0 = \frac{d}{du} v \cdot &\left[(\ell_1(u) - \tau)\sigma^*(\ell_1) + \lambda(\log \sigma^*(\ell_1))^2\right] \\ &+ (1 - v) \cdot \left[(\ell_{-1}(u) - \tau)\sigma^*(\ell_{-1}) + \lambda(\log \sigma^*(\ell_{-1}))^2\right]. \end{aligned} \tag{B.50}$$

Similar to the computation of the focal loss, we need

$$\frac{\frac{d}{du}(\ell_1(u) - \tau)\sigma^*(\ell_1) + \lambda(\log \sigma^*(\ell_1))^2}{\frac{d}{du}(\ell_{-1}(u) - \tau)\sigma^*(\ell_{-1}) + \lambda(\log \sigma^*(\ell_{-1}))^2} = -\frac{(1 - v)}{v} = -K, \tag{B.51}$$

and this needs to hold (via the choice of parameters τ, λ) for any $u \in [0, 1)$ and $K > 0$. To save notations, define

$$\beta_b(u) \quad = \quad \frac{\ell_b(u) - \tau}{2\lambda}.$$

Remark that *if* $\ell_b(u) > \tau - (2\lambda)/e$, we have $\sigma^*(\ell_b(u)) = \exp(-W(\beta_b(u)))$ and so

$$\begin{aligned}
&\frac{d}{du}(\ell_b(u) - \tau)\sigma^*(\ell_b(u)) + \lambda(\log \sigma^*(\ell_b(u)))^2 && \text{(B.52)}\\
&= \quad 2\lambda \cdot \frac{d}{du}\left[\beta_b(u)\exp(-W(\beta_b(u))) + \frac{W^2(\beta_b(u))}{2}\right] && \text{(B.53)}\\
&= \quad 2\lambda\cdot\left[\beta_b'(u)\exp(-W(\beta_b(u)))-\beta_b(u)\beta_b'(u)\cdot\frac{\exp(-W(\beta_b(u)))}{\beta_b(u)+\exp(W(\beta_b(u)))}+\frac{\beta_b'(u)W(\beta_b(u))}{\beta_b(u)+\exp(W(\beta_b(u)))}\right] && \text{(B.54)}\\
&= \quad 2\lambda\beta_b'(u)\cdot\frac{1+\beta_b(u)\exp(-W(\beta_b(u)))-\beta_b(u)\exp(-W(\beta_b(u)))+W(\beta_b(u))}{\beta_b(u)+\exp(W(\beta_b(u)))} && \text{(B.55)}\\
&= \quad \ell_b'(u)\cdot\frac{1+W(\beta_b(u))}{\beta_b(u)+\exp(W(\beta_b(u)))} && \text{(B.56)}\\
&= \quad \ell_b'(u)\cdot\exp(-W(\beta_b(u))), && \text{(B.57)}
\end{aligned}$$

since indeed it comes from (B.45),

$$\frac{1+W(z))}{z+\exp(W(z))} \quad = \quad \exp(-W(z)); \tag{B.58}$$

also, *if* $\ell_b(u) \leq \tau - (2\lambda)/e$, we have $\sigma^*(\ell_b(u)) = \exp(-W(-1/e)) = e$ and so

$$\frac{d}{du}(\ell_b(u) - \tau)\sigma^*(\ell_b(u)) + \lambda(\log \sigma^*(\ell_b(u)))^2 \quad = \quad \ell_b'(u)\cdot e. \tag{B.59}$$

Since $\lim_{z\to -1/e^+}\exp(-W(z)) = e$, we can summarize both (B.57) and (B.59) as

$$\begin{aligned}
&\frac{d}{du}(\ell_b(u) - \tau)\sigma^*(\ell_b(u)) + \lambda(\log \sigma^*(\ell_b(u)))^2\\
&\qquad = \ell_b'(u)\cdot\exp(-W(\max\{-1/e, \beta_b(u)\})).
\end{aligned}\tag{B.60}$$

Now consider a loss ℓ satisfying ℓ_{-1} strictly increasing and ℓ_1 strictly decreasing. Pick u so that we have simultaneously

$$\begin{aligned}
\ell_1(u) \quad &> \quad \tau - \frac{2\lambda}{e}, && \text{(B.61)}\\
\ell_{-1}(u) \quad &\leq \quad \tau - \frac{2\lambda}{e}, && \text{(B.62)}
\end{aligned}$$

which, assuming both inequalities fit in the range of the respective partial loss, that $u \in [0, \gamma]$ for some $\gamma > 0$. Rewriting (B.51), we need to show that for any such $\gamma > 0$ and $u \in [0, \gamma]$ and $K > 0$, there exists a choice of τ, λ such that

$$\frac{\ell_1'(u)\cdot\exp(-W(\beta_1(u)))}{\ell_{-1}'(u)\cdot e} \quad = \quad -K, \tag{B.63}$$

which rewrites conveniently as

$$\exp(-W(\beta_1(u))) \quad = \quad -Ke\cdot\frac{\ell_{-1}'(u)}{\ell_1'(u)}, \tag{B.64}$$

or,

$$\exp(-W(\beta_1(u))) \quad = \quad K', \tag{B.65}$$

for any $K' > 0$ (the RHS of (B.64) is indeed always strictly positive). But indeed we have that $\sup \exp(-W(z)) = e$ (B.45), so (B.65) cannot hold and the Super Loss is not twist proper. □

B.1.5 Proof of Lemma 9

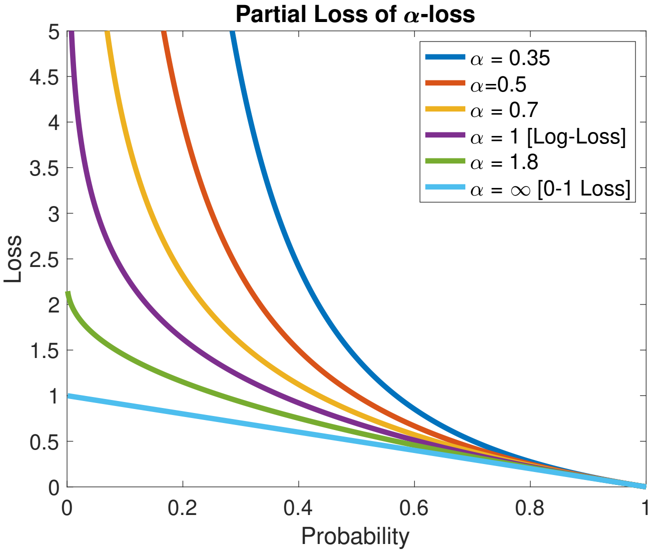


Figure B.1: A Plot of $\ell_1^\alpha(u)$ for $\alpha > 0$ as given in Definition 7.

For part **(a)**: we cite Sypherd *et al.* (2022a) which demonstrates **(M)**, **(D)**, **(S)** for $\alpha > 0$. With our extension of α-loss, these can also be readily shown for $\alpha < 0$, since they are mapped back to the $\alpha > 0$ losses.

For part **(b)**: we know from Lemma 6 that α-loss, for $\alpha \in \mathbb{R} \setminus \{0, \pm\infty\}$, due to strict convexity, returns a singleton, i.e., $|t_{\ell^\alpha}(\eta_{\mathrm{t}})| = 1$. With regards to that singleton, we know from (Sypherd *et al.*, 2022a; Liao *et al.*, 2018b) for $\alpha > 0$ that $t_{\ell^\alpha}(\eta_{\mathrm{t}}) = \frac{\eta_{\mathrm{t}}^\alpha}{\eta_{\mathrm{t}}^\alpha + (1-\eta_{\mathrm{t}})^\alpha}$. A very similar calculation recovers $t_{\ell^\alpha}(\eta_{\mathrm{t}}) = \frac{\eta_{\mathrm{t}}^{-\alpha}}{\eta_{\mathrm{t}}^{-\alpha} + (1-\eta_{\mathrm{t}})^{-\alpha}}$ for $\alpha < 0$. Multiplying the numerator and denominator of this expression by $(1-\eta_{\mathrm{t}})^\alpha$, we can simply write both expressions using $\frac{\eta_{\mathrm{t}}^\alpha}{\eta_{\mathrm{t}}^\alpha + (1-\eta_{\mathrm{t}})^\alpha}$. Regarding the limit as $\alpha \to \pm\infty$ yielding $t_{\ell^{\pm\infty}}(\eta_{\mathrm{t}}) = \pm 1$ or ∓ 1, this was also already shown by Sypherd *et al.* (2022a) for $+\infty$ and is similarly (readily) extended for the $\alpha \to -\infty$ case.

For part **(c)**: here, we break entirely new ground. Let $\alpha > 0$. To obtain twist-properness as stipulated in Definition 6, we seek to know for what α the following holds

$$\eta_{\mathrm{c}} = \frac{\eta_{\mathrm{t}}^\alpha}{\eta_{\mathrm{t}}^\alpha + (1-\eta_{\mathrm{t}})^\alpha}. \tag{B.66}$$

Solving for α, we obtain

$$\eta_{\mathrm{c}} = \frac{\eta_{\mathrm{t}}^{\alpha}}{\eta_{\mathrm{t}}^{\alpha} + (1-\eta_{\mathrm{t}})^{\alpha}} \tag{B.67}$$

$$\frac{1}{\eta_{\mathrm{c}}} = 1 + \left(\frac{1}{\eta_{\mathrm{t}}} - 1\right)^{\alpha} \tag{B.68}$$

$$\frac{1}{\eta_{\mathrm{c}}} - 1 = \left(\frac{1}{\eta_{\mathrm{t}}} - 1\right)^{\alpha} \tag{B.69}$$

$$\log\left(\frac{1}{\eta_{\mathrm{c}}} - 1\right) = \alpha \log\left(\frac{1}{\eta_{\mathrm{t}}} - 1\right) \tag{B.70}$$

$$\alpha^{*} = \frac{\log\left(\frac{1-\eta_{\mathrm{c}}}{\eta_{\mathrm{c}}}\right)}{\log\left(\frac{1-\eta_{\mathrm{t}}}{\eta_{\mathrm{t}}}\right)}. \tag{B.71}$$

After multiplying the numerator and denominator by -1, we obtain the desired result. Namely, α-loss is twist-proper for

$$\alpha^{*} = \frac{\iota(\eta_{\mathrm{c}})}{\iota(\eta_{\mathrm{t}})}. \tag{B.72}$$

Interestingly, (B.72) is the ratio of the logits (which is the link function $-\underline{L}'$ of the log-loss, $\alpha = 1$) evaluated at the clean posterior and twisted posterior, in essence, a kind of ratio test.

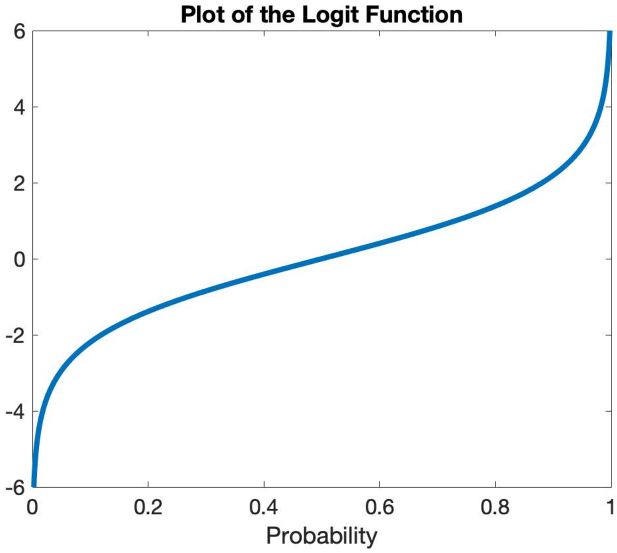


Figure B.2: A Plot of the Logit $\iota(u) \doteq \log(u/(1-u))$.

For part **(d)**: we recall the definition of Bayes blunting twist from Definition 5: a twist $\eta_{\mathrm{c}} \mapsto \eta_{\mathrm{t}}$ is Bayes blunting iff $(\eta_{\mathrm{c}} \leq \eta_{\mathrm{t}} \leq 1/2) \vee (\eta_{\mathrm{c}} \geq \eta_{\mathrm{t}} \geq 1/2)$. Also, recall that α^* is given in (B.72), and see Figure B.2 for a plot of the logit function. Let $\eta_c \geq 1/2$. The Bayes blunting twist can take η_t from $\eta_c \geq \eta_t \geq 1/2$. If $\eta_t = \eta_c$, then $\alpha^* = 1$. If $\eta_t \to 1/2$, as can be seen in the figure, the sign crossover point is 1/2, so $\alpha^* \to \infty$. Thus, by continuity, we have that $\alpha^* \geq 1$. Finally, the case where $\eta_c < 1/2$ follows, *mutatis mutandis.*

B.1.6 Proof of Theorem 7

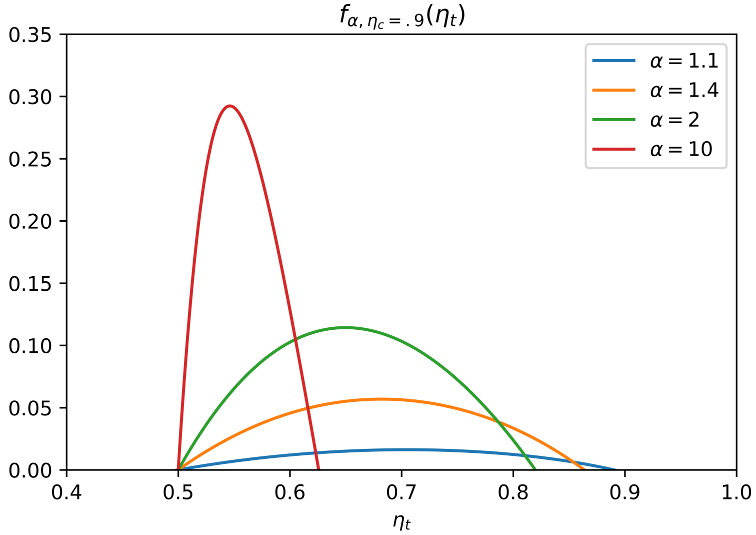


Figure B.3: Characteristic Plot of the Non-negative Part of $f_{\alpha,\eta_c}(\eta_t)$, Where $\eta_c = .9$, as a Function of η_t for Several Values of α. Recall That the Non-negative Region of f Indicates Where Using Bayes Tilted α-estimate, as Measured with the Cross Entropy for α given In (3.7), Is Strictly Less than the $\alpha = 1$ Cross Entropy. Also Recall That a Bayes Blunting Twist Has the Capability to Shift η_t Anywhere in $[.5, \eta_c = .9]$. We See That for Small α, More Twisted Probabilities Are "covered", Whereas for Large α, Less Twisted Probabilities Are "covered", However, the Large α's Induce a Large Positive Magnitude (Ultimately Measured by the Kl-divergence) Increase over the Proper $\alpha = 1$. A Key Takeaway Is That a Fixed α (Small Enough) Can Correct a Bayes Blunting Twist for Almost All $x \in \mathcal{X}$. However, It Is Not Necessarily Optimal as a Perfectly Tuned α-mapping Will Use Larger α's to Optimally Correct Strongly Twisted Posteriors, Inducing More Gains over the $\alpha = 1$ Cross Entropy.

We want to show that for any strictly Bayes blunting twist $\eta_c \mapsto \eta_t$, there exists a fixed $\alpha_0 > 1$ and an optimal $\alpha^\star$-mapping, $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$, which induces the following ordering

$$D_{\mathrm{KL}}(\eta_c, \eta_t; 1) > D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha_0) \geq D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha^\star). \quad \text{(B.73)}$$

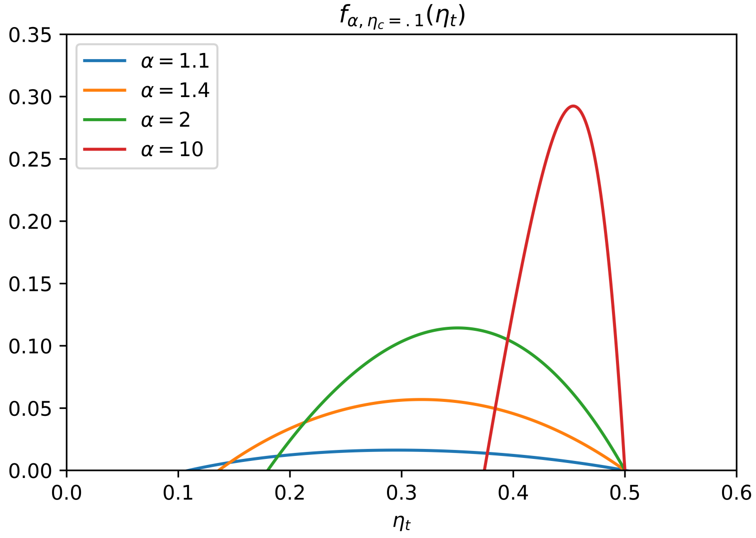


Figure B.4: Symmetric Image of Figure B.3 for $\eta_{\mathrm{c}} = .1$.

Recalling (3.9), which is the identity

$$\begin{aligned} D_{\mathrm{KL}}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha) &= \mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha) - H(\eta_{\mathrm{c}}) && \text{(B.74)} \\ &= \mathbb{E}_{\mathsf{X}\sim\mathrm{M}}\left[\eta_{\mathrm{c}}(X)\log\left(\frac{\eta_{\mathrm{c}}(X)}{t_{\ell^\alpha}(\eta_{\mathrm{t}}(X))}\right) + (1-\eta_{\mathrm{c}}(X))\log\left(\frac{1-\eta_{\mathrm{c}}(X)}{1-t_{\ell^\alpha}(\eta_{\mathrm{t}}(X))}\right)\right], && \text{(B.75)} \end{aligned}$$

by subtracting $H(\eta_{\mathrm{c}})$ from both sides, we rewrite the desired statement (3.10) (also given here in (B.73)) as

$$\mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; 1) > \mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha_0) \geq \mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha^\star). \tag{B.76}$$

In essence, we want to show that

$$\mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha) < \mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; 1) \quad \text{OR} \quad 0 < \mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; 1) - \mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha), \tag{B.77}$$

for some $\alpha_0 > 1$. Continuing with the right-hand-side of (B.77), we have

$$\begin{aligned} &\mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; 1) - \mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha) && \text{(B.78)} \\ &= \mathbb{E}_{\mathsf{X}\sim\mathrm{M}}[\eta_{\mathrm{c}}(X)\cdot -\log\eta_{\mathrm{t}}(X) + (1-\eta_{\mathrm{c}}(X))\cdot -\log(1-\eta_{\mathrm{t}}(X))] \\ &\quad - \mathbb{E}_{\mathsf{X}\sim\mathrm{M}}[\eta_{\mathrm{c}}(X)\cdot -\log t_{\ell^\alpha}(\eta_{\mathrm{t}}(X)) + (1-\eta_{\mathrm{c}}(X))\cdot -\log(1-t_{\ell^\alpha}(\eta_{\mathrm{t}}(X))] && \text{(B.79)} \\ &= \mathbb{E}_{\mathsf{X}\sim\mathrm{M}}\left[\eta_{\mathrm{c}}(X)\log\left(\frac{t_{\ell^\alpha}(\eta_{\mathrm{t}}(X))}{\eta_{\mathrm{t}}(X)}\right) + (1-\eta_{\mathrm{c}}(X))\log\left(\frac{1-t_{\ell^\alpha}(\eta_{\mathrm{t}}(X))}{1-\eta_{\mathrm{t}}(X)}\right)\right] && \text{(B.80)} \\ &= \mathbb{E}_{\mathsf{X}\sim\mathrm{M}}\left[\eta_{\mathrm{c}}(X)\log\left(\tfrac{\eta_{\mathrm{t}}(X)^{\alpha-1}}{\eta_{\mathrm{t}}(X)^\alpha+(1-\eta_{\mathrm{t}}(X))^\alpha}\right)+(1-\eta_{\mathrm{c}}(X))\log\left(\tfrac{(1-\eta_{\mathrm{t}}(X))^{\alpha-1}}{(1-\eta_{\mathrm{t}}(X))^\alpha+\eta_{\mathrm{t}}(X)^\alpha}\right)\right], && \text{(B.81)} \end{aligned}$$

where we used the linearity of the expectation and some algebra to combine the expressions. We want to show that the expression in brackets in (B.81) is strictly positive as this implies that $0 < \textsc{ce}(\eta_\mathrm{c}, \eta_\mathrm{t}; 1) - \textsc{ce}(\eta_\mathrm{c}, \eta_\mathrm{t}; \alpha)$, in words, that the α-Bayes tilted estimate untwists the Bayes blunting twist. Continuing, we examine the expression in brackets in (B.81)

$$\begin{aligned} f_{\alpha,\eta_\mathrm{c}}(\eta_\mathrm{t}) &= \eta_\mathrm{c} \log \frac{\eta_\mathrm{t}^{\alpha-1}}{\eta_\mathrm{t}^{\alpha} + (1-\eta_\mathrm{t})^{\alpha}} + (1-\eta_\mathrm{c}) \log \frac{(1-\eta_\mathrm{t})^{\alpha-1}}{\eta_\mathrm{t}^{\alpha} + (1-\eta_\mathrm{t})^{\alpha}} && \text{(B.82)} \\ &= (\alpha-1)\eta_\mathrm{c} \log \eta_\mathrm{t} + (\alpha-1)(1-\eta_\mathrm{c}) \log (1-\eta_\mathrm{t}) - \log (\eta_\mathrm{t}^{\alpha} + (1-\eta_\mathrm{t})^{\alpha}), && \text{(B.83)} \end{aligned}$$

where we implicitly fix $X = x$ and consider scalar-valued $\eta_\mathrm{c}, \eta_\mathrm{t} \in [0,1]$ and $\alpha \in \mathbb{R}_+/\{1\}$. We note that $\alpha < 0$ does not need to be considered in this analysis since that regime of α is primarily useful for very strong twists due to its symmetry property (recall in Lemma 9 that t_{ℓ^α} is symmetric upon permuting $(\eta_\mathrm{t}, \alpha)$ and $(1-\eta_\mathrm{t}, -\alpha)$), i.e., *not* useful for Bayes blunting twists which reduce confidence in the posterior but do not flip its sign across $\eta_\mathrm{t} - 1/2$. To build intuition of f, see Figure B.3 for a plot of this function. Formally, we take note of the following observations/properties of f:

1. **CONTINUITY.** From (B.83), it can be readily shown that for any fixed $\eta_\mathrm{c}, \eta_\mathrm{t} \in [0,1]$, $f_{\alpha,\eta_\mathrm{c}}(\eta_\mathrm{t})$ is continuous in $\alpha \geq 1$.

2. **CONCAVITY.** For arbitrarily fixed η_c and for any $\alpha > 1$, $f_{\alpha,\eta_\mathrm{c}}(\cdot)$ is concave in η_t, since (from (B.82)) the composition of a concave function with a non-decreasing concave function yields a concave function. As a side note, observe that $f_{\alpha,\eta_\mathrm{c}}(\cdot)$ is convex for $0 < \alpha < 1$, thus this regime of α does not untwist Bayes blunting twists. Regarding (increa/decrea)sing concavity of $f_{\alpha,\eta_\mathrm{c}}(\cdot)$ for any fixed $\eta_\mathrm{c} \in [0,1]$ as a function of α, traditionally a second derivative argument could indicate whether concavity is increasing or decreasing as a function of α. Unfortunately, $\frac{d^2}{d\eta_\mathrm{t}^2} f_{\alpha,\eta_\mathrm{c}}(\eta_\mathrm{t})$ is an unwieldy analytical expression. However, using a Taylor series approximation of $\frac{d^2}{d\eta_\mathrm{t}^2} f_{\alpha,\eta_\mathrm{c}}(\eta_\mathrm{t})$ near $\eta_\mathrm{t} = 1/2$, we find that the dominating term is $\approx -\alpha^2$. Thus, while not a proof, this *indicates* that concavity of $f_{\alpha,\eta_\mathrm{c}}(\cdot)$ increases as α increases greater than 1, which is sufficient for our purposes in the sequel.

3. **ZEROES.** It can be readily shown that for every $\eta_\mathrm{c} \in [0,1]$, $f_{\alpha,\eta_\mathrm{c}}(1/2) = 0$ for any $\alpha > 1$. Further, it can be shown that for any $\eta_\mathrm{c} \in [0,1]$, $\lim\limits_{\alpha \to 1^+} f_{\alpha,\eta_\mathrm{c}}(\eta_\mathrm{c}) \to 0^-$. Thus, the exact values of η_t for the other zero of $f_{\alpha,\eta_\mathrm{c}}(\cdot)$ (not $\eta_\mathrm{t} = 1/2$), for each $\alpha > 1$, are given by the solution to the following transcendental equation:

$$\eta_\mathrm{t} = \left(\left((1-\eta_\mathrm{t})^{\alpha-1} \right)^{1-\frac{1}{\eta_c}} \left(\eta_\mathrm{t}^{\alpha} + (1-\eta_\mathrm{t})^{\alpha} \right)^{\frac{1}{\eta_c}} \right)^{\frac{1}{\alpha-1}}, \tag{B.84}$$

which can be rewritten as

$$\log \left(\frac{\eta_\mathrm{t}}{(1-\eta_\mathrm{t})^{1-\frac{1}{\eta_c}}} \right) = \frac{\alpha}{\alpha-1} \log \left(\eta_\mathrm{t}^{\frac{1}{\eta_c}} \right) + \frac{1}{\eta_\mathrm{c}(\alpha-1)} \log \left(1 + \left(\frac{1}{\eta_t} - 1 \right)^{\alpha} \right), \tag{B.85}$$

and can also be rewritten as

$$\left(\frac{\eta_{\mathrm{t}}}{1-\eta_{\mathrm{t}}}\right)^{\eta_{\mathrm{c}}(\alpha-1)} = \eta_{\mathrm{t}}\left(\frac{\eta_{\mathrm{t}}}{1-\eta_{\mathrm{t}}}\right)^{\alpha-1} + (1-\eta_{\mathrm{t}}). \tag{B.86}$$

Suppose $\eta_{\mathrm{c}} > 1/2$, then since we have a Bayes blunting twist, $\eta_{\mathrm{c}} \geq \eta_{\mathrm{t}} \geq 1/2$. Letting $\alpha \to \infty$, note that the second term on the right-hand-side is 0. Thus, we can solve for the zeroes from (B.85) when $\alpha = \infty$ by examining

$$\log\left(\frac{\eta_{\mathrm{t}}}{(1-\eta_{\mathrm{t}})^{1-1/\eta_{\mathrm{c}}}}\right) = \log\left(\eta_{\mathrm{t}}^{\frac{1}{\eta_{\mathrm{c}}}}\right). \tag{B.87}$$

After some manipulations, we obtain $\log\left(\frac{1}{\eta_{\mathrm{t}}} - 1\right) = 0$, which is only satisfied when $\eta_{\mathrm{t}} = 1/2$. Thus, for $\eta_{\mathrm{c}} > 1/2$ and $\alpha \to \infty$, both zeroes of (B.82) converge at $\eta_{\mathrm{t}} = 1/2$. For $\eta_{\mathrm{c}} < 1/2$, the same argument holds, *mutatis mutandis*. Lastly, from (B.86), it can be shown that given fixed η_{c} and η_{t} under a Bayes blunting twist, a solution $\alpha > 1$ must exist, through reasoning about the rate of increase of $\left(\frac{\eta_{\mathrm{t}}}{1-\eta_{\mathrm{t}}}\right)^{\alpha-1}$ as a function of α, which is common to both sides.

4. **MAXIMUM.** It can also be shown that the maximum of $f_{\alpha,\eta_{\mathrm{c}}}(\cdot)$ for each $\alpha > 1$ as a function of η_{t} is given by the following transcendental equation

$$\frac{\alpha(1-\eta_{\mathrm{c}}) + \eta_{\mathrm{c}} - \eta_{\mathrm{t}}}{\alpha\eta_{\mathrm{c}} - \eta_{\mathrm{c}} + \eta_{\mathrm{t}}} = \left(\frac{1}{\eta_{\mathrm{t}}} - 1\right)^{\alpha}. \tag{B.88}$$

One key observation we can make from (B.88) is that as α increases, the term on the right-hand-side grows (or decays) exponentially with α, whereas the term on the left-hand-side is linear in α. With case-by-case analysis, i.e., for $\eta_{\mathrm{c}} > 1/2$ or $\eta_{\mathrm{c}} < 1/2$, it can be reasoned that as α increases, the solution to (B.88), η_{t}, approaches $1/2$. A second key observation we can make from (B.88) is that as $\alpha \to 1^{+}$, the solution to (B.88), η_{t}, approaches $\eta_{\mathrm{c}}/2 + 1/4$. This is readily observed by setting $\alpha = 1+\epsilon$, for some $\epsilon > 0$, and $\eta_{\mathrm{t}} = \frac{\eta_{\mathrm{c}}-1/2}{2} + 1/2 = \eta_{\mathrm{c}}/2 + 1/4$, along with a Taylor series approximation of $(1/\eta_{\mathrm{t}} - 1)^{1+\epsilon}$ for ϵ near 0.

Intuitively, the remainder of the proof consists of a "covering" argument. In words, we choose the least twisted η_{c}, i.e. η_{c}^{*}, under $\eta_{\mathrm{c}} \to \eta_{\mathrm{t}}$, via its associated $\alpha_0 > 1$ (as given in Lemma 9**(d)**), then we notice that this induces non-negativity of $f_{\alpha_0,\eta_{\mathrm{c}}^{*}}(\eta_{\mathrm{t}})$ given in (B.82). Next, we argue that this choice of $\alpha_0 > 1$ implies that all η_{c} are "covered" - in the sense of inducing non-negativity of (B.82), i.e., $f_{\alpha_0,\eta_{\mathrm{c}}}(\eta_{\mathrm{t}}) > 0$ for all η_{c} under $\eta_{\mathrm{c}} \to \eta_{\mathrm{t}}$. Finally, we use the non-negativity of the expectation to achieve the desired result, i.e., the left-hand-side of (B.73).

Continuing, let $\eta_{\mathrm{c}} \to \eta_{\mathrm{t}}$ be a *strictly* Bayes blunting twist. Thus, we have that there exists $\epsilon > 0$ such that either $(\eta_{\mathrm{c}} + \epsilon \leq \eta_{\mathrm{t}} \leq 1/2)$ or $(\eta_{\mathrm{c}} - \epsilon \geq \eta_{\mathrm{t}} \geq 1/2)$ for all η_{c}. By ZEROES of f, we have that for every $\eta_{\mathrm{c}} \in [0,1]$, $f_{\alpha,\eta_{\mathrm{c}}}(1/2) = 0$ for any $\alpha > 1$ and $\lim_{\alpha\to 1^{+}} f_{\alpha,\eta_{\mathrm{c}}}(\eta_{\mathrm{c}}) \to 0^{-}$. We also have that as $\alpha \to \infty$, both zeroes of (B.82) converge at $\eta_{\mathrm{t}} = 1/2$. Thus, for every η_{c}, the second zero (the first one is at $\eta_{\mathrm{t}} = 1/2$)

continuously shifts (CONTINUITY) from being located at $\eta_{\text{t}} = \eta_{\text{c}}$ (as $\alpha \to 1$) to being located at $\eta_{\text{t}} = 1/2$ (as $\alpha \to \infty$). In order to identify the least twisted η_{c} under $\eta_{\text{c}} \to \eta_{\text{t}}$, let

$$\alpha^*(\eta_{\text{c}}, \eta_{\text{t}}) := \frac{\iota(\eta_{\text{c}})}{\iota(\eta_{\text{t}})}, \tag{B.89}$$

as stated in Lemma 9**(c)**. By Lemma 9**(d)**, we know that $\alpha^*(\eta_{\text{c}}, \eta_{\text{t}}) \geq 1$; furthermore, due to *strictness* of the Bayes blunting twist $\eta_{\text{c}} \to \eta_{\text{t}}$, we indeed have a *strict inequality*, i.e., $\alpha^*(\eta_{\text{c}}, \eta_{\text{t}}) > 1$ for all η_{c} under $\eta_{\text{c}} \to \eta_{\text{t}}$. Choose $\alpha_0 := \inf\{\alpha^*(\eta_{\text{c}}, \eta_{\text{t}}) : \forall \eta_{\text{c}} \text{ under } \eta_{\text{c}} \to \eta_{\text{t}}\}$, where we break ties arbitrarily, and note that $\alpha_0 > 1$, again by strictness. Also note that there exists η_{c}^* associated with α_0 such that $\alpha_0 = \iota(\eta_{\text{c}}^*)/\iota(\eta_{\text{t}})$ under $\eta_{\text{c}} \to \eta_{\text{t}}$, in other words, $f_{\alpha_0, \eta_{\text{c}}^*}(\eta_{\text{t}}) > 0$ (due to $t_{\ell^\alpha}(\eta_{\text{t}})$ reversing the effects of the Bayes blunting twist and tuning back to η_{c}^*). Thus, by CONTINUITY, CONCAVITY, and ZEROES of f above and this choice of $\alpha_0 > 1$, we have that for all η_{c}, $f_{\alpha_0, \eta_{\text{c}}}(\eta_{\text{t}}) > 0$ under $\eta_{\text{c}} \to \eta_{\text{t}}$, i.e., that all η_{c} are "covered", due to the ordering of the relative zeroes of f induced by identifying η_{c}^* through α_0. Therefore, we obtain from (B.81) and (B.82) (i.e., the non-negativity of the expectation) that for the chosen $\alpha_0 > 1$, we have that $0 < \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; 1) - \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha_0)$, i.e.,

$$\text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; 1) > \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha_0), \tag{B.90}$$

as desired.

We now show that $\text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha_0) \geq \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha^\star)$, where $\alpha^\star$ is a mapping such that $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$, i.e., returning an $\alpha > 1$ for every $x \in \mathcal{X}$. By MAXIMUM above, we note that for a given η_{c}, the maximum of $f_{\alpha, \eta_{\text{c}}}(\cdot)$ moves from being achieved at $\eta_{\text{t}} = \eta_{\text{c}}/2 + 1/4$, to being achieved at $\eta_{\text{t}} = 1/2$ in the limit as α increases greater than 1. By CONCAVITY above, we also observe that $f_{\alpha, \eta_{\text{c}}}(\cdot)$ for a fixed η_{c} *appears* (which is sufficient for the inequality) to become more strongly convex in general as α increases greater than 1. We also note by CONTINUITY of f above that the maximums are continuous in $\alpha > 1$. Thus, under the *strictly* Bayes blunting twist $\eta_{\text{c}} \to \eta_{\text{t}}$, for every η_{c}, there may exist an $\alpha > 1$ which induces a larger (positive) magnitude in $f_{\alpha, \eta_{\text{c}}}(\eta_{\text{t}})$ than for the fixed $\alpha_0 > 1$ we found previously (for (B.90)). Thus, there exists an optimal mapping $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$, such that

$$\text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha_0) \geq \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha^\star). \tag{B.91}$$

Note that in the degenerate case, $\alpha^\star = \alpha_0$ for every $x \in \mathcal{X}$. Therefore, combining (B.90) and (B.91), we obtain

$$\text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; 1) > \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha_0) \geq \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha^\star), \tag{B.92}$$

which is the desired result. Lastly, note from Lemma 9**(c)** and (B.75) that $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$ is indeed given by $\alpha^\star(x) := \iota(\eta_{\text{c}}(x))/\iota(\eta_{\text{t}}(x))$, for every $x \in \mathcal{X}$, and hence note that $\text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha^\star) = H(\eta_{\text{c}})$, i.e., from (3.9) that $D_{\text{KL}}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha^\star) = 0$.

B.1.7 Proof of Theorem 8

As explained in the main body, we prove a result more general than Theorem 8. However, briefly note that (3.13), like the statement provided in Theorem 7 in (3.10), is proved for CE, but the statement provided in the main body as KL is readily obtained by subtracting $H(\eta_c)$ from both sides of the inequality.

First, we need a simple technical Lemma.

Lemma 14. *For any $B > 0$, $\forall |z| \leq B, \forall \alpha \in \mathbb{R}$,*

$$\log(1+\exp(\alpha z)) \quad \leq \quad \log(1+\exp(\alpha B)) - \frac{B-z}{2} \cdot \alpha. \tag{B.93}$$

Proof. We first note that $\forall |z| \leq 1, \forall \alpha \in \mathbb{R}$,

$$\begin{aligned} \log(1+\exp(\alpha z)) \quad &\leq \quad \frac{1+z}{2} \cdot \log(1+\exp(\alpha)) + \frac{1-z}{2} \cdot \log(1+\exp(-\alpha)) && \text{(B.94)} \\ &= \log(1+\exp(\alpha)) - \frac{1-z}{2} \cdot \alpha, && \text{(B.95)} \end{aligned}$$

which indeed holds as the LHS of (B.94) is convex and the RHS is the equation of a line passing through the points $(-1, \log(1+\exp(-\alpha)))$ and $(1, \log(1+\exp(\alpha)))$. In (B.95), we use $\log(1+\exp(-\alpha)) = \log(\exp(-\alpha) \cdot (1+\exp(\alpha)))$ on the second term in the RHS. Hence if instead $|z| \leq B$, then

$$\begin{aligned} \log(1+\exp(\alpha z)) \quad &= \quad \log\left(1+\exp\left(\alpha B \cdot \frac{z}{B}\right)\right) && \text{(B.96)} \\ &\leq \quad \log(1+\exp(\alpha B)) - \frac{B-z}{2} \cdot \alpha, && \text{(B.97)} \end{aligned}$$

as claimed. □

We now show another Lemma which bounds the log quantities appearing in the cross-entropy in (3.7), recalling that $\iota(u) \doteq \log(u/(1-u))$.

Lemma 15. *Fix $B > 0$. For any $\boldsymbol{x} \in \mathcal{X}$ such that*

$$\frac{1}{1+\exp B} \leq \eta_t(\boldsymbol{x}) \leq \frac{\exp B}{1+\exp B}, \tag{B.98}$$

the following properties hold for the Bayes tiltes estimate t_ℓ of α-loss:

$$\begin{aligned} -\log t_{\ell^\alpha}(\eta_t(\boldsymbol{x})) \quad &\leq \quad \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B+\iota(\eta_t(\boldsymbol{x}))}{2}, \\ -\log(1-t_{\ell^\alpha}(\eta_t(\boldsymbol{x}))) \quad &\leq \quad \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B-\iota(\eta_t(\boldsymbol{x}))}{2}. \end{aligned}$$

Proof. We note, using $z \doteq -\log\left(\frac{1-\eta_t(\boldsymbol{x})}{\eta_t(\boldsymbol{x})}\right)$, which satisfies $|z| \leq B$ from (B.98) and

Lemma 14,

$$\begin{aligned} -\log t_{\ell^\alpha}(\eta_{\mathrm{t}}(\boldsymbol{x})) &= -\log\left(\frac{\eta_{\mathrm{t}}(\boldsymbol{x})^\alpha}{\eta_{\mathrm{t}}(\boldsymbol{x})^\alpha + (1-\eta_{\mathrm{t}}(\boldsymbol{x}))^\alpha}\right) \\ &= -\log\left(\frac{1}{1+\left(\frac{1-\eta_{\mathrm{t}}(\boldsymbol{x})}{\eta_{\mathrm{t}}(\boldsymbol{x})}\right)^\alpha}\right) \\ &= \log\left(1+\left(\frac{1-\eta_{\mathrm{t}}(\boldsymbol{x})}{\eta_{\mathrm{t}}(\boldsymbol{x})}\right)^\alpha\right) \\ &= \log\left(1+\exp\left(\alpha\log\left(\frac{1-\eta_{\mathrm{t}}(\boldsymbol{x})}{\eta_{\mathrm{t}}(\boldsymbol{x})}\right)\right)\right) && \text{(B.99)} \\ &\leq \log(1+\exp(\alpha B)) - \alpha\cdot\frac{B-\log\left(\frac{1-\eta_{\mathrm{t}}(\boldsymbol{x})}{\eta_{\mathrm{t}}(\boldsymbol{x})}\right)}{2} && \text{(B.100)} \\ &= \log(1+\exp(\alpha B)) - \alpha\cdot\frac{B+\iota(\eta_{\mathrm{t}}(\boldsymbol{x}))}{2}, && \text{(B.101)} \end{aligned}$$

and similarly,

$$\begin{aligned} -\log(1-t_{\ell^\alpha}(\eta_{\mathrm{t}}(\boldsymbol{x}))) &= \log\left(1+\exp\left(-\alpha\log\left(\frac{1-\eta_{\mathrm{t}}(\boldsymbol{x})}{\eta_{\mathrm{t}}(\boldsymbol{x})}\right)\right)\right) && \text{(B.102)} \\ &\leq \log(1+\exp(-\alpha B)) + \alpha\cdot\frac{B+\iota(\eta_{\mathrm{t}}(\boldsymbol{x}))}{2} \\ &= \log(1+\exp(\alpha B)) - \alpha B + \alpha\cdot\frac{B+\iota(\eta_{\mathrm{t}}(\boldsymbol{x}))}{2} \\ &= \log(1+\exp(\alpha B)) - \alpha\cdot\frac{B-\iota(\eta_{\mathrm{t}}(\boldsymbol{x}))}{2}, && \text{(B.103)} \end{aligned}$$

as claimed. □

Denote M(B) the distribution restricted to the support for which we have a.s. (B.98) and let $p(B)$ be the weight of this support in M. Let M($\overline{B}$) denote the restriction of M to the complement of this support. We let D(B) is the product distribution on examples ($\mathcal{X}\times\mathcal{Y}$) over the support of M($B$) induced by marginal M($B$) and posterior η_{c} (see Reid and Williamson (2011)). We are now in a position to show our generalization to Theorem 8.

Theorem 16. *For any fixed $B > 0$, let*

$$\mathsf{e}(B) \doteq \frac{\mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim \mathrm{D}(B)}\left[\mathsf{Y}\cdot\iota(\eta_{\mathrm{t}}(\mathsf{X}))\right]}{B} \in [-1,1]. \quad \text{(B.104)}$$

and suppose we fix the scalar $\alpha \doteq \alpha^{**}$ with

$$\alpha^{**} \doteq \frac{\iota\left(\frac{1+\mathsf{e}(B)}{2}\right)}{B}. \quad \text{(B.105)}$$

then the following bound holds on the cross-entropy of the Bayes tilted estimate of the α-loss:

$$\begin{aligned}\mathrm{CE}(\eta_c, \eta_t; \alpha) &\leq p(B) \cdot H\left(\frac{1+\mathsf{e}(B)}{2}\right) \\ &\quad +(1-p(B)) \cdot \left(\mathsf{e}(\overline{B}) \cdot \log\left(\frac{1+|\mathsf{e}(B)|}{1-|\mathsf{e}(B)|}\right) + \frac{1-|\mathsf{e}(B)|}{1+|\mathsf{e}(B)|}\right),\end{aligned}$$

where $\mathsf{e}(\overline{B}) \doteq \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim \mathrm{D}(\overline{B})}\left[\max\left\{0, -\mathrm{sign}(\alpha^{**}) \cdot \mathsf{Y} \cdot \iota(\eta_t(\mathsf{X}))\right\}\right]/B$ *and* $\mathrm{D}(\overline{B})$ *is is defined analogously to* $\mathrm{D}(B)$ *with respect to* $\mathrm{M}(\overline{B})$*.*

Remark: we notice this is indeed a generalization of Theorem 8, which corresponds to case $p(B) = 1$. We also note $|\mathsf{e}(\overline{B})| \geq 1$.

Proof. We remark that the cross-entropy (3.7) can be split as:

$$\begin{aligned}\mathrm{CE}(\eta_c, \eta_t; \alpha) &\doteq \mathbb{E}_{\mathsf{X}\sim\mathrm{M}}\left[\begin{array}{c}\eta_c(\mathsf{X}) \cdot -\log t_{\ell^\alpha}(\eta_t(\mathsf{X})) \\ +(1-\eta_c(\mathsf{X})) \cdot -\log(1-t_{\ell^\alpha}(\eta_t(\mathsf{X})))\end{array}\right] \\ &= p(B) \cdot K(\alpha) + (1-p(B)) \cdot L(B), \qquad (B.106)\end{aligned}$$

with

$$\begin{aligned}K(\alpha) &\doteq \mathbb{E}_{\mathsf{X}\sim\mathrm{M}(B)}\left[\begin{array}{c}\eta_c(\mathsf{X}) \cdot -\log t_{\ell^\alpha}(\eta_t(\mathsf{X})) \\ +(1-\eta_c(\mathsf{X})) \cdot -\log(1-t_{\ell^\alpha}(\eta_t(\mathsf{X})))\end{array}\right], \qquad (B.107) \\ J(B) &\doteq \mathbb{E}_{\mathsf{X}\sim\mathrm{M}(\overline{B})}\left[\begin{array}{c}\eta_c(\mathsf{X}) \cdot -\log t_{\ell^\alpha}(\eta_t(\mathsf{X})) \\ +(1-\eta_c(\mathsf{X})) \cdot -\log(1-t_{\ell^\alpha}(\eta_t(\mathsf{X})))\end{array}\right]. \qquad (B.108)\end{aligned}$$

We now focus on a bound on $K(\alpha)$, which we achieve via Lemma 15:

$$\begin{aligned}K(\alpha) &\leq \mathbb{E}_{\mathsf{X}\sim\mathrm{M}(B)}\left[\begin{array}{c}\eta_c(\mathsf{X}) \cdot \left(\log(1+\exp(\alpha B)) - \alpha \cdot \frac{B+\iota(\eta_t(\mathsf{X}))}{2}\right) \\ +(1-\eta_c(\mathsf{X})) \cdot \left(\log(1+\exp(\alpha B)) - \alpha \cdot \frac{B-\iota(\eta_t(\mathsf{X}))}{2}\right)\end{array}\right] \\ &= \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B+\mathbb{E}_{\mathsf{X}\sim\mathrm{M}(B)}[\eta_c(\mathsf{X})\iota(\eta_t(\mathsf{X}))+(1-\eta_c(\mathsf{X}))\cdot-\iota(\eta_t(\mathsf{X}))]}{2} \\ &= \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B+\mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathrm{D}(B)}\left[\mathsf{Y}\cdot\iota(\eta_t(\mathsf{X}))\right]}{2} \qquad (B.109) \\ &\doteq \underbrace{\log(1+\exp(\alpha B)) - \alpha \cdot \frac{B + B\cdot\mathsf{e}(B)}{2}}_{\doteq L(\alpha)},\end{aligned}$$

where we recall

$$\mathsf{e}(B) \doteq \frac{\mathbb{E}_{\mathsf{X}\sim\mathrm{D}(B)}\left[\mathsf{Y}\cdot\iota(\eta_t(\mathsf{X}))\right]}{B} \quad \in [-1, 1]. \qquad (B.110)$$

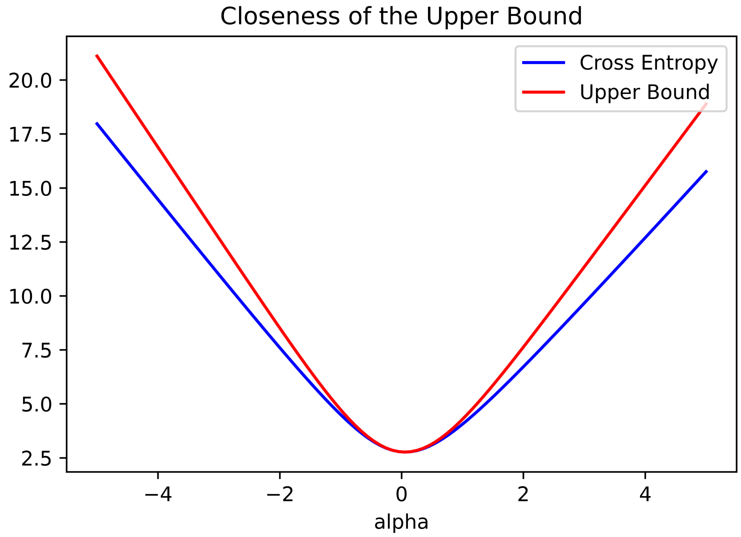


Figure B.5: A Plot Illustrating the Closeness Of (B.109) Where $k(\alpha)$ Is given in Blue and $l(\alpha)$ Is given in Red for a Toy Distribution: $x \sim \text{Uniform}([-b, B])$ (Recall $b > 0$ Is the Clipping Threshold and We Set $b = 2$ Here) Where $\eta_{\text{c}}(X) = (1+\exp{(-x/a)})^{-1}$ and $\eta_{\text{t}}(X) = (1+\exp{(-x/b)})^{-1}$ Such That $a = 10$ and $b = .6$.

Notice the change in distribution in (B.109), where $\mathrm{D}(B)$ is the product distribution on examples $(\mathcal{X} \times \mathcal{Y})$ over the support of $\mathrm{M}(B)$ induced by marginal $\mathrm{M}(B)$ and posterior η_{c} (see Reid and Williamson (2011)). We have

$$L'(\alpha) \quad = \quad B \cdot \left(\frac{\exp(B\alpha)}{1+\exp(B\alpha)} - \frac{1+\mathsf{e}(B)}{2}\right), \tag{B.111}$$

which zeroes for

$$\alpha^{**} \quad = \quad \frac{1}{B} \cdot \log\left(\frac{1+\mathsf{e}(B)}{1-\mathsf{e}(B)}\right) = \frac{\iota(q(B))}{B}. \tag{B.112}$$

Further, we have that

$$L''(\alpha) = \frac{d}{d\alpha}L'(\alpha) = \frac{B^2\exp{(\alpha B)}}{(\exp{(\alpha B)}+1)^2}, \tag{B.113}$$

and plugging in (B.112) yields

$$L''(\alpha^{**}) = B^2\frac{1-\mathsf{e}^2}{4}. \tag{B.114}$$

Note that for fixed $B > 0$, as $|\alpha|$ increases in (B.113), $L''(\alpha)$ decreases. Thus, when the magnitude of α^* is large (due to the distribution and twist), this implies that there is more "flatness" near the choice of α^*. Hence, in these regimes, a choice of α_0

"close-enough" to α^* should have similar performance in practice. Continuing with the main line, plugging in (B.112) into (B.109) yields

$$\begin{aligned}
K(\alpha^{**}) &\leq \log(1+\exp(B\alpha^{**})) - B\cdot\frac{1+\mathsf{e}(B)}{2}\cdot\alpha^{**} && \text{(B.115)}\\
&= -\log\left(\frac{1-\mathsf{e}(B)}{2}\right) - \frac{1+\mathsf{e}(B)}{2}\cdot\log\left(\frac{1+\mathsf{e}(B)}{1-\mathsf{e}(B)}\right) && \text{(B.116)}\\
&= -\frac{1+\mathsf{e}(B)}{2}\log\left(\frac{1+\mathsf{e}(B)}{2}\right) - \frac{1-\mathsf{e}(B)}{2}\log\left(\frac{1-\mathsf{e}(B)}{2}\right) && \text{(B.117)}\\
&= H\left(\frac{1+\mathsf{e}(B)}{2}\right), && \text{(B.118)}
\end{aligned}$$

which is the statement of Theorem 8. We now focus on $J(B)$. Since $\log(1+\exp(-z)) \leq \exp(-z), \forall z$ via an order-1 Taylor expansion, it follows that if $z \geq C$ for some $C > 0$, then $\log(1+\exp(-z)) \leq \exp(-C)$. Equivalently, we get

$$z \geq C \quad\Rightarrow\quad \log(1+\exp(z)) \leq z + \exp(-C). \tag{B.119}$$

By symmetry, we have

$$z \leq -C \quad\Rightarrow\quad \log(1+\exp(z)) \leq \exp(-C), \tag{B.120}$$

so we get

$$|z| \geq C \quad\Rightarrow\quad \log(1+\exp(z)) \leq \max\{0, z\} + \exp(-C). \tag{B.121}$$

By definition, we have for any $\boldsymbol{x}$ in the support of $\mathrm{M}(\overline{B})$,

$$\left|\log\left(\frac{1-\eta_{\mathsf{t}}(\boldsymbol{x})}{\eta_{\mathsf{t}}(\boldsymbol{x})}\right)\right| \quad\geq\quad B, \tag{B.122}$$

so have, considering $C \doteq B\cdot|\alpha^*|$, from (B.99) and (B.102),

$$\begin{aligned}
J(B) &\doteq \mathbb{E}_{\mathsf{X}\sim\mathrm{M}(\overline{B})}\left[\begin{array}{c}\eta_{\mathsf{c}}(\mathsf{X})\cdot -\log t_{\ell^\alpha}(\eta_{\mathsf{t}}(\mathsf{X}))\\ +(1-\eta_{\mathsf{c}}(\mathsf{X}))\cdot -\log(1-t_{\ell^\alpha}(\eta_{\mathsf{t}}(\mathsf{X})))\end{array}\right] && \text{(B.123)}\\
&= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathrm{D}(\overline{B})}\left[\log\left(1+\exp\left(\mathsf{Y}\alpha^*\log\left(\frac{1-\eta_{\mathsf{t}}(\mathsf{X})}{\eta_{\mathsf{t}}(\mathsf{X})}\right)\right)\right)\right]\\
&\leq \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathrm{D}(\overline{B})}\left[\max\left\{0, \mathsf{Y}\alpha^{**}\log\left(\frac{1-\eta_{\mathsf{t}}(\mathsf{X})}{\eta_{\mathsf{t}}(\mathsf{X})}\right)\right\}\right] + \exp\left(-B\cdot|\alpha^{**}|\right)\\
&= |\alpha^{**}|\cdot\mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathrm{D}(\overline{B})}\left[\max\left\{0, -\mathrm{sign}(\alpha^{**})\cdot\mathsf{Y}\cdot\iota(\eta_{\mathsf{t}}(\mathsf{X}))\right\}\right] + \frac{1-|\mathsf{e}(B)|}{1+|\mathsf{e}(B)|}\\
&= \mathsf{e}(\overline{B})\log\left(\frac{1+|\eta(B)|}{1-|\mathsf{e}(B)|}\right) + \frac{1-|\mathsf{e}(B)|}{1+|\mathsf{e}(B)|}, && \text{(B.124)}
\end{aligned}$$

which completes the proof of Theorem 16 after replacing the expression of α^*. □

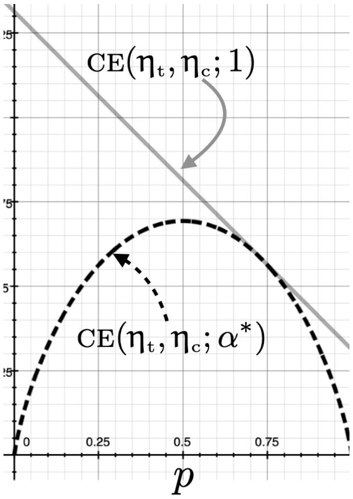


Figure B.6: Comparison Between the Cross-entropy of the Logistic Loss ($\alpha = 1$) and That of the α-loss for the Scalar Correction in (B.126) in Theorem 16.

Remarks: Theorem 16 calls for several remarks:

Suppose $p(B) = 1$ so the cross-entropy $\mathrm{CE}(\eta_c, \eta_t; \alpha)$ in (B.106) reduces to $K(.)$, $B = 1$ and all logits take ±1 value a.e.,

$$z(\boldsymbol{x}) \doteq \log\left(\frac{\eta_t(\boldsymbol{x})}{1 - \eta_t(\boldsymbol{x})}\right) = \pm 1, \tag{B.125}$$

which can be achieved by clamping, and $p \doteq \mathbb{P}_{(\mathsf{X},\mathsf{Y})\sim \mathrm{M}(1)}[\mathsf{Y}z(\mathsf{X}) = 1]$, which gives $\mathsf{e}(1) = 2p - 1$,

$$\alpha^{**} = \log\left(\frac{p}{1-p}\right),$$

and

$$\mathrm{CE}(\eta_c, \eta_t; \alpha^{**}) \leq H(p) \tag{B.126}$$

from Theorem 16. The properness choice $\alpha^* = 1$ however gives

$$\begin{aligned} \mathrm{CE}(\eta_c, \eta_t; 1) = K(1) &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim \mathrm{M}(1)}\left[\log\left(1 + \exp\left(-\mathsf{Y}z(\mathsf{X})\right)\right)\right] &(\mathrm{B.127})\\ &= p\log(1+\exp(-1)) + (1-p)\log(1+\exp(1)). &(\mathrm{B.128})\\ &= \log(1+e) - p. &(\mathrm{B.129}) \end{aligned}$$

Figure B.6 plots $\mathrm{CE}(\eta_c, \eta_t; \alpha^{**})$ (B.126) vs $\mathrm{CE}(\eta_c, \eta_t; 1)$ (B.129). We remark that $\mathrm{CE}(\eta_c, \eta_t; \alpha^{**}) \leq \mathrm{CE}(\eta_c, \eta_t; 1)$, and the difference is especially large as $p \to \{0, 1\}$, for which $\mathrm{CE}(\eta_c, \eta_t; \alpha^{**}) \to 0$ while we always have $\mathrm{CE}(\eta_c, \eta_t; 1) > 0.3, \forall p$.

Incidence of computing α^* on an *estimate* of $\mathsf{e}(B)$: Theorem 16 can be refined if, instead of the true value $\mathsf{e}(B)$ we have access to an estimate $\hat{\mathsf{e}}(B)$. In this case, we can refine the proof of the Theorem from the series of eqs in (B.118). We remark that

$$H'\left(\frac{1+z}{2}\right) \quad = \quad \frac{1}{2}\cdot\log\left(\frac{1-z}{1+z}\right), \tag{B.130}$$

so since H is concave, we have for any $\mathsf{e}(B), \hat{\mathsf{e}}(B)$,

$$\begin{aligned} H\left(\frac{1+\mathsf{e}(B)}{2}\right) &\leq H\left(\frac{1+\hat{\mathsf{e}}(B)}{2}\right) + \left(\frac{1+\mathsf{e}(B)}{2} - \frac{1+\hat{\mathsf{e}}(B)}{2}\right)\cdot\frac{1}{2}\cdot\log\left(\frac{1-\hat{\mathsf{e}}(B)}{1+\hat{\mathsf{e}}(B)}\right) \\ &= H\left(\frac{1+\hat{\mathsf{e}}(B)}{2}\right) + \frac{\mathsf{e}(B)-\hat{\mathsf{e}}(B)}{4}\cdot\log\left(\frac{1-\hat{\mathsf{e}}(B)}{1+\hat{\mathsf{e}}(B)}\right) \\ &\leq H\left(\frac{1+\hat{\mathsf{e}}(B)}{2}\right) + \frac{|\mathsf{e}(B)-\hat{\mathsf{e}}(B)|}{4}\cdot\log\left(\frac{1+|\hat{\mathsf{e}}(B)|}{1-|\hat{\mathsf{e}}(B)|}\right) \\ &= H\left(\frac{1+\hat{\mathsf{e}}(B)}{2}\right) + \frac{|\mathsf{e}(B)-\hat{\mathsf{e}}(B)|}{4}\cdot\log\left(1+\frac{2|\hat{\mathsf{e}}(B)|}{1-|\hat{\mathsf{e}}(B)|}\right) \\ &\leq H\left(\frac{1+\hat{\mathsf{e}}(B)}{2}\right) + \frac{|\mathsf{e}(B)-\hat{\mathsf{e}}(B)||\hat{\mathsf{e}}(B)|}{2(1-|\hat{\mathsf{e}}(B)|)}, \end{aligned} \tag{B.131}$$

where we have used $\log(1+z) \leq z$ for the last inequality.

Polarity of α^{**}: as presented in the main body, the state of the art defines the α-loss only for $\alpha \geq 0$. The proof of Theorem 16, and more specifically its proof, hints at why alleviating this constraint is important and corresponds to especially difficult cases. We have the general rule $\alpha^{**} \leq 0$ iff $\mathsf{e}(B) \leq 0$, which indicates that the twisted posterior tends to be small when the clean posterior tends to be large. Since the Bayes tilted estimate is symmetric if we switch the couple $(\alpha, \eta_{\mathrm{t}})$ for $(-\alpha, 1-\eta_{\mathrm{t}})$, $\alpha^{**} \leq 0$ provokes a change of polarity in the Bayes tilted estimate compared to the twisted posterior. It thus corrects the twisted posterior. We emphasize that such a situation happens for especially damaging twists (in particular, *not* Bayes blunting).

B.1.8 Pseudo-Inverse Link

Derivation of (3.16): From the definition of F in (3.3), we have that

$$F(z) := (-\underline{L})^{\star}(-z), \forall z \in \mathbb{R}. \tag{B.132}$$

Given $\underline{L}(u)$ associated with an underlying CPE loss, we have that

$$(-\underline{L})^{\star}(z) = \sup_{u\in[0,1]} [zu + \underline{L}(u)]. \tag{B.133}$$

Taking the derivative of the expression in brackets and solving for u, we obtain (assuming strictly concave $\underline{L}$)

$$z + \underline{L}'(u) = 0 \tag{B.134}$$

$$u^* = (\underline{L}')^{-1}(-z). \tag{B.135}$$

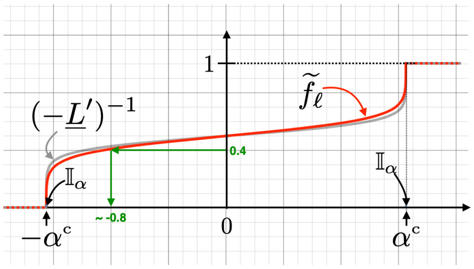


Figure B.7: The Pseudo-inverse Link Vs Inverse Link for the $\alpha = 5$ Loss. Notice the Quality of the Approximation.

Plugging (B.135) back into the expression in brackets in (B.133), we obtain

$$(-\underline{L})^\star(z) = z(\underline{L}')^{-1}(-z) + \underline{L}((\underline{L}')^{-1}(-z)), \tag{B.136}$$

and

$$F(z) = (-\underline{L})^\star(-z) = -z(\underline{L}')^{-1}(z) + \underline{L}((\underline{L}')^{-1}(z)). \tag{B.137}$$

Then, we have by the chain rule that

$$\frac{d}{dz}F(z) = \frac{d}{dz}\left[-z(\underline{L}')^{-1}(z) + \underline{L}((\underline{L}')^{-1}(z))\right] \tag{B.138}$$
$$= -(\underline{L}')^{-1}(z) - z\left((\underline{L}')^{-1}(z)\right)' + \underline{L}'((\underline{L}')^{-1}(z))\left((\underline{L}')^{-1}(z)\right)' \tag{B.139}$$
$$= -(\underline{L}')^{-1}(z), \tag{B.140}$$

where the last step is obtained since $\underline{L}'((\underline{L}')^{-1}(z))\left((\underline{L}')^{-1}(z)\right)' = z\left((\underline{L}')^{-1}(z)\right)'$. Thus, we have that

$$F'(z) := -(\underline{L}')^{-1}(z) = -(-\underline{L}')^{-1}(-z). \tag{B.141}$$

From property **(D)** of Lemma 12 in Appendix B.1.3 in (B.28), namely that $\underline{L}'(u) = \ell_1(t_\ell(u)) - \ell_{-1}(t_\ell(u))$, and from (B.141), we get that

$$-F'(z) = (\ell_{-1} \circ t_\ell - \ell_1 \circ t_\ell)^{-1}(-z), \tag{B.142}$$

as desired.

Pil Approximation: Let $\alpha \in [-\infty, \infty]$, and define the conjugate α^{c} such that $1/\alpha^{\mathrm{c}} + 1/\alpha = 1$, using by extension $\alpha^{\mathrm{c}}(\infty) = 1, \alpha^{\mathrm{c}}(1) = \infty$. If we were to exactly implement a boosting algorithm for the α-loss, we would have to find the *exact* inverse of (3.16), which would require inverting $-\underline{L}'(v) \doteq \alpha^{\mathrm{c}} \cdot t_\ell(v)^{\alpha^{\mathrm{c}}} - \alpha^{\mathrm{c}} \cdot t_\ell(1-v)^{\alpha^{\mathrm{c}}}$. Owing to the difficulty to carry out this step, we choose a sidestep that makes inversion straightforward and can fall in the conditions to apply Theorem 9, thus making PilBoost a boosting algorithm for the α-loss of interest. The trick does not just hold for the α-loss, so we describe it for a general loss ℓ assuming for simplicity that $\ell_1(1) = \ell_{-1}(0) = 0$ and $t_\ell, \ell_1, \ell_{-1}$ are invertible with ℓ_1, ℓ_{-1} non-negative, conditions that would hold for many popular losses (log, square, Matusita, etc.), and the α-loss.

We then approximate the link $-\underline{L}'$ by using just one of ℓ_{-1} or ℓ_1 depending on their argument, while ensuring functions match in $0, 1/2, 1$. We name $\widetilde{f_\ell}$ the *clipped inverse link*, CIL. Letting $a_\ell^- \doteq \ell_1(0)/(\ell_1(0) - \ell_1(1/2))$ and $a_\ell^+ \doteq \ell_{-1}(1)/(\ell_{-1}(1) - \ell_{-1}(1/2))$, our link approximation makes use of the following function: $f_\ell(u) \doteq f_\ell^-(u)$ if $u \leq 1/2$ and $f_\ell(u) \doteq f_\ell^+(u)$ otherwise, with the shorthands $f_\ell^-(u) \doteq a_\ell^- \cdot (\ell_1(1/2) - \ell_1(t_\ell(u)))$, $f_\ell^+(u) \doteq a_\ell^+ \cdot (\ell_{-1}(t_\ell(u)) - \ell_{-1}(1/2))$. The following Lemma shows, in addition to properties of f_ℓ, the expression obtained for the clipped inverse link for a general CPE loss.

Lemma 16. $f_\ell(u) = -\underline{L}'(u), \forall u \in \{0, 1/2, 1\}$*; furthermore, the clipped inverse link* $\widetilde{f_\ell} \doteq f_\ell^{-1}$ *is: (i)* $\widetilde{f_\ell}(z) = 0$ *if* $z < -\ell_1(0)$*; (ii)* $\widetilde{f_\ell}(z) = t_\ell^{-1} \circ \ell_1^{-1}\left(\frac{\ell_1(1/2)-\ell_1(0)}{\ell_1(0)} \cdot z + \ell_1(1/2)\right)$ *if* $-\ell_1(0) \leq z < 0$*; (iii)* $\widetilde{f_\ell}(z) = t_\ell^{-1} \circ \ell_{-1}^{-1}\left(\frac{\ell_{-1}(1)-\ell_{-1}(1/2)}{\ell_{-1}(1)} \cdot z + \ell_{-1}(1/2)\right)$ *if* $0 \leq z < \ell_{-1}(1)$*; (iv)* $\widetilde{f_\ell}(z) = 1$ *if* $z \geq \ell_{-1}(1)$*. Furthermore,* $\widetilde{f_\ell}$ *is continuous and if* ***(S)*** *and* ***(D)*** *hold, then* $\widetilde{f_\ell}$ *is derivable on* $\mathbb{R}$ *(with the only possible exceptions of* $\{-\ell_1(0), \ell_{-1}(1)\}$*).*

The proof is immediate once we remark that $\ell_1(1) = \ell_{-1}(0) = 0$ bring ”properness for the extremes”, *i.e.* $0 \in t_\ell(0), 1 \in t_\ell(1)$. We now give the expression of the formulas of interest regarding Lemma 16 for the α-loss.

Lemma 17. *We have for the α-loss,*

$$f_\ell(u) = \alpha^c \cdot \begin{cases} \left(\frac{2\cdot u^\alpha}{u^\alpha+(1-u)^\alpha}\right)^{\frac{1}{\alpha^c}} - 1 & \text{if} \quad u \leq 1/2, \\ 1 - \left(\frac{2\cdot(1-u)^\alpha}{u^\alpha+(1-u)^\alpha}\right)^{\frac{1}{\alpha^c}} & \text{if} \quad u \geq 1/2 \end{cases}, \tag{B.143}$$

$$\widetilde{f}(z) = \begin{cases} 0 & \text{if} \quad z \leq -\alpha^c, \\ \frac{(\alpha^c+z)^{\frac{\alpha^c}{\alpha}}}{(\alpha^c+z)^{\frac{\alpha^c}{\alpha}} + \left(2\alpha^{c\alpha^c} - (\alpha^c+z)^{\alpha^c}\right)^{\frac{1}{\alpha}}} & \text{if} \quad -\alpha^c \leq z \leq 0, \\ \frac{\left(2\alpha^{c\alpha^c} - (\alpha^c-z)^{\alpha^c}\right)^{\frac{1}{\alpha}}}{(\alpha^c-z)^{\frac{\alpha^c}{\alpha}} + \left(2\alpha^{c\alpha^c} - (\alpha^c-z)^{\alpha^c}\right)^{\frac{1}{\alpha}}} & \text{if} \quad 0 \leq z \leq \alpha^c, \\ 1 & \text{if} \quad z \geq \alpha^c. \end{cases} \tag{B.144}$$

Rewritten, we have that

$$\widetilde{f}(z) = \begin{cases} 0 & z \leq -\frac{\alpha}{\alpha-1} \\ \frac{\left(\frac{\alpha}{\alpha-1}+z\right)^{\frac{1}{\alpha-1}}}{\left(\frac{\alpha}{\alpha-1}+z\right)^{\frac{1}{\alpha-1}} + \left(2\left(\frac{\alpha}{\alpha-1}\right)^{\frac{\alpha}{\alpha-1}} - \left(\frac{\alpha}{\alpha-1}+z\right)^{\frac{\alpha}{\alpha-1}}\right)^{\frac{1}{\alpha}}} & -\frac{\alpha}{\alpha-1} \leq z \leq 0 \\ \frac{\left(2\left(\frac{\alpha}{\alpha-1}\right)^{\frac{\alpha}{\alpha-1}} - \left(\frac{\alpha}{\alpha-1}-z\right)^{\frac{\alpha}{\alpha-1}}\right)^{\frac{1}{\alpha}}}{\left(\frac{\alpha}{\alpha-1}-z\right)^{\frac{1}{\alpha-1}} + \left(2\left(\frac{\alpha}{\alpha-1}\right)^{\frac{\alpha}{\alpha-1}} - \left(\frac{\alpha}{\alpha-1}-z\right)^{\frac{\alpha}{\alpha-1}}\right)^{\frac{1}{\alpha}}} & 0 \leq z \leq \frac{\alpha}{\alpha-1} \\ 1 & z \geq \frac{\alpha}{\alpha-1} \end{cases} \tag{B.145}$$

Figure B.8 plots (B.145) for several values of α.

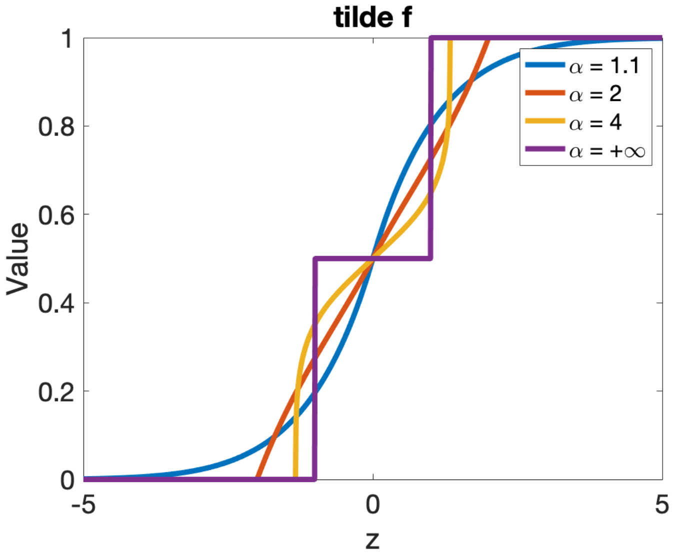


Figure B.8: A Plot of $\tilde{f}(z)$ as a Function of α as given In (B.145).

Remark 5. *It could be tempting to think that the clipped inverse link trivially comes from clipping the partial losses themselves such as replacing $\ell_1(u)$ by 0 if $u \geq 1/2$ and symmetrically for $\ell_{-1}(u)$. This is not the case as it would lead to $\underline{L}$ piecewise constant and therefore $-\underline{L}' = 0$ when defined.*

We turn to a result that authorizes us to use Thm 9 while virtually not needing **(E)** and **(C)** for α-loss. Denote $\mathbb{I}_\alpha \doteq \pm\alpha^{\mathrm{c}} \cdot [1 - (1/\alpha^4), 1]$ (See Fig. B.7).

Lemma 18. *Suppose $\alpha \geq 1.2$. For $\widetilde{f}_\ell$ defined as in Lemma 16, $\exists K \geq 0.133$ such that α-loss satisfies:*

$$\forall z \not\in \mathbb{I}_\alpha, |(\widetilde{f}_\ell - (-\underline{L}')^{-1})(z)| \quad \lesssim \quad K/\alpha. \tag{B.146}$$

Remark the necessity of a trick as we do not compute $(-\underline{L}')^{-1}$ in (B.146). The proof, in Section B.1.9, bypasses the difficulty by bounding the *horizontal* distance between the *inverses*. The Lemma can be read as: with the exception of an interval vanishing rapidly with α, the difference between $\widetilde{f}_\ell$ (that we can easily compute for the α-loss) and $(-\underline{L}')^{-1}$ (that we do not compute for the α-loss), in order or just pointwise (typically for $\alpha < 10$) is at most $0.14/\alpha$. We now show how we can virtually "get rid of" (**E**) and (**C**) in such a context to apply Theorem 9. Consider the following assumptions: (i) no edge falls in $\mathbb{I}_\alpha$, (ii) the weak learner guarantees $\gamma = 0.14$, (iii) the average weights, $\overline{w}_j \doteq \mathbf{1}^\top \boldsymbol{w}_j/m$, satisfies $\overline{w}_j \geq 0.4$. Looking at Figure B.7, we see that (i) is virtually not limiting at all; (ii) is a reasonable assumption on WL; remembering that a weight has the form $w = \widetilde{f}_\ell(-yH(\boldsymbol{x}))$, we see that (iii) requires H to be not "too good", see for example Figure B.7 in which case $w = 0.4$ implies an edge $yH \leq 0.8$. We now observe that given (i), it is trivial to find a_f to satisfy (**C**) since we focus only on one α-loss. Suppose $\alpha \geq 2.7$, which approaches the average value of the αs in our experiments, and finally let $\zeta \doteq 2.5/2.7 \approx 0.926$. Then we get the chain of inequalities:

$$\begin{aligned}\Delta(F) \underset{\text{Lem. 18}}{\leq} M \cdot \frac{0.14}{\alpha} \underset{\text{(ii)}}{=} \gamma M \cdot \frac{1}{\alpha} \underset{\textbf{(WLA)}}{\leq} \frac{1}{\alpha} \cdot \tilde{\mathsf{e}}_j \\ \doteq \frac{1}{\alpha \overline{w}_j} \cdot \mathsf{e}_j \underset{\text{(iii)}}{\leq} \frac{2.5}{\alpha} \cdot \mathsf{e}_j \leq \frac{2.5}{2.7} \cdot \mathsf{e}_j \doteq \zeta \cdot \mathsf{e}_j, \end{aligned} \tag{B.147}$$

and so $(\mathbf{E})$ is implied by the weak learning assumption. To summarise, PilBoost boosts the convex surrogate of the α-loss without either computing it or its derivative, and achieves boosting compliant convergence using only the classical assumptions of boosting, **(R, WLA)**. The proof of Lemma 18 being very conservative, we can expect that the smallest value of K of interest is smaller than the one we use, indicating that (B.147) should hold for substantially smaller limit values in (ii, iii).

B.1.9 Proof of Lemma 18

Define for short

$$F(u) \doteq \left(\frac{u^\alpha}{u^\alpha + (1-u)^\alpha}\right)^{\alpha^{\mathrm{c}}} - \left(\frac{(1-u)^\alpha}{u^\alpha + (1-u)^\alpha}\right)^{\alpha^{\mathrm{c}}} \tag{B.148}$$

$$G(u) \doteq 1 - \left(\frac{2 \cdot (1-u)^\alpha}{u^\alpha + (1-u)^\alpha}\right)^{\frac{1}{\alpha^{\mathrm{c}}}}, \tag{B.149}$$

that we study for $u \geq 1/2$ (the bound also holds by construction for $u < 1/2$). Define the following functions:

$$g(u) \doteq 1 - (2u)^{\frac{1}{\alpha^{\mathrm{c}}}}, \tag{B.150}$$

$$h(z) \doteq \frac{1}{1+z}, \tag{B.151}$$

$$i_\alpha(u) \doteq u^\alpha, \tag{B.152}$$

and $u_\alpha \doteq h \circ i_{-\alpha} \circ h^{-1}(u)$, $f(u) \doteq i_{\alpha^{\mathrm{c}}}(1-u) - i_{\alpha^{\mathrm{c}}}(u)$. We remark that g is convex if $\alpha \geq 1$ while f is concave. Both derivatives match in $1/2$ if

$$(\alpha^{\mathrm{c}})^2 2^{1-\alpha^{\mathrm{c}}} = 1, \tag{B.153}$$

whose roots are $\alpha^{\mathrm{c}} < 6$. It means if $\alpha \geq 6/5 = 1.2$, then $(g-f)' \geq 0$, and so if we measure

$$k^* \doteq \arg\sup_k \sup_{x,x' : g(x) = f(x') = k} |x - x'|,$$

then k^* is obtained for $x = 1$, for which $g(x) = 1 - 2^{\frac{1}{\alpha^{\mathrm{c}}}} = k^*$. We then need to lowerbound x' such that $f(x') = 1 - 2^{\frac{1}{\alpha^{\mathrm{c}}}}$, which amounts to finding x^* such that $f(x^*) \geq 1 - 2^{\frac{1}{\alpha^{\mathrm{c}}}}$, since f is strictly decreasing. Fix

$$x^* \doteq 1 - \frac{K}{\alpha}, \tag{B.154}$$

A series expansion reveals that for $x = x^*$ and $K = \log 2$,

$$f(x^*) \;=\; g(x^*) + O\left(\frac{1}{\alpha^2}\right), \tag{B.155}$$

and we thus get that there exists $K \geq \log 2$ such that

$$\sup_k \sup_{x,x':g(x)=f(x')=k} |x - x'| \;\leq\; \frac{K}{\alpha}, \tag{B.156}$$

or similarly for any ordinate value, the difference between the abscissae giving the value for f and g are distant by at most K/α. The exact value of the constant is not so much important than the dependence in $1/\alpha$: we now plug this in the u_αs notation and ask the following question: suppose $f(u_\alpha) = g(v_\alpha) = k$. Since $|u_\alpha - v_\alpha| \leq K/\alpha$, what is the maximum difference $|u - v|$ as a function of α ? We observe

$$\frac{\partial}{\partial u} u_\alpha \;=\; -\frac{\alpha(u(1-u))^{\alpha-1}}{(u^\alpha + (1-u)^\alpha)^2}, \tag{B.157}$$

$$\frac{\partial^2}{\partial u^2} u_\alpha \;=\; \alpha \cdot \frac{(u(1-u))^{\alpha-2}((\alpha - 2u + 1)u^\alpha - (\alpha + 2u - 1)(1-u)^\alpha)}{(u^\alpha + (1-u)^\alpha)^3}, \tag{B.158}$$

which shows the convexity of u_α as long as

$$\left(\frac{u}{1-u}\right)^\alpha \;\geq\; \frac{\alpha + 2u - 1}{\alpha - 2u + 1}, \tag{B.159}$$

a sufficient condition for which – given the RHS increases with u – is

$$u \;\geq\; \frac{\left(\frac{4}{\alpha-1}\right)^{\frac{1}{\alpha}}}{1 + \left(\frac{4}{\alpha-1}\right)^{\frac{1}{\alpha}}}. \tag{B.160}$$

Since $u \geq 1/2$, we note the constraint quickly vanishes. In particular, if $\alpha \geq 5$, the RHS is $\leq 1/2$, so u_α is strictly convex. Otherwise, scrutinising the maximal values of the derivative for $\alpha \geq 1$ reveals that if we suppose $v \leq \delta$ for some δ, then $|u - v|$ is maximal for $v = \delta$. So, suppose $v_\alpha = \epsilon$ and we solve for $u_\alpha = K/\alpha + \varepsilon$, which yields

$$u \;=\; \frac{\left(1 - \frac{K}{\alpha} - \varepsilon\right)^{\frac{1}{\alpha}}}{\left(\frac{K}{\alpha} + \varepsilon\right)^{\frac{1}{\alpha}} + \left(1 - \frac{K}{\alpha} - \varepsilon\right)^{\frac{1}{\alpha}}} \tag{B.161}$$

$$\;=\; \frac{((1-\varepsilon)\alpha - K)^{\frac{1}{\alpha}}}{(K + \varepsilon\alpha)^{\frac{1}{\alpha}} + ((1-\varepsilon)\alpha - K)^{\frac{1}{\alpha}}}, \tag{B.162}$$

while the v producing the largest $|u - v|$ is:

$$v \;=\; \frac{(1-\varepsilon)^{\frac{1}{\alpha}}}{\varepsilon^{\frac{1}{\alpha}} + (1-\varepsilon)^{\frac{1}{\alpha}}}, \tag{B.163}$$

so

$$\begin{aligned} |v-u| &= (v-u)(\varepsilon) && \text{(B.164)} \\ &= \frac{(1-\varepsilon)^{\frac{1}{\alpha}}}{\varepsilon^{\frac{1}{\alpha}} + (1-\varepsilon)^{\frac{1}{\alpha}}} - \frac{((1-\varepsilon)\alpha - K)^{\frac{1}{\alpha}}}{(K+\varepsilon\alpha)^{\frac{1}{\alpha}} + ((1-\varepsilon)\alpha - K)^{\frac{1}{\alpha}}}. && \text{(B.165)} \end{aligned}$$

If we fix

$$\varepsilon \quad = \quad \frac{1}{\alpha^4}, \tag{B.166}$$

then we get after separate series are computed in $\alpha \to +\infty$,

$$\begin{aligned} |v-u| = (v-u)(\varepsilon) &= \frac{\log(1+\log K)}{4\alpha} + O\left(\frac{1}{\alpha^2}\right) && \text{(B.167)} \\ &\lesssim \frac{0.133}{\alpha}. && \text{(B.168)} \end{aligned}$$

The "forbidden interval" for v is then

$$\left[\frac{(\alpha^4-1)^{\frac{1}{\alpha}}}{1+(\alpha^4-1)^{\frac{1}{\alpha}}}, 1\right] \quad \approx \quad \left[\frac{1}{2} + \frac{\log\alpha}{\alpha}, 1\right]; \tag{B.169}$$

what is more interesting for us is the corresponding forbidden images for $g(v_\alpha)$, which are thus

$$g \quad \not\in \quad \alpha^{\mathrm{c}} \cdot \left[1 - \frac{1}{\alpha^4}, 1\right] \doteq \mathbb{I}_\alpha, \tag{B.170}$$

where we use shorthand $z \cdot [a, b] \doteq [az, bz]$. This, we note, translates directly into observable edges since g is the function we invert. Summarizing, we have shown that if (i) $\alpha \geq 1.2$ then for any u, v such that $F(u) = G(v) \not\in \mathbb{I}_\alpha$, then $|u - v| \lesssim 0.133/\alpha$. It suffices to remark that $\mathbb{I}_\alpha$ represents the set of forbidden weights to get the statement of the Lemma.

B.1.10 Proof of Theorem 9

Remark 6. *Consider the following setup: $h_. \in [-1, 1]$, use the logistic loss surrogate for F and the derivative of Matusita's loss (just for illustration as the pseudo-inverse-link approximation of the logistic loss) for the weights (see e.g., Nock and Nielsen (2008) Table 1), we get in the worst case that* $\Delta_j(F) \leq 2{\cdot}\sup\left|\frac{1}{2} - \frac{x}{2\sqrt{1+x^2}} - \frac{1}{1+\exp(x)}\right| <$ 0.24707. *So, all we need is $|\mathsf{e}_j| > 0.24707$ to get $\zeta < 1$ constant. Since $|\mathsf{e}_j| \in [0, 1]$, this constraint is more than reasonable and turns into a very reasonable penalty in $Q(F)$ (Theorem 9).*

Remark 7. PilBoost *and its convergence analysis bring a side contribution of ours: it is impossible to exactly encode in standard machine types the inverse link of* *losses like the log loss, so the implementation of classical boosting algorithms (Friedman, 2001; Friedman* et al., *2000) can only rely on approximations of the inverse link function. Our results yield convergence guarantees for the implementations of such algorithms, and* ***(E)*** *can be interpreted and checked in the context of machine encoding. Two additional remarks hold regarding convergence rate: first, the* $1/\gamma^2$ *dependence meets the general optimum for boosting (Alon* et al.*, 2021); second, the* $1/\varepsilon^2$ *dependence parallels classical training convergence of convex optimization (Thekumparampil* et al.*, 2020) (and references therein). There is however a major difference with such work:* PilBoost *requires no function oracles for* F *(function values, (sub)gradients, etc.). This "sideways" fork to minimizing* F *pays (only) a* $1/(1-\zeta)^2$ *factor in convergence.*

We proceed in two steps, assuming **(WLA)** holds for WL and **(R)** holds for the weak classifiers.

In Step 1, we show that for any loss defined by F twice differentiable, convex and non-increasing, for any $z^* \in \mathbb{R}$, as long as F satisfies assumptions **(E)** and **(C)** for T iterations such that

$$\sum_{t=0}^{T} \tilde{w}_t^2 \quad \geq \quad \frac{2F^*(F(0)-F(z^*))}{\gamma^2(1-\zeta)^2(1-\pi^2)}, \tag{B.171}$$

we have the guarantee on the risk defined by F:

$$\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_T(\boldsymbol{x}_i))\right] \quad \leq \quad F(z^*). \tag{B.172}$$

Let F be any twice differentiable, convex and non-increasing function. We wish to find a lowerbound $\triangle$ on the decrease of the expected loss computed using F:

$$\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_t(\boldsymbol{x}_i))\right] - \mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_{t+1}(\boldsymbol{x}_i))\right] \quad \geq \quad \triangle, \tag{B.173}$$

where with a slight abuse of notation we let H_t denote the learned real-valued classifier at iteration t. We make use of a similar proof technique as in Nock and Williamson (2019). Suppose

$$H_{t+1} \quad = \quad H_t + \beta_j \cdot h_j, \tag{B.174}$$

index j being returned by WL at iteration t. For any such index j, any $g : \mathbb{R} \to \mathbb{R}_+$ and any $H \in \mathbb{R}^{\mathcal{X}}$, let

$$\mathsf{e}(j, g, H) \quad \doteq \quad \mathbb{E}_{i\sim\mathcal{S}}\left[y_i h_j(\boldsymbol{x}_i) \cdot g(y_i H(\boldsymbol{x}_i))\right], \tag{B.175}$$

denote the expected edge of h_j on weights defined by the couple (g, H). Furthermore, let

$$\Delta(g_1, g_2) \quad \doteq \quad \left|\mathsf{e}(j, g_1, H_t) - \mathsf{e}(j, g_2, H_t)\right|, \tag{B.176}$$

denote the discrepancy between two expected edges defined by g_1, g_2, respectively.

There are two quantities we consider. First, let

$$\begin{align} X &\doteq \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i))F'(y_i H_t(\boldsymbol{x}_i))\right] \tag{B.177}\\ &= \beta_j \cdot \mathbb{E}_{i\sim\mathcal{S}}\left[y_i h_j(\boldsymbol{x}_i) \cdot -F'(y_i H_t(\boldsymbol{x}_i))\right] \tag{B.178}\\ &= \beta_j \cdot \mathsf{e}(j, -F', H_t) \tag{B.179}\\ &\geq \beta_j \cdot \mathbb{E}_{i\sim\mathcal{S}}\left[y_i h_j(\boldsymbol{x}_i) \cdot \widetilde{f}(-y_i H_t(\boldsymbol{x}_i))\right] - \beta_j \cdot \Delta(-F', \widetilde{f}_s) \tag{B.180}\\ &= a_f \mathsf{e}^2(j, \widetilde{f}_s, H_t) - a_f \mathsf{e}(j, \widetilde{f}_s, H_t) \cdot \Delta(-F', \widetilde{f}_s) \tag{B.181}\\ &\geq a_f(1-\zeta)\mathsf{e}^2(j, \widetilde{f}_s, H_t), \tag{B.182} \end{align}$$

where $\widetilde{f}_s(z) \doteq \widetilde{f}_s(-z)$ and finally (B.182) makes use of assumption **(E)**. The second quantity we define is:

$$Y(\mathcal{Z}) \doteq \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i))^2 F''(z_i)\right], \tag{B.183}$$

where $\mathcal{Z} \doteq \{z_1, z_2, ..., z_m\} \subset \mathbb{R}^m$. Using assumption **(R)** and letting F^* be any real such that $F^* \geq \sup F''(z)$, we obtain:

$$\begin{align} Y(\mathcal{Z}) &\leq F^* \cdot \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i))^2\right] \notag\\ &= F^* \cdot \beta_j^2 \cdot \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i h_j(\boldsymbol{x}_i))^2\right] \notag\\ &\leq F^* \cdot \beta_j^2 \cdot M^2 \notag\\ &= F^* a^2 M^2 \cdot \mathsf{e}^2(j, \widetilde{f}_s, H_t). \tag{B.184} \end{align}$$

A second order Taylor expansion on F brings that there exists $\mathcal{Z} \doteq \{z_1, z_2, ..., z_m\} \subset \mathbb{R}^m$ such that:

$$\begin{align} \mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_t(\boldsymbol{x}_i))\right] &= \mathbb{E}_{i\sim\mathcal{S}}[F(y_i H_{t+1}(\boldsymbol{x}_i))] + \mathbb{E}_{i\sim\mathcal{S}}[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i))F'(y_i H_t(\boldsymbol{x}_i))] \notag\\ &\quad + \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i))^2 \cdot \frac{F''(z_i)}{2}\right], \tag{B.185} \end{align}$$

So,

$$\begin{align} &\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_t(\boldsymbol{x}_i))\right] - \mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_{t+1}(\boldsymbol{x}_i))\right] = X - \frac{Y(\mathcal{Z})}{2} \notag\\ &\geq \underbrace{\left(1 - \zeta - \frac{F^* a M^2}{2}\right) a}_{\doteq Z(a)} \cdot \mathsf{e}^2(j, \widetilde{f}_s, H_t). \tag{B.186} \end{align}$$

Suppose we fix $\pi \in [0, 1]$ and choose any

$$a \in \frac{1-\zeta}{F^* M^2} \cdot [1-\pi, 1+\pi]. \tag{B.187}$$

We can check that

$$Z(a) \geq \frac{(1-\zeta)^2(1-\pi^2)}{2F^* M^2}, \tag{B.188}$$

and so

$$\mathbb{E}_{i\sim\mathcal{S}}[F(y_iH_t(\boldsymbol{x}_i))]-\mathbb{E}_{i\sim\mathcal{S}}[F(y_iH_{t+1}(\boldsymbol{x}_i))] \quad \geq \quad \frac{(1-\zeta)^2(1-\pi^2)}{2F^*M^2}\cdot \mathsf{e}^2(j,\widetilde{f}_s,H_t). \quad \text{(B.189)}$$

So, taking into account that for the first classifier, we have $\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_iH_0(\boldsymbol{x}_i))\right] = F(0)$, if we take any $z^*\in\mathbb{R}$ and we boost for a number of iterations T satisfying (we use notation e_t as a summary for $\mathsf{e}^2(j,\widetilde{f}_s,H_t)$ with respect to PilBoost):

$$\sum_{t=1}^{T}\mathsf{e}_t^2 \quad \geq \quad \frac{2F^*M^2(F(0)-F(z^*))}{(1-\zeta)^2(1-\pi^2)}, \quad \text{(B.190)}$$

then $\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_iH_T(\boldsymbol{x}_i))\right]\leq F(z^*)$. We now assume **(WLA)** holds, the LHS of (B.190) is $\geq T\gamma^2$. Given that we choose $a=a_f$ in PilBoost, we need to make sure (B.187) is satisfied for the loss defined by F, which translates to

$$F^* \quad \in \quad \frac{1-\zeta}{a_fM^2}\cdot\left[1-\pi,1+\pi\right], \quad \text{(B.191)}$$

and defines assumption **(C)**. To complete Step 1, we normalize the edge. Letting $\tilde{w}_i \doteq w_i/\sum_k w_k$, $\tilde{w}_t\doteq \mathbf{1}^\top\boldsymbol{w}_t/m$ and

$$\tilde{\mathsf{e}}_t \quad \doteq \quad \frac{\mathsf{e}_t}{\tilde{w}_t}\in[-M,M], \quad \text{(B.192)}$$

which is then properly normalized and such that (B.190) becomes equivalently:

$$\sum_{t=0}^{T}\tilde{w}_t^2\tilde{\mathsf{e}}_t^2 \quad \geq \quad \frac{2F^*M^2(F(0)-F(z^*))}{(1-\zeta)^2(1-\pi^2)}, \quad \text{(B.193)}$$

and so under the (weak learning) assumption on $\tilde{\mathsf{e}}_t$ that $|\tilde{\mathsf{e}}_t|\geq\gamma\cdot M$, a sufficient condition for (B.193) is then

$$\sum_{t=0}^{T}\tilde{w}_t^2 \quad \geq \quad \frac{2F^*(F(0)-F(z^*))}{\gamma^2(1-\zeta)^2(1-\pi^2)}, \quad \text{(B.194)}$$

completing step 1 of the proof.

In Step 2, we show a result on the distribution of edges, *i.e.* margins. (B.194) contains all the intuition about how the rest of the proof unfolds, as we have two major steps: in step 2.1, we translate the guarantee of (B.194) on margins, and in step 2.2, we translate the "margin" based (B.194) in a readable guarantee in the boosting framework (we somehow "get rid" of the $\tilde{w}_t^2$ in the LHS of (B.194)).

Step 2.1. Let $\mathcal{Z}\doteq\{z_1,z_2,...,z_m\}\subset\mathbb{R}$ a set of reals. Since F is non-increasing, we have $\forall u\in[0,1],\forall\theta\geq 0$,

$$\begin{aligned}\mathbb{P}_i[z_i\leq\theta]>u\Rightarrow\mathbb{E}_i[F(z_i)] \quad &> \quad (1-u)\inf_z F(z)+uF(\theta)\\ &\doteq (1-u)F^\circ+uF(\theta),\end{aligned} \quad \text{(B.195)}$$

so if we pick z^* in (B.194) such that

$$F(z^*) \doteq (1-u)F^\circ + uF(\theta), \tag{B.196}$$

then (B.194) implies $\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_T(\boldsymbol{x}_i))\right] \leq (1-u)F^\circ + uF(\theta)$ and so by the contraposition of (B.195) yields:

$$\mathbb{P}_{i\sim\mathcal{S}}\left[y_i H_T(\boldsymbol{x}_i) \leq \theta\right] \leq u, \tag{B.197}$$

which yields our margin based guarantee.

Step 2.2. At this point, the key (in)equalities are (B.194) (for boosting) and (B.197) (for margins). Fix $\kappa > 0$. We have two cases:

- Case 1: $\tilde{w}_t$ never gets too small, say $\tilde{w}_t \geq \kappa, \forall t \geq 0$. In this case, granted the weak learning assumption holds on $\tilde{\mathsf{e}}_t$, (B.194) yields a direct lowerbound on iteration number T to get $\mathbb{P}_{i\sim\mathcal{S}}\left[y_i H_T(\boldsymbol{x}_i) \leq \theta\right] \leq u$;
- Case 2: $\tilde{w}_t \leq \kappa$ at some iteration t. Since the smaller it is, the better classified are the examples, if we pick κ small enough, then we can get $\mathbb{P}_{i\sim\mathcal{S}}\left[y_i H_T(\boldsymbol{x}_i) \leq \theta\right] \leq u$ "straight".

This suggests our use of the notion of "denseness" for weights (Bun *et al.*, 2020).

Definition 9. *The weights at iteration t is called κ-dense iff $\tilde{w}_t \geq \kappa$.*

We now have the following Lemma.

Lemma 19. *For any $t \geq 0, \theta \in \mathbb{R}, \kappa > 0$, if weights produced in Step 2.1 of* PilBoost *fail to be κ-dense, then*

$$\mathbb{P}_{i\sim\mathcal{S}}\left[y_i H_T(\boldsymbol{x}_i) \leq \theta\right] \leq \frac{\kappa}{\widetilde{f}(-\theta)}. \tag{B.198}$$

Proof. Let $\mathcal{Z} \doteq \{z_1, z_2, ..., z_m\} \subset \mathbb{R}$ a set of reals. Since $\widetilde{f}$ is non-decreasing, we have $\forall \theta \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}_i[\widetilde{f}(z_i)] &\geq \mathbb{P}_i[z_i < -\theta] \cdot \inf_z \widetilde{f}(z) + \mathbb{P}_i[z_i \geq -\theta] \cdot \widetilde{f}(-\theta) \\ &= \mathbb{P}_i[z_i \geq -\theta] \cdot \widetilde{f}(-\theta) \end{aligned} \tag{B.199}$$

since by assumption $\inf \widetilde{f} = 0$. Pick $z_i \doteq -y_i H_T(\boldsymbol{x}_i)$. We get that if $\mathbb{P}_{i\sim\mathcal{S}}[-y_i H_T(\boldsymbol{x}_i) \geq -\theta] = \mathbb{P}_{i\sim\mathcal{S}}[y_i H_T(\boldsymbol{x}_i) \leq \theta] \geq \xi$, then $\tilde{w}_t \doteq \mathbb{E}_{i\sim\mathcal{S}}[\widetilde{f}(-y_i H_T(\boldsymbol{x}_i))] \geq \xi \cdot \widetilde{f}(-\theta)$. If we fix

$$\xi = \frac{\kappa}{\widetilde{f}(-\theta)}, \tag{B.200}$$

then $\tilde{w}_t < \kappa$ implies (B.198), which ends the proof of Lemma 19. □

From Lemma 19, we let $\kappa \doteq \xi_* \cdot \widetilde{f}(-\theta)$ and $u \doteq \xi_*$ in (B.197). If at any iteration, H_T fails to be κ-dense, then $\mathbb{P}_{i\sim\mathcal{S}}\left[y_i H_{\boldsymbol{\beta}}(\boldsymbol{x}_i) \leq \theta\right] \leq \xi_*$ and classifier $H_{\boldsymbol{\beta}}$ satisfies the conditions of Theorem 9 (this is our Case 2 above).

Otherwise, suppose it is always κ-dense (this is our Case 1 above). We then have at any iteration T $\sum_{t<T} \tilde{w}_t^2 \geq T\xi_*^2 \cdot \widetilde{f}^2(-\theta)$ and so a sufficient condition to get (B.194) is then $T \geq \frac{2F^*(F(0)-F(z^*))}{\xi_*^2 \widetilde{f}^2(-\theta)\gamma^2(1-\zeta)^2(1-\pi^2)}$, where we recall z^* is chosen so that $F(z^*) = (1-\xi_*)F^\circ + \xi_* F(\theta)$. This ends the proof of Theorem 9 (with the change of notation $\xi_* \leftrightarrow \varepsilon$).

B.2 Additional Experimental Results

In this section, we provide additional experimental results and discussion to accompany Section 4.5 in the main text. The code for all of our experiments (including the implementation of PilBoost) can be found at the following github repository link:

```
https://github.com/SankarLab/Being-Properly-Improper
```

B.2.1 General Details

Most of the experiments were performed over the course of a month on a 2015 MacBook Pro with a 2.2 GHz Quad-Core Intel Core *i*7 processor and 16GB of memory. The Adaptive α experiments were performed on a computing cluster and each required about 30 minutes of compute time. Code can be found in *PILBoostExperiments.py*, *AdaptiveAlphaMenon.py*, *AdaptiveAlphaALL.py*. Averaged experiments employed 10-fold cross validation, and when twisters were present, randomization occurred over the twisted samples as well. All algorithms across all experiments ran for 1000 iterations.

B.2.2 Discussion of a_f and α

In general, we found that for most experiments, $0.1 \leq a_f \leq 15$. From the theory, we know that if a_f is too small, boosting needs to occur for a very long time, and if a_f is too large, almost no loss fits to (**C**) (equivalently, (**C**) fails for us). We also generally found that PilBoost was not particularly sensitive to the choice of a_f as long as it was in the "right ballpark", hence our use of integer or rational values of a_f for all experiments. When there is twist present, we found that $\alpha > 1$ performed best, where α^* increased as the amount of twist increased (both observations are conistent with our theory, see for example Lemma 3.4). Regarding the relationship between a_f and α, this appeared to depend on the dataset and depth of the decision trees.

B.2.3 Random Class Noise Twister

Dataset	Algorithm	Random Class Noise Twister		
		$p = 0$	0.15	0.3
cancer	AdaBoost	0.97 ± 0.02	0.91 ± 0.03	0.86 ± 0.03
	us ($\alpha = 1.1$)	0.94 ± 0.03	0.91 ± 0.01	0.86 ± 0.04
	us ($\alpha = 2.0$)	0.96 ± 0.02	0.94 ± 0.02	0.91 ± 0.04
	us ($\alpha = 4.0$)	0.96 ± 0.01	0.92 ± 0.01	0.92 ± 0.03
	XGBoost	0.97 ± 0.01	0.86 ± 0.03	0.73 ± 0.03

Table B.1: Cancer Feature Random Class Noise. Accuracies Reported for Each Algorithm and Level of Twister. Depth One Trees. For $\alpha = 1.1$, $a_f = 7$, for $\alpha = 2$, $a_f = 2$, and for $\alpha = 4$, $a_f = 1$.

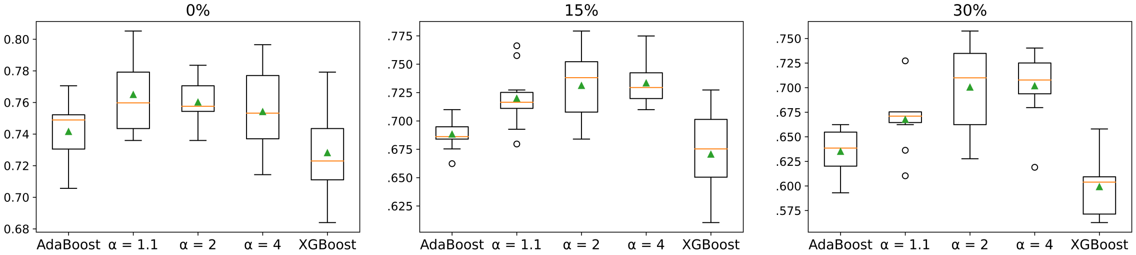


Figure B.9: Random Class Noise Twister on the Diabetes Dataset. Depth 3 Trees. $a_f = .1$ for All α.

Dataset	Algorithm	Random Class Noise Twister		
		$p = 0$	0.15	0.3
xd6	AdaBoost	1.000 ± 0.000	0.949 ± 0.016	0.830 ± 0.043
	us ($\alpha = 1.1$)	1.000 ± 0.000	0.981 ± 0.013	0.886 ± 0.033
	us ($\alpha = 2.0$)	1.000 ± 0.000	0.992 ± 0.009	0.900 ± 0.027
	us ($\alpha = 4.0$)	1.000 ± 0.000	0.999 ± 0.003	0.927 ± 0.023
	XGBoost	1.000 ± 0.000	0.912 ± 0.016	0.776 ± 0.041

Table B.2: Xd6 Random Class Noise. Accuracies Reported for Each Algorithm and Level of Twister. Depth Three Trees. $a_f = 8$ for All α. Note That for 0% Noise $\alpha = 4$ Used $a_f = .1$.

Dataset	Algorithm	Random Class Noise Twister			
		$p = 0$	0.10	0.20	0.30
Online Shopping	AdaBoost	0.90 ± 0.00	0.90 ± 0.00	0.90 ± 0.01	0.89 ± 0.00
	us ($\alpha = 1.1$)	0.90 ± 0.01	0.90 ± 0.00	0.90 ± 0.00	0.89 ± 0.00
	us ($\alpha = 2.0$)	0.90 ± 0.00	0.90 ± 0.00	0.90 ± 0.00	0.89 ± 0.00
	us ($\alpha = 4.0$)	0.90 ± 0.00	0.87 ± 0.01	0.89 ± 0.01	0.89 ± 0.01
	XGBoost	0.89 ± 0.01	0.87 ± 0.00	0.84 ± 0.01	0.78 ± 0.01

Table B.3: Accuracies Reported for Each Algorithm and Level of Twister. Random Training Sample Selected with Probability p. Then, for Selected Training Sample, Boolean Feature Flipped with Probability p for Each Feature, Independently. Depth Three Trees. For $\alpha = 1.1$, $a_f = 7$, for $\alpha = 2$, $a_f = 8$, and for $\alpha = 4$, $a_f = 15$.

B.2.4 Insider Twister

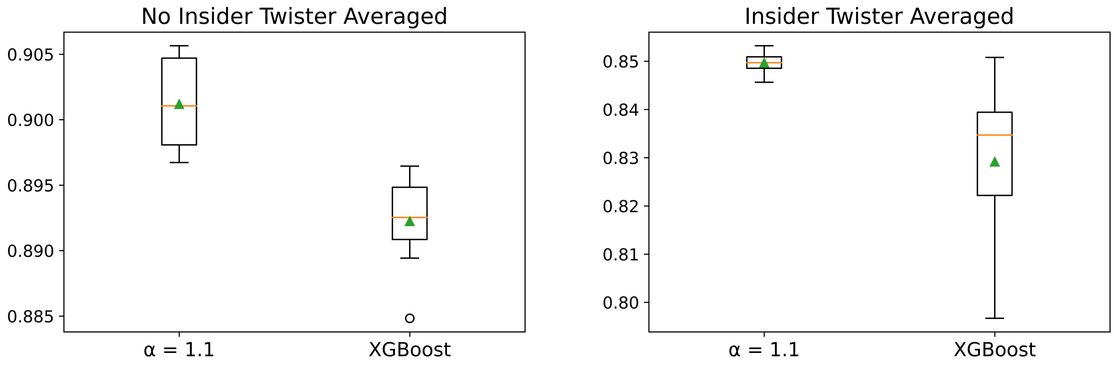


Figure B.10: Box and Whisker Visualization of Scores Associated with Figure 3.3. For All Insider Twister Results, We Fixed $a_f = 7$. Under No Twister, $\alpha = 1.1$, Has Accuracy $0.901 \pm .003$, and Xgboost Has Accuracy $0.892 \pm .003$. Under the Insider Twister, $\alpha = 1.1$, Has Accuracy $0.850 \pm .002$, and Xgboost Has Accuracy $0.829 \pm .016$; Under the Welch t-test, the Results Have a p-value of 0.004.

B.2.5 Discussion of XGBoost

Algorithm	Average Compute Times			
	cancer	xd6	diabetes	shoppers
AdaBoost	1.41	0.75	1.11	13.68
us ($\alpha = 1.1$)	2.19	2.01	2.19	30.88
us ($\alpha = 2.0$)	1.11	0.79	2.09	21.85
us ($\alpha = 4.0$)	0.96	1.35	1.82	13.01
XGBoost	0.29	0.28	0.46	3.16

Table B.4: Average Compute times per Run (10 Runs) in Seconds Across the Datasets. Note That the Values of a_F Are Chosen Identically to Choices in Section B.2.3.

XGBoost is a very fast, very well engineered boosting algorithm. It employs many different hyperparameters and customizations. In order to report the fairest comparison between AdaBoost, PilBoost , and XGBoost, we opted to keep as many hyperparameters fixed (and similar, e.g., depth of decision trees) as possible. That being said, it appears that XGBoost inherently uses pruning, so the algorithm pruned while the other two did not. Further details regarding three other important points related to XGBoost:

1. Please refer to Table B.4 for averaged compute times for the three different algorithms. In general, XGBoost had the far faster computation time among the three. However, note that PilBoost was not particularly engineered for speed. Indeed, we estimate that the computation of $\tilde{f}$ accounts for $40 - 50\%$ of the total computation time, which we believe can be improved. Thus, we leave the further computational optimization of PilBoost for future work.

2. For details regarding regularization, refer to Figure B.11, where we report a comparison of regularized XGBoost and PilBoost such that the training data suffers from the insider twister. We find that regularization improves the ability of XGBoost to combat the twister, but it is not as effective as PilBoost.

3. For details regarding early stopping, refer to Figure B.13, where we report a comparison of early-stopped XGBoost (on un-twisted validation data, i.e., cheating) and PilBoost such that the training data suffers from the insider twister. We find that even early-stopping does not improve XGBoost's ability to combat the insider twister as effectively as PilBoost.

Early stopping - on an untwisted hold-out set contradicts our experiment. With early stopping enabled on a twisted hold-out set, XGBoost generally did not early stop.

B.2.6 Adaptive α Experiment

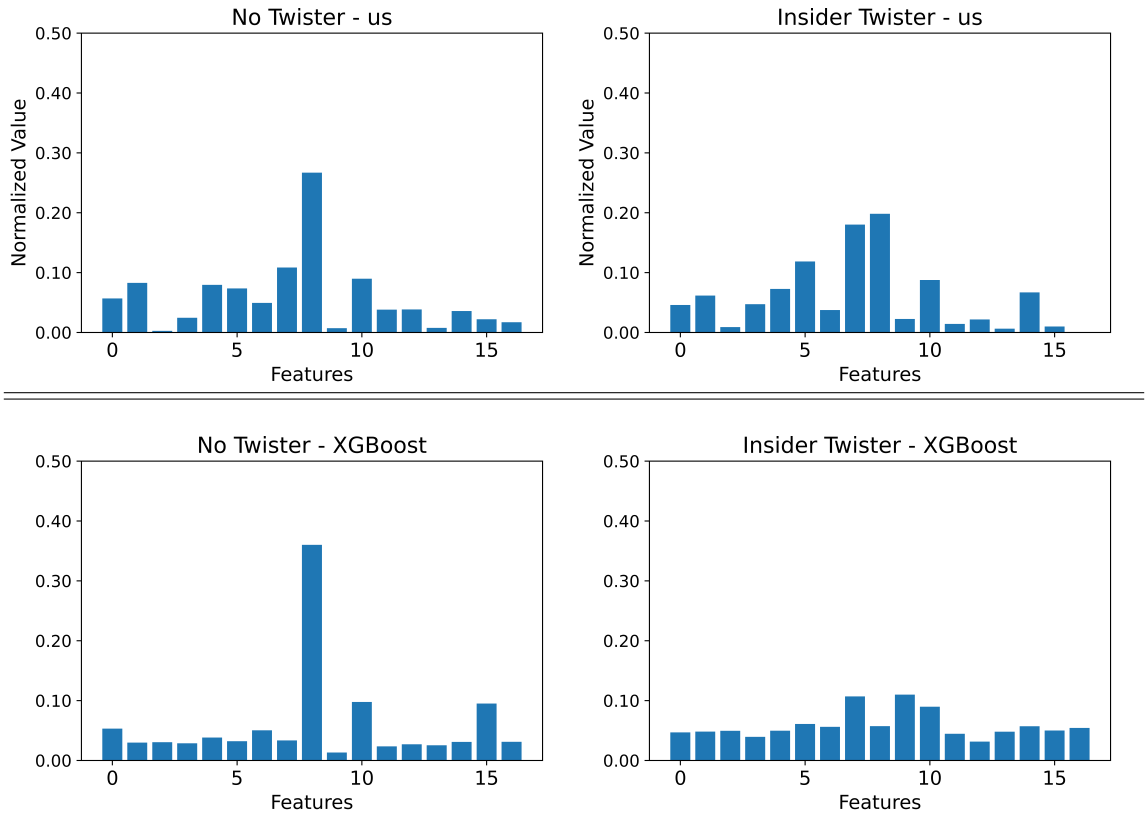


Figure B.11: With Regularization (Where $\lambda = 20$), We Still Observe That the Feature Importance of Xgboost Is Perturbed. Note That PILBOOST is Not Regularized.

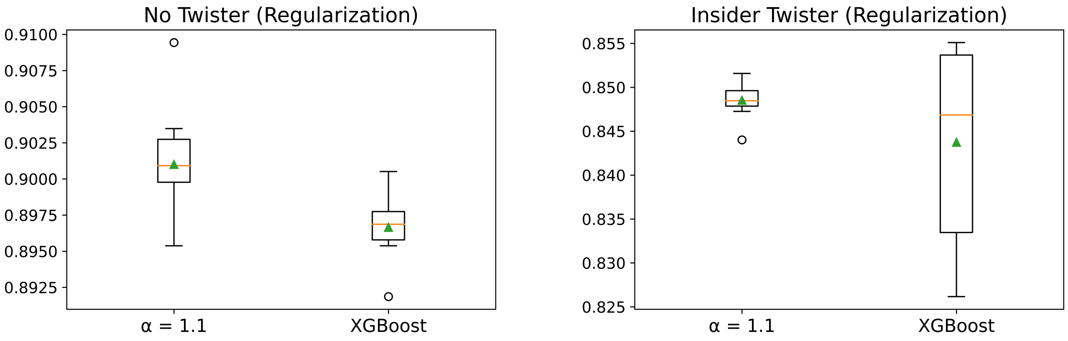


Figure B.12: Scores Associated with Figure B.11.

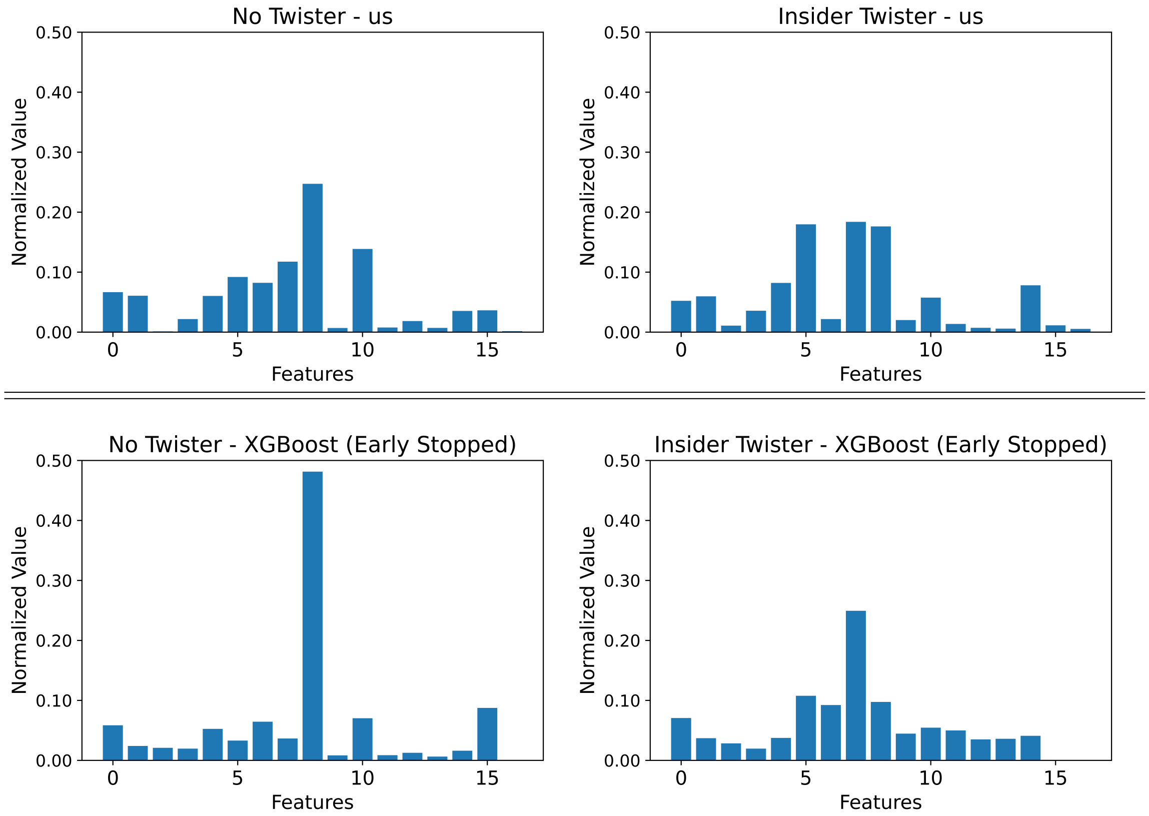


Figure B.13: With Early Stopping (Where Xgboost Has Access to Clean Validation Data - Cheating Scenario), We Still Observe That the Feature Importance of Xgboost Is Perturbed. Note That PilBoost is Not Early Stopped.

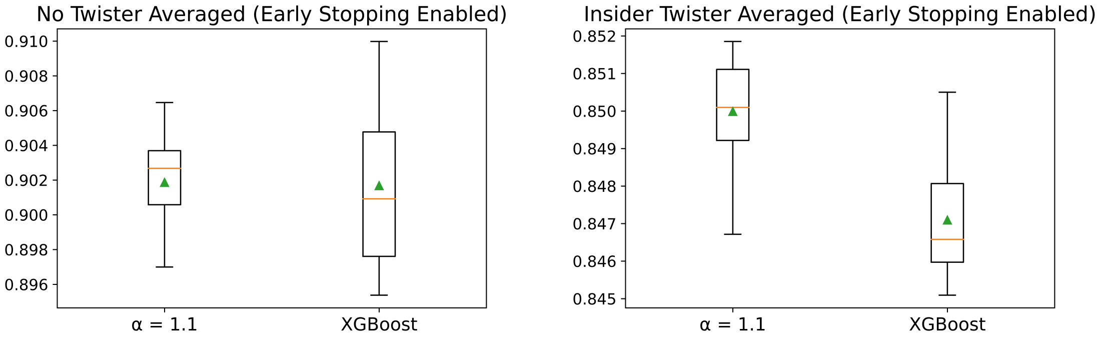


Figure B.14: Scores Associated with Figure B.13.

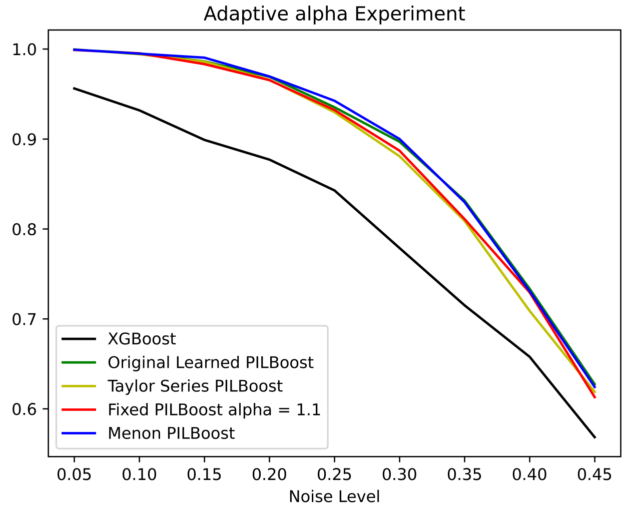


Figure B.15: Extended Version of Figure 3.2 with Two Additional Adaptive α Methods. Original Learned PilBoost estimates Its Choice of α by Using Xgboost as an Oracle. That Is, the Method Trains Xgboost on the Noisy Data, Then Computes Its Confusion Matrix on a *Clean* Validation Set. From the Confusion Matrix, the Label Flip Probability p Is Estimated Using $p = \text{Avg}\left(\frac{\text{Fp}}{\text{Tp+Fp}}, \frac{\text{Fn}}{\text{Fn+Tn}}\right)$. Next, We Estimate η_{c} and η_{t} with $\eta_{\text{c}} = \frac{\text{Fn+Tp}}{\text{Fp+Tp+Fn+Tn}}$ and $\eta_{\text{t}} = \frac{\text{Fp+Tp}}{\text{Fp+Tp+Fn+Tn}}$, Respectively. Lastly Similar to Menon PilBoost, Using the Estimates of p, η_{c}, and η_{t}, We Apply the Formula in Lemma 9(c) and the SLN Formula given Just Before Lemma 7 to Obtain an Estimate for α^*. Taylor Series PilBoost is Identical to Original Learned PilBoost except at the Last Step, Where a Taylor Series Approximation of the Formula in Lemma 9(c) Is Used Instead. We Find That Menon's Method Also Outperforms Both of These Methods on the Xd6 Dataset, Except for When Original Learned PilBoost slightly Outperforms Menon's Method in the Very High Noise Regime. Even Stronger, Note That Both Original Learned PilBoost and Taylor Series PilBoost assume More Information than "Menon's Method", Which Only Uses the Noisy Training Data.

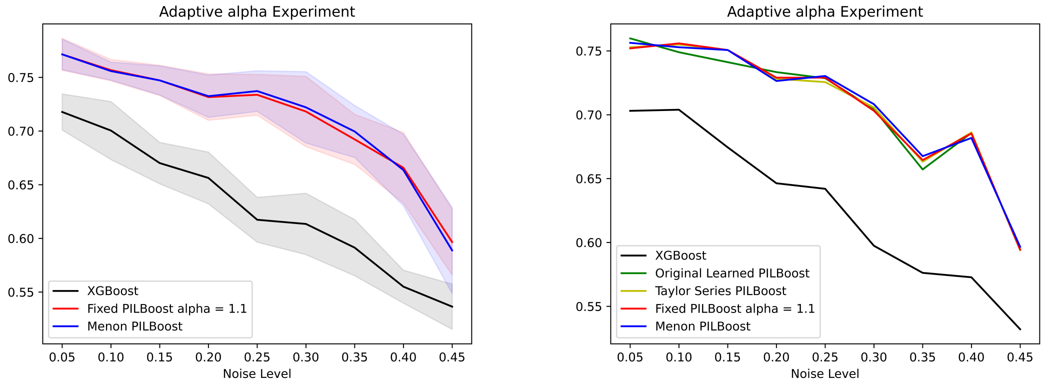


Figure B.16: Companion Figure to Figures 3.2 And B.15 on the *Diabetes* Dataset.

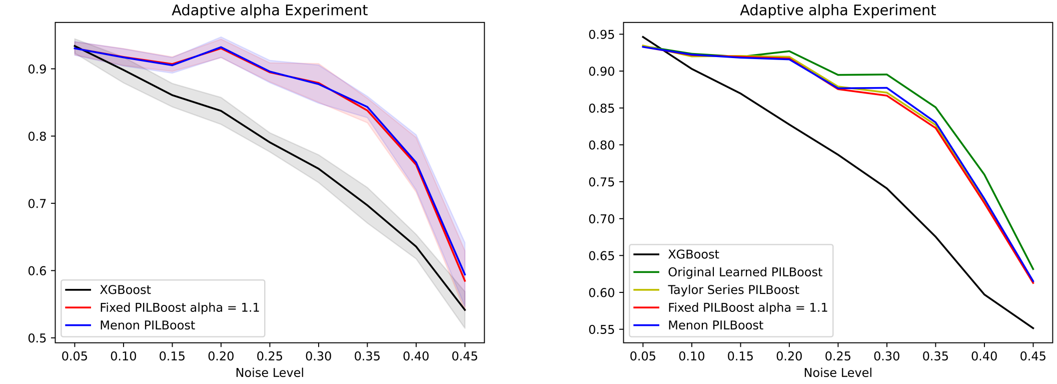


Figure B.17: Companion Figure to Figures 3.2 And B.15 on the *Cancer* Dataset.

APPENDIX C

APPENDIX TO CHAPTER 4

Lemma 20. *For $\alpha \in (0, \infty]$, the first derivative of $\tilde{l}^{\alpha}$ with respect to the margin is given by*

$$\tilde{l}^{\alpha\prime}(z) := \frac{d}{dz}\tilde{l}^{\alpha}(z) = -\sigma'(z)\sigma(z)^{-\frac{1}{\alpha}}, \tag{C.1}$$

its second derivative is given by

$$\tilde{l}^{\alpha\prime\prime}(z) := \frac{d^2}{dz^2}\tilde{l}^{\alpha}(z) = \frac{e^z\left(\alpha e^z - \alpha + 1\right)}{\alpha(e^{-z}+1)^{-\frac{1}{\alpha}}(e^z+1)^3}, \tag{C.2}$$

and its third derivative is given by

$$\tilde{l}^{\alpha\prime\prime\prime}(z) := \frac{d^3}{dz^3}\tilde{l}^{\alpha}(z) = \frac{-e^{2z} + 4e^z - 1 - \frac{3e^z-2}{\alpha} - \frac{1}{\alpha^2}}{e^{-z}\left(1+e^{-z}\right)^{-\frac{1}{\alpha}}\left(e^z+1\right)^4}. \tag{C.3}$$

Discussion of Algorithm 2 The weighting used for the weak learner in Algorithm 2, namely that $\theta_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$, is the expression commonly used in vanilla AdaBoost ($\alpha = 1/2$ for AdaBoost.α) (Schapire and Freund, 2013). However, there are several other possibilities of θ_t for AdaBoost.α, due to its interpolating characteristics. One possibility is to use $\theta_t = \alpha\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$, for $\alpha \in (0, \infty]$, which is the optimal classification function of the margin-based α-loss (Sypherd *et al.*, 2022b). Another possibility is to use a Wolfe line search (Telgarsky, 2013). Consideration of the weighting of the weak learners, and the ensuing convergence (and consistency) characteristics for Algorithm 2, is left for future work.

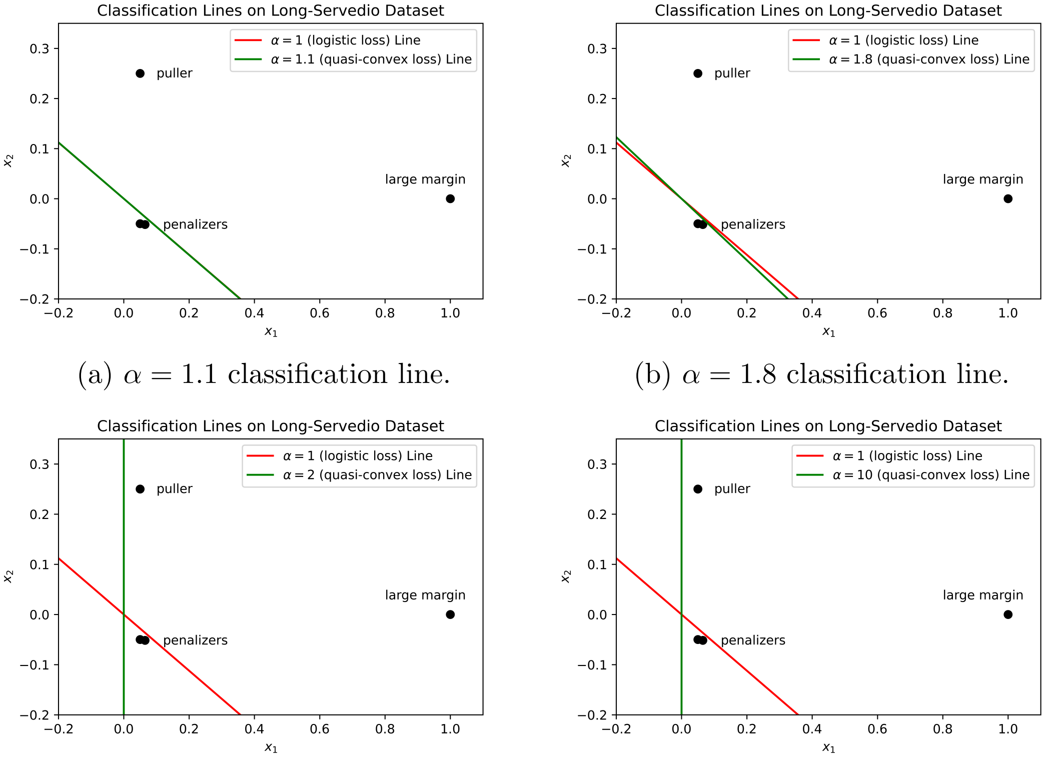
C.1.1 Proof of Theorem 10

The strategy of the proof is as follows:

1. First, we quantify what a perfect classification solution on the Long-Servedio dataset looks like, namely, inequality requirements involving θ_1 and θ_2 derived from the interaction of the "penalizers", "puller", and "large margin" examples and the linear hypothesis class.

2. Next, we invoke the pathological result of (Long and Servedio, 2010), which yields a "bad" margin γ for any noise level and the margin-based α-loss with $\alpha \leq 1$ (i.e., convex losses as articulated in Proposition 7).

3. Then, we reduce the first order equation of the margin-based α-loss evaluated at the four examples over the linear weights for $\alpha \in (1, \infty)$, and through a cancellation yield an equation which has a function of θ_1 on the LHS and a similar function of both θ_1 and θ_2 on the RHS, i.e., an asymmetric equation not allowing full analytical solution but allowing reasoning about possible solutions.



(a) $\alpha = 1.1$ classification line.

(b) $\alpha = 1.8$ classification line.

(c) $\alpha = 2$ classification line.

(d) $\alpha = 10$ classification line.

Figure C.1: Companion Figure of Figure 4.2 Where $n = 2$ ($p = 1/3$) and $\gamma = 1/20$ for $\alpha \in \{1.1, 1.8, 2, 10\}$.

4. Finally, using continuity arguments exploiting the giving up properties of the quasi-convex margin-based α-losses for $\alpha \in (1, \infty)$, we guarantee the existence of a solution (θ_1^*, θ_2^*) with perfect classification accuracy on the *clean* Long-Servedio dataset under the given pathological margin γ.

By the construction of the hypothesis class (Long and Servedio, 2010), namely that $\mathcal{H} = \{h_1(\mathbf{x}) = x_1, h_2(\mathbf{x}) = x_2\}$, notice that the classification lines (constructed by the boosting algorithm in this pathological example) are given by $\theta_1 x_1 + \theta_2 x_2 = 0$ and must pass through the origin. Rewriting this classification line, we have that $x_2 = -\frac{\theta_1}{\theta_2} x_1$. Reasoning about perfect classification weights (θ_1^*, θ_2^*), notice (see Figure 4.2) that the "large margin" example forces $\theta_1^* > 0$. Further, reasoning about the "penalizers", we find that we require $\theta_1^* > \theta_2^*$, and reasoning about the "puller", we also find that we require $\theta_1^* > -5\theta_2^*$. Thus, perfect classification weights on the Long-Servedio dataset must satisfy all of the following:

$$\theta_1^* > 0 \quad \text{and} \quad \theta_1^* > \theta_2^* \quad \text{and} \quad \theta_1^* > -5\theta_2^*. \tag{C.4}$$

We now examine the solutions to the first-order equation for $\alpha \in (0, \infty]$.

As in (Long and Servedio, 2010), let $1 < N < \infty$ be the noise parameter such that the noise rate $p = \frac{1}{N+1}$, and hence $1 - p = \frac{N}{N+1}$. Under the Long-Servedio setup with the margin-based α-loss (and recalling that all four examples have classification label $y = 1$), we have that

$$R_{\alpha}^{p}(\theta_1, \theta_2) = \frac{1}{4} \sum_{x \in S} \left[(1-p)\tilde{l}^{\alpha}(\theta_1 x_1 + \theta_2 x_2) + p\tilde{l}^{\alpha}(-\theta_1 x_1 - \theta_2 x_2) \right]. \tag{C.5}$$

It is clear that minimizing $4(N+1)R_{\alpha}^{p}$ is the same as minimizing R_{α}^{p} so we shall henceforth work with $4(N+1)R_{\alpha}^{p}$ since it gives rise to cleaner expressions. We have that

$$\begin{aligned} 4(N+1)R_{\alpha}^{p}(\theta_1, \theta_2) &= \sum_{x \in S} [N\tilde{l}^{\alpha}(\theta_1 x_1 + \theta_2 x_2) + \tilde{l}^{\alpha}(-\theta_1 x_1 - \theta_2 x_2)] & \text{(C.6)} \\ &= N\tilde{l}^{\alpha}(\theta_1) + \tilde{l}^{\alpha}(-\theta_1) + 2N\tilde{l}^{\alpha}(\theta_1\gamma - \theta_2\gamma) + 2\tilde{l}^{\alpha}(-\theta_1\gamma + \theta_2\gamma) \\ &\quad + N\tilde{l}^{\alpha}(\theta_1\gamma + 5\theta_2\gamma) + \tilde{l}^{\alpha}(-\theta_1\gamma - 5\theta_2\gamma). & \text{(C.7)} \end{aligned}$$

See Figure C.2 for a visualization of (C.7).

Again following notation in (Long and Servedio, 2010), let $P_1^{\alpha}(\theta_1, \theta_2)$ and $P_2^{\alpha}(\theta_1, \theta_2)$ be defined as follows:

$$P_1^{\alpha}(\theta_1, \theta_2) := \frac{\partial}{\partial \theta_1} 4(N+1)R_{\alpha}^{p}(\theta_1, \theta_2) \quad \text{and} \quad P_2^{\alpha}(\theta_1, \theta_2) := \frac{\partial}{\partial \theta_2} 4(N+1)R_{\alpha}^{p}(\theta_1, \theta_2). \tag{C.8}$$

Thus, differentiating (C.7) by θ_1 and θ_2 respectively, we have

$$\begin{aligned} P_1^{\alpha}(\theta_1, \theta_2) &= N\tilde{l}^{\alpha'}(\theta_1) - \tilde{l}^{\alpha'}(-\theta_1) + 2\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma - \theta_2\gamma) \\ &\quad - 2\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma + \theta_2\gamma) + N\gamma\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma) - \gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma), \end{aligned} \tag{C.9}$$

and

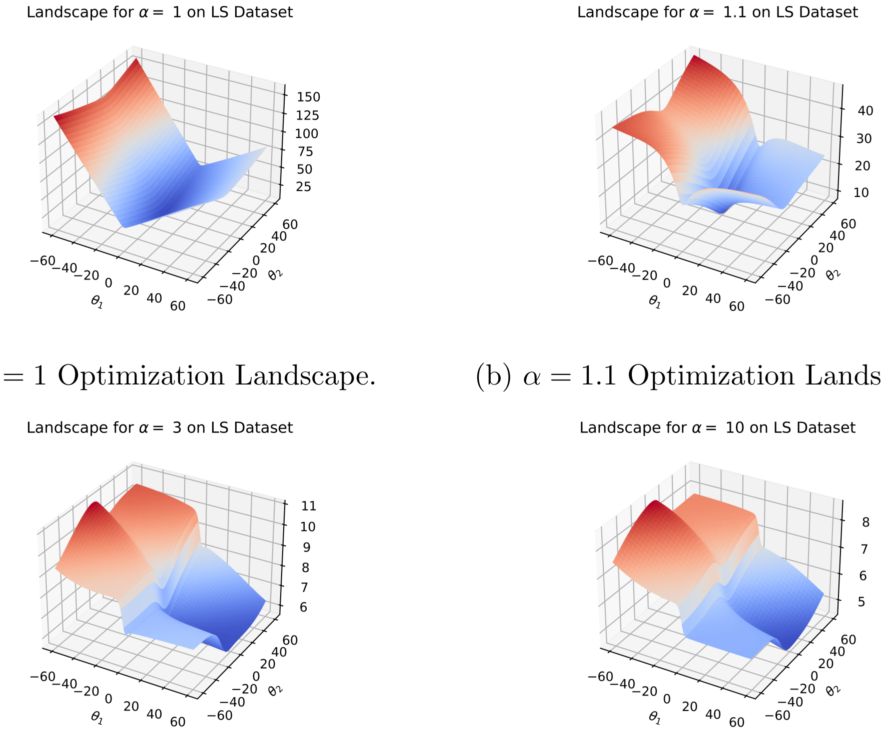
$$\begin{aligned} P_2^{\alpha}(\theta_1, \theta_2) &= -2\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma - \theta_2\gamma) + 2\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma + \theta_2\gamma) \\ &\quad + 5\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma) - 5\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma). \end{aligned} \tag{C.10}$$

In order to reason about the quality of the solutions to (C.7) for $\alpha \in (1, \infty)$, we want to find where $P_1^{\alpha}(\theta_1, \theta_2) = P_2^{\alpha}(\theta_1, \theta_2) = 0$ for the margin-based α-loss. So, rewriting $P_1^{\alpha}(\theta_1, \theta_2) = 0$, we obtain

$$\begin{aligned} &N\tilde{l}^{\alpha'}(\theta_1) + 2\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma - \theta_2\gamma) + N\gamma\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma) \\ &\quad = \tilde{l}^{\alpha'}(-\theta_1) + 2\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma + \theta_2\gamma) + \gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma), \end{aligned} \tag{C.11}$$

and rewriting $P_2^{\alpha}(\theta_1, \theta_2) = 0$, we obtain

$$2\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma + \theta_2\gamma) + 5\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma) = 2\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma - \theta_2\gamma) + 5\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma). \tag{C.12}$$



(a) $\alpha = 1$ Optimization Landscape.

(b) $\alpha = 1.1$ Optimization Landscape.

(c) $\alpha = 3$ Optimization Landscape.

(d) $\alpha = 10$ Optimization Landscape.

Figure C.2: Plots of Optimization Landscapes on the Long-servedio Dataset, I.E. (C.7), for $\alpha \in \{1, 1.1, 3, 10\}$. Aligning with Figure 4.2, $n = 2$ and $\gamma = 1/20$. For $\alpha = 1$, the Landscape Is Convex, Which Was Formally Proved (for Any Distribution) In (Sypherd *et al.*, 2020). For $\alpha = 1.1$, the Landscape Is Non-convex, but Not Too Much, Which Was Also Quantified In (Sypherd *et al.*, 2020). For $\alpha = 3$, the Landscape Is More Non-convex, and Notice That the Quality of the Solutions (in the Sense Of (C.4)) Is Significantly Better for $\alpha = 3$. For $\alpha = 10$, the Landscape Strongly Resembles the $\alpha = 3$, but Is "flatter".

Substituting (C.12) into (C.11), we are able to cancel a term and recover

$$N\tilde{l}^{\alpha'}(\theta_1) - \tilde{l}^{\alpha'}(-\theta_1) = 6\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma) - 6N\gamma\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma). \tag{C.13}$$

Rewriting, we obtain

$$N\tilde{l}^{\alpha'}(\theta_1) - \tilde{l}^{\alpha'}(-\theta_1) = -6\gamma\left[N\tilde{l}^{\alpha'}(\gamma(\theta_1 + 5\theta_2)) - \tilde{l}^{\alpha'}(-\gamma(\theta_1 + 5\theta_2))\right]. \tag{C.14}$$

Notice that $B_N^\alpha(x) = N\tilde{l}^{\alpha'}(x) - \tilde{l}^{\alpha'}(-x)$, with $x \in \mathbb{R}$, is common on both sides. From Lemma 20, we have that $\tilde{l}^{\alpha'}(x) := -\sigma'(x)\sigma(x)^{-1/\alpha}$ for $\alpha \in (0, \infty]$. Plugging this into

B_N^α (and using the fact that $\sigma'(x)$ is an even function), we have that

$$B_N^\alpha(x) = N\left(-\sigma'(x)\sigma(x)^{-1/\alpha}\right) - \left(-\sigma'(-x)\sigma(-x)^{-1/\alpha}\right) \tag{C.15}$$
$$= \sigma'(x)\sigma(-x)^{-1/\alpha} - N\sigma'(x)\sigma(x)^{-1/\alpha} \tag{C.16}$$
$$= \sigma'(x)\left(\sigma(-x)^{-1/\alpha} - N\sigma(x)^{-1/\alpha}\right). \tag{C.17}$$

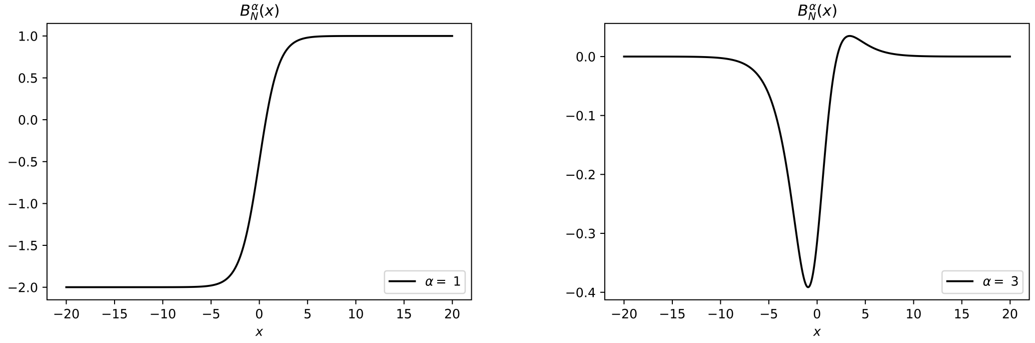
Using this, we can rewrite (C.14) as

$$B_N^\alpha(\theta_1) = -6\gamma B_N^\alpha(\gamma(\theta_1 + 5\theta_2)), \tag{C.18}$$

which is equivalent to

$$\begin{aligned}&\sigma'(\theta_1)\left(\sigma(-\theta_1)^{-1/\alpha} - N\sigma(\theta_1)^{-1/\alpha}\right)\\ &\quad = -6\gamma\sigma'(\gamma(\theta_1+5\theta_2))\left(\sigma(-\gamma(\theta_1+5\theta_2))^{-1/\alpha} - N\sigma(\gamma(\theta_1+5\theta_2))^{-1/\alpha}\right),\end{aligned} \tag{C.19}$$

and both quantify solutions (θ_1^*, θ_2^*). Notice that it is unfortunately not possible to analytically reduce (C.19) for general $\alpha \in (1, \infty)$ because it is a difference of α power expressions, i.e., a transcendental equation. However, while we cannot analytically recover solutions (θ_1^*, θ_2^*) for $\alpha \in (1, \infty)$, we can reason about the solutions themselves (from the perspective of (C.4)), because we can utilize nice properties of B_N^α. For instance, one key thing to notice in (C.18) is that B_N^α on the LHS depends only on one component of the solution vector, namely θ_1, whereas the RHS depends on both components of the solution vector (θ_1, θ_2).



(a) $B_N^\alpha(x)$ for $\alpha = 1$.

(b) $B_N^\alpha(x)$ for $\alpha = 3$.

Figure C.3: Plots of $B_n^\alpha(x)$ for $\alpha = 1$ and 3, Where $n = 2$. For $\alpha = 1$, Notice That $B_n^\alpha(x)$ Is Non-decreasing in x. On the Other Hand, Notice That for $\alpha = 3$, $B_n^\alpha(x)$ Is Not Non-decreasing. One Can Also See Other Properties of B_n^α in Figure (b) as Articulated in Lemma 21.

To this end, we take a detour from the main thread to aggregate some nice properties of B_N^α for $\alpha > 1$. See Figure C.3 for a plot of B_N^α.

Lemma 21. *Consider for $\alpha \in (0, \infty]$ and $1 < N < \infty$,*

$$B_N^\alpha(x) := \sigma'(x)\left(\sigma(-x)^{-1/\alpha} - N\sigma(x)^{-1/\alpha}\right), \tag{C.20}$$

where $x \in \mathbb{R}$. The following are properties of B_N^α:

1. *For $\alpha \leq 1$, $B_N^\alpha(x)$ is non-decreasing in x.*

2. *For $\alpha > 1$, $B_N^\alpha(x)$ is <u>not</u> non-decreasing in x.*

3. *Note that $\lim\limits_{\alpha \to \infty} B_N^\alpha(x) = \sigma'(x)(1-N)$.*

4. *For $\alpha > 1$, $\lim\limits_{x \to +\infty} B_N^\alpha(x) \to 0^+$ and $\lim\limits_{x \to -\infty} B_N^\alpha(x) \to 0^-$.*

5. *For $\alpha > 1$, the resulting limits of the previous property are reversed for $-B_N^\alpha$.*

6. *For $\alpha > 1$, $B_N^\alpha(x) > 0$ if and only if $x > \alpha \ln N$.*

The proof of the first property is obtained by invoking one of the results of Long and Servedio (2010) for convex, classification-calibrated loss functions. The remaining properties can be readily shown using standard techniques.

With these nice properties of B_N^α in hand, we now return to the main thread. Using the properties in Lemma 21, we want to reason about the solutions of (C.18), i.e.,

$$B_N^\alpha(\theta_1) = -6\gamma B_N^\alpha(\gamma(\theta_1 + 5\theta_2)), \tag{C.21}$$

as a function of $\alpha \in (0, \infty]$. From Propositions 7 and (Sypherd *et al.*, 2022b), we know that $\tilde{l}^\alpha$ is classification-calibrated for all $\alpha \in (0, \infty]$, convex for $\alpha \leq 1$, and quasi-convex for $\alpha > 1$. Thus, via (Long and Servedio, 2010), for each $\hat{\alpha} \leq 1$, there exists some $0 < \gamma_{\hat{\alpha}} < 1/6$ such that there exists a solution $(\theta_1^{\hat{\alpha}}, \theta_2^{\hat{\alpha}})$ of (C.21) which has classification accuracy of 0.5 (fair coin) on the Long-Servedio dataset. Without loss of generality, fix $\hat{\alpha} \leq 1$ and its associated pathological $0 < \gamma_{\hat{\alpha}} < 1/6$.

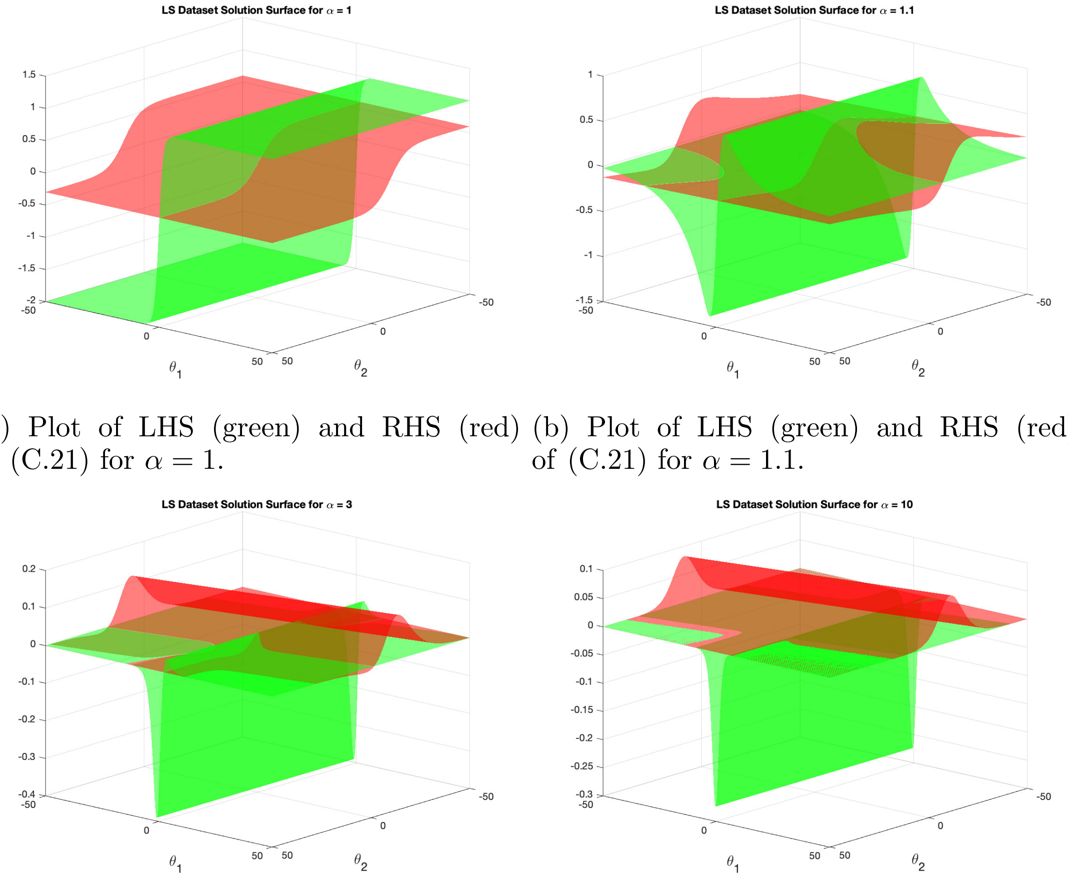
For $\alpha = \infty$, notice that there are *no solutions* to (C.21) since via the third property in Lemma 21, (C.21) reduces to

$$\sigma'(\theta_1)(1-N) = -6\gamma_{\hat{\alpha}}\sigma'(\gamma_{\hat{\alpha}}(\theta_1 + 5\theta_2))(1-N), \tag{C.22}$$

which is not satisfied because $\sigma'(\theta_1)(1-N) < 0$ and $-6\gamma_{\hat{\alpha}}\sigma'(\gamma_{\hat{\alpha}}(\theta_1 + 5\theta_2))(1-N) > 0$ for all (θ_1, θ_2); intuitively, the LHS and RHS in (C.22) look like mirrored $\sigma'(x)$ type functions.

Now, we consider $\alpha \in (1, \infty)$ in (C.21), which is the key region of α for the proof. Examining the LHS of (C.21), i.e. $B_N^\alpha(\theta_1)$, we note from the fourth property of Lemma 21 that $\lim\limits_{\theta_1 \to +\infty} B_N^\alpha(\theta_1) \to 0^+$. Furthermore, we note via the sixth property in Lemma 21 that $B_N^\alpha(\theta_1) > 0$ if and only if $\theta_1 > \alpha \ln N$. So, tuning $\alpha \in (1, \infty)$ greater moves the crossover (from negative to positive) of B_N^α further in θ_1.

We now examine the RHS of (C.21), i.e., $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}(\theta_1 + 5\theta_2))$. Set $\theta_2 = 0$, so we reduce $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}(\theta_1 + 5\theta_2))$ to $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}\theta_1)$. From the fifth property of Lemma 21, we have that $\lim\limits_{\theta_1 \to \infty} -6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}\theta_1) \to 0^-$. Furthermore, we note via the sixth property in Lemma 21 that $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}\theta_1) < 0$ if and only if $\theta_1 > \frac{\alpha \ln N}{\gamma_{\hat{\alpha}}}$. So, tuning $\alpha \in (1, \infty)$ greater moves the crossover (from positive to negative) of $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}\theta_1)$ further in θ_1.



(a) Plot of LHS (green) and RHS (red) of (C.21) for $\alpha = 1$.

(b) Plot of LHS (green) and RHS (red) of (C.21) for $\alpha = 1.1$.

(c) Plot of LHS (green) and RHS (red) of (C.21) for $\alpha = 3$.

(d) Plot of LHS (green) and RHS (red) of (C.21) for $\alpha = 10$.

Figure C.4: Plots of LHS (Green) and RHS (Red) Of (C.21) for $\alpha \in \{1, 1.1, 3, 10\}$, and $N = 2$ and $\gamma = 1/20$. The Intersections of the Surfaces Indicate Solutions Of (C.21). One Can See That the Solutions for $\alpha = 1$ Are Not "good" in the Sense Of (C.4) Because θ_1 Is Small and Fixed; This Phenomenon Was Proved By (Long and Servedio, 2010) since $\alpha = 1$ Is a Convex Loss. For $\alpha = 1.1$, One Can See the Resemblance of $\alpha = 1$ and $\alpha = 3$, and the Fact That "good" Solutions Are Starting to Accumulate. For $\alpha = 3$, There Are Many Solutions with Diverse (θ_1, θ_2) Values, since the Loss Is No Longer Convex. "good" Solutions for $\alpha = 3$ Can Be Seen Where θ_1 Is Positive and Large with Respect to θ_2, I.E., In the Middle/Right Side of the Plot. For $\alpha = 10$, One Can See That the "good" Solutions Have Been Pushed out Further in the Parameter Space and the Two Surfaces Are Starting to Separate (Reflecting the Fact That $\alpha = \infty$ Has No Solutions). Viewing All Four Plots Together, One Observes Smooth Transitions in α, Indicating That Finding a Good Solution Is Not Difficult.

Taking the limit and crossover behaviors in θ_1 of $B_N^\alpha(\theta_1)$ (the LHS of (C.21)) and $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}\theta_1)$ (the reduced RHS of (C.21)) together, we have *by continuity* that there must exist some $\tilde{\theta}_1 > 0$ which satisfies

$$B_N^\alpha(\tilde{\theta}_1) = -6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}\tilde{\theta}_1), \tag{C.23}$$

for each $\alpha \in (1, \infty)$. Furthermore, the choice of $\alpha \in (1, \infty)$ directly influences the magnitude of $\tilde{\theta}_1 > 0$, with larger α increasing the value of $\tilde{\theta}_1$ because of the crossover points, particularly that we require $\tilde{\theta}_1 > \frac{\alpha \ln N}{\gamma_{\hat{\alpha}}}$, which is more restrictive than the requirement that $\tilde{\theta}_1 > \alpha \ln N$, since $0 < \gamma_{\hat{\alpha}} < 1/6$, i.e., B_N^α is more "expansive" when its argument is multiplied by $\gamma_{\hat{\alpha}} < 1/6$. See Figure C.5 for a plot.

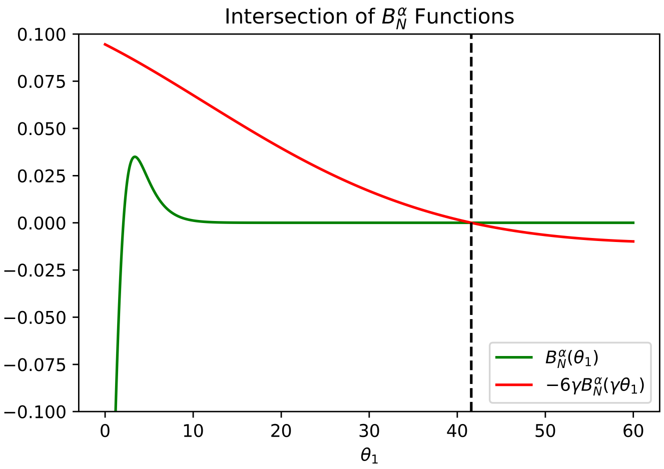
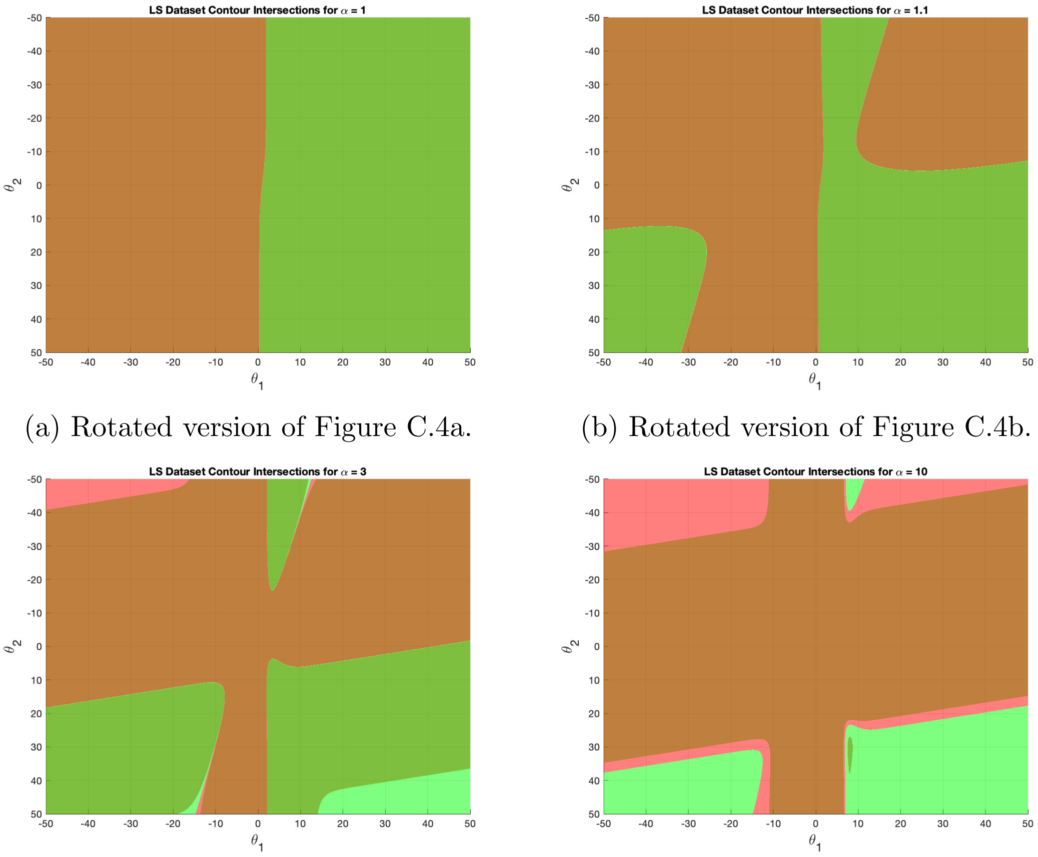


Figure C.5: A Plot of Lhs (Green) and Rhs (Red) Of (C.23) for $\alpha = 3$, Where $N = 2$, $\gamma = 1/20$. Notice That the Intersection Point $\tilde{\theta}_1$ (Dotted Line) Is *Very* Close to the Crossover Point $\frac{\alpha \ln N}{\gamma_{\hat{\alpha}}} \approx \frac{3 \times .69}{1/20} \approx 41.59$, and Also Notice That This Solution Nicely Coincides with the Grid-search Solution Presented in Figure 4.2.

Therefore, for each $\alpha \in (1, \infty)$, there exists a solution $(\theta_1^\alpha, \theta_2^\alpha)$ to (C.21), where $\theta_1^\alpha = \tilde{\theta}_1 > 0$ (indeed, we have that $\theta_1^\alpha = \mathcal{O}\left(\alpha \gamma_{\hat{\alpha}}^{-1} \ln\left(p^{-1} - 1\right)\right)$) and $\theta_2^\alpha = 0$, which is a good solution in the sense of (C.4) and thus has perfect classification accuracy on the *clean* LS dataset.

Next, while not necessary for the proof of Theorem 10, we also argue for the existence of other optima near $(\theta_1^\alpha, \theta_2^\alpha)$. Reconsidering the full (with θ_2 included) expression, $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}(\theta_1 + 5\theta_2))$ in (C.21), we take $\alpha \in (1, \infty)$ large enough in (C.23) and thus $\tilde{\theta}_1 > \frac{\alpha \ln N}{\gamma_{\hat{\alpha}}}$ is large enough such that $B_N^\alpha(\tilde{\theta}_1) \approx 0$ and is locally very "flat" (as given by the third property in Lemma 21). Hence, perturbing $\tilde{\theta}_1$ slightly induces an extremely slight movement in $B_N^\alpha(\tilde{\theta}_1)$. Now, considering $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}(\tilde{\theta}_1 + 5\theta_2))$, we fix θ_2^* to be *very* small (either positive or negative). We then "wiggle" $\tilde{\theta}_1$ slightly to (potentially) recover a solution θ_1^* to

$$B_N^\alpha(\theta_1^*) = -6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}(\theta_1^* + 5\theta_2^*)), \tag{C.24}$$



(a) Rotated version of Figure C.4a.

(b) Rotated version of Figure C.4b.

(c) Rotated version of Figure C.4c.

(d) Rotated version of Figure C.4d.

Figure C.6: Companion Figures of Figure C.4 for $\alpha \in \{1, 1.1, 3, 10\}$, and $N = 2$ and $\gamma = 1/20$. The Contours Indicate Solutions Of (C.21). In Figure C.6c, One Can See a Contour of “good” LS Solutions near Where $\theta_1 \approx 41.59$ and θ_2 Is Very Small.

which (might) exist by continuity. See Figure C.6 for a plot; intuitively, the fact that the LHS and RHS of (C.23) intersect, not merely "touch", suggests the existence of (θ_1^*, θ_2^*), indeed a "strip" of good solutions.

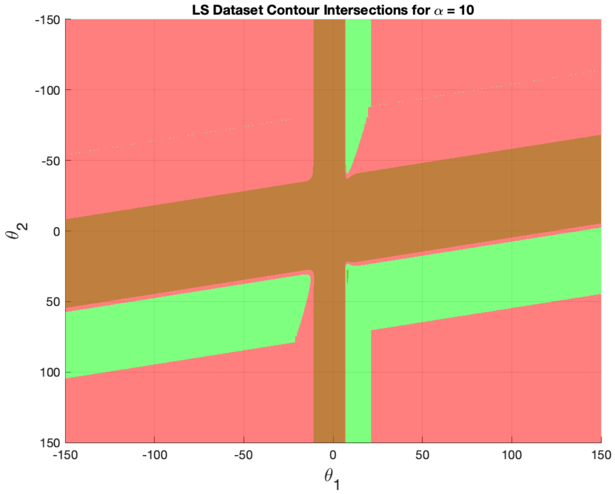


Figure C.7: Companion Figure of Figure C.6d, Again for $\alpha = 10$, Where the Parameter Space Has Been *Increased*. One Can Again See "good" LS Solutions for Large θ_1 and Small θ_2. This Is Indicative of a Trade off Between the Value of α and the Range of the Parameter Space for the LS Dataset.

C.1.2 Proof of Theorem 11

In this section, we provide the proof of Theorem 11. First, however, we provide lemmas useful in the proof of Theorem 11, which indicate useful bounds for $\alpha = 1$ and ∞, and their respective proofs.

Lemma 22. *For all $z \in \mathbb{R}$, we have that*

$$\left|\frac{d^2}{dz^2}\tilde{l}^1(z)\right| \geq \left|\frac{d^2}{dz^2}\tilde{l}^\infty(z)\right|, \tag{C.25}$$

Proof. Examining $\left|\frac{d^2}{dz^2}\tilde{l}^1(z)\right| = \left|\frac{d^2}{dz^2}\tilde{l}^\infty(z)\right|$, we have that

$$\left|\frac{d^2}{dz^2}\tilde{l}^1(z)\right| = \left|\frac{d^2}{dz^2}\tilde{l}^\infty(z)\right| \tag{C.26}$$

$$\left|\frac{e^z}{(e^z+1)^2}\right| = \left|\frac{e^z(e^z-1)}{(e^z+1)^3}\right| \tag{C.27}$$

$$e^z = \left|\frac{e^z(e^z-1)}{e^z+1}\right|, \tag{C.28}$$

however, there are no *real* solutions to this equation. Thus, $\left|\frac{d^2}{dz^2}\tilde{l}^1(z)\right|$ and $\left|\frac{d^2}{dz^2}\tilde{l}^\infty(z)\right|$ do not intersect.

Considering the large $z > 0$ regime, we find that

$$e^z \geq e^z - 1, \tag{C.29}$$

for all $z \in \mathbb{R}$, where we used the fact that $\lim_{z\to\infty} \frac{e^z(e^z-1)}{e^z+1} = e^z - 1$. Thus, by the Intermediate Value Theorem, we have the desired conclusion. □

Lemma 23. *For* $|z| > \ln(2)$*, we have that*

$$\left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right| \leq \left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right|. \tag{C.30}$$

Proof. Consider

$$\left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right| = \left|\frac{e^z - e^{2z}}{(e^z+1)^3}\right| \tag{C.31}$$

and

$$\left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right| = \left|\frac{-e^{3z} + 4e^{2z} - e^z}{(e^z+1)^4}\right|. \tag{C.32}$$

Setting

$$\left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right| = \left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right|, \tag{C.33}$$

after some algebra, we find that $z^* = \pm\ln(2)$. Furthermore, considering the large $z > 0$ regime, we find that

$$\begin{aligned}
\left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right| &\overset{?}{\leq} \left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right| &\quad (C.34)\\
\left|\frac{-e^{3z} + 4e^{2z} - e^z}{(e^z+1)^4}\right| &\overset{?}{\leq} \left|\frac{e^z - e^{2z}}{(e^z+1)^3}\right| &\quad (C.35)\\
\frac{e^{3z} - 4e^{2z} + e^z}{e^z+1} &\overset{?}{\leq} e^{2z} - e^z &\quad (C.36)\\
e^{2z} - 4e^z &\leq e^{2z} - e^z, &\quad (C.37)
\end{aligned}$$

thus by the IVT and symmetry, we have the desired result. □

With Lemmas 22 and 23 in hand, we now present the proof of Theorem 11. Recall from (4.11) that for $\alpha \in (0, \infty]$

$$\nabla_\theta \tilde{l}^\alpha(\langle YX, \theta\rangle) = -\sigma(\langle YX, \theta\rangle)^{1-\frac{1}{\alpha}}\sigma(-\langle YX, \theta\rangle)YX = \tilde{l}^{\alpha\prime}(\langle YX, \theta\rangle)YX, \tag{C.38}$$

since for each $i \in [d]$, $\frac{\partial}{\partial \theta_i}\tilde{l}^\alpha(\langle YX, \theta\rangle) = \tilde{l}^{\alpha\prime}(\langle YX, \theta\rangle)YX_i$.

Hence, the gradient of the noisy α-risk from (4.12) is

$$\begin{aligned}\nabla_\theta R_\alpha^p(\theta) &= \mathbb{E}_{X,Y}\left[(1-p)\nabla_\theta \tilde{l}^\alpha(\langle YX, \theta\rangle) + p\nabla_\theta \tilde{l}^\alpha(\langle -YX, \theta\rangle)\right] &\text{(C.39)}\\ &= \mathbb{E}_{X,Y}\left[\left((1-p)\tilde{l}^{\alpha\prime}(\langle YX, \theta\rangle) - p\tilde{l}^{\alpha\prime}(\langle -YX, \theta\rangle)\right)YX\right], &\text{(C.40)}\end{aligned}$$

where we expanded the expression for clarity. Notice that for $\alpha = 1$ (from Lemma 20),

$$\tilde{l}^{1\prime}(-z) = -\tilde{l}^{1\prime}(z) - 1, \tag{C.41}$$

namely that $\tilde{l}^{1\prime}$ is *almost* an **odd** function, and for $\alpha = \infty$,

$$\tilde{l}^{\infty\prime}(-z) = \tilde{l}^{\infty\prime}(z), \tag{C.42}$$

namely that $\tilde{l}^{\infty\prime}$ is an **even** function.

Thus, we have by the definition of $\hat{\theta}^1$ and $\hat{\theta}^\infty$ that for $\alpha = 1$

$$\begin{aligned}\mathbf{0} = \nabla_\theta R_1^p(\hat{\theta}^1) &= \mathbb{E}_{X,Y}\left[\left((1-p)\tilde{l}^{1\prime}(\langle YX, \hat{\theta}^1\rangle) - p\tilde{l}^{1\prime}(\langle -YX, \hat{\theta}^1\rangle)\right)YX\right] &\text{(C.43)}\\ &= (1-p)\mathbb{E}_{X,Y}\left[\tilde{l}^{1\prime}(\langle YX, \hat{\theta}^1\rangle)YX\right] - p\mathbb{E}_{X,Y}\left[\tilde{l}^{1\prime}(\langle -YX, \hat{\theta}^1\rangle)YX\right] &\text{(C.44)}\\ &= (1-p)\mathbb{E}_{X,Y}\left[\tilde{l}^{1\prime}(\langle YX, \hat{\theta}^1\rangle)YX\right] - p\mathbb{E}_{X,Y}\left[\left(-\tilde{l}^{1\prime}(\langle YX, \hat{\theta}^1\rangle) - 1\right)YX\right] &\text{(C.45)}\\ &= \mathbb{E}_{X,Y}\left[\tilde{l}^{1\prime}(\langle YX, \hat{\theta}^1\rangle)YX\right] + p\mathbb{E}_{X,Y}[YX], &\text{(C.46)}\end{aligned}$$

and for $\alpha = \infty$

$$\begin{aligned}\mathbf{0} = \nabla_\theta R_\infty^p(\hat{\theta}^\infty) &= \mathbb{E}_{X,Y}\left[\left((1-p)\tilde{l}^{\infty\prime}(\langle YX, \hat{\theta}^\infty\rangle) - p\tilde{l}^{\infty\prime}(\langle -YX, \hat{\theta}^\infty\rangle)\right)YX\right] &\text{(C.47)}\\ &= (1-p)\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty\prime}(\langle YX, \hat{\theta}^\infty\rangle)YX\right] - p\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty\prime}(\langle -YX, \hat{\theta}^\infty\rangle)YX\right] &\text{(C.48)}\\ &= (1-2p)\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty\prime}(\langle YX, \hat{\theta}^\infty\rangle)YX\right]. &\text{(C.49)}\end{aligned}$$

And, thus we have that for each $i \in [d]$,

$$\mathbb{E}_{X,Y}\left[\tilde{l}^{1\prime}(\langle YX, \hat{\theta}^1\rangle)YX_i\right] + p\mathbb{E}_{X,Y}[YX_i] = 0, \tag{C.50}$$

and

$$\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty\prime}(\langle YX, \hat{\theta}^\infty\rangle)YX_i\right] = 0. \tag{C.51}$$

In order to evaluate the efficacy of the gradient of the noisy α-risk at recovering the data generating vector $\theta^* \in \mathbb{B}_d(r)$, we seek to upper bound $\|\nabla_\theta R_1^p(\theta^*)\|_\infty$ and $\|\nabla_\theta R_\infty^p(\theta^*)\|_\infty$. To this end, recall the Taylor-Lagrange equality (Kline, 1998) for a twice continuously differentiable $f : \mathbb{R} \to \mathbb{R}$,

$$f(b) = f(a) + (b-a)f'(a) + \frac{(b-a)^2}{2} f''(c), \tag{C.52}$$

where $c \in [a, b]$.

Let $i \in [d]$ be arbitrary, but fixed. From (C.40) (and the reductions from (C.46) and (C.49)) we have that at $\theta^* \in \mathbb{B}_d(r)$

$$\frac{\partial}{\partial \theta_i} R_1^p(\theta^*) = \mathbb{E}_{X,Y}\left[\tilde{l}^{1\prime}(\langle YX, \theta^* \rangle) YX_i\right] + p\mathbb{E}_{X,Y}[YX_i], \tag{C.53}$$

and

$$\frac{\partial}{\partial \theta_i} R_\infty^p(\theta^*) = (1-2p)\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty\prime}(\langle YX, \theta^* \rangle) YX_i\right]. \tag{C.54}$$

Using the Taylor-Lagrange equality, we let $f = \tilde{l}^{\alpha\prime}$ (where $\alpha = 1$ or ∞ for simplicity for the time being), and thus we have that for each $(X, Y) \in \mathcal{X} \times \{-1, +1\}$,

$$\tilde{l}^{\alpha\prime}(b_{(X,Y)}) = \tilde{l}^{\alpha\prime}(a_{(X,Y)}) + (b_{(X,Y)} - a_{(X,Y)})\tilde{l}^{\alpha\prime\prime}(a_{(X,Y)}) + \frac{(b_{(X,Y)} - a_{(X,Y)})^2}{2}\tilde{l}^{\alpha\prime\prime\prime}(c^\alpha_{(X,Y)}), \tag{C.55}$$

where $b_{(X,Y)} = \langle YX, \theta^* \rangle$ and $a_{(X,Y)} = \langle YX, \hat{\theta}^\alpha \rangle$, hence $c^\alpha_{(X,Y)} \in [\langle YX, \hat{\theta}^\alpha \rangle, \langle YX, \theta^* \rangle]$. Examining each of (C.53) (first term) and (C.54) (without coefficient) with the Taylor-Lagrange equality, we have that

$$\mathbb{E}_{X,Y}\left[\tilde{l}^{\alpha\prime}(\langle YX,\theta^* \rangle)YX_i\right] = \mathbb{E}_{X,Y}\left[\left(\tilde{l}^{\alpha\prime}(a_{(X,Y)}) + (b_{(X,Y)} - a_{(X,Y)})\tilde{l}^{\alpha\prime\prime}(a_{(X,Y)}) + \frac{(b_{(X,Y)} - a_{(X,Y)})^2}{2}\tilde{l}^{\alpha\prime\prime\prime}(c^\alpha_{(X,Y)})\right)YX_i\right]. \tag{C.56}$$

Thus, for $\alpha = 1$, we have that

$$\begin{aligned}\frac{\partial}{\partial \theta_i} R_1^p(\theta^*) &= \mathbb{E}_{X,Y}\left[\tilde{l}^{1\prime}(\langle YX, \theta^* \rangle) YX_i\right] + p\mathbb{E}_{X,Y}[YX_i] \\ &= p\mathbb{E}_{X,Y}[YX_i] + \mathbb{E}_{X,Y}\left[\left(\tilde{l}^{1\prime}(\langle YX,\hat{\theta}^1 \rangle) + (\langle YX,\theta^* \rangle - \langle YX,\hat{\theta}^1 \rangle)\tilde{l}^{1\prime\prime}(\langle YX,\hat{\theta}^1 \rangle) + \frac{(\langle YX,\theta^* \rangle - \langle YX,\hat{\theta}^1 \rangle)^2}{2}\tilde{l}^{1\prime\prime\prime}(c^1_{(X,Y)})\right)YX_i\right],\end{aligned} \tag{C.57}$$

where $c^1_{(X,Y)} \in [\langle YX, \hat{\theta}^1 \rangle, \langle YX, \theta^* \rangle]$. Noticing that

$$\mathbb{E}_{X,Y}\left[\tilde{l}^{1\prime}(\langle YX, \hat{\theta}^1 \rangle) YX_i\right] + p\mathbb{E}_{X,Y}[YX_i] = 0, \tag{C.58}$$

we thus obtain

$$\frac{\partial}{\partial\theta_i}R_1^p(\theta^*)=\mathbb{E}_{X,Y}\left[\left((\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)+\frac{(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right)YX_i\right]. \tag{C.59}$$

Using similar steps, we can also obtain

$$\frac{\partial}{\partial\theta_i}R_\infty^p(\theta^*)=(1-2p)\mathbb{E}_{X,Y}\left[\left((\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^\infty\rangle)\tilde{l}^{\infty''}(\langle YX,\hat{\theta}^\infty\rangle)+\frac{(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^\infty\rangle)^2}{2}\tilde{l}^{\infty'''}(c^\infty_{(X,Y)})\right)YX_i\right], \tag{C.60}$$

where $c^\infty_{(X,Y)} \in [\langle YX, \hat{\theta}^\infty\rangle, \langle YX, \theta^*\rangle]$ and we note a difference between (C.59) and (C.60), i.e. the latter has the $1-2p$ coefficient.

Now, we consider $\left|\frac{\partial}{\partial\theta_i}R_1^p(\theta^*)\right|$ and seek an upperbound. We have that from (C.59)

$$\left|\frac{\partial}{\partial\theta_i}R_1^p(\theta^*)\right|=\left|\mathbb{E}_{X,Y}\left[\left((\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)+\frac{(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right)YX_i\right]\right| \tag{C.61}$$

$$\leq\mathbb{E}_{X,Y}\left[\left|\left((\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)+\frac{(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right)YX_i\right|\right] \tag{C.62}$$

$$=\mathbb{E}_{X,Y}\left[|X_i|\left|(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)+\frac{(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right] \tag{C.63}$$

$$\leq\mathbb{E}_{X,Y}\left[|X_i|\left(\left|(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|+\left|\frac{(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right], \tag{C.64}$$

where we used Jensen's inequality via the absolute value, the triangle inequality, and the fact that $|ab| = |a| \cdot |b|$. Continuing,

$$\mathbb{E}_{X,Y}\left[|X_i|\left(\left|(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|+\left|\frac{(\langle YX,\theta^*\rangle-\langle YX,\hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right] \tag{C.65}$$

$$=\mathbb{E}_{X,Y}\left[|X_i|\left(\left|\langle YX,\theta^*-\hat{\theta}^1\rangle\right|\left|\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|+\frac{\langle YX,\theta^*-\hat{\theta}^1\rangle^2}{2}\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right] \tag{C.66}$$

$$\leq\mathbb{E}_{X,Y}\left[|X_i|\left(\|YX\|\left\|\theta^*-\hat{\theta}^1\right\|\left|\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|+\frac{\|YX\|^2\left\|\theta^*-\hat{\theta}^1\right\|^2}{2}\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right], \tag{C.67}$$

where we used the Cauchy-Schwarz inequality on both inner products. Next, we use the observation that $X \in [0,1]^d$, and thus $\|X\| \leq \sqrt{d}$, and that $\theta^* - \theta \in \mathbb{B}_d(2r)$, for all $\theta \in \mathbb{B}_d(r)$. Thus, we have that

$$\mathbb{E}_{X,Y}\left[|X_i|\left(\|YX\|\left\|\theta^*-\hat{\theta}^1\right\|\left|\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|+\frac{\|YX\|^2\left\|\theta^*-\hat{\theta}^1\right\|^2}{2}\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right] \tag{C.68}$$

$$\leq \mathbb{E}_{X,Y}\left[\sqrt{d}2r\left|\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|+\frac{4dr^2}{2}\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right] \tag{C.69}$$

$$= 2d^{1/2}r\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|\right]+2dr^2\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right]. \tag{C.70}$$

Thus, we obtain that

$$\left|\frac{\partial}{\partial\theta_i}R_1^p(\theta^*)\right| \leq 2d^{1/2}r\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|\right]+2dr^2\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right]. \tag{C.71}$$

For $\alpha = \infty$, the exact same steps go through, so we also have that

$$\left|\frac{\partial}{\partial \theta_i} R^p_\infty(\theta^*)\right| \leq (1-2p)\left(2d^{1/2} r \mathbb{E}_{X,Y}\left[\left|\tilde{l}^{\infty''}(\langle YX, \hat{\theta}^\infty\rangle)\right|\right] + 2dr^2 \mathbb{E}_{X,Y}\left[\left|\tilde{l}^{\infty'''}(c^\infty_{(X,Y)})\right|\right]\right). \tag{C.72}$$

Considering $\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle)\right|\right]$ in (C.71), we let

$$z_1^* = \underset{z \in \{\langle yx, \hat{\theta}^1\rangle | (x,y) \in \mathcal{X} \times \{-1,+1\}\}}{\arg\max} \left|\tilde{l}^{1''}(z)\right|, \tag{C.73}$$

and we thus obtain $\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle)\right|\right] \leq \left|\tilde{l}^{1''}(z_1^*)\right|$, where we note that $z_1^* > \ln(2+\sqrt{3})$ by assumption. Similarly, considering $\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{\infty''}(\langle YX, \hat{\theta}^\infty\rangle)\right|\right]$ in (C.72), we let

$$z_\infty^* = \underset{z \in \{\langle yx, \hat{\theta}^\infty\rangle | (x,y) \in \mathcal{X} \times \{-1,+1\}\}}{\arg\max} \left|\tilde{l}^{\infty''}(z)\right|, \tag{C.74}$$

and we thus obtain $\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{\infty''}(\langle YX, \hat{\theta}^\infty\rangle)\right|\right] \leq \left|\tilde{l}^{\infty''}(z_\infty^*)\right|$, where $z_\infty^* \geq z_1^* > \ln(2+\sqrt{3})$ again by assumption.

Indeed, since $\left|\tilde{l}^{1'''}(z)\right|$ and $\left|\tilde{l}^{\infty'''}(z)\right|$ are monotonically decreasing for $z > \ln(2+\sqrt{3})$ we also have that

$$\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right] \leq \left|\tilde{l}^{1'''}(z_1^*)\right|, \tag{C.75}$$

and

$$\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{\infty'''}(c^\infty_{(X,Y)})\right|\right] \leq \left|\tilde{l}^{\infty'''}(z_\infty^*)\right|. \tag{C.76}$$

Next, we invoke Lemma 22, i.e., that for all $z \in \mathbb{R}$,

$$\left|\frac{d^2}{dz^2}\tilde{l}^1(z)\right| \geq \left|\frac{d^2}{dz^2}\tilde{l}^\infty(z)\right|, \tag{C.77}$$

and Lemma 23, i.e., that for $z > \ln 2$,

$$\left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right| \leq \left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right|. \tag{C.78}$$

Thus, we have that (also by monotonically decreasing)

$$\left|\tilde{l}^{\infty''}(z_\infty^*)\right| \leq \left|\tilde{l}^{1''}(z_1^*)\right|, \tag{C.79}$$

and

$$\left|\tilde{l}^{\infty'''}(z_\infty^*)\right| \leq \left|\tilde{l}^{1'''}(z_1^*)\right|. \tag{C.80}$$

Hence, since the bounds on (C.71) and (C.72) hold for all $i \in [d]$, we obtain the desired result.

C.1.3 Proof of Theorem 12

The strategy of the proof is to upperbound and lowerbound $\|\nabla_\theta R_\alpha^p(\theta) - \mathbb{E}[X^{[1]}]\|$. For the lowerbound, we use the reverse triangle inequality. Combining the upper and lowerbounds, we then rewrite the bounded expressions to induce a lowerbound on $\|\nabla_\theta R_\alpha^p(\theta)\|$ itself. For notational convenience, we used $\gamma = C_{p,r\sqrt{d},\alpha}$ in the main body.

Now, for each $y \in \{-1, 1\}$, let $X^{[y]}$ denote the random variable having the distribution of X conditioned on $Y = y$. We further assume that $X^{[1]} \overset{\mathrm{d}}{=} -X^{[-1]}$, $\mathbb{E}[X^{[1]}] \neq 0$, namely, a skew-symmetric distribution. Examining the gradient of the noisy α-risk (under the skew-symmetric distribution), we have that ($P_1 = \mathbb{P}[Y = 1]$)

$$\begin{aligned}
&\nabla_\theta R_\alpha^p(\theta) \\
&= \mathbb{E}_{X,Y}\left[\Big(pYg_\theta(-YX)^{1-1/\alpha}g_\theta(YX) - (1-p)Yg_\theta(YX)^{1-1/\alpha}g_\theta(-YX)\Big)X\right] \quad (\text{C.81})\\
&= P_1\mathbb{E}_{X^{[1]}}\left[\left(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]})\right)X^{[1]}\right]\\
&+ P_{-1}\mathbb{E}_{X^{[-1]}}\left[\Big(-pg_\theta(X^{[-1]})^{1-\frac{1}{\alpha}}g_\theta(-X^{[-1]}) + (1-p)g_\theta(-X^{[-1]})^{1-\frac{1}{\alpha}}g_\theta(X^{[-1]})\Big)X^{[-1]}\right] \quad (\text{C.82})\\
&= \mathbb{E}_{X^{[1]}}\left[\left(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]})\right)X^{[1]}\right]. \quad (\text{C.83})
\end{aligned}$$

First considering the upperbound on $\|\nabla_\theta R_\alpha^p(\theta) - \mathbb{E}[X^{[1]}]\|$, we have that

$$\begin{aligned}
&\|\mathbb{E}_{X^{[1]}}\left[\left(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]})\right)X^{[1]}\right] - \mathbb{E}[X^{[1]}]\| \quad (\text{C.84})\\
&= \|\mathbb{E}_{X^{[1]}}\left[\left(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - 1\right)X^{[1]}\right]\| \quad (\text{C.85})\\
&= \|\mathbb{E}_{X^{[1]}}\left[\Big(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - p - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - (1-p)\Big)X^{[1]}\right]\| \quad (\text{C.86})\\
&= \|\mathbb{E}_{X^{[1]}}\left[\Big(p\Big(g_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - 1\Big) - (1-p)\Big(g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - 1\Big)\Big)X^{[1]}\right]\| \quad (\text{C.87})\\
&\leq \mathbb{E}_{X^{[1]}}\left[\Big|p\Big(g_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - 1\Big) - (1-p)\Big(g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - 1\Big)\Big|\|X^{[1]}\|\right], \quad (\text{C.88})
\end{aligned}$$

where we used Jensen's inequality due to the convexity of the norm.

We now consider the term in absolute value above, which we rewrite for simplicity as

$$f_{\alpha,p}(x) := p\left(\sigma(-x)^{1-\frac{1}{\alpha}}\sigma(x) - 1\right) - (1-p)\left(\sigma(x)^{1-\frac{1}{\alpha}}\sigma(-x) - 1\right). \quad (\text{C.89})$$

We examine

$$\frac{\partial}{\partial\alpha}f_{\alpha,p}(x) = (1-p)\frac{\sigma(x)^{1-\frac{1}{\alpha}}\log\left(e^{-x}+1\right)}{(e^x+1)\alpha^2} - p\frac{\sigma(-x)^{1-\frac{1}{\alpha}}\log\left(e^{x}+1\right)}{(e^{-x}+1)\alpha^2}, \quad (\text{C.90})$$

which follows from the fact that

$$\frac{\partial}{\partial\alpha}\sigma(x)^{1-\frac{1}{\alpha}}\sigma(-x) = \frac{\sigma(x)^{1-\frac{1}{\alpha}}\log\left(\sigma(x)\right)}{(e^x+1)\alpha^2}. \quad (\text{C.91})$$

Considering $x > 0$ and $0 < p < 1/2$, one can show that

$$\frac{\partial}{\partial \alpha} f_{\alpha,p}(x) > 0 \tag{C.92}$$

is equivalent to

$$\left(\frac{1}{p} - 1\right) > e^{\frac{x}{\alpha}} \frac{\log\left(e^{x}+1\right)}{\log\left(e^{-x}+1\right)}, \tag{C.93}$$

and it can be shown that the term on the right-hand-side is monotonically increasing in $x > 0$ for $\alpha \in [1, \infty]$. Hence choosing $x > 0$ (i.e., $r > 0$) small enough ensures that $f_{\alpha,p}(x)$ is monotonically increasing in $\alpha \in [1, \infty]$. Furthermore, since $\dfrac{\partial}{\partial x} f_{\alpha,p}(x) > 0$ for $x > 0$, $p < 1/2$, and $\alpha \in [1, \infty]$, and $X \in [0,1]^d$, $\theta \in \mathbb{B}_d(r)$, we have by the Cauchy-Schwarz inequality (i.e., $\langle \theta, X\rangle \leq r\sqrt{d}$) that

$$p\left(g_\theta(-X^{[1]})^{1-1/\alpha} g_\theta(X^{[1]}) - 1\right) - (1-p)\left(g_\theta(X^{[1]})^{1-1/\alpha} g_\theta(-X^{[1]}) - 1\right) \tag{C.94}$$

$$\leq p\left(\sigma(-r\sqrt{d})^{1-\frac{1}{\alpha}}\sigma(r\sqrt{d}) - 1\right) - (1-p)\left(\sigma(r\sqrt{d})^{1-\frac{1}{\alpha}}\sigma(-r\sqrt{d}) - 1\right) =: C_{p,r\sqrt{d},\alpha}. \tag{C.95}$$

Note that $C_{p,r\sqrt{d},1} := \sigma(r\sqrt{d}) - p > 0$ (since $r\sqrt{d} > 0$ and $p < 1/2$), and $C_{p,r\sqrt{d},\infty} := (1-2p)(1-\sigma'(r\sqrt{d}))$, and by the restriction on $r > 0$ (C.93), we have that for $\alpha \in (1, \infty)$

$$0 < C_{p,r\sqrt{d},1} \leq C_{p,r\sqrt{d},\alpha} \leq C_{p,r\sqrt{d},\infty}. \tag{C.96}$$

Thus, considering the upperbound on $\|\nabla_\theta R_\alpha^p(\theta) - \mathbb{E}[X^{[1]}]\|$ in (C.88), we have that

$$\|\nabla_\theta R_\alpha^p(\theta) - \mathbb{E}[X^{[1]}]\| \leq C_{p,r\sqrt{d},\alpha}\mathbb{E}_{X^{[1]}}[\|X^{[1]}\|], \tag{C.97}$$

where $C_{p,r\sqrt{d},\alpha}$ is given in (C.95).

Now, considering a lowerbound on $\|\nabla_\theta R_\alpha^p(\theta) - \mathbb{E}[X^{[1]}]\|$, via the reverse triangle inequality we have that

$$\|\nabla_\theta R_\alpha^p(\theta) - \mathbb{E}[X^{[1]}]\| \geq \|\mathbb{E}[X^{[1]}]\| - \|\nabla_\theta R_\alpha^p(\theta)\|. \tag{C.98}$$

Combining this with our derived upperbound (C.97), we have that

$$C_{p,r\sqrt{d},\alpha}\mathbb{E}[\|X^{[1]}\|] \geq \|\mathbb{E}[X^{[1]}]\| - \|\nabla_\theta R_\alpha^p(\theta)\|. \tag{C.99}$$

Rewriting, we have that

$$\|\nabla_\theta R_\alpha^p(\theta)\| \geq \|\mathbb{E}[X^{[1]}]\| - C_{p,r\sqrt{d},\alpha}\mathbb{E}[\|X^{[1]}\|]. \tag{C.100}$$

Using our observation earlier regarding the monotonically increasing property of $C_{p,r\sqrt{d},\alpha}$ in $\alpha \in [1, \infty]$, we can write that

$$\begin{aligned}\|\nabla_\theta R^p_\alpha(\theta)\| &\geq \|\mathbb{E}[X^{[1]}]\| - C_{p,r\sqrt{d},\alpha}\mathbb{E}[\|X^{[1]}\|] \\ &\geq \|\mathbb{E}[X^{[1]}]\| - (1-2p)\left(1 - \sigma'(r\sqrt{d})\right)\mathbb{E}[\|X^{[1]}\|] > 0, \end{aligned} \tag{C.101}$$

which is nonnegative by distributional assumption on the skew-symmetric distribution itself, namely we assume that

$$(1-2p)(1-\sigma'(r\sqrt{d})) < \frac{\|\mathbb{E}(X^{[1]})\|}{\mathbb{E}(\|X^{[1]}\|)}. \tag{C.102}$$

C.2 Further Experimental Results and Details

C.2.1 Boosting Experiments

Long-Servedio

Dataset The Long-Servedio dataset is a synthetic dataset which was first suggested in (Long and Servedio, 2010) and also considered in (Cheamanunkul *et al.*, 2014). The dataset has input $x \in \mathbb{R}^{21}$ (which *differs* from the two-dimensional theoretical version in Section 4.3.2) with binary features $x_i \in \{-1, +1\}$ and label $y \in \{-1, +1\}$. Each instance is generated as follows. First, the label y is chosen to be -1 or $+1$ with equal probability. Given y, the features x_i are chosen according to the following mixture distribution:

- Large margin: with probability 1/4, we choose $x_i = y$ for all $1 \leq i \leq 21$.
- Pullers: with probability 1/4, we choose $x_i = y$ for $1 \leq i \leq 11$ and $x_i = -y$ for $12 \leq i \leq 21$.
- Penalizers: with probability 1/2, we choose 5 random coordinates from the first 11 and 6 from the last 10 to be equal to the label y. The remaining 10 coordinates are equal to $-y$.

Breast Cancer

Dataset The Wisconsin Breast Cancer dataset (Wolberg *et al.*, 1995) is a widely used medical dataset in the boosting community.

C.2.2 Logistic Model Experiments

GMM Setup

Dataset In order to evaluate the effect of generalizing log-loss with α-loss in the logistic model, we first analyze its performance learning on a two-dimensional dataset with Gaussian class-conditional distributions. The data was distributed as follows:

$$\begin{aligned} Y = 1 :& X \sim \mathcal{N}[\mu_1, \sigma^2\mathbb{I}], \\ Y = -1 :& X \sim \mathcal{N}[\mu_2, \sigma^2\mathbb{I}], \end{aligned}$$

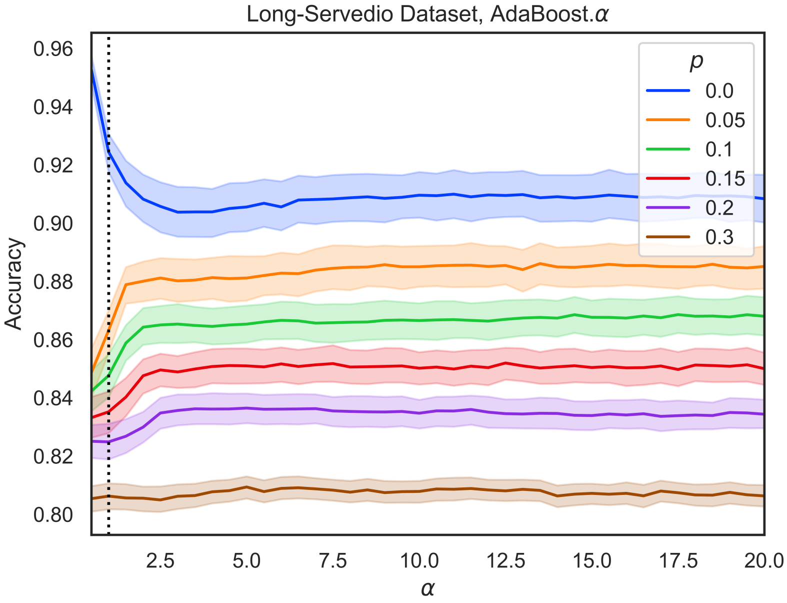


Figure C.8: Accuracy of Adaboost.α on the Long-Servedio Dataset. We See That Accuracy Levels off as α Increases, Implying That Tuning α Can Be as Simple as Choosing $\alpha \approx 5$. The Thresholding Behavior Is Supported by Figure C.2.

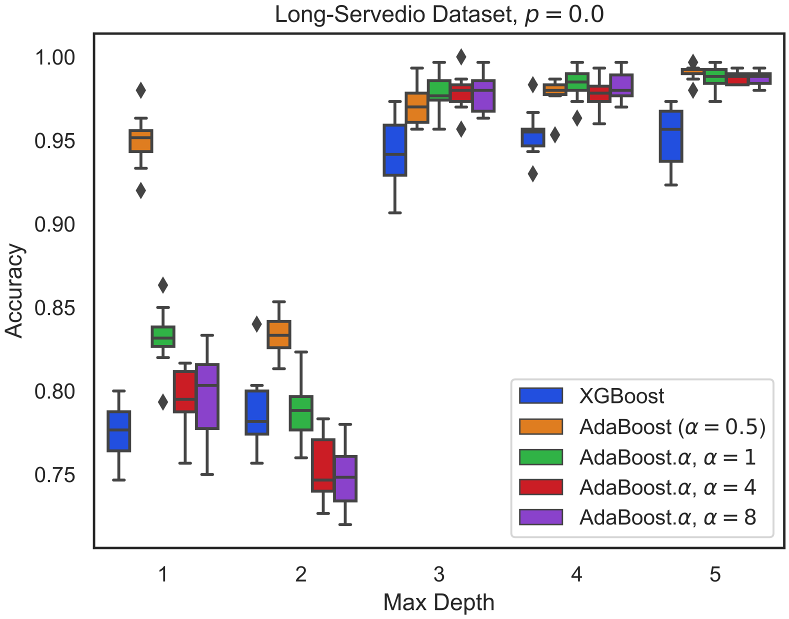


Figure C.9: Clean Test Accuracy of Various Models on the Long-servedio Dataset with No Added Label Noise. Models Trained for 100 Iterations. Vanilla Adaboost Performs Well Here, but Note That Figure C.12 Implies That with a Larger Number of Iterations, $\alpha = 1, 2$ Would Have Similar Performance.

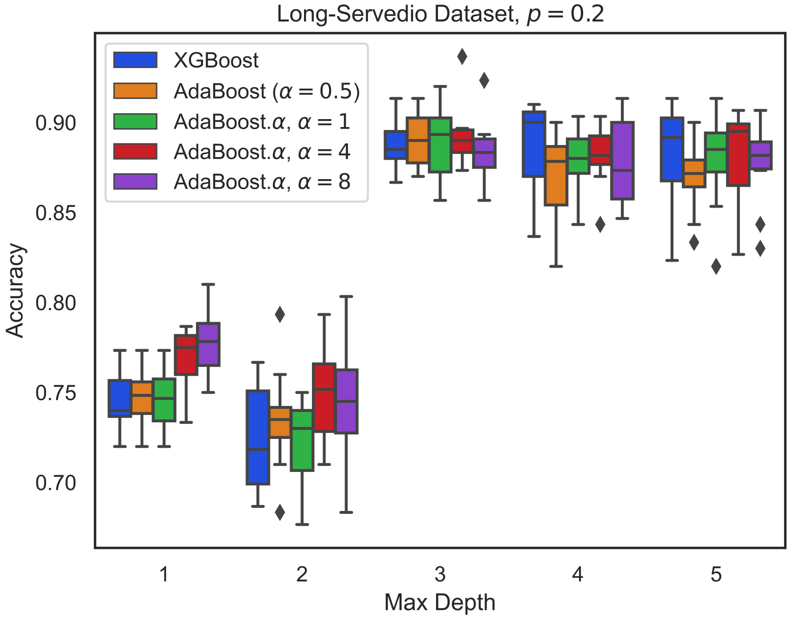


Figure C.10: Clean Test Accuracy Vs the Depth of Weak Learners on the Long-servedio Dataset with SLN. 100 Iterations of Boosting. We See That That for Low Depth Weak Learners, $\alpha > 1$ Outperforms Convex α in Terms of Clean Classification Accuracy. This Benefit Diminishes with Growing Depth.

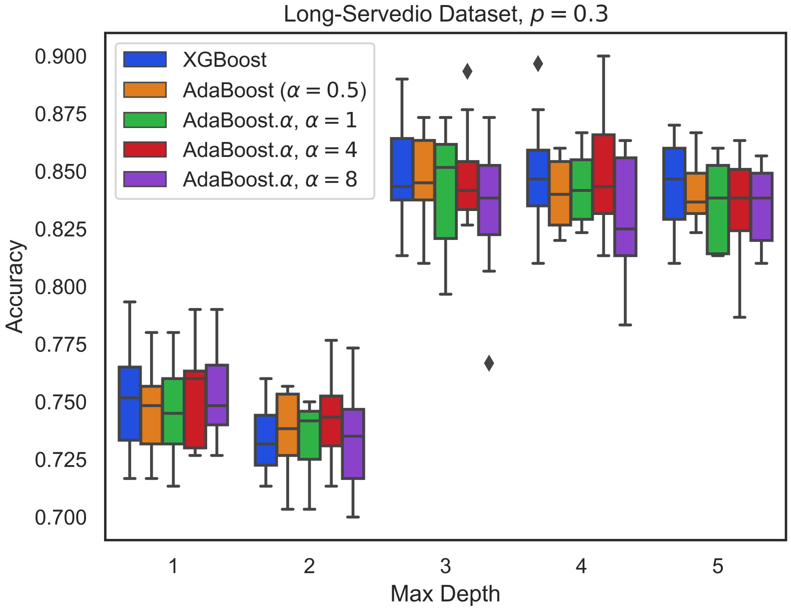


Figure C.11: Clean Test Accuracy Vs the Depth of Weak Learners on the Long-Servedio Dataset with Sln. 100 Iterations of Boosting. In This Higher Noise Setting, α Has Little Effect on the Clean Test Accuracy.

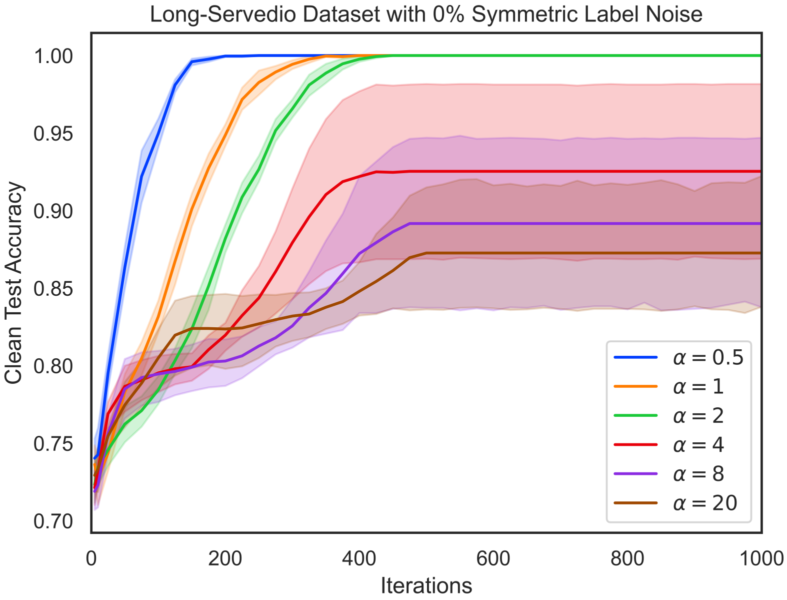


Figure C.12: Clean Test Accuracy of Adaboost.α on the Long-servedio Dataset with No Added Label Noise. In This Zero Noise Setting, Convex α Values Perform Well. Performance Gains Slow with Increasing α Which Corresponds to Increasing Non-convexity in the Optimization.

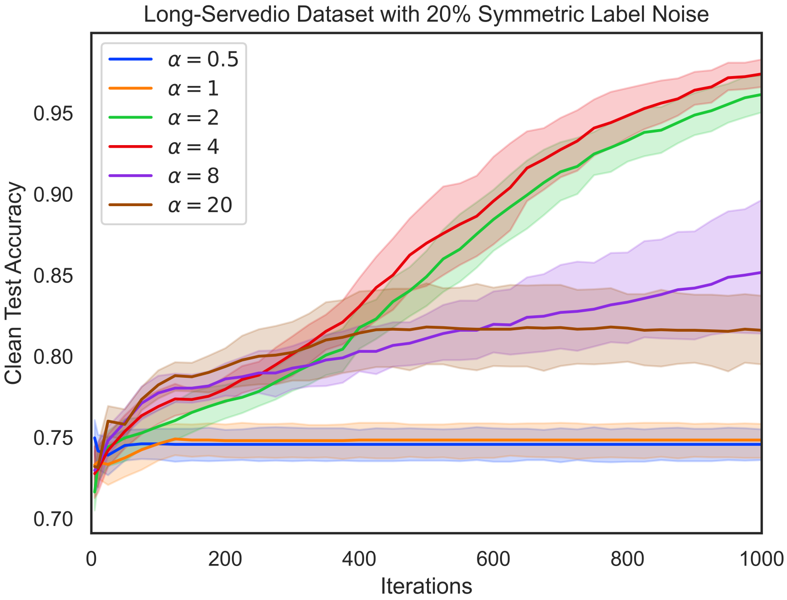


Figure C.13: Accuracy of Adaboost.α on the Long-servedio Dataset. We See That Convex $\alpha < 1$, Is Unable to Learn by Increasing the Number of Weak Learners, Likely Because It Is Getting Stuck Trying to Learn on Large-margin Example. $\alpha > 1$ Continues to Learn with Increasing Iterations, Though Growth Is Slower than in Smaller Noise Levels.

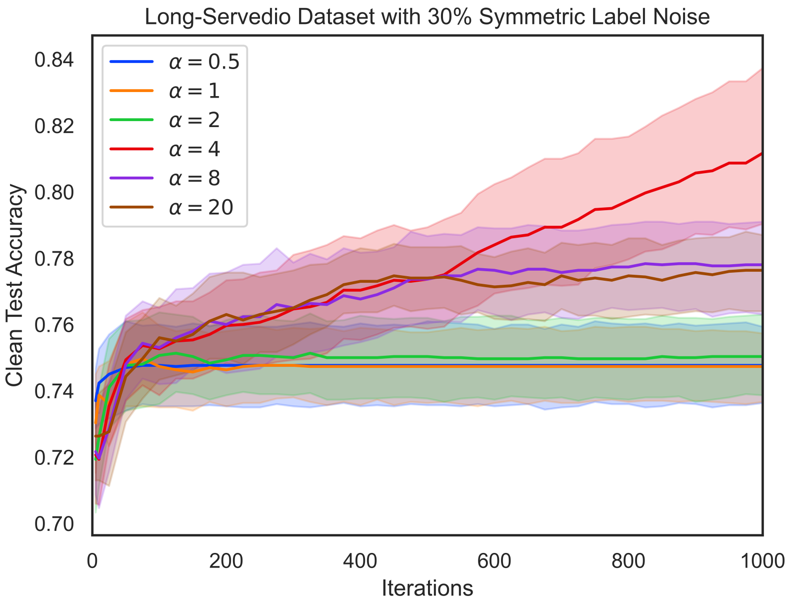


Figure C.14: Accuracy of Adaboost.α on the Long-servedio Dataset. We See That Convex $\alpha < 1$, Is Unable to Learn by Increasing the Number of Weak Learners, Likely Because It Is Getting Stuck Trying to Learn on Large-margin Example. $\alpha > 1$ Continues to Learn with Increasing Iterations, Though Growth Is Slower than in Smaller Noise Levels.

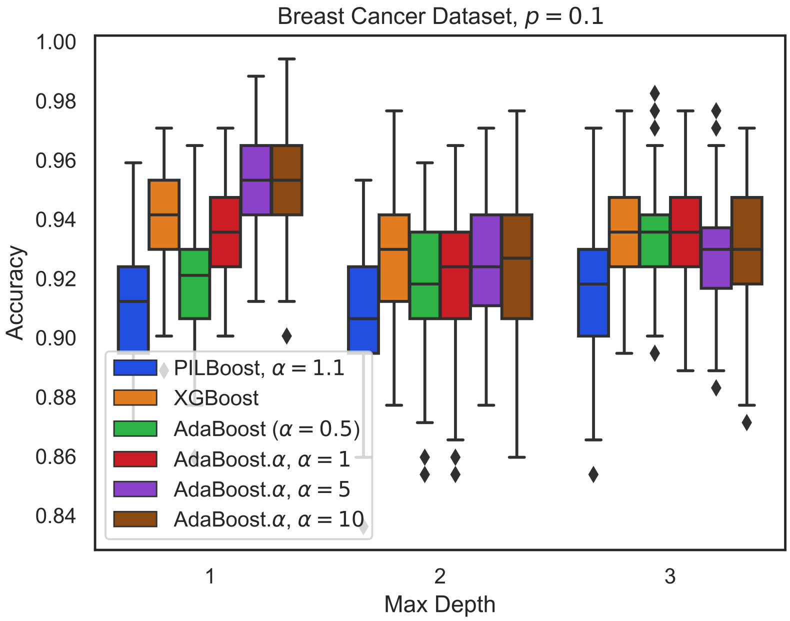


Figure C.15: Accuracy of Various Models on the Breast Cancer Dataset. We See That with Low Depth (and Thus Low Complexity) Weak Learners, the Use of a Non-convex Loss, Namely $\alpha > 1$, Permits Some Gains in Accuracy. These Diminish for More Complex Weak Learners.

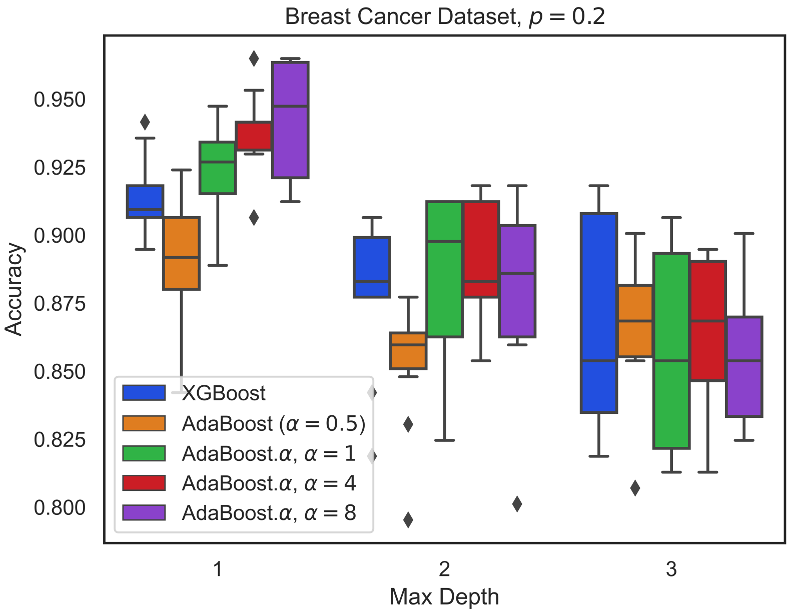


Figure C.16: Accuracy of Various Models on the Breast Cancer Dataset. We See That with Low Depth (and Thus Low Complexity) Weak Learners, the Use of a Nonconvex Loss, Namely $\alpha > 1$, Permits Some Gains in Accuracy. These Diminish for More Complex Weak Learners.

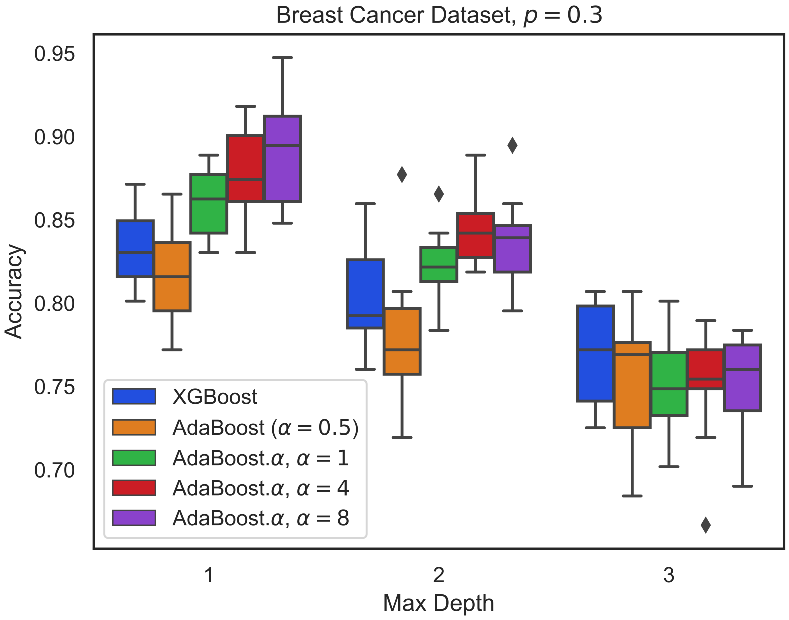


Figure C.17: Accuracy of Various Models on the Breast Cancer Dataset. We See That with Low Depth (and Thus Low Complexity) Weak Learners, the Use of a Nonconvex Loss, Namely $\alpha > 1$, Permits Some Gains in Accuracy. These Diminish for More Complex Weak Learners.

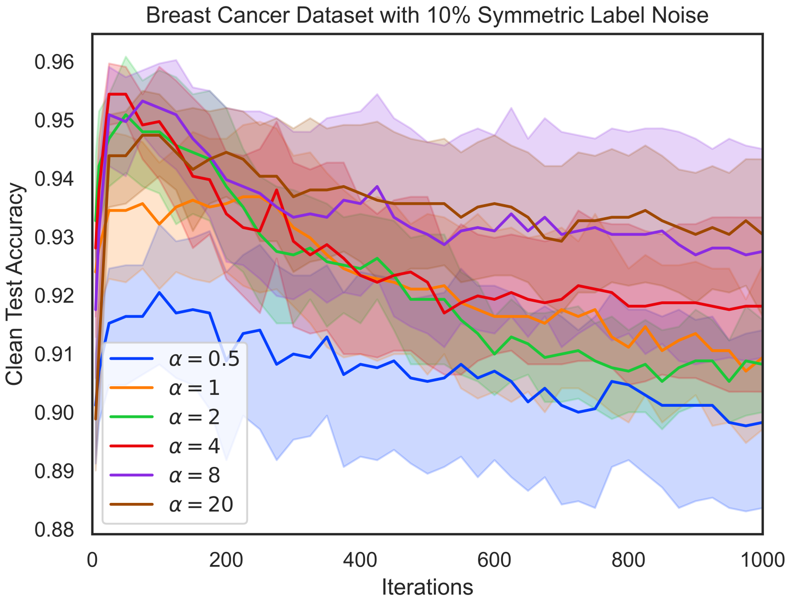


Figure C.18: Accuracy of Adaboost.α on the Wisconsin Breast Cancer Dataset. Nonconvex α Values Perform Significantly Better than Convex α Values. Unlike the Longservedio Dataset, Convex α Values Are Still Able to Learn as the Iterations Increase, Though There Appears to Be Some Overfitting.

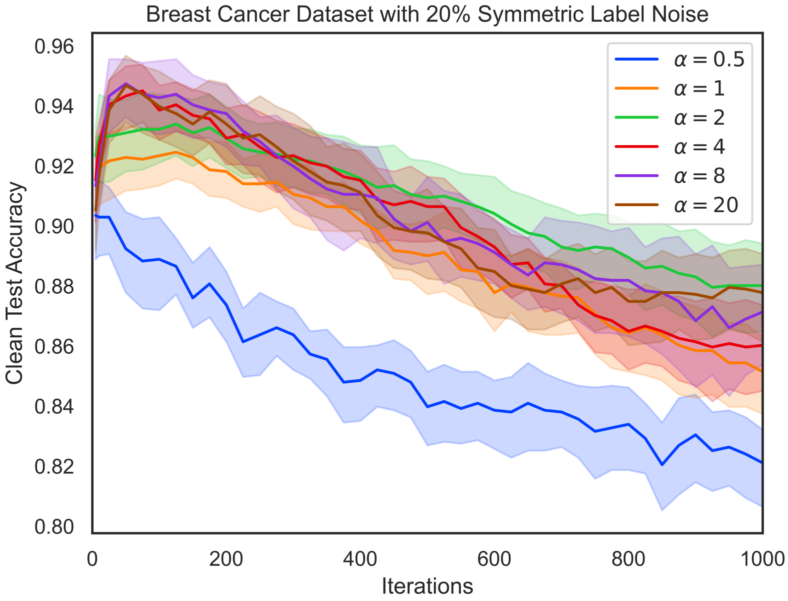


Figure C.19: Accuracy of Adaboost.α on the Wisconsin Breast Cancer Dataset. Nonconvex α Values Perform Significantly Better than Convex α Values. Unlike the Longservedio Dataset, Convex α Values Are Still Able to Learn as the Iterations Increase, Though There Appears to Be Some Overfitting.

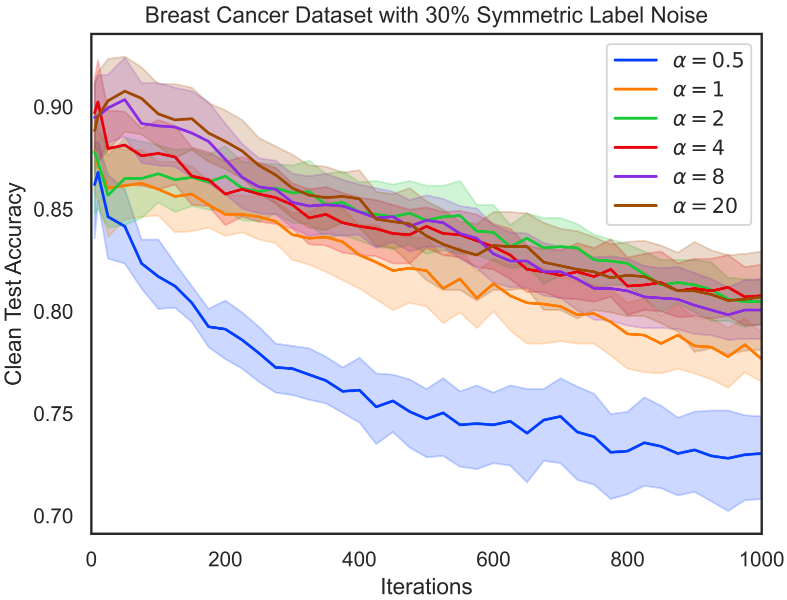


Figure C.20: Accuracy of Adaboost.α on the Wisconsin Breast Cancer Dataset. Nonconvex α Values Perform Significantly Better than Convex α Values. Unlike the Long-servedio Dataset, Convex α Values Are Still Able to Learn as the Iterations Increase, Though There Appears to Be Some Overfitting.

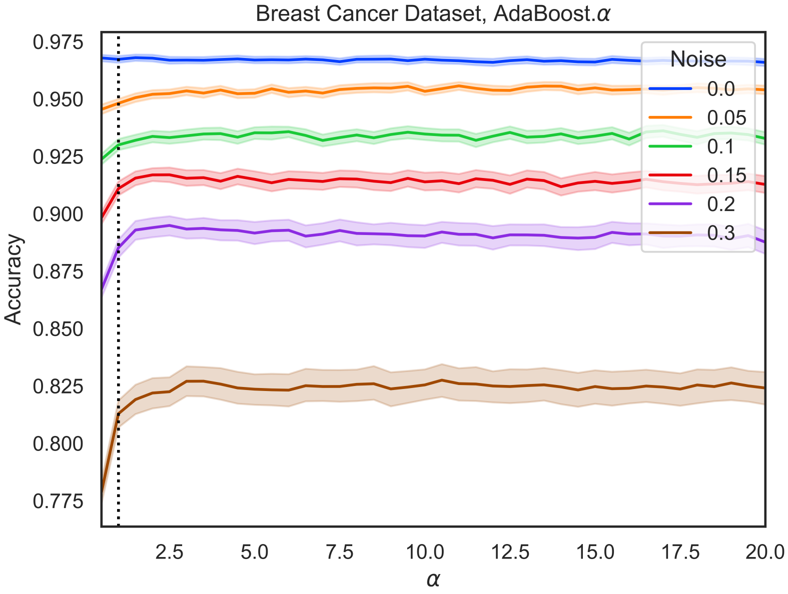


Figure C.21: Accuracy of Adaboost.α on the Breast Cancer Dataset with 100 Iterations. We See That Tuning $\alpha > 1$ Permits Significant Gains of Convex α Values, but That It Is Not Necessary to Tune α Too Large. Most of the Gains Are Realized with Small α.

where $\mu_i \in \mathbb{R}^2$, $\sigma \in \mathbb{R}$, and $\mathbb{I}$ is the 2×2 identity matrix.

We evaluate this simple two-dimensional equivariant case for reasons of interpretability and visualization. Additionally, we tune the prior of Y in order to control the level of class imbalance in the dataset to demonstrate that α-loss works well even under class imbalance conditions. Symmetric label noise is then added to this clean data.

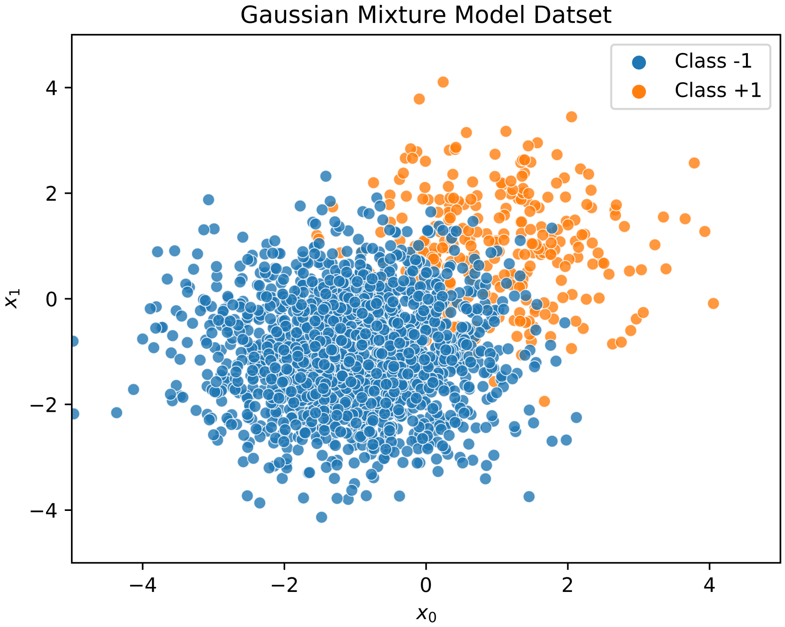


Figure C.22: Sample Dataset Generated with Gaussian Class-conditional Distributions with $p(Y = 1) = 0.14$ and $\mu_1 = [1, 1]^t, \mu_2 = [-1, -1]^t$; We Use a Spherical Covariance with $\sigma = 1$ for Both Classes.

Under this scenario, the Bayes-optimal classifier is linear because the variances of the two modes are equal and the features are uncorrelated. We can see this directly through the likelihood ratio test. Thus, we can compare the separating line given by training with α-loss on the logistic model and the optimal classifier.

Model A logistic model was trained on noisy data, then tested on clean data from the same data generating distribution. Models were trained over a grid of possible noise values, $p \in [0, 0.4]$, and $\alpha \in [0.5, 10]$. Learning rate was selected as 1e−2 and models were trained until convergence. For each pair, 30 models were trained with different noise seeds, and metrics were then averaged across models.

COVID-19 Logistic Setup

Model For better accuracy and a simpler, interpretable logistic model, we restrict the model to predict using a smaller set of 8 features; we choose these as the features with the largest odds ratio on the validation set and they are enumerated in Table 4.1. The learning rate was selected as 1e−3 and models were trained until convergence.

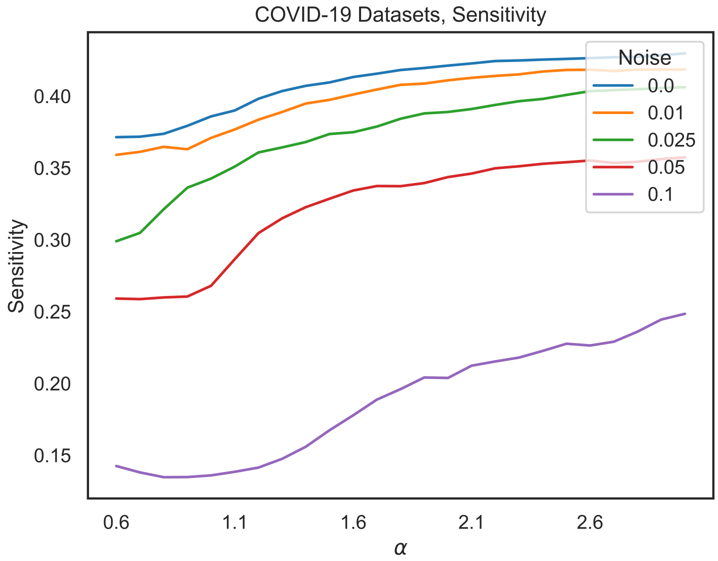


Figure C.23: Sensitivity of the Classifiers Trained on Noisy Covid-19 Data. We See That $\alpha > 1$ Yields Gains in Sensitivity. This Is Important to Note as the Mse Results Do Not Come at the Cost of Sensitivity. Recall That Sensitivity $= \frac{\text{Tp}}{\text{Tp+Fn}}$.

Models were trained over a grid of possible noise values, p, and α values, $(p, \alpha) \in [0, 0.15] \times [0.6, 3]$. For each pair (p, α), 5 models were trained with a different random noise seed and results were averaged across these samples for every metric.

Baseline Because the underlying true statistics are not available as a ground truth, a "clean" model is selected for a baseline comparison. We select this model to be one with no added noise ($p = 0$) and log-loss ($\alpha = 1$). Because log-loss ($\alpha = 1$) is calibrated, the "clean" posterior distribution will be the distribution with the smallest KL divergence to the data-generating distribution.

APPENDIX D

APPENDIX TO CHAPTER 5

D.1 Proof of Theorem 13

For a fixed generator, G_θ, we first solve the optimization problem

$$\sup_{\omega \in \Omega} \int_{\mathcal{X}} \frac{\alpha}{\alpha-1} \left(p_r(x) D_\omega(x)^{\frac{\alpha-1}{\alpha}} + p_{G_\theta}(x)(1-D_\omega(x))^{\frac{\alpha-1}{\alpha}} \right). \tag{D.1}$$

Consider the function

$$g(y) = \frac{\alpha}{\alpha-1} \left(a y^{\frac{\alpha-1}{\alpha}} + b(1-y)^{\frac{\alpha-1}{\alpha}} \right), \tag{D.2}$$

for $a, b \in \mathbb{R}_+$ and $y \in [0,1]$. To show that the optimal discriminator is given by the expression in (5.15), it suffices to show that $g(y)$ achieves its maximum in $[0,1]$ at $y^* = \frac{a^\alpha}{a^\alpha + b^\alpha}$. Notice that for $\alpha > 1$, $y^{\frac{\alpha-1}{\alpha}}$ is a concave function of y, meaning the function g is concave. For $0 < \alpha < 1$, $y^{\frac{\alpha-1}{\alpha}}$ is a convex function of y, but since $\frac{\alpha}{\alpha-1}$ is negative, the overall function g is again concave. Consider the derivative $g'(y^*) = 0$, which gives us

$$y^* = \frac{a^\alpha}{a^\alpha + b^\alpha}. \tag{D.3}$$

This gives (5.15). With this, the optimization problem in (5.14) can be written as $\inf_{\theta \in \Theta} C(G_\theta)$, where

$$C(G_\theta) = \frac{\alpha}{\alpha-1} \left[\int_{\mathcal{X}} \left(p_r(x) D_{\omega^*}(x)^{\frac{\alpha-1}{\alpha}} + p_{G_\theta}(x)(1-D_{\omega^*}(x))^{\frac{\alpha-1}{\alpha}} \right) dx - 2 \right] \tag{D.4}$$

$$= \tfrac{\alpha}{\alpha-1} \left[\int_{\mathcal{X}} \left(p_r(x) \left(\tfrac{p_r(x)^\alpha}{p_r(x)^\alpha + p_{G_\theta}(x)^\alpha} \right)^{\frac{\alpha-1}{\alpha}} + p_{G_\theta}(x) \left(\tfrac{p_r(x)^\alpha}{p_r(x)^\alpha + p_{G_\theta}(x)^\alpha} \right)^{\frac{\alpha-1}{\alpha}} \right) dx - 2 \right] \tag{D.5}$$

$$= \frac{\alpha}{\alpha-1} \left(\int_{\mathcal{X}} (p_r(x)^\alpha + p_{G_\theta}(x)^\alpha)^{\frac{1}{\alpha}} \, dx - 2 \right) \tag{D.6}$$

$$= D_{f_\alpha}(P_r || P_{G_\theta}) + \frac{\alpha}{\alpha-1} \left(2^{\frac{1}{\alpha}} - 2 \right), \tag{D.7}$$

where for the convex function f_α in (5.17),

$$D_{f_\alpha}(P_r || P_{G_\theta}) = \int_{\mathcal{X}} p_{G_\theta}(x) f_\alpha \left(\frac{p_r(x)}{p_{G_\theta}(x)} \right) dx \tag{D.8}$$

$$= \frac{\alpha}{\alpha-1} \left(\int_{\mathcal{X}} (p_r(x)^\alpha + p_{G_\theta}(x)^\alpha)^{\frac{1}{\alpha}} \, dx - 2^{\frac{1}{\alpha}} \right). \tag{D.9}$$

This gives us (5.16). Since $D_{f_\alpha}(P_r || P_{G_\theta}) \geq 0$ with equality if and only if $P_r = P_{G_\theta}$, we have $C(G_\theta) \geq \frac{\alpha}{\alpha-1} \left(2^{\frac{1}{\alpha}} - 2 \right)$ with equality if and only if $P_r = P_{G_\theta}$. Proof of Theorem 14 First, using L'Hôpital's rule we can verify that, for $a, b > 0$,

$$\lim_{\alpha \to 1} \frac{\alpha}{\alpha-1} \left((a^\alpha + b^\alpha)^{\frac{1}{\alpha}} - 2^{\frac{1}{\alpha}-1}(a+b) \right) = a \log \left(\frac{a}{\frac{a+b}{2}} \right) + b \log \left(\frac{b}{\frac{a+b}{2}} \right). \tag{D.10}$$

Using this, we have

$$D_{f_1}(P_r||P_{G_\theta}) \triangleq \lim_{\alpha\to 1} D_{f_\alpha}(P_r||P_{G_\theta}) \tag{D.11}$$

$$= \lim_{\alpha\to 1} \frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}} \left(p_r(x)^\alpha + p_{G_\theta}(x)^\alpha\right)^{\frac{1}{\alpha}} dx - 2^{\frac{1}{\alpha}}\right) \tag{D.12}$$

$$= \lim_{\alpha\to 1}\left[\frac{\alpha}{\alpha-1} \times \int_{\mathcal{X}} \left(\left(p_r(x)^\alpha + p_{G_\theta}(x)^\alpha\right)^{\frac{1}{\alpha}} - 2^{\frac{1}{\alpha}-1}(p_r(x)+p_{G_\theta}(x))\right) dx\right] \tag{D.13}$$

$$= \int_{\mathcal{X}} p_r(x)\log\frac{p_r(x)}{\left(\frac{p_r(x)+p_{G_\theta}(x)}{2}\right)}dx + \int_{\mathcal{X}} p_{G_\theta}(x)\log\frac{p_{G_\theta}(x)}{\left(\frac{p_r(x)+p_{G_\theta}(x)}{2}\right)}dx \tag{D.14}$$

$$=: 2D_{\mathrm{JS}}(P_r||P_{G_\theta}), \tag{D.15}$$

where $D_{\mathrm{JS}}(\cdot||\cdot)$ is the Jensen-Shannon divergence. Now, as $\alpha \to 1$, (5.16) equals $\inf_{\theta\in\Theta} 2D_{\mathrm{JS}}(P_r||P_{G_\theta}) - \log 4$ recovering vanilla GAN.

Substituting $\alpha = \frac{1}{2}$ in (5.18), we get

$$D_{f_{\frac{1}{2}}}(P_r||P_{G_\theta}) = -\int_{\mathcal{X}}\left(\sqrt{p_r(x)} + \sqrt{p_{G_\theta}(x)}\right)^2 dx + 4 \tag{D.16}$$

$$= \int_{\mathcal{X}}\left(\sqrt{p_r(x)} - \sqrt{p_{G_\theta}(x)}\right)^2 dx \tag{D.17}$$

$$=: 2D_{\mathrm{H}^2}(P_r||P_{G_\theta}), \tag{D.18}$$

where $D_{\mathrm{H}^2}(P_r||P_{G_\theta})$ is the squared Hellinger distance. For $\alpha = \frac{1}{2}$, (5.16) gives $2\inf_{\theta\in\Theta} D_{\mathrm{H}^2}(P_r||P_{G_\theta}) - 2$ recovering Hellinger GAN (up to a constant).

Noticing that, for $a, b > 0$, $\lim_{\alpha\to\infty}(a^\alpha + b^\alpha)^{\frac{1}{\alpha}} = \max\{a,b\}$ and defining $\mathcal{A} := \{x \in \mathcal{X} : p_r(x) \geq p_{G_\theta}(x)\}$, we have

$$D_{f_1}(P_r||P_{G_\theta}) \triangleq \lim_{\alpha\to\infty} D_{f_\alpha}(P_r||P_{G_\theta}) \tag{D.19}$$

$$= \lim_{\alpha\to\infty} \frac{\alpha}{\alpha-1}\left(\int_{\mathcal{X}} \left(p_r(x)^\alpha + p_{G_\theta}(x)^\alpha\right)^{\frac{1}{\alpha}} dx - 2^{\frac{1}{\alpha}}\right) \tag{D.20}$$

$$= \int_{\mathcal{X}} \max\{p_r(x), p_{G_\theta}(x)\}\ dx - 1 \tag{D.21}$$

$$= \int_{\mathcal{X}} \max\{p_r(x) - p_{G_\theta}(x), 0\}\ dx \tag{D.22}$$

$$= \int_{\mathcal{A}} (p_r(x) - p_{G_\theta}(x))\ dx \tag{D.23}$$

$$= \int_{\mathcal{A}} \frac{p_r(x) - p_{G_\theta}(x)}{2}\ dx + \int_{\mathcal{A}^c} \frac{p_{G_\theta}(x) - p_r(x)}{2}\ dx \tag{D.24}$$

$$= \frac{1}{2}\int_{\mathcal{X}} |p_r(x) - p_{G_\theta}(x)|\ dx \tag{D.25}$$

$$=: D_{\mathrm{TV}}(P_r||P_{G_\theta}), \tag{D.26}$$

where $D_{\text{TV}}(P_r||P_{G_\theta})$ is the total variation distance between P_r and P_{G_θ}. Thus, as $\alpha \to \infty$, (5.16) equals $\inf_{\theta\in\Theta} D_{\text{TV}}(P_r||P_{G_\theta}) - 1$ recovering TV-GAN (modulo a constant).

D.2 Proof of Theorem 15

We know from (Sypherd *et al.*, 2019, Corollary 1) that for $\eta \in [0,1]$,

$$\inf_t \eta\tilde{\ell}_\alpha(t) + (1-\eta)\tilde{\ell}_\alpha(-t) = \frac{\alpha}{\alpha-1}\left(1 - (\eta^\alpha + (1-\eta)^\alpha)^{\frac{1}{\alpha}}\right).$$

This implies that

$$\inf_t \frac{\eta}{1-\eta}\tilde{\ell}_\alpha(t) + \tilde{\ell}_\alpha(-t) = \frac{\alpha}{\alpha-1}\left(1 + \frac{\eta}{1-\eta} - \left(\left(\frac{\eta}{1-\eta}\right)^\alpha + 1\right)^{\frac{1}{\alpha}}\right). \tag{D.27}$$

Now substituting u for $\frac{\eta}{1-\eta}$ and taking negation in (D.27), we get

$$-\inf_t u\tilde{\ell}_\alpha(t) + \tilde{\ell}_\alpha(-t) = \frac{\alpha}{\alpha-1}\left((u^\alpha+1)^{\frac{1}{\alpha}} - (1+u)\right) \quad \text{for } u \geq 0, \tag{D.28}$$

giving us (5.20).