Smoothly Giving up: Robustness for Simple Models

Tyler Sypherd[1] **Nathan Stromberg**[1] **Richard Nock**[2] **Visar Berisha**[1] **Lalitha Sankar**[1]
[1]Arizona State University [2]Google Research

Abstract

There is a growing need for models that are interpretable and have reduced energy/computational cost (e.g., in health care analytics and federated learning). Examples of algorithms to train such models include logistic regression and boosting. However, one challenge facing these algorithms is that they provably suffer from label noise; this has been attributed to the joint interaction between oft-used convex loss functions and simpler hypothesis classes, resulting in too much emphasis being placed on outliers. In this work, we use the margin-based α-loss, which continuously tunes between canonical convex and quasi-convex losses, to robustly train simple models. We show that the α hyperparameter smoothly introduces non-convexity and offers the benefit of "giving up" on noisy training examples. We also provide results on the Long-Servedio dataset for boosting and a COVID-19 survey dataset for logistic regression, highlighting the efficacy of our approach across multiple relevant domains.

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., 2019; 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).

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

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 (2008) 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 (2008) 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.

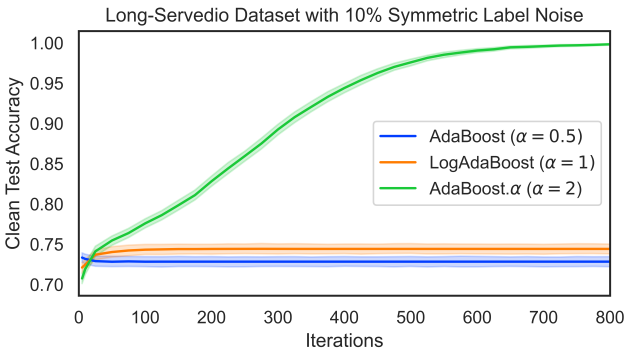


Figure 1: Quasi-convex α-loss booster ($\alpha = 2$) vs. convex boosters ($\alpha \leq 1$) on decision stumps for the Long-Servedio dataset. Full version and details in Section 5.

Recent work argues that the negative result of Long and Servedio (2008) could perhaps be circumvented by increasing the complexity of the weak learner (Mansour et al., 2022), 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. To this end, 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., 2022a). 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 2 and see the exponential vs. sigmoid losses). Hence, "giving up" on noisy training examples reduces the sensitivity of a simple hypothesis class (see Figure 1).

Our contributions are as follows:

1. In Theorem 1, we show that there exist robust solutions of the margin-based α-loss for $\alpha > 1$ to the problem of Long and Servedio (2008); we verify this result with simulation (Figure 3) and experimental results (Section 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 1 in Section 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 (2008) 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 2) and dominating terms also appearing in a lowerbound (Theorem 3). In essence, we provide guarantees on the quality of optima, showing with upper and lower bounds on the noisy gradient that $\alpha > 1$ is better for "good solutions" than $\alpha \leq 1$.
4. Finally, in Section 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.

1.1 Related Work

Convex and Non-Convex Losses While a small amount of carefully introduced label noise has been observed to improve the generalization capabilities of a model (Li et al., 2020), in general label noise during training is very detrimental to learning and thus represents an important problem for the community (Frénay and Verleysen, 2013; Rauscher et al., 2008; Gorber et al., 2009). In an effort to address this, 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., 2017; Lee et al., 2006; Lin et al., 2017; Leng et al., 2022). Other approaches include abstention (Thulasidasan et al., 2019; Ziyin et al., 2020; Cortes et al., 2016) 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., 2006)) 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, 2018; 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 *implicitly* assign less weight to misclassified training examples, and thus algorithms optimizing such losses are often less perturbed by outliers during training. This contrasts with another set of approaches (Lee et al., 2016; Yao et al., 2021; Maas et al., 2019; Bootkrajang and Kabán, 2012; Lee et al., 2016) which seek to *explicitly* estimate the noise transition matrix, sometimes requiring many parameters to do so.

α-loss The α-loss, where $\alpha \in (0, \infty]$, arose in information theory (Liao et al., 2018; Arimoto, 1971), and was recently introduced to ML (Sypherd et al., 2019). It smoothly tunes through several important losses (exponential for $\alpha = 1/2$, log for $\alpha = 1$, 0-1 for $\alpha = \infty$), and has statistical, optimization, and generalization tradeoffs dependent on α (Sypherd et al., 2022a). 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 (although there is a saturation effect), hence creating an optimization/robustness tradeoff (Sypherd et al., 2020). The α-loss is equivalent (under appropriate hyperparameter restriction) to the Generalized Cross Entropy loss (Zhang and Sabuncu, 2018), which was motivated by the Box-Cox transformation in statistics (Box and Cox, 1964). Also, the α-loss was recently shown to satisfy a statistical notion of robustness for loss functions in the class probability estimation setting (Sypherd et al., 2022b). More broadly, α-loss and related quantities have been used in Generative Adversarial Networks (Kurri et al., 2021, 2022) and in robust Bayesian posterior estimation (Zecchin et al., 2022).

Convex and Non-Convex Boosting AdaBoost (Freund and Schapire, 1997) (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. (2022b) 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, 2010), and presented experi-

mental results on the robustness of PILBoost.

However, the seminal work of Long and Servedio (2008) showed that convex boosters provably suffer from label noise, particularly for simple weak learners (Mansour et al., 2022). Van Rooyen et al. (2015) proposed relaxing the non-negativity 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; Miao et al., 2015), but there remains a large gap between the convex and non-convex realms. Therefore, we propose directly using the margin-based α-loss (rather than a "proper" convex approximation as in PILBoost), which smoothly tunes through several canonical convex and quasi-convex losses, for boosting. Our AdaBoost.α (generalizing vanilla AdaBoost with $\alpha = 1/2$ and LogAdaBoost with $\alpha = 1$), "gives up" on noisy outliers during training (for $\alpha > 1$), thus allowing practitioners to continue using simpler models (e.g., for interpretability or energy efficiency) in noisy settings.

2 PRELIMINARIES

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., 2006). 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

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

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, 2010). Noticing that $\sigma(-z) = 1 - \sigma(z)$, we have that

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

and note that σ' is an even function.

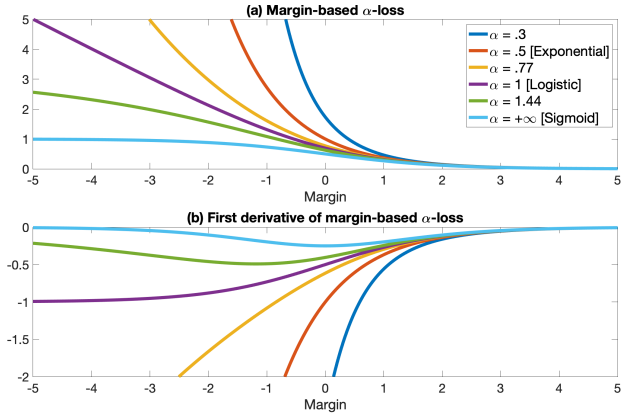


Figure 2: **(a)** Margin-based α-loss (3) as a function of the margin ($z := yH(x)$) for $\alpha \in \{0.3, 0.5, 0.77, 1, 1.44, \infty\}$; **(b)** its first derivative (see Lemma 1 in Appendix A) with respect to the margin for the same set of α. The "giving up" ability of the margin-based α-loss for $\alpha > 1$ can be seen from its first derivative, where it is more constrained (than $\alpha \leq 1$) for large negative values of the margin.

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., 2022a).

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

$$\tilde{l}^\alpha(z) := \frac{\alpha}{\alpha - 1}\left(1 - \sigma(z)^{1-1/\alpha}\right), \tag{3}$$

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 2(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. (2022a) showed that the margin-based α-loss is classification-calibrated for all $\alpha \in (0, \infty]$ (Bartlett et al., 2006). 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. (2022a) presented the following result regarding the convexity characteristics of $\tilde{l}^\alpha$.

Proposition 1. *$\tilde{l}^\alpha : \mathbb{R} \to \mathbb{R}_+$ 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, 2004)).

In light of Proposition 1, consider Figure 2(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. sub-linear; this is similarly observed in Figure 2(b) with the first derivative of $\tilde{l}^\alpha$ (see Lemma 1 in Appendix A). 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 3) vs. more nuanced function classes (Rolnick et al., 2017).

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

$$H_T(\cdot) := \sum_{t \in [T]} \theta_t h_t(\cdot), \tag{4}$$

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

$$D_t(i) := -\tilde{l}'(y^i H_{t-1}(x^i)), \forall i \in [m]. \tag{5}$$

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 1 for the general procedure).

In the next section, we show that using the derivative of the margin-based α-loss in (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 (2008).

3 ROBUSTNESS FOR BOOSTING

3.1 AdaBoost.α: Boosting with a Give Up Option

Algorithm 1 AdaBoost.α

1: **Given:** $(x^1, y^1), \ldots, (x^m, y^m)$ where $x^i \in \mathcal{X}$, $y^i \in \{-1, +1\}$, and $\alpha \in (0, \infty]$
2: Initialize: $H_0 = 0$.
3: **for** $t = 1, 2, \ldots, T$:
4: Update, for $i = 1, \ldots, m$:

$$D_t(i) = \frac{\sigma'(y^i H_{t-1}(x^i))\sigma(y^i H_{t-1}(x^i))^{-\frac{1}{\alpha}}}{\mathcal{Z}_t}, \tag{6}$$

where $\mathcal{Z}_t$ is a normalization factor.
5: Return h_t, weakly learned on D_t.
6: Compute error of weak hypothesis h_t:

$$\epsilon_t = \sum_{i : h_t(x^i) \neq y^i} D_t(i). \tag{7}$$

7: Let $\theta_t = \frac{1}{2} \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$.
8: Update: $H_t = H_{t-1} + \theta_t h_t$
9: **Return** $\overline{H}(\cdot) = \text{sign}\left(H_T(\cdot)\right)$

Using the smooth tuning of the margin-based α-loss, we present a novel robust boosting algorithm, AdaBoost.α in Algorithm 1, 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 1 in Appendix A) in (6), we obtain a hyperparameterized family of "AdaBoost-type" algorithms.

Indeed, AdaBoost.α also recovers LogAdaBoost (see Section 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 1). As argued in Section 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 1.1). However, the novelty 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). Lastly, we note that because of the modularity of the α hyperparameter generalization, a multiclass extension of

AdaBoost.α readily follows from standard approaches of multiclass AdaBoost, e.g., Hastie et al. (2009).

3.2 Robustness on the Long-Servedio Dataset

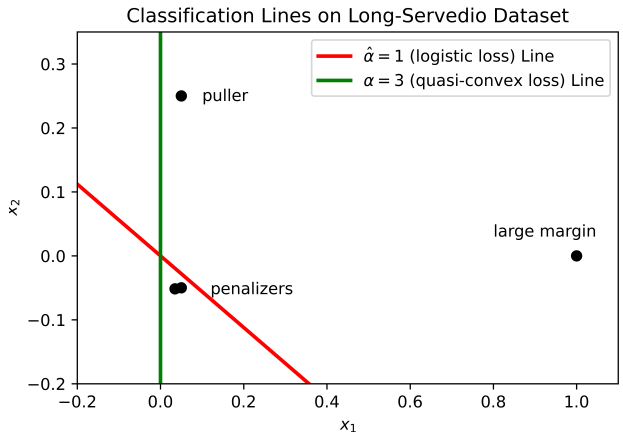


Figure 3: A plot depicting optimal classification lines of $\hat{\alpha} = 1$ and $\alpha = 3$ for the *clean* Long-Servedio dataset $\mathcal{S}$, where the penalizer examples are slightly separated for display. The $\hat{\alpha}, \alpha$ optima are obtained by grid-search on the noisy Long-Servedio 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 (2008) to "kick-in". The $\hat{\alpha} = 1$ (logistic loss) line (red) is given by $(\theta_1^{\hat{\alpha}}, \theta_2^{\hat{\alpha}}) = (0.79, 1.41)$ for (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 1 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 A.1.

In Long and Servedio (2008), 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 "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 3 for a plot of the dataset. Notice that $H(x) = \text{sign}(x_1)$ (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),

$$H_{\tilde{l},\gamma}(x_1, x_2) := \theta_1 x_1 + \theta_2 x_2, \tag{8}$$

achieves perfect classification accuracy on $\mathcal{S}$, where the dependency on the loss $\tilde{l}$ and data margin γ is clear. Note that (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., 2022), 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 (2008) 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, 2008). 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 1. *Let $0 < p < 1/2$ for $\hat{\mathcal{S}}$, and $\hat{\alpha} \leq 1$ for $\tilde{l}^{\alpha}$. By Long and Servedio (2008), 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 1 (in Appendix A.1) is nontrivial since $\alpha > 1$ has a non-convex optimization landscape. In Figure 3 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 1, 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 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 ROBUSTNESS FOR LINEAR MODELS

Taking inspiration from the boosting setup in Section 3.2, where the weak learner recovered a 2D linear model in (8), we now consider a generalization of the 2D linear hypothesis class to $d \in \mathbb{N}$ dimensions, which in binary classification is equivalent to the logistic model (Sypherd et al.,

2022a). Ideally, one would like to give *direct* expressions of gradient optimizers $\hat{\theta}^{\alpha}$ as we do for the Long-Servedio setup in Theorem 1, however, the logistic model has sigmoid non-linearities that make this difficult for general data distributions. Instead, we take an *indirect* approach where we provide guarantees on the quality of gradient optima, showing with upper and lower bounds that the noisy gradient for $\alpha > 1$ is smaller for "good solutions" than when $\alpha = 1$ (logistic regression). Thus, the motivation for Theorems 2 and 3 is to argue that a gradient optimizer is more likely to converge near a "good solution" when $\alpha > 1$ than when $\alpha = 1$; indeed, this is another way to view how the $\alpha > 1$ "give up" on the noise in the training data.

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 $\langle yx, \theta\rangle \leq r\sqrt{d}$.

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

$$R_{\alpha}(\theta) := \mathbb{E}_{X,Y}\left[\tilde{l}^{\alpha}(\langle YX, \theta\rangle)\right], \tag{9}$$

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

$$R^{p}_{\alpha}(\theta) := \mathbb{E}_{X,Y}\left[\mathbb{E}_{\tau\sim\text{Rad}(p)}\left(\tilde{l}^{\alpha}(\langle -\tau YX, \theta\rangle)\right)\right], \tag{10}$$

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, 2004). Thus, the gradient of the α-risk in (9) is

$$\nabla_{\theta} R_{\alpha}(\theta) := \mathbb{E}_{X,Y}\left[\nabla_{\theta}\tilde{l}^{\alpha}(\langle YX, \theta\rangle)\right], \tag{11}$$

$\nabla_{\theta}\tilde{l}^{\alpha}(\langle YX, \theta\rangle) := -\sigma'(\langle YX, \theta\rangle)\sigma(\langle YX, \theta\rangle)^{-\frac{1}{\alpha}}YX$ for $\alpha \in (0,\infty]$ from Lemma 1 in Appendix A. Hence, the gradient of the noisy α-risk (10) is given by

$$\nabla_{\theta} R^{p}_{\alpha}(\theta) := \mathbb{E}_{X,Y}\left[\mathbb{E}_{\tau\sim\text{Rad}(p)}\left(\nabla_{\theta}\tilde{l}^{\alpha}(\langle -\tau YX, \theta\rangle)\right)\right]. \tag{12}$$

We now present a result in the realizable setting, indicating (12) is smaller for $\alpha = \infty$ (soft 0-1 loss) at any data generating vector $\theta^* \in \mathbb{B}_d(r)$ than for $\alpha = 1$ (logistic loss).

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

$$\langle yx, \hat{\theta}^{\infty}\rangle \geq \langle yx, \hat{\theta}^1\rangle > \ln(2+\sqrt{3}). \tag{13}$$

If for any $\theta^* \in \mathbb{B}_d(r)$ we have $\langle yx, \theta^*\rangle \geq \langle yx, \hat{\theta}^{\infty}\rangle$ for all $(x,y) \in \mathcal{X} \times \{-1,+1\}$, then we have that for $\alpha \in \{1,\infty\}$,

$$\frac{\|\nabla_{\theta} R^{p}_{\alpha}(\theta^*)\|_{\infty}}{C_{\alpha}} \leq d^{\frac{1}{2}} r\left|\tilde{l}^{\alpha\prime\prime}(z^*_{\alpha})\right| + dr^2\left|\tilde{l}^{\alpha\prime\prime\prime}(z^*_{\alpha})\right|, \tag{14}$$

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

$$1 - 2p < \frac{d^{\frac{1}{2}} r\left|\tilde{l}^{1\prime\prime}(z^*_1)\right| + dr^2\left|\tilde{l}^{1\prime\prime\prime}(z^*_1)\right|}{d^{\frac{1}{2}} r\left|\tilde{l}^{\infty\prime\prime}(z^*_{\infty})\right| + dr^2\left|\tilde{l}^{\infty\prime\prime\prime}(z^*_{\infty})\right|}. \tag{15}$$

Theorem 2 uses symmetries of the first derivative of $\tilde{l}^{\alpha}$ for $\alpha \in \{1,\infty\}$; see Appendix A.2 for proof details. Intuitively, (15) indicates that there is a significant discrepancy between the two upper bounds as the noise rate $p \to 1/2$, suggesting that $\nabla_{\theta} R^{p}_{\alpha}(\cdot)$ is smaller at any data generating θ^* for $\alpha = \infty$ than for $\alpha = 1$ (logistic regression). Note that the assumption in (13) is mild because *both* vectors (rather than just $\hat{\theta}^{\infty}$) are assumed to achieve perfect accuracy on the clean data distribution.

In support of the upper bounds in Theorem 2, we now present a uniform lower bound on the norm of (12) for the skew-symmetric family of distributions (e.g., GMMs).

Theorem 3. *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 $X^{[1]} \overset{\text{d}}{=} -X^{[-1]}$, and $\mathbb{E}[X^{[1]}] \neq \mathbf{0}$. We also assume that $r > 0$ is small enough such that both of the following hold:*

$$(1-2p)(1-\sigma'(r\sqrt{d})) < \frac{\|\mathbb{E}(X^{[1]})\|_2}{\mathbb{E}(\|X^{[1]}\|_2)}, \tag{16}$$

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

$$e^{\frac{r\sqrt{d}}{\alpha}}\log\left(e^{r\sqrt{d}}+1\right) < \left(p^{-1}-1\right)\log\left(e^{-r\sqrt{d}}+1\right). \tag{17}$$

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

$$\|\nabla_{\theta} R^{p}_{\alpha}(\theta)\|_2 \geq \|\mathbb{E}[X^{[1]}]\|_2 - \chi\mathbb{E}[\|X^{[1]}\|_2] > 0, \tag{18}$$

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

$$\chi := \begin{cases} \sigma(r\sqrt{d}) - p & \alpha = 1 \\ p\tilde{\chi} - (1-p)\tilde{\chi} & \alpha \in (1,\infty) \\ (1-2p)(1-\sigma'(r\sqrt{d})) & \alpha = \infty, \end{cases} \tag{19}$$

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

The proof of Theorem 3 (in Appendix A.3) is inspired by the Morse landscape analysis in (Sypherd et al., 2019). Intuitively, (19) implies that the RHS in (18) is monotonically *decreasing* in $\alpha \in [1,\infty]$, which aligns with the ordering given by the upper bounds in Theorem 2. Regarding the

assumptions in (16) and (17), they are both more easily satisfied for smaller $r > 0$, indicating alignment with the underlying optimization landscape phenomena. Taken together, Theorems 2 and 3 suggest that larger $\alpha > 1$ are more robust than $\alpha = 1$ (logistic regression); also, notice the $1-2p$ coefficient for $\alpha = \infty$ appearing in both bounds.

5 EXPERIMENTS

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

5.1 Boosting

For the boosting experiments, we utilize the *experiment version* of the Long-Servedio dataset (Long and Servedio, 2008; Cheamanunkul et al., 2014), where the feature vectors are 21D, which differs from the theory version presented in Section 3.2, where the feature vectors are 2D. A full description of the dataset is presented in Appendix B.1. We introduce symmetric label noise in the training data with flip probability $0 < p < 1/2$.

Robustness for simple models In Figure 4, 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 1.1). For lower maximum tree depth[1] 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 1. When the maximum depth is increased, all of the algorithms perform roughly the same (Mansour et al., 2022).

Giving up In Figure 5, 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 16 in Appendix B.1 for consideration on the Long-Servedio dataset. Sypherd et al. (2022a) 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 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.

[1]Increasing the maximum tree depth exponentially increases the number of parameters for the weak learner, which impacts energy consumption, interpretability, and generalization (e.g., via VC dimension).

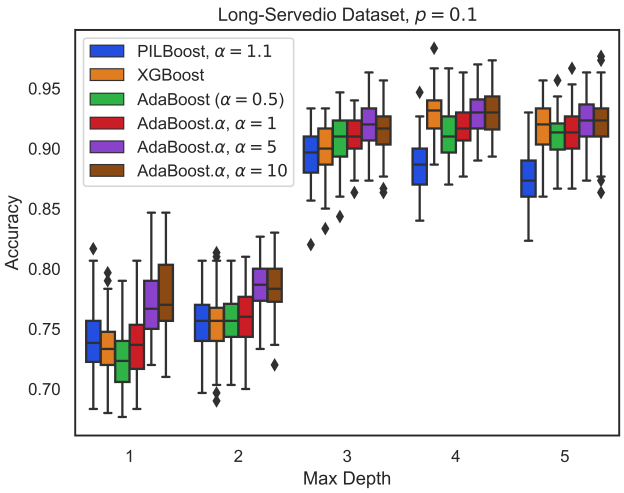


Figure 4: Box and whisker plots of the clean test accuracies of several boosters with 100 decision trees of varying maximum depth on the Long-Servedio dataset for $p = 0.1$ symmetric label noise. The boxes are the interquartile ranges, the lines in the boxes are the medians, and the diamonds are the 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 5.1, and more noise levels are in Appendix B.1.

We see similar behavior in Figure 6 on the breast cancer dataset (Wolberg et al., 1995), namely that for every non-zero level of symmetric label noise, an $\alpha > 1$ is able to achieve greater accuracy on a clean test set, and we note the smoothness of the gains with α, implying that tuning α is simple for this dataset as well. In Appendix B.1, we present full results of AdaBoost.α on the breast cancer dataset, similarly observing gains for smaller maximum tree depths.

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

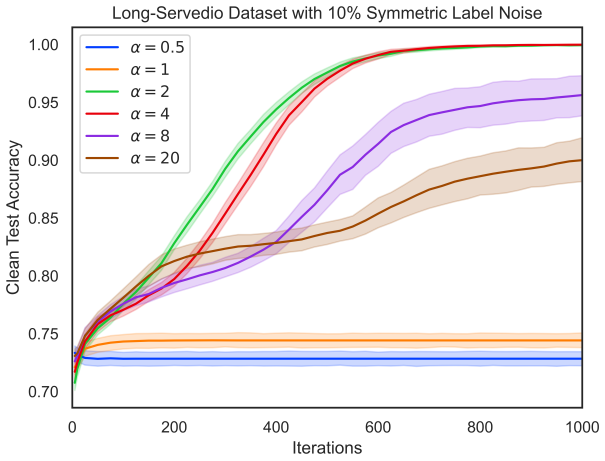


Figure 5: Full version of Figure 1, where we plot clean test accuracies vs. the number of iterations of AdaBoost.α boosting decision stumps for several values of α on the Long-Servedio dataset with $p = 0.1$ symmetric label noise. Note that the solid curves correspond to mean accuracy and shaded areas 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 5.1, and more noise levels are in Appendix B.1.

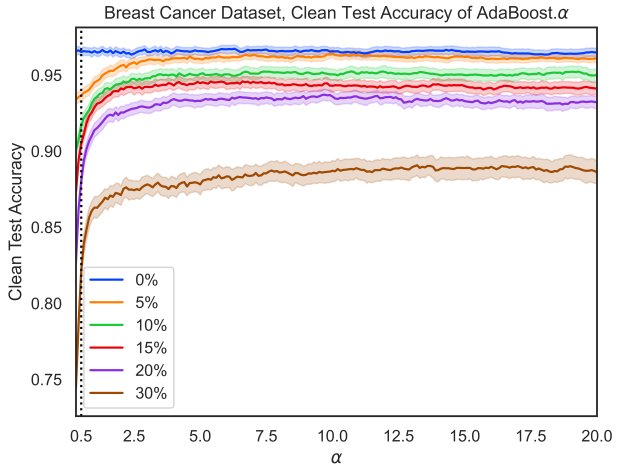


Figure 6: Clean test accuracy of AdaBoost.α with 100 depth 1 decision trees on the breast cancer dataset suffering from various levels of symmetric label noise during training. We see that vanilla AdaBoost ($\alpha = 0.5$) struggles at large noise levels, but simply tuning α larger gives significant performance gains. For lower noise levels, tuning α large does not impact classification accuracy for this dataset, implying that tuning α is simple for this dataset.

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 7, 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 2 and 3. Tuning on this simple dataset is quite easy as the MSE is fairly flat for $\alpha > 1$, see Appendix B.2 for more details.

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 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 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 8, 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, 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 30 in Appendix B we show that the sensitivity of the model increases with increasing α.

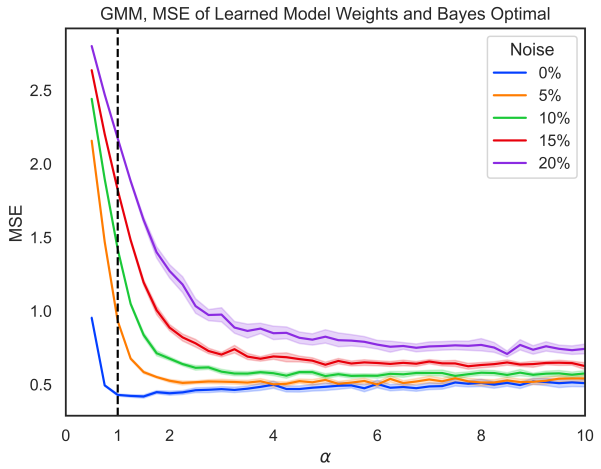


Figure 7: MSE of Bayes optimal line and the parameters learned by α-loss, on a 2D GMM with $86:14$ class imbalance and varying label noise levels. We see that $\alpha > 1$ is able to more closely approximate the clean parameters than $\alpha \leq 1$, and the MSE is fairly flat in the large α regime, 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).

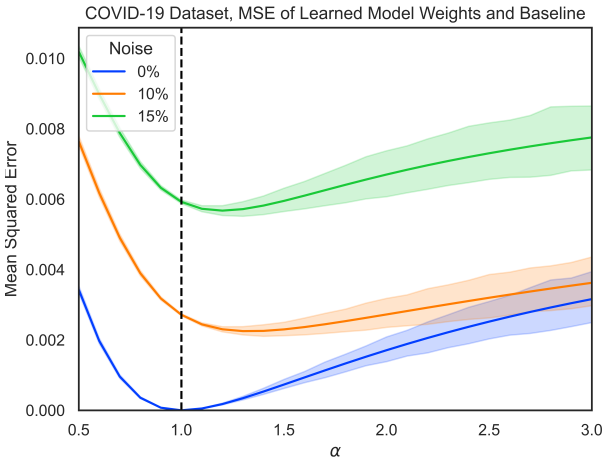


Figure 8: A US COVID-19 survey dataset (Salomon et al., 2021), plotting MSE of logistic regression ($\alpha = 1$) baseline parameters on clean data and model parameters learned using α-loss on noisy data vs. α. For non-zero noise the MSE is minimized for $\alpha > 1$, but some care is required in increasing $\alpha \gg 1$ as the confidence intervals widen, likely due to this being non-realizable and highly imbalanced data. Also, note that we do not consider noise beyond 15% because of stratification, i.e., higher noise would completely overwhelm the class imbalance.

6 CONCLUSION

In this work, we have presented results indicating that the margin-based α-loss is able to "give up" on noisy training data and robustly train simple models. For boosting, we have shown, theoretically and experimentally, how tuning $\alpha > 1$ can address the negative result of Long and Servedio (2008), in the process presenting a novel robust boosting algorithm called AdaBoost.α, which may be of independent interest. For linear models, we have also presented theoretical and experimental results, notably showing robustness for a highly imbalanced COVID-19 survey dataset (Salomon et al., 2021). Additionally, we have presented straightforward tuning characteristics for α in both settings. Lastly, regarding societal impacts, we argue that it is important to consider simple models, since they are more interpretable and have reduced energy cost; we have shown for multiple relevant domains that one can robustly train simple models with a single α hyperparameter.

Acknowledgements

We thank the anonymous reviewers for their comments, and Monica Welfert at Arizona State University for her contributions to the preliminary code. This work is supported in part by NSF grants SCH-2205080, CIF-1901243, CIF-2134256, CIF-2007688, CIF-1815361, a Google AI for Social Good grant, and an Office of Naval Research grant N00014-21-1-2615. This research is based on survey results from Carnegie Mellon University's Delphi Group.

References

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, May 2019.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie 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, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2788613.

Harsha Nori, Rich Caruana, Zhiqi Bu, Judy Hanwen Shen, and Janardhan Kulkarni. Accuracy, interpretability, and differential privacy via explainable boosting. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8227–8237. PMLR, 18–24 Jul 2021.

Zhi Chen, Sarah Tan, Harsha Nori, Kori Inkpen, Yin Lou, and Rich Caruana. Using explainable boosting machines (ebms) to detect common flaws in data. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 534–551, Cham,

2021. Springer International Publishing. ISBN 978-3-030-93736-2.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Paul Viola and Michael 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*, volume 1, pages I–I. Ieee, 2001.

Shai Ben-David, David Loker, Nathan Srebro, and Karthik Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. *arXiv preprint arXiv:1206.6442*, 2012.

Ziwei Ji, Kwangjun Ahn, Pranjal Awasthi, Satyen Kale, and Stefani Karp. Agnostic learnability of halfspaces via logistic loss. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10068–10103. PMLR, 17–23 Jul 2022.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.

Philip M Long and Rocco A Servedio. Random classification noise defeats all convex potential boosters. In *Proceedings of the 25th international conference on Machine learning*, pages 608–615, 2008.

Yishay Mansour, Richard Nock, and Robert C. Williamson. What killed the convex booster ?, 2022. URL `https://arxiv.org/abs/2205.09628`.

Tyler Sypherd, Mario Diaz, John Kevin Cava, Gautam Dasarathy, Peter Kairouz, and Lalitha Sankar. A tunable loss function for robust classification: Calibration, landscape, and generalization. *IEEE Transactions on Information Theory*, 68(9):6021–6051, 2022a. doi: 10.1109/TIT.2022.3169440.

Robert E Schapire and Yoav Freund. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

Joshua A. Salomon, Alex Reinhart, Alyssa Bilinski, Eu Jing Chua, Wichada La Motte-Kerr, Minttu M. Rönn, Marissa B. Reitsma, Katherine A. Morris, Sarah LaRocca, Tamer H. Farag, Frauke Kreuter, Roni Rosenfeld, and Ryan 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. ISSN 0027-8424. doi: 10.1073/pnas.2111454118.

Weizhi Li, Gautam Dasarathy, and Visar Berisha. Regularization via structural label smoothing. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1453–1463. PMLR, 26–28 Aug 2020.

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

Garth H. Rauscher, Timothy P. Johnson, Young Ik Cho, and Jennifer A. Walk. Accuracy of Self-Reported Cancer-Screening Histories: A Meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention*, 17(4):748–757, 04 2008. ISSN 1055-9965. doi: 10.1158/1055-9965.EPI-07-2629.

Sarah Connor Gorber, Sean Schofield-Hurwitz, Jill Hardt, Geneviève Levasseur, and Mark 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, 01 2009. ISSN 1462-2203. doi: 10.1093/ntr/ntn010.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26:1196–1204, 2013.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pages 6543–6553. PMLR, 2020.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pages 6226–6236. PMLR, 2020.

Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen 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*, pages 1944–1952, 2017.

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient l˜ 1 regularized logistic regression. In *Aaai*, volume 6, pages 401–408, 2006.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Xiaojie Shi, Shuyang Cheng, and Dragomir Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. *arXiv preprint arXiv:2204.12511*, 2022.

Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*, 2019.

Liu Ziyin, Blair Chen, Ru Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Learning not to learn in the presence of noisy labels. *arXiv preprint arXiv:2002.06541*, 2020.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. *Advances in Neural Information Processing Systems*, 29, 2016.

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403, 2021.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

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

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Eyal Beigman and Beata Beigman 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*, pages 280–287, 2009.

Naresh Manwani and PS Sastry. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3): 1146–1151, 2013.

Tan Nguyen and Scott Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *International Conference on Machine Learning*, pages 1085–1093, 2013.

Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4331–4339, 2019.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Lei Zhao, Musa Mammadov, and John Yearwood. From convex to nonconvex: a loss function analysis for binary classification. In *2010 IEEE International Conference on Data Mining Workshops*, pages 1281–1288. IEEE, 2010.

T. Sypherd, M. Diaz, L. Sankar, and P. Kairouz. A tunable loss function for binary classification. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2479–2483, July 2019. doi: 10.1109/ISIT.2019.8849796.

Olivier Chapelle, Choon Teo, Quoc Le, Alex Smola, et al. Tighter bounds for structured estimation. *Advances in neural information processing systems*, 21, 2008.

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

Sunsern Cheamanunkul, Evan Ettinger, and Yoav Freund. Non-convex boosting overcomes random label noise. *arXiv preprint arXiv:1409.2905*, 2014.

Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and SavageBoost. In *Advances in Neural Information Processing Systems*, pages 1049–1056, 2009.

Seokho Lee, Hyejin Shin, and Sang Han Lee. Label-noise resistant logistic regression for functional data classification with an application to alzheimer's disease study. *Biometrics*, 72(4):1325–1335, 2016.

Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34:4409–4420, 2021.

Alina E Maas, Franz Rottensteiner, and Christian Heipke. A label noise tolerant random forest for the classification of remote sensing data based on outdated maps for training. *Computer Vision and Image Understanding*, 188: 102782, 2019.

Jakramate Bootkrajang and Ata Kabán. Label-noise robust logistic regression and its applications. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part I 23*, pages 143–158. Springer, 2012.

Jiachun Liao, Oliver Kosut, Lalitha Sankar, and Flavio P Calmon. A tunable measure for information leakage. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 701–705. IEEE, 2018.

Suguru Arimoto. Information-theoretical considerations on estimation problems. *Information and control*, 19(3): 181–194, 1971.

T. Sypherd, 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)*, pages 2700–2705, 2020. doi: 10.1109/ISIT44484.2020.9174356.

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

Tyler Sypherd, Richard Nock, and Lalitha Sankar. Being properly improper. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20891–20932. PMLR, 17–23 Jul 2022b.

Gowtham R. Kurri, Tyler Sypherd, and Lalitha Sankar. Realizing gans via a tunable loss function. In *2021 IEEE Information Theory Workshop (ITW)*, pages 1–6, 2021. doi: 10.1109/ITW48936.2021.9611499.

Gowtham R. Kurri, Monica Welfert, Tyler Sypherd, and Lalitha Sankar. alpha-gan: Convergence and estimation guarantees. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 276–281, 2022. doi: 10.1109/ISIT50566.2022.9834890.

Matteo Zecchin, Sangwoo Park, Osvaldo Simeone, Marios Kountouris, and David Gesbert. Robust pac-m: Training ensemble models under model misspecification and outliers. *arXiv preprint arXiv:2203.01859*, 2022.

Yoav Freund and Robert 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, 1997.

Michael Collins, Robert E Schapire, and Yoram Singer. Logistic regression, adaboost and bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

Ross A McDonald, David J Hand, and Idris A Eckley. An empirical comparison of three boosting algorithms on real data sets with artificial class noise. In *International Workshop on Multiple Classifier Systems*, pages 35–44. Springer, 2003.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

R. Nock and R.-C. Williamson. Lossless or quantized boosting with integer arithmetic. pages 4829–4838, 2019.

Mark D Reid and Robert C Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11: 2387–2422, 2010.

Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems*, 28, 2015.

Philip M Long and Rocco 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.

Qiguang Miao, Ying Cao, Ge Xia, Maoguo Gong, Jiachen Liu, and Jianfeng Song. Rboost: Label noise-robust boosting algorithm based on a nonconvex loss function and the numerically stable base learners. *IEEE transactions on neural networks and learning systems*, 27(11): 2216–2228, 2015.

Shai Ben-David, Nadav Eiron, and Philip M Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

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

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.

Jeff Cooper and Lev Reyzin. Improved algorithms for distributed boosting. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 806–813. IEEE, 2017.

Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multiclass adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

Dr. William Wolberg, Nick Street, and Olvi Mangasarian. Breast cancer wisconsin (diagnostic) data set, 1995.

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

Matus Telgarsky. Boosting with the logistic loss is consistent. *arXiv preprint arXiv:1305.2648*, 2013.

Morris Kline. *Calculus: an intuitive and physical approach*. Courier Corporation, 1998.

Appendices

A Further Theoretical Results, Commentary, and Proofs

Classification-Calibration Regarding the statistical efficacy of $\tilde{l}^{\alpha}$, Sypherd et al. (2022a) showed that the margin-based α-loss is classification-calibrated for all $\alpha \in (0, \infty]$, which is a necessary minimum condition for a "good" margin-based loss function to satisfy. In words, a margin-based loss function is classification-calibrated if for each feature vector, the minimizer of its conditional risk agrees in sign with the Bayes optimal predictor; this is a pointwise form of Fisher consistency from the perspective of classification (Lin, 2004; Bartlett et al., 2006).

Lemma 1. *For $\alpha \in (0, \infty]$, the first derivative of $\tilde{l}^{\alpha}$ with respect to the margin is given by*

$$\tilde{l}^{\alpha\prime}(z) := \frac{d}{dz}\tilde{l}^{\alpha}(z) = -\sigma'(z)\sigma(z)^{-\frac{1}{\alpha}}, \tag{20}$$

its second derivative is given by

$$\tilde{l}^{\alpha\prime\prime}(z) := \frac{d^2}{dz^2}\tilde{l}^{\alpha}(z) = \frac{e^z\left(\alpha e^z - \alpha + 1\right)}{\alpha(e^{-z}+1)^{-\frac{1}{\alpha}}(e^z+1)^3}, \tag{21}$$

and its third derivative is given by

$$\tilde{l}^{\alpha\prime\prime\prime}(z) := \frac{d^3}{dz^3}\tilde{l}^{\alpha}(z) = \frac{-e^{2z} + 4e^z - 1 - \frac{3e^z-2}{\alpha} - \frac{1}{\alpha^2}}{e^{-z}\left(1+e^{-z}\right)^{-\frac{1}{\alpha}}\left(e^z+1\right)^4}. \tag{22}$$

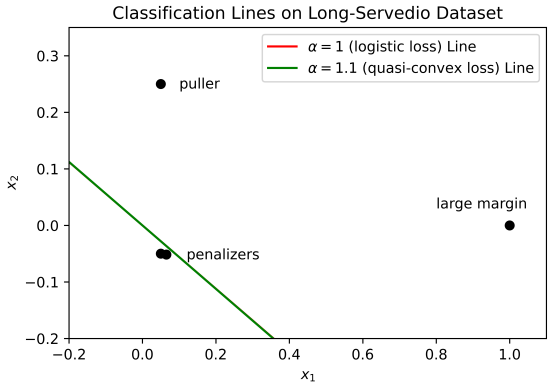
Discussion of Algorithm 1 The weighting used for the weak learner in Algorithm 1, namely that $\theta_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$, is the expression commonly used in vanilla AdaBoost ($\alpha = 1/2$ for AdaBoost.α) (Schapire and Freund, 2013). However, there are several 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., 2022a). 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 1, is left for future work.

A.1 Proof of Theorem 1

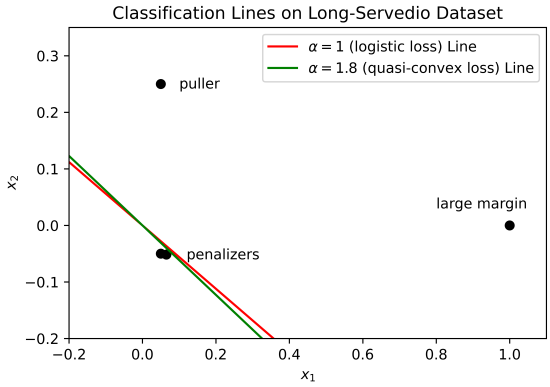
The strategy of the proof is as follows:

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, 2008), 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 1).
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.
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.

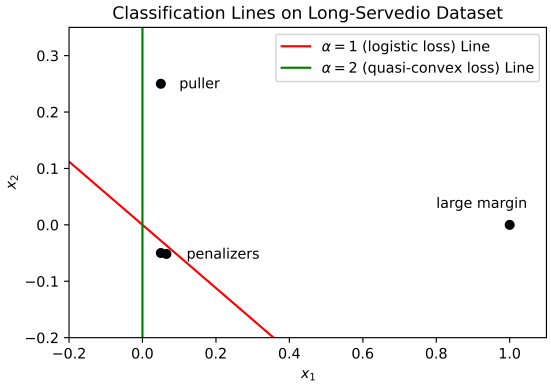
By the construction of the hypothesis class (Long and Servedio, 2008), namely that $\mathcal{H} = \{h_1(\mathbf{x}) = x_1, h_2(\mathbf{x}) = x_2\}$, notice that the classification lines (constructed by the boosting algorithm in this pathological example) are given by $\theta_1 x_1 + \theta_2 x_2 = 0$ 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 3) that the "large margin" example forces $\theta_1^* > 0$. Further,



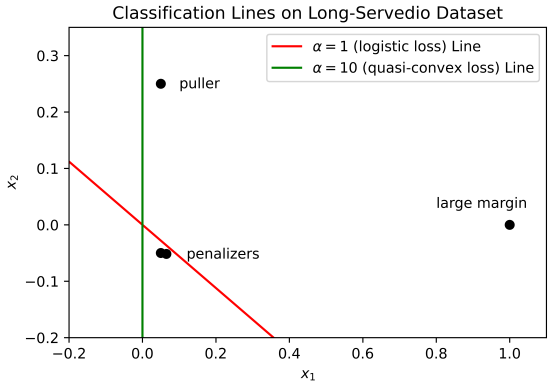
(a) $\alpha = 1.1$ classification line.



(b) $\alpha = 1.8$ classification line.



(c) $\alpha = 2$ classification line.



(d) $\alpha = 10$ classification line.

Figure 9: Companion figure of Figure 3 where $N = 2$ ($p = 1/3$) and $\gamma = 1/20$ for $\alpha \in \{1.1, 1.8, 2, 10\}$.

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:

$$\theta_1^* > 0 \quad \text{and} \quad \theta_1^* > \theta_2^* \quad \text{and} \quad \theta_1^* > -5\theta_2^*. \tag{23}$$

We now examine the solutions to the first-order equation for $\alpha \in (0, \infty]$.

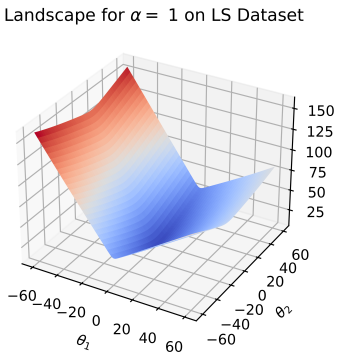
As in (Long and Servedio, 2008), let $1 < N < \infty$ be the noise parameter such that the noise rate $p = \frac{1}{N+1}$, and hence $1 - p = \frac{N}{N+1}$. Under the Long-Servedio setup with the margin-based α-loss (and recalling that all four examples have classification label $y = 1$), we have that

$$R_\alpha^p(\theta_1, \theta_2) = \frac{1}{4} \sum_{x \in S} \left[(1-p)\tilde{l}^\alpha(\theta_1 x_1 + \theta_2 x_2) + p\tilde{l}^\alpha(-\theta_1 x_1 - \theta_2 x_2) \right]. \tag{24}$$

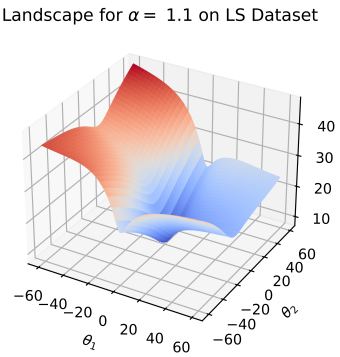
It is clear that minimizing $4(N+1)R_\alpha^p$ is the same as minimizing R_α^p so we shall henceforth work with $4(N+1)R_\alpha^p$ since it gives rise to cleaner expressions. We have that

$$\begin{aligned} 4(N+1)R_\alpha^p(\theta_1, \theta_2) &= \sum_{x \in S} [N\tilde{l}^\alpha(\theta_1 x_1 + \theta_2 x_2) + \tilde{l}^\alpha(-\theta_1 x_1 - \theta_2 x_2)] \qquad (25) \\ &= N\tilde{l}^\alpha(\theta_1) + \tilde{l}^\alpha(-\theta_1) + 2N\tilde{l}^\alpha(\theta_1\gamma - \theta_2\gamma) + 2\tilde{l}^\alpha(-\theta_1\gamma + \theta_2\gamma) \\ &\quad + N\tilde{l}^\alpha(\theta_1\gamma + 5\theta_2\gamma) + \tilde{l}^\alpha(-\theta_1\gamma - 5\theta_2\gamma). \qquad (26) \end{aligned}$$

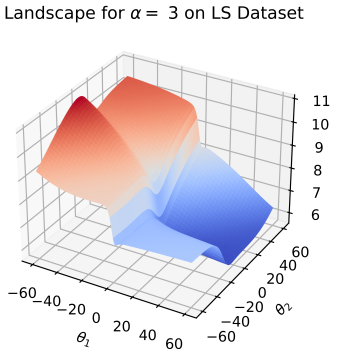
See Figure 10 for a visualization of (26).



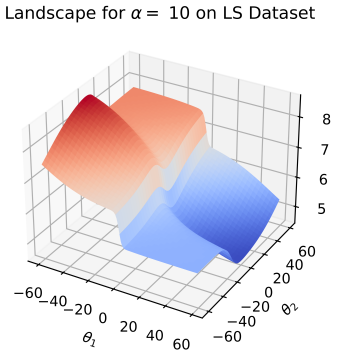
(a) $\alpha = 1$ Optimization Landscape.



(b) $\alpha = 1.1$ Optimization Landscape.



(c) $\alpha = 3$ Optimization Landscape.



(d) $\alpha = 10$ Optimization Landscape.

Figure 10: Plots of optimization landscapes on the Long-Servedio dataset, i.e. (26), for $\alpha \in \{1, 1.1, 3, 10\}$. Aligning with Figure 3, $N = 2$ 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 (23)) is significantly better for $\alpha = 3$. For $\alpha = 10$, the landscape strongly resembles the $\alpha = 3$, but is "flatter".

Again following notation in (Long and Servedio, 2008), let $P_1^\alpha(\theta_1, \theta_2)$ and $P_2^\alpha(\theta_1, \theta_2)$ be defined as follows:

$$P_1^\alpha(\theta_1, \theta_2) := \frac{\partial}{\partial \theta_1} 4(N+1) R_\alpha^p(\theta_1, \theta_2) \quad \text{and} \quad P_2^\alpha(\theta_1, \theta_2) := \frac{\partial}{\partial \theta_2} 4(N+1) R_\alpha^p(\theta_1, \theta_2). \tag{27}$$

Thus, differentiating (26) by θ_1 and θ_2 respectively, we have

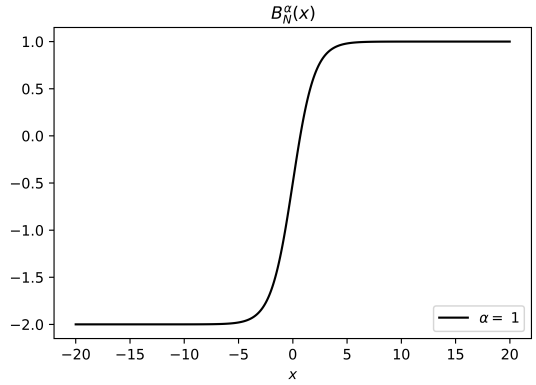
$$\begin{aligned} P_1^\alpha(\theta_1, \theta_2) = N\tilde{l}^{\alpha'}(\theta_1) - \tilde{l}^{\alpha'}(-\theta_1) + 2\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma - \theta_2\gamma) \\ - 2\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma + \theta_2\gamma) + N\gamma\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma) - \gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma), \end{aligned} \tag{28}$$

and

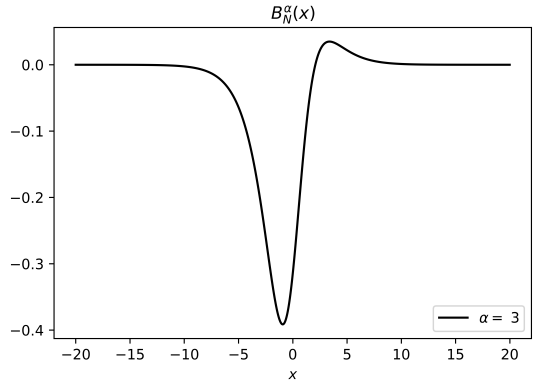
$$P_2^\alpha(\theta_1, \theta_2) = -2\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma - \theta_2\gamma) + 2\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma + \theta_2\gamma) + 5\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma) - 5\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma). \tag{29}$$

In order to reason about the quality of the solutions to (26) 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

$$\begin{aligned} N\tilde{l}^{\alpha'}(\theta_1) + 2\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma - \theta_2\gamma) + N\gamma\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma) \\ = \tilde{l}^{\alpha'}(-\theta_1) + 2\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma + \theta_2\gamma) + \gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma), \end{aligned} \tag{30}$$



(a) $B_N^\alpha(x)$ for $\alpha = 1$.



(b) $B_N^\alpha(x)$ for $\alpha = 3$.

Figure 11: 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 <u>not</u> non-decreasing. One can also see other properties of B_N^α in figure (b) as articulated in Lemma 2.

and rewriting $P_2^\alpha(\theta_1, \theta_2) = 0$, we obtain

$$2\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma + \theta_2\gamma) + 5\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma) = 2\gamma N\tilde{l}^{\alpha'}(\theta_1\gamma - \theta_2\gamma) + 5\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma). \tag{31}$$

Substituting (31) into (30), we are able to cancel a term and recover

$$N\tilde{l}^{\alpha'}(\theta_1) - \tilde{l}^{\alpha'}(-\theta_1) = 6\gamma\tilde{l}^{\alpha'}(-\theta_1\gamma - 5\theta_2\gamma) - 6N\gamma\tilde{l}^{\alpha'}(\theta_1\gamma + 5\theta_2\gamma). \tag{32}$$

Rewriting, we obtain

$$N\tilde{l}^{\alpha'}(\theta_1) - \tilde{l}^{\alpha'}(-\theta_1) = -6\gamma\left[N\tilde{l}^{\alpha'}(\gamma(\theta_1 + 5\theta_2)) - \tilde{l}^{\alpha'}(-\gamma(\theta_1 + 5\theta_2))\right]. \tag{33}$$

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 1, 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

$$B_N^\alpha(x) = N\left(-\sigma'(x)\sigma(x)^{-1/\alpha}\right) - \left(-\sigma'(-x)\sigma(-x)^{-1/\alpha}\right) \tag{34}$$

$$= \sigma'(x)\sigma(-x)^{-1/\alpha} - N\sigma'(x)\sigma(x)^{-1/\alpha} \tag{35}$$

$$= \sigma'(x)\left(\sigma(-x)^{-1/\alpha} - N\sigma(x)^{-1/\alpha}\right). \tag{36}$$

Using this, we can rewrite (33) as

$$B_N^\alpha(\theta_1) = -6\gamma B_N^\alpha(\gamma(\theta_1 + 5\theta_2)), \tag{37}$$

which is equivalent to

$$\sigma'(\theta_1)\left(\sigma(-\theta_1)^{-1/\alpha} - N\sigma(\theta_1)^{-1/\alpha}\right) = -6\gamma\sigma'(\gamma(\theta_1 + 5\theta_2))\left(\sigma(-\gamma(\theta_1 + 5\theta_2))^{-1/\alpha} - N\sigma(\gamma(\theta_1 + 5\theta_2))^{-1/\alpha}\right), \tag{38}$$

and both quantify solutions (θ_1^*, θ_2^*). Notice that it is unfortunately not possible to analytically reduce (38) 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 (23)), because we can utilize nice properties of B_N^α. For instance, one key thing to notice in (37) 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).

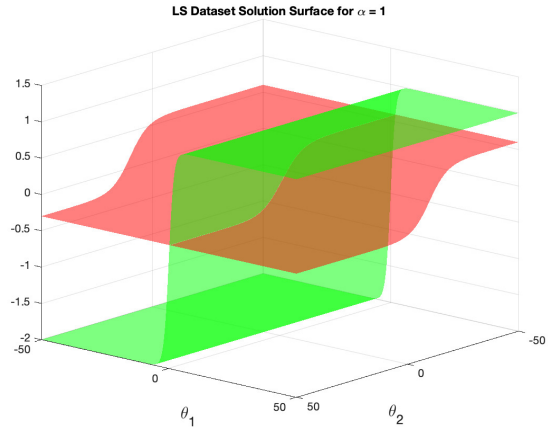
To this end, we take a detour from the main thread to aggregate some nice properties of B_N^α for $\alpha > 1$. See Figure 11 for a plot of B_N^α.

Lemma 2. *Consider for $\alpha \in (0, \infty]$ and $1 < N < \infty$,*

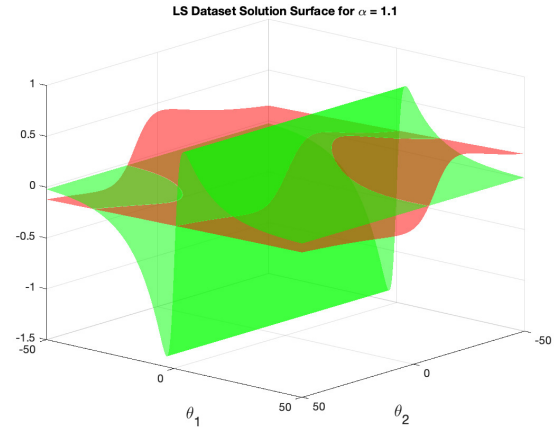
$$B_N^\alpha(x) := \sigma'(x) \left(\sigma(-x)^{-1/\alpha} - N\sigma(x)^{-1/\alpha} \right), \tag{39}$$

where $x \in \mathbb{R}$. The following are properties of B_N^α:

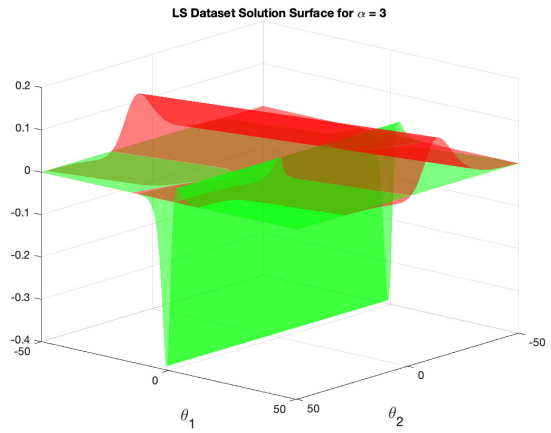
1. *For $\alpha \leq 1$, $B_N^\alpha(x)$ is non-decreasing in x.*
2. *For $\alpha > 1$, $B_N^\alpha(x)$ is not non-decreasing in x.*
3. *Note that $\lim_{\alpha \to \infty} B_N^\alpha(x) = \sigma'(x)(1 - N)$.*
4. *For $\alpha > 1$, $\lim_{x \to +\infty} B_N^\alpha(x) \to 0^+$ and $\lim_{x \to -\infty} B_N^\alpha(x) \to 0^-$.*
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 $x > \alpha \ln N$.*



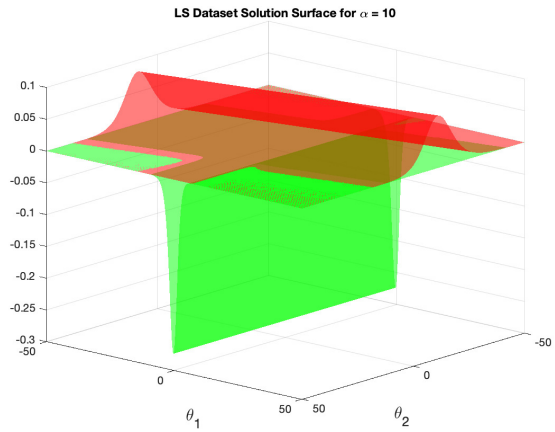
(a) Plot of LHS (green) and RHS (red) of (40) for $\alpha = 1$.



(b) Plot of LHS (green) and RHS (red) of (40) for $\alpha = 1.1$.



(c) Plot of LHS (green) and RHS (red) of (40) for $\alpha = 3$.



(d) Plot of LHS (green) and RHS (red) of (40) for $\alpha = 10$.

Figure 12: Plots of LHS (green) and RHS (red) of (40) for $\alpha \in \{1, 1.1, 3, 10\}$, and $N = 2$ and $\gamma = 1/20$. The intersections of the surfaces indicate solutions of (40). One can see that the solutions for $\alpha = 1$ are not "good" in the sense of (23) because θ_1 is small and fixed; this phenomenon was proved by (Long and Servedio, 2008) 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.

The proof of the first property is obtained by invoking one of the results of Long and Servedio (2008) 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 2, we want to reason about the solutions of (37), i.e.,

$$B_N^\alpha(\theta_1) = -6\gamma B_N^\alpha(\gamma(\theta_1 + 5\theta_2)), \tag{40}$$

as a function of $\alpha \in (0, \infty]$. From Propositions 1 and (Sypherd et al., 2022a), 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, 2008), 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 (40) 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 (40) since via the third property in Lemma 2, (40) reduces to

$$\sigma'(\theta_1)(1 - N) = -6\gamma_{\hat{\alpha}}\sigma'(\gamma_{\hat{\alpha}}(\theta_1 + 5\theta_2))(1 - N), \tag{41}$$

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 (41) look like mirrored $\sigma'(x)$ type functions.

Now, we consider $\alpha \in (1, \infty)$ in (40), which is the key region of α for the proof. Examining the LHS of (40), i.e. $B_N^\alpha(\theta_1)$, we note from the fourth property of Lemma 2 that $\lim_{\theta_1 \to +\infty} B_N^\alpha(\theta_1) \to 0^+$. Furthermore, we note via the sixth property in Lemma 2 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 (40), 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 2, we have that $\lim_{\theta_1 \to \infty} -6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}\theta_1) \to 0^-$. Furthermore, we note via the sixth property in Lemma 2 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.

Taking the limit and crossover behaviors in θ_1 of $B_N^\alpha(\theta_1)$ (the LHS of (40)) and $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}\theta_1)$ (the reduced RHS of (40)) together, we have *by continuity* that there must exist some $\tilde{\theta}_1 > 0$ which satisfies

$$B_N^\alpha(\tilde{\theta}_1) = -6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}\tilde{\theta}_1), \tag{42}$$

for each $\alpha \in (1, \infty)$.

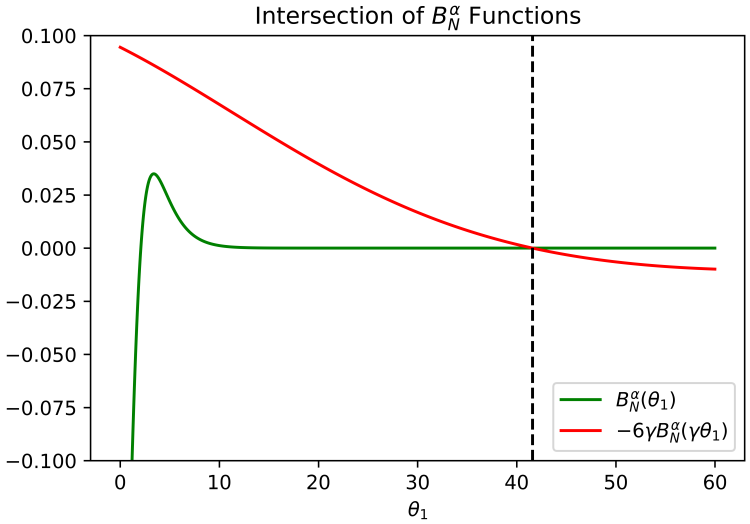


Figure 13: A plot of LHS (green) and RHS (red) of (42) for $\alpha = 3$, where $N = 2$, $\gamma = 1/20$. Notice that the intersection point $\tilde{\theta}_1$ (dotted line) is *very* close to the crossover point $\frac{\alpha \ln N}{\gamma_{\hat{\alpha}}} \approx \frac{3 \times 0.69}{1/20} \approx 41.59$, and also notice that this solution nicely coincides with the grid-search solution presented in Figure 3.

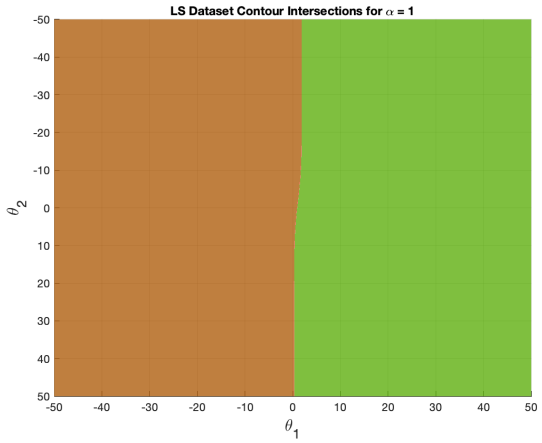
Furthermore, the choice of $\alpha \in (1, \infty)$ directly influences the magnitude of $\tilde{\theta}_1 > 0$, with larger α increasing the value of $\tilde{\theta}_1$ because of the crossover points, particularly that we require $\tilde{\theta}_1 > \frac{\alpha \ln N}{\gamma_{\hat{\alpha}}}$, which 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 13 for a plot.

Therefore, for each $\alpha \in (1, \infty)$, there exists a solution $(\theta_1^\alpha, \theta_2^\alpha)$ to (40), 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 (23) and thus has perfect classification accuracy on the *clean* LS dataset.

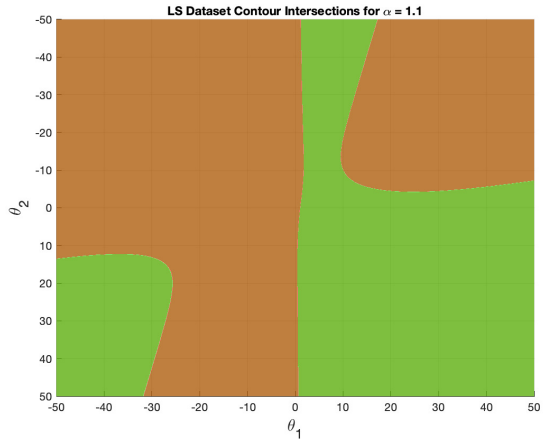
Next, while not necessary for the proof of Theorem 1, we also argue for the existence of other optima near $(\theta_1^\alpha, \theta_2^\alpha)$. Reconsidering the full (with θ_2 included) expression, $-6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}(\theta_1 + 5\theta_2))$ in (40), we take $\alpha \in (1, \infty)$ large enough in (42) 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 2). 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

$$B_N^\alpha(\theta_1^*) = -6\gamma_{\hat{\alpha}} B_N^\alpha(\gamma_{\hat{\alpha}}(\theta_1^* + 5\theta_2^*)), \tag{43}$$

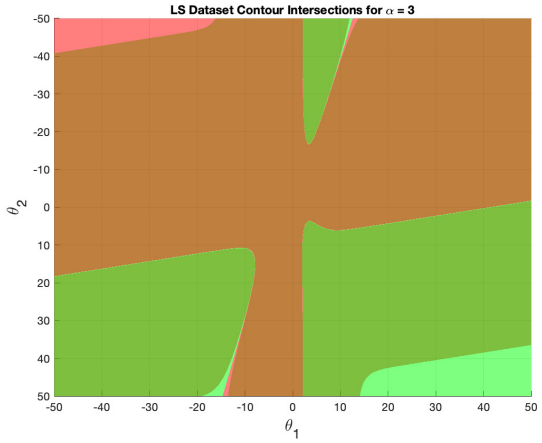
which (might) exist by continuity. See Figure 14 for a plot; intuitively, the fact that the LHS and RHS of (42) intersect, not merely "touch", suggests the existence of (θ_1^*, θ_2^*), indeed a "strip" of good solutions.



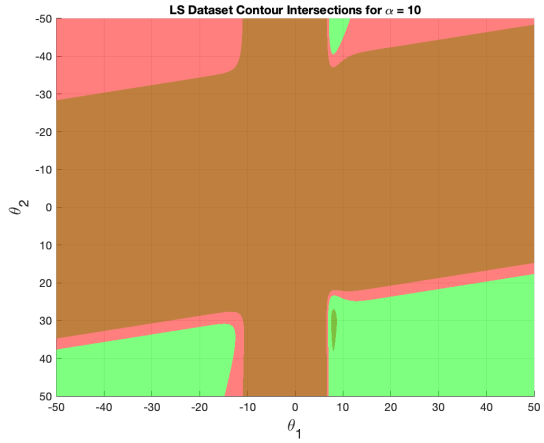
(a) Rotated version of Figure 12a.



(b) Rotated version of Figure 12b.



(c) Rotated version of Figure 12c.



(d) Rotated version of Figure 12d.

Figure 14: Companion figures of Figure 12 for $\alpha \in \{1, 1.1, 3, 10\}$, and $N = 2$ and $\gamma = 1/20$. The contours indicate solutions of (40). In Figure 14c, one can see a contour of "good" LS solutions near where $\theta_1 \approx 41.59$ and θ_2 is very small.

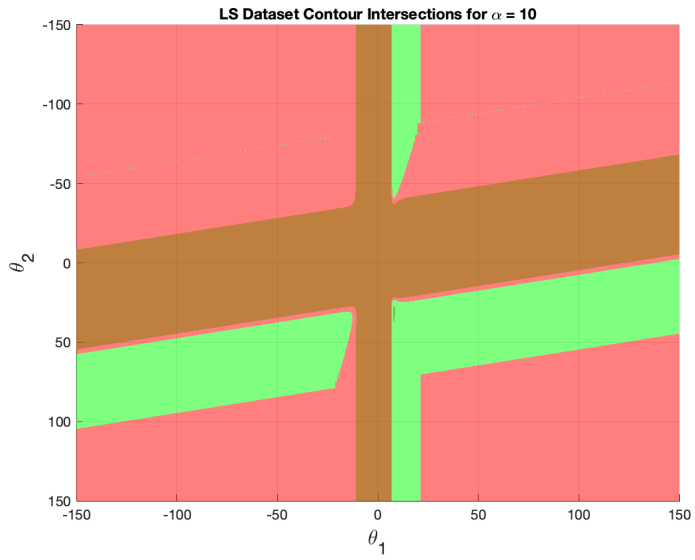


Figure 15: Companion figure of Figure 14d, again for $\alpha = 10$, where the parameter space has been *increased*. One can again see "good" LS solutions for large θ_1 and small θ_2. This is indicative of a trade off between the value of α and the range of the parameter space for the LS dataset.

A.2 Proof of Theorem 2

In this section, we provide the proof of Theorem 2. First, however, we provide lemmas useful in the proof of Theorem 2, which indicate useful bounds for $\alpha = 1$ and ∞, and their respective proofs.

Lemma 3. *For all $z \in \mathbb{R}$, we have that*

$$\left|\frac{d^2}{dz^2}\tilde{l}^1(z)\right| \geq \left|\frac{d^2}{dz^2}\tilde{l}^\infty(z)\right|, \tag{44}$$

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

$$\left|\frac{d^2}{dz^2}\tilde{l}^1(z)\right| = \left|\frac{d^2}{dz^2}\tilde{l}^\infty(z)\right| \tag{45}$$

$$\left|\frac{e^z}{(e^z+1)^2}\right| = \left|\frac{e^z(e^z-1)}{(e^z+1)^3}\right| \tag{46}$$

$$e^z = \left|\frac{e^z(e^z-1)}{e^z+1}\right|, \tag{47}$$

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

$$e^z \geq e^z - 1, \tag{48}$$

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 4. *For $|z| > \ln(2)$, we have that*

$$\left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right| \leq \left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right|. \tag{49}$$

Proof. Consider

$$\left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right| = \left|\frac{e^z - e^{2z}}{(e^z+1)^3}\right| \tag{50}$$

and

$$\left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right| = \left|\frac{-e^{3z} + 4e^{2z} - e^z}{(e^z+1)^4}\right|. \tag{51}$$

Setting

$$\left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right| = \left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right|, \tag{52}$$

after some algebra, we find that $z^* = \pm\ln(2)$. Furthermore, considering the large $z > 0$ regime, we find that

$$\left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right| \overset{?}{\leq} \left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right| \tag{53}$$

$$\left|\frac{-e^{3z} + 4e^{2z} - e^z}{(e^z+1)^4}\right| \overset{?}{\leq} \left|\frac{e^z - e^{2z}}{(e^z+1)^3}\right| \tag{54}$$

$$\frac{e^{3z} - 4e^{2z} + e^z}{e^z+1} \overset{?}{\leq} e^{2z} - e^z \tag{55}$$

$$e^{2z} - 4e^z \leq e^{2z} - e^z, \tag{56}$$

thus by the IVT and symmetry, we have the desired result. □

With Lemmas 3 and 4 in hand, we now present the proof of Theorem 2.

Recall from (11) that for $\alpha \in (0, \infty]$

$$\nabla_\theta \tilde{l}^\alpha(\langle YX, \theta\rangle) = -\sigma(\langle YX,\theta\rangle)^{1-\frac{1}{\alpha}}\sigma(-\langle YX,\theta\rangle)YX = \tilde{l}^{\alpha'}(\langle YX,\theta\rangle)YX, \tag{57}$$

since for each $i \in [d]$, $\dfrac{\partial}{\partial\theta_i}\tilde{l}^\alpha(\langle YX,\theta\rangle) = \tilde{l}^{\alpha'}(\langle YX,\theta\rangle)YX_i$.

Hence, the gradient of the noisy α-risk from (12) is

$$\nabla_\theta R^p_\alpha(\theta) = \mathbb{E}_{X,Y}\left[(1-p)\nabla_\theta\tilde{l}^\alpha(\langle YX,\theta\rangle) + p\nabla_\theta\tilde{l}^\alpha(\langle -YX,\theta\rangle)\right] \tag{58}$$

$$= \mathbb{E}_{X,Y}\left[\left((1-p)\tilde{l}^{\alpha'}(\langle YX,\theta\rangle) - p\tilde{l}^{\alpha'}(\langle -YX,\theta\rangle)\right)YX\right], \tag{59}$$

where we expanded the expression for clarity. Notice that for $\alpha = 1$ (from Lemma 1),

$$\tilde{l}^{1'}(-z) = -\tilde{l}^{1'}(z) - 1, \tag{60}$$

namely that $\tilde{l}^{1'}$ is *almost* an **odd** function, and for $\alpha = \infty$,

$$\tilde{l}^{\infty'}(-z) = \tilde{l}^{\infty'}(z), \tag{61}$$

namely that $\tilde{l}^{\infty'}$ is an **even** function.

Thus, we have by the definition of $\hat{\theta}^1$ and $\hat{\theta}^\infty$ that for $\alpha = 1$

$$\mathbf{0} = \nabla_\theta R^p_1(\hat{\theta}^1) = \mathbb{E}_{X,Y}\left[\left((1-p)\tilde{l}^{1'}(\langle YX,\hat{\theta}^1\rangle) - p\tilde{l}^{1'}(\langle -YX,\hat{\theta}^1\rangle)\right)YX\right] \tag{62}$$

$$= (1-p)\mathbb{E}_{X,Y}\left[\tilde{l}^{1'}(\langle YX,\hat{\theta}^1\rangle)YX\right] - p\mathbb{E}_{X,Y}\left[\tilde{l}^{1'}(\langle -YX,\hat{\theta}^1\rangle)YX\right] \tag{63}$$

$$= (1-p)\mathbb{E}_{X,Y}\left[\tilde{l}^{1'}(\langle YX,\hat{\theta}^1\rangle)YX\right] - p\mathbb{E}_{X,Y}\left[\left(-\tilde{l}^{1'}(\langle YX,\hat{\theta}^1\rangle) - 1\right)YX\right] \tag{64}$$

$$= \mathbb{E}_{X,Y}\left[\tilde{l}^{1'}(\langle YX,\hat{\theta}^1\rangle)YX\right] + p\mathbb{E}_{X,Y}[YX], \tag{65}$$

and for $\alpha = \infty$

$$\mathbf{0} = \nabla_\theta R^p_\infty(\hat{\theta}^\infty) = \mathbb{E}_{X,Y}\left[\left((1-p)\tilde{l}^{\infty'}(\langle YX, \hat{\theta}^\infty\rangle) - p\tilde{l}^{\infty'}(\langle -YX, \hat{\theta}^\infty\rangle)\right)YX\right] \tag{66}$$

$$= (1-p)\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty'}(\langle YX, \hat{\theta}^\infty\rangle)YX\right] - p\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty'}(\langle -YX, \hat{\theta}^\infty\rangle)YX\right] \tag{67}$$

$$= (1-2p)\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty'}(\langle YX, \hat{\theta}^\infty\rangle)YX\right]. \tag{68}$$

And, thus we have that for each $i \in [d]$,

$$\mathbb{E}_{X,Y}\left[\tilde{l}^{1'}(\langle YX, \hat{\theta}^1\rangle)YX_i\right] + p\mathbb{E}_{X,Y}[YX_i] = 0, \tag{69}$$

and

$$\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty'}(\langle YX, \hat{\theta}^\infty\rangle)YX_i\right] = 0. \tag{70}$$

In order to evaluate the efficacy of the gradient of the noisy α-risk at recovering any data generating vector $\theta^* \in \mathbb{B}_d(r)$, we seek to upper bound $\|\nabla_\theta R^p_1(\theta^*)\|_\infty$ and $\|\nabla_\theta R^p_\infty(\theta^*)\|_\infty$. To this end, recall the Taylor-Lagrange equality (Kline, 1998) for a twice continuously differentiable $f : \mathbb{R} \to \mathbb{R}$,

$$f(b) = f(a) + (b-a)f'(a) + \frac{(b-a)^2}{2}f''(c), \tag{71}$$

where $c \in [a, b]$.

Let $i \in [d]$ be arbitrary, but fixed. From (59) (and the reductions from (65) and (68)) we have that at $\theta^* \in \mathbb{B}_d(r)$

$$\frac{\partial}{\partial \theta_i}R^p_1(\theta^*) = \mathbb{E}_{X,Y}\left[\tilde{l}^{1'}(\langle YX, \theta^*\rangle)YX_i\right] + p\mathbb{E}_{X,Y}[YX_i], \tag{72}$$

and

$$\frac{\partial}{\partial \theta_i}R^p_\infty(\theta^*) = (1-2p)\mathbb{E}_{X,Y}\left[\tilde{l}^{\infty'}(\langle YX, \theta^*\rangle)YX_i\right]. \tag{73}$$

Using the Taylor-Lagrange equality, we let $f = \tilde{l}^{\alpha'}$ (where $\alpha = 1$ or ∞ for simplicity for the time being), and thus we have that for each $(X, Y) \in \mathcal{X} \times \{-1, +1\}$,

$$\tilde{l}^{\alpha'}(b_{(X,Y)}) = \tilde{l}^{\alpha'}(a_{(X,Y)}) + (b_{(X,Y)} - a_{(X,Y)})\tilde{l}^{\alpha''}(a_{(X,Y)}) + \frac{(b_{(X,Y)} - a_{(X,Y)})^2}{2}\tilde{l}^{\alpha'''}(c^\alpha_{(X,Y)}), \tag{74}$$

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 (72) (first term) and (73) (without coefficient) with the Taylor-Lagrange equality, we have that

$$\mathbb{E}_{X,Y}\left[\tilde{l}^{\alpha'}(\langle YX, \theta^*\rangle)YX_i\right]$$

$$= \mathbb{E}_{X,Y}\left[\left(\tilde{l}^{\alpha'}(a_{(X,Y)}) + (b_{(X,Y)} - a_{(X,Y)})\tilde{l}^{\alpha''}(a_{(X,Y)}) + \frac{(b_{(X,Y)} - a_{(X,Y)})^2}{2}\tilde{l}^{\alpha'''}(c^\alpha_{(X,Y)})\right)YX_i\right]. \tag{75}$$

Thus, for $\alpha = 1$, we have that

$$\frac{\partial}{\partial \theta_i}R^p_1(\theta^*) = \mathbb{E}_{X,Y}\left[\tilde{l}^{1'}(\langle YX, \theta^*\rangle)YX_i\right] + p\mathbb{E}_{X,Y}[YX_i] \tag{76}$$

$$= p\mathbb{E}_{X,Y}[YX_i]$$

$$+ \mathbb{E}_{X,Y}\left[\left(\tilde{l}^{1'}(\langle YX, \hat{\theta}^1\rangle) + (\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle) + \frac{(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right)YX_i\right], \tag{77}$$

where $c^1_{(X,Y)} \in [\langle YX, \hat{\theta}^1\rangle, \langle YX, \theta^*\rangle]$. Noticing that $\mathbb{E}_{X,Y}\left[\tilde{l}^{1'}(\langle YX, \hat{\theta}^1\rangle)YX_i\right] + p\mathbb{E}_{X,Y}[YX_i] = 0$, we thus obtain

$$\frac{\partial}{\partial\theta_i}R_1^p(\theta^*) = \mathbb{E}_{X,Y}\left[\left((\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle) + \frac{(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right)YX_i\right]. \quad (78)$$

Using similar steps, we can also obtain

$$\begin{aligned}&\frac{\partial}{\partial\theta_i}R_\infty^p(\theta^*)\\ &= (1-2p)\mathbb{E}_{X,Y}\left[\left((\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^\infty\rangle)\tilde{l}^{\infty''}(\langle YX, \hat{\theta}^\infty\rangle) + \frac{(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^\infty\rangle)^2}{2}\tilde{l}^{\infty'''}(c^\infty_{(X,Y)})\right)YX_i\right],\end{aligned} \quad (79)$$

where $c^\infty_{(X,Y)} \in [\langle YX, \hat{\theta}^\infty\rangle, \langle YX, \theta^*\rangle]$ and we note a difference between (78) and (79), 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 (78)

$$\left|\frac{\partial}{\partial\theta_i}R_1^p(\theta^*)\right| = \left|\mathbb{E}_{X,Y}\left[\left((\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle) + \frac{(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right)YX_i\right]\right| \quad (80)$$

$$\leq \mathbb{E}_{X,Y}\left[\left|\left((\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle) + \frac{(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right)YX_i\right|\right] \quad (81)$$

$$= \mathbb{E}_{X,Y}\left[|X_i|\left|(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle) + \frac{(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right] \quad (82)$$

$$\leq \mathbb{E}_{X,Y}\left[|X_i|\left(\left|(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle)\right| + \left|\frac{(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right], \quad (83)$$

where we used Jensen's inequality via the absolute value, the triangle inequality, and the fact that $|ab| = |a|\cdot|b|$. Continuing,

$$\mathbb{E}_{X,Y}\left[|X_i|\left(\left|(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle)\right| + \left|\frac{(\langle YX, \theta^*\rangle - \langle YX, \hat{\theta}^1\rangle)^2}{2}\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right] \quad (84)$$

$$= \mathbb{E}_{X,Y}\left[|X_i|\left(\left|\langle YX, \theta^* - \hat{\theta}^1\rangle\right|\left|\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle)\right| + \frac{\langle YX, \theta^* - \hat{\theta}^1\rangle^2}{2}\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right] \quad (85)$$

$$\leq \mathbb{E}_{X,Y}\left[|X_i|\left(\|YX\|\left\|\theta^* - \hat{\theta}^1\right\|\left|\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle)\right| + \frac{\|YX\|^2\left\|\theta^* - \hat{\theta}^1\right\|^2}{2}\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right], \quad (86)$$

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

$$\mathbb{E}_{X,Y}\left[|X_i|\left(\|YX\|\left\|\theta^* - \hat{\theta}^1\right\|\left|\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle)\right| + \frac{\|YX\|^2\left\|\theta^* - \hat{\theta}^1\right\|^2}{2}\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right)\right] \quad (87)$$

$$\leq \mathbb{E}_{X,Y}\left[\sqrt{d}2r\left|\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle)\right| + \frac{4dr^2}{2}\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right] \quad (88)$$

$$= 2d^{1/2}r\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1''}(\langle YX, \hat{\theta}^1\rangle)\right|\right] + 2dr^2\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right]. \quad (89)$$

Thus, we obtain that

$$\left|\frac{\partial}{\partial\theta_i}R_1^p(\theta^*)\right| \leq 2d^{1/2}r\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|\right] + 2dr^2\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right]. \tag{90}$$

For $\alpha=\infty$, the exact same steps go through, so we also have that

$$\left|\frac{\partial}{\partial\theta_i}R_\infty^p(\theta^*)\right| \leq (1-2p)\left(2d^{1/2}r\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{\infty''}(\langle YX,\hat{\theta}^\infty\rangle)\right|\right] + 2dr^2\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{\infty'''}(c^\infty_{(X,Y)})\right|\right]\right). \tag{91}$$

Considering $\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1''}(\langle YX,\hat{\theta}^1\rangle)\right|\right]$ in (90), we let

$$z_1^* = \underset{z\in\{\langle yx,\hat{\theta}^1\rangle|(x,y)\in\mathcal{X}\times\{-1,+1\}\}}{\arg\max}\left|\tilde{l}^{1''}(z)\right|, \tag{92}$$

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 (91), we let

$$z_\infty^* = \underset{z\in\{\langle yx,\hat{\theta}^\infty\rangle|(x,y)\in\mathcal{X}\times\{-1,+1\}\}}{\arg\max}\left|\tilde{l}^{\infty''}(z)\right|, \tag{93}$$

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

$$\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{1'''}(c^1_{(X,Y)})\right|\right] \leq \left|\tilde{l}^{1'''}(z_1^*)\right|, \tag{94}$$

and

$$\mathbb{E}_{X,Y}\left[\left|\tilde{l}^{\infty'''}(c^\infty_{(X,Y)})\right|\right] \leq \left|\tilde{l}^{\infty'''}(z_\infty^*)\right|. \tag{95}$$

Next, we invoke Lemma 3, i.e., that for all $z\in\mathbb{R}$,

$$\left|\frac{d^2}{dz^2}\tilde{l}^1(z)\right| \geq \left|\frac{d^2}{dz^2}\tilde{l}^\infty(z)\right|, \tag{96}$$

and Lemma 4, i.e., that for $z > \ln 2$,

$$\left|\frac{d^3}{dz^3}\tilde{l}^\infty(z)\right| \leq \left|\frac{d^3}{dz^3}\tilde{l}^1(z)\right|. \tag{97}$$

Thus, we have that (also by monotonically decreasing)

$$\left|\tilde{l}^{\infty''}(z_\infty^*)\right| \leq \left|\tilde{l}^{1''}(z_1^*)\right|, \tag{98}$$

and

$$\left|\tilde{l}^{\infty'''}(z_\infty^*)\right| \leq \left|\tilde{l}^{1'''}(z_1^*)\right|. \tag{99}$$

Hence, since the bounds on (90) and (91) hold for all $i\in[d]$, we obtain the desired result.

A.3 Proof of Theorem 3

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 ($P_1 = \mathbb{P}[Y = 1]$)

$$\nabla_\theta R_\alpha^p(\theta) = \mathbb{E}_{X,Y}\left[\left(pYg_\theta(-YX)^{1-1/\alpha}g_\theta(YX) - (1-p)Yg_\theta(YX)^{1-1/\alpha}g_\theta(-YX)\right)X\right] \tag{100}$$

$$= P_1\mathbb{E}_{X^{[1]}}\left[\left(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]})\right)X^{[1]}\right]$$

$$+ P_{-1}\mathbb{E}_{X^{[-1]}}\left[\left(-pg_\theta(X^{[-1]})^{1-\frac{1}{\alpha}}g_\theta(-X^{[-1]}) + (1-p)g_\theta(-X^{[-1]})^{1-\frac{1}{\alpha}}g_\theta(X^{[-1]})\right)X^{[-1]}\right] \tag{101}$$

$$= \mathbb{E}_{X^{[1]}}\left[\left(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]})\right)X^{[1]}\right]. \tag{102}$$

First considering the upperbound on $\|\nabla_\theta R_\alpha^p(\theta) - \mathbb{E}[X^{[1]}]\|$, we have that

$$\|\mathbb{E}_{X^{[1]}}\left[\left(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]})\right)X^{[1]}\right] - \mathbb{E}[X^{[1]}]\| \tag{103}$$

$$= \|\mathbb{E}_{X^{[1]}}\left[\left(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - 1\right)X^{[1]}\right]\| \tag{104}$$

$$= \|\mathbb{E}_{X^{[1]}}\left[\left(pg_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - p - (1-p)g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - (1-p)\right)X^{[1]}\right]\| \tag{105}$$

$$= \|\mathbb{E}_{X^{[1]}}\left[\left(p\left(g_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - 1\right) - (1-p)\left(g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - 1\right)\right)X^{[1]}\right]\| \tag{106}$$

$$\leq \mathbb{E}_{X^{[1]}}\left[\left|p\left(g_\theta(-X^{[1]})^{1-1/\alpha}g_\theta(X^{[1]}) - 1\right) - (1-p)\left(g_\theta(X^{[1]})^{1-1/\alpha}g_\theta(-X^{[1]}) - 1\right)\right|\|X^{[1]}\|\right], \tag{107}$$

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

$$f_{\alpha,p}(x) := p\left(\sigma(-x)^{1-\frac{1}{\alpha}}\sigma(x) - 1\right) - (1-p)\left(\sigma(x)^{1-\frac{1}{\alpha}}\sigma(-x) - 1\right). \tag{108}$$

We examine

$$\frac{\partial}{\partial\alpha}f_{\alpha,p}(x) = (1-p)\frac{\sigma(x)^{1-\frac{1}{\alpha}}\log\left(e^{-x}+1\right)}{(e^x+1)\alpha^2} - p\frac{\sigma(-x)^{1-\frac{1}{\alpha}}\log\left(e^{x}+1\right)}{(e^{-x}+1)\alpha^2}, \tag{109}$$

which follows from the fact that

$$\frac{\partial}{\partial\alpha}\sigma(x)^{1-\frac{1}{\alpha}}\sigma(-x) = \frac{\sigma(x)^{1-\frac{1}{\alpha}}\log\left(\sigma(x)\right)}{(e^x+1)\alpha^2}. \tag{110}$$

Considering $x > 0$ and $0 < p < 1/2$, one can show that

$$\frac{\partial}{\partial\alpha}f_{\alpha,p}(x) > 0 \tag{111}$$

is equivalent to

$$\left(\frac{1}{p} - 1\right) > e^{\frac{x}{\alpha}}\frac{\log\left(e^x+1\right)}{\log\left(e^{-x}+1\right)}, \tag{112}$$

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 $\frac{\partial}{\partial x} f_{\alpha,p}(x) > 0$ 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

$$p\left(g_\theta(-X^{[1]})^{1-1/\alpha} g_\theta(X^{[1]}) - 1\right) - (1-p)\left(g_\theta(X^{[1]})^{1-1/\alpha} g_\theta(-X^{[1]}) - 1\right) \tag{113}$$

$$\leq p\left(\sigma(-r\sqrt{d})^{1-\frac{1}{\alpha}} \sigma(r\sqrt{d}) - 1\right) - (1-p)\left(\sigma(r\sqrt{d})^{1-\frac{1}{\alpha}} \sigma(-r\sqrt{d}) - 1\right) =: C_{p,r\sqrt{d},\alpha}. \tag{114}$$

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 $r > 0$ (112), we have that for $\alpha \in (1, \infty)$

$$0 < C_{p,r\sqrt{d},1} \leq C_{p,r\sqrt{d},\alpha} \leq C_{p,r\sqrt{d},\infty}. \tag{115}$$

Thus, considering the upperbound on $\|\nabla_\theta R^p_\alpha(\theta) - \mathbb{E}[X^{[1]}]\|$ in (107), we have that

$$\|\nabla_\theta R^p_\alpha(\theta) - \mathbb{E}[X^{[1]}]\| \leq C_{p,r\sqrt{d},\alpha} \mathbb{E}_{X^{[1]}}[\|X^{[1]}\|], \tag{116}$$

where $C_{p,r\sqrt{d},\alpha}$ is given in (114).

Now, considering a lowerbound on $\|\nabla_\theta R^p_\alpha(\theta) - \mathbb{E}[X^{[1]}]\|$, via the reverse triangle inequality we have that

$$\|\nabla_\theta R^p_\alpha(\theta) - \mathbb{E}[X^{[1]}]\| \geq \|\mathbb{E}[X^{[1]}]\| - \|\nabla_\theta R^p_\alpha(\theta)\|. \tag{117}$$

Combining this with our derived upperbound (116), we have that

$$C_{p,r\sqrt{d},\alpha} \mathbb{E}[\|X^{[1]}\|] \geq \|\mathbb{E}[X^{[1]}]\| - \|\nabla_\theta R^p_\alpha(\theta)\|. \tag{118}$$

Rewriting, we have that

$$\|\nabla_\theta R^p_\alpha(\theta)\| \geq \|\mathbb{E}[X^{[1]}]\| - C_{p,r\sqrt{d},\alpha} \mathbb{E}[\|X^{[1]}\|]. \tag{119}$$

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

$$\begin{aligned} \|\nabla_\theta R^p_\alpha(\theta)\| &\geq \|\mathbb{E}[X^{[1]}]\| - C_{p,r\sqrt{d},\alpha} \mathbb{E}[\|X^{[1]}\|] \\ &\geq \|\mathbb{E}[X^{[1]}]\| - (1-2p)\left(1 - \sigma'(r\sqrt{d})\right) \mathbb{E}[\|X^{[1]}\|] > 0, \end{aligned} \tag{120}$$

which is nonnegative by distributional assumption on the skew-symmetric distribution itself, namely we assume that

$$(1-2p)(1-\sigma'(r\sqrt{d})) < \frac{\|\mathbb{E}(X^{[1]})\|}{\mathbb{E}(\|X^{[1]}\|)}. \tag{121}$$

B Further Experimental Results and Details

B.1 Boosting Experiments

B.1.1 Long-Servedio

Dataset The Long-Servedio dataset is a synthetic dataset which was first suggested in (Long and Servedio, 2008) 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 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 $-y$.

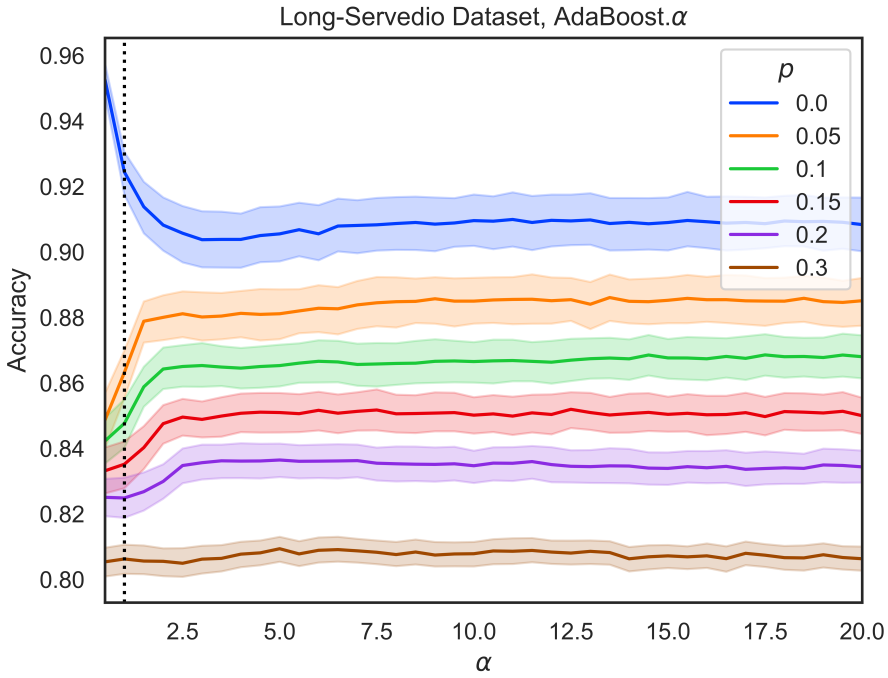


Figure 16: Accuracy of AdaBoost.α on the Long-Servedio dataset. We see that accuracy levels off as α increases, implying that tuning α can be as simple as choosing $\alpha \approx 5$. The thresholding behavior is supported by Figure 10

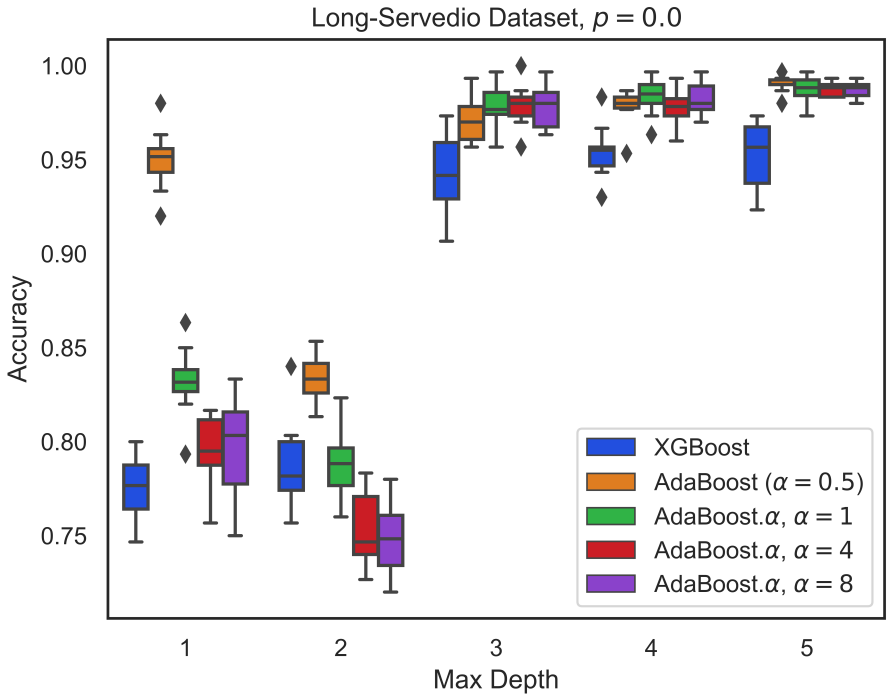


Figure 17: Clean test accuracy of various models on the Long-Servedio dataset with no added label noise. Models trained for 100 iterations. Vanilla AdaBoost performs well here, but note that Figure 20 implies that with a larger number of iterations, $\alpha = 1, 2$ would have similar performance.

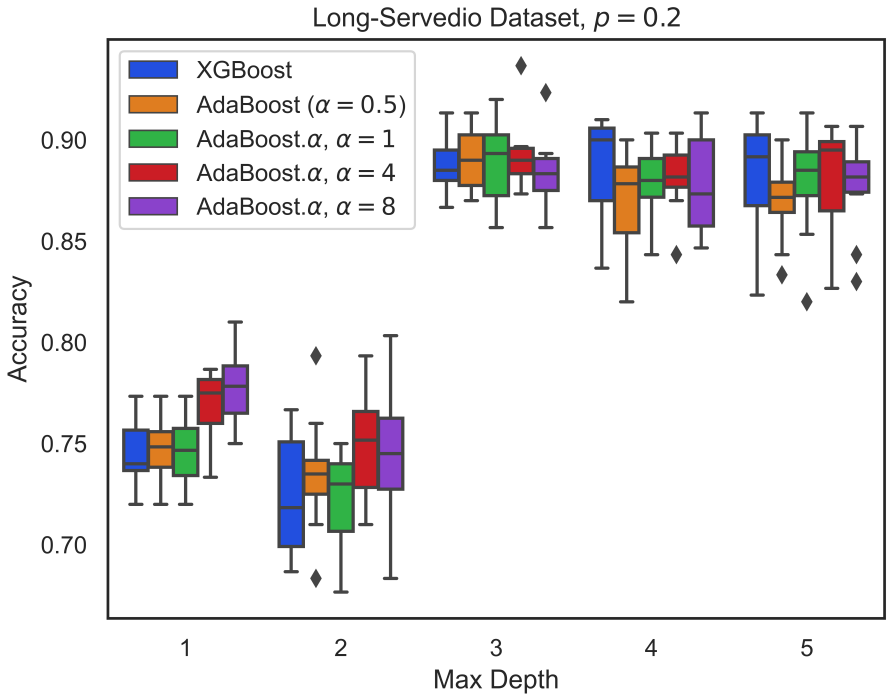


Figure 18: Clean test accuracy vs the depth of weak learners on the Long-Servedio dataset with SLN. 100 iterations of boosting. We see that that for low depth weak learners, $\alpha > 1$ outperforms convex α in terms of clean classification accuracy. This benefit diminishes with growing depth.

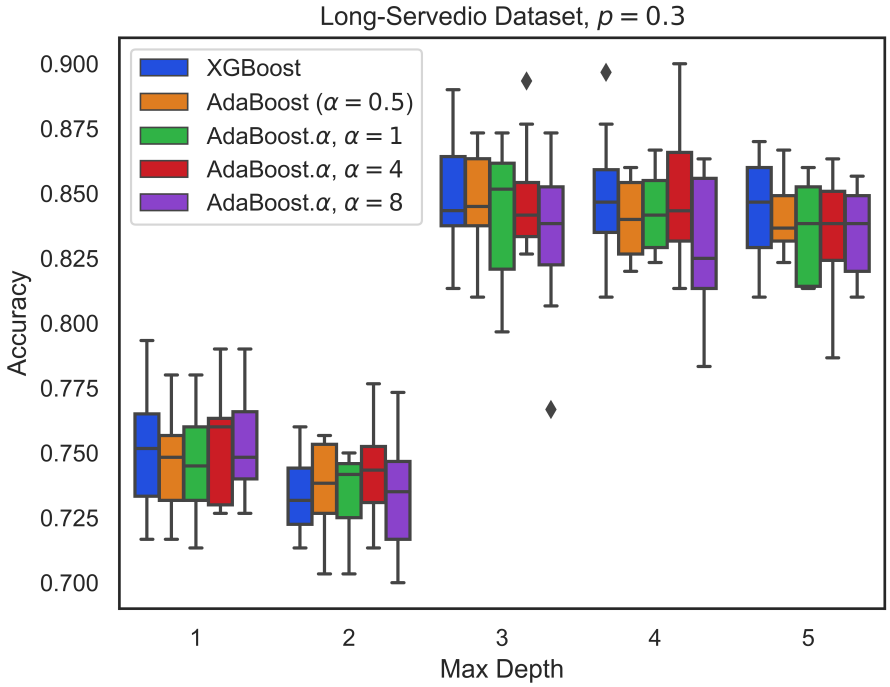


Figure 19: Clean test accuracy vs the depth of weak learners on the Long-Servedio dataset with SLN. 100 iterations of boosting. In this higher noise setting, α has little effect on the clean test accuracy.

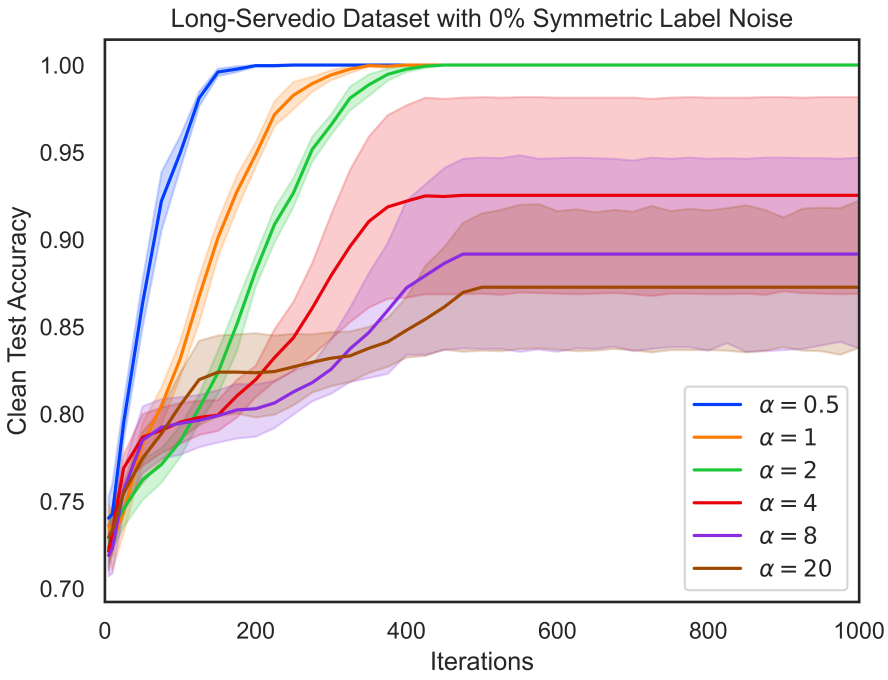


Figure 20: Clean test accuracy of AdaBoost.α on the Long-Servedio dataset with no added label noise. In this zero noise setting, convex α values perform well. Performance gains slow with increasing α which corresponds to increasing non-convexity in the optimization.

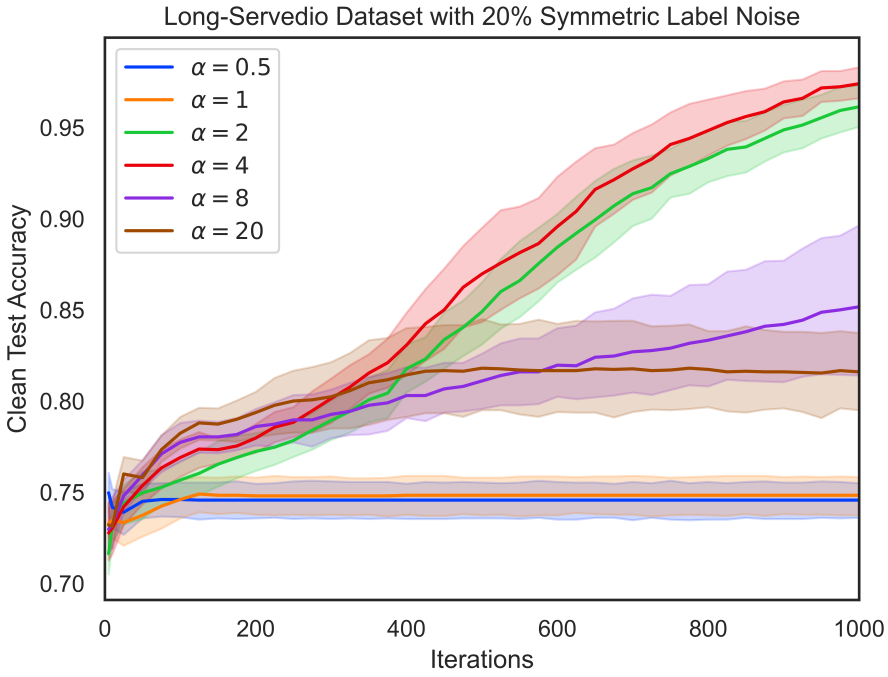


Figure 21: Accuracy of AdaBoost.α on the Long-Servedio dataset. We see that convex $\alpha < 1$, is unable to learn by increasing the number of weak learners, likely because it is getting stuck trying to learn on large-margin example. $\alpha > 1$ continues to learn with increasing iterations, though growth is slower than in smaller noise levels.

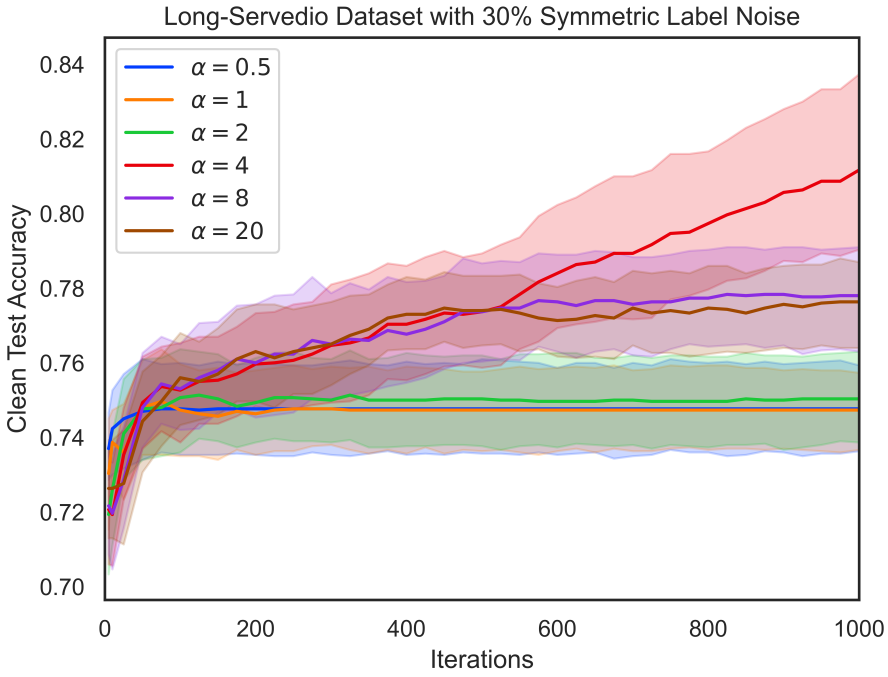


Figure 22: Accuracy of AdaBoost.α on the Long-Servedio dataset. We see that convex $\alpha < 1$, is unable to learn by increasing the number of weak learners, likely because it is getting stuck trying to learn on large-margin example. $\alpha > 1$ continues to learn with increasing iterations, though growth is slower than in smaller noise levels.

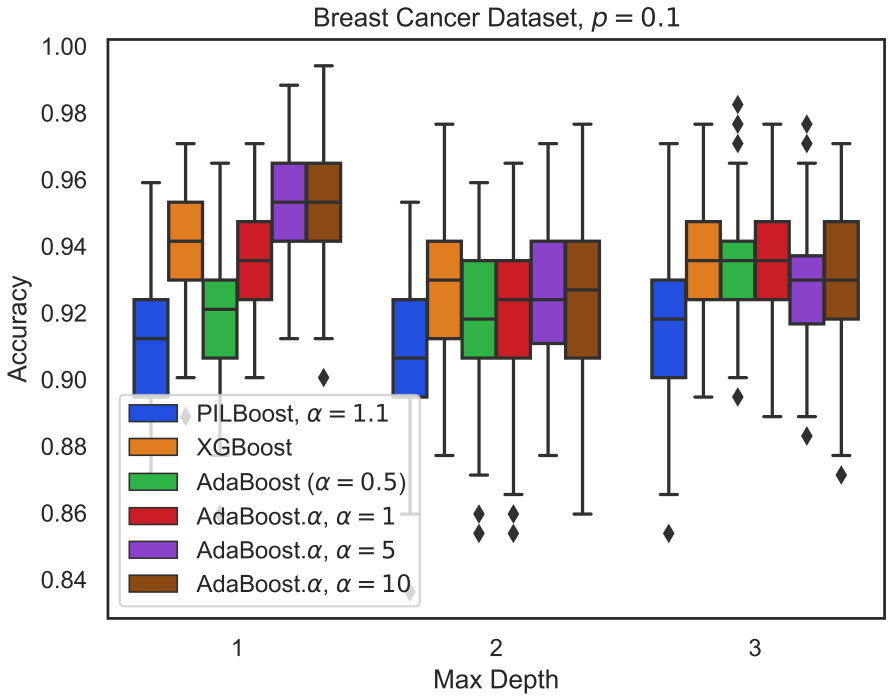


Figure 23: Accuracy of various models on the breast cancer dataset. We see that with low depth (and thus low complexity) weak learners, the use of a non-convex loss, namely $\alpha > 1$, permits some gains in accuracy. These diminish for more complex weak learners.

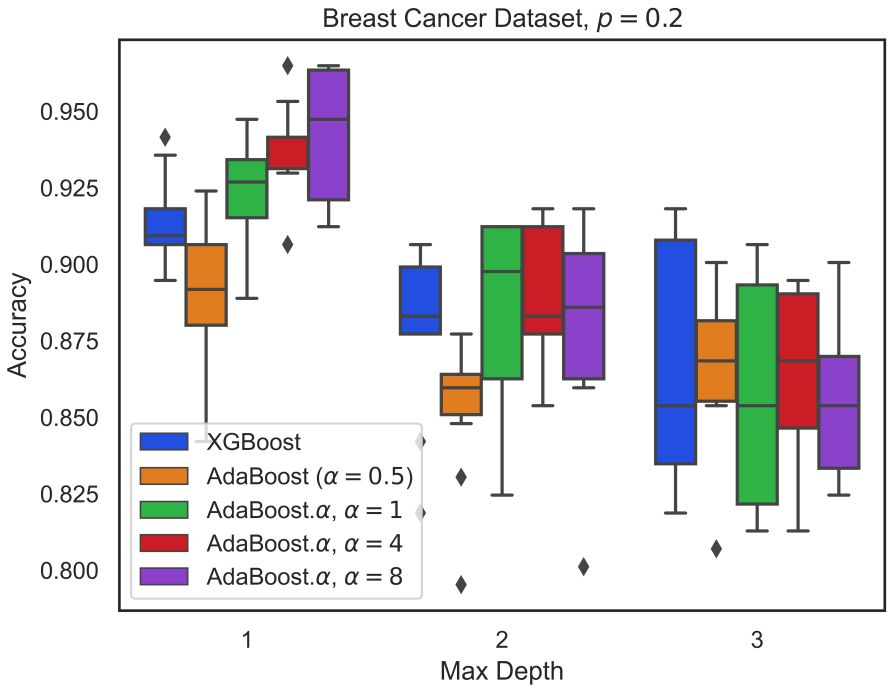


Figure 24: Accuracy of various models on the breast cancer dataset. We see that with low depth (and thus low complexity) weak learners, the use of a non-convex loss, namely $\alpha > 1$, permits some gains in accuracy. These diminish for more complex weak learners.

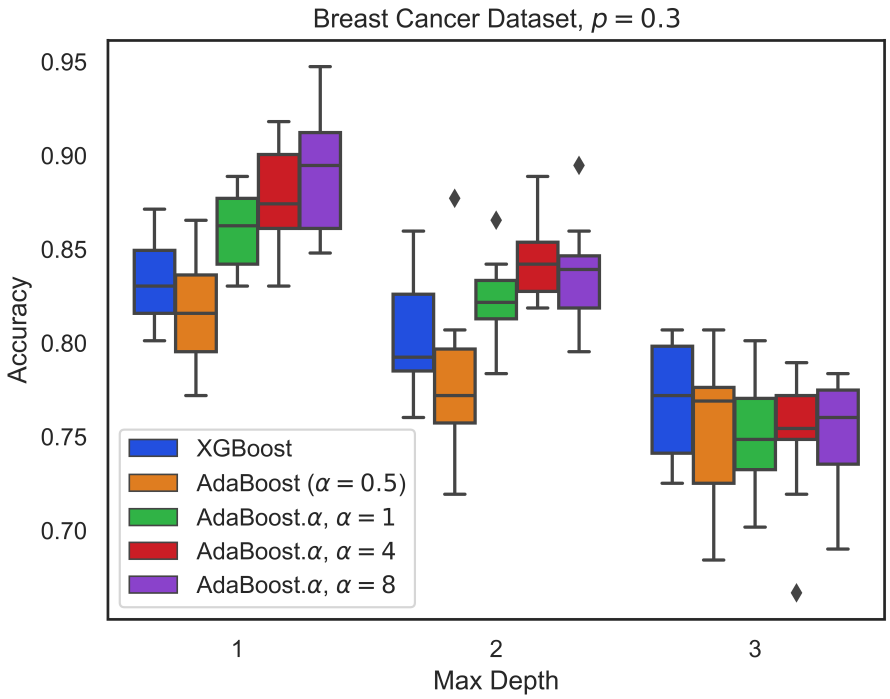


Figure 25: Accuracy of various models on the breast cancer dataset. We see that with low depth (and thus low complexity) weak learners, the use of a non-convex loss, namely $\alpha > 1$, permits some gains in accuracy. These diminish for more complex weak learners.

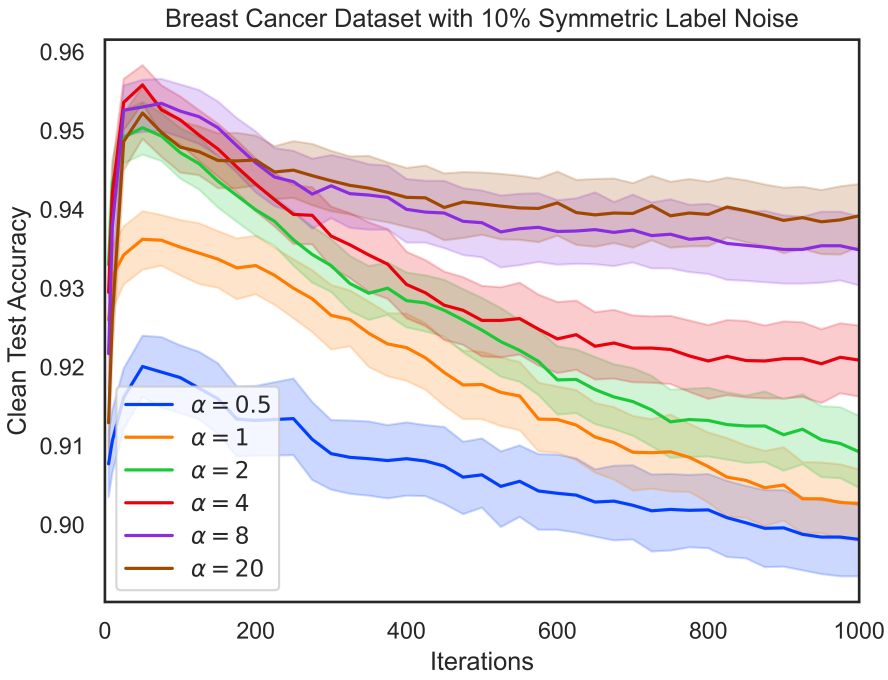


Figure 26: Accuracy of AdaBoost.α on the Wisconsin Breast Cancer dataset. Non-convex α values perform significantly better than convex α values. Unlike the Long-Servedio dataset, convex α values are still able to learn as the iterations increase, though there appears to be some overfitting.

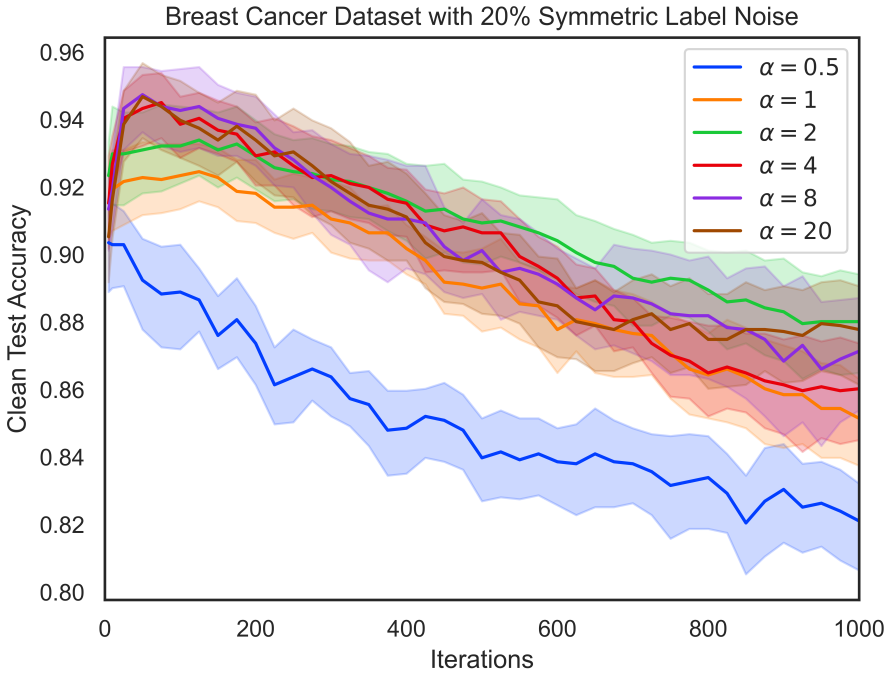


Figure 27: Accuracy of AdaBoost.α on the Wisconsin Breast Cancer dataset. Non-convex α values perform significantly better than convex α values. Unlike the Long-Servedio dataset, convex α values are still able to learn as the iterations increase, though there appears to be some overfitting.

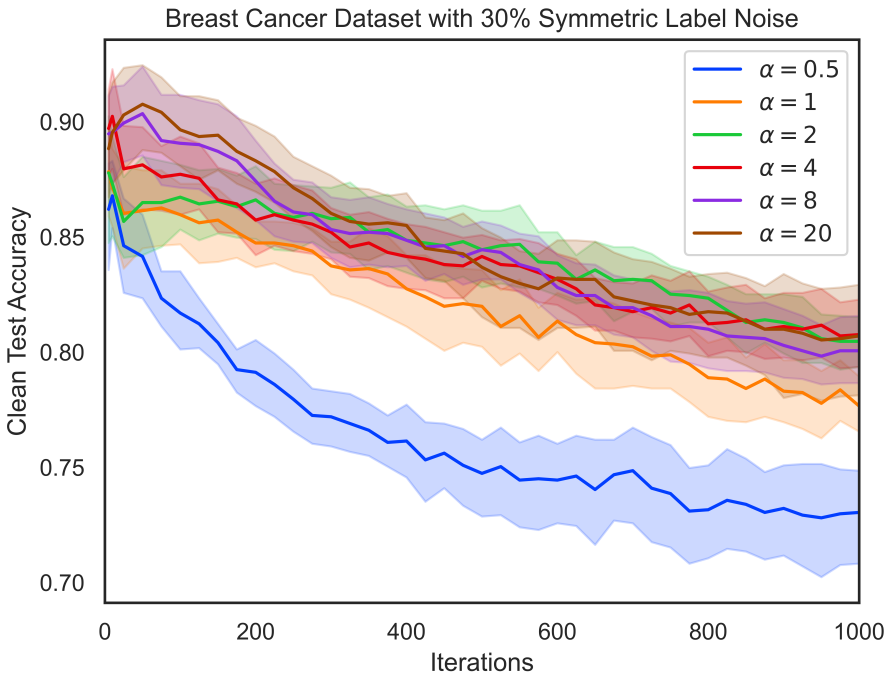


Figure 28: Accuracy of AdaBoost.α on the Wisconsin Breast Cancer dataset. Non-convex α values perform significantly better than convex α values. Unlike the Long-Servedio dataset, convex α values are still able to learn as the iterations increase, though there appears to be some overfitting.

B.1.2 Breast Cancer

Dataset The Wisconsin Breast Cancer dataset (Wolberg et al., 1995) is a widely used medical dataset in the boosting community.

B.2 Logistic Model Experiments

B.2.1 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:

$$Y = 1 : X \sim \mathcal{N}[\mu_1, \sigma^2 \mathbb{I}],$$
$$Y = -1 : X \sim \mathcal{N}[\mu_2, \sigma^2 \mathbb{I}],$$

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.

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 $1\text{e}{-2}$ and models were trained until convergence. For each pair, 30 models were trained with different noise seeds, and metrics were then averaged across models.

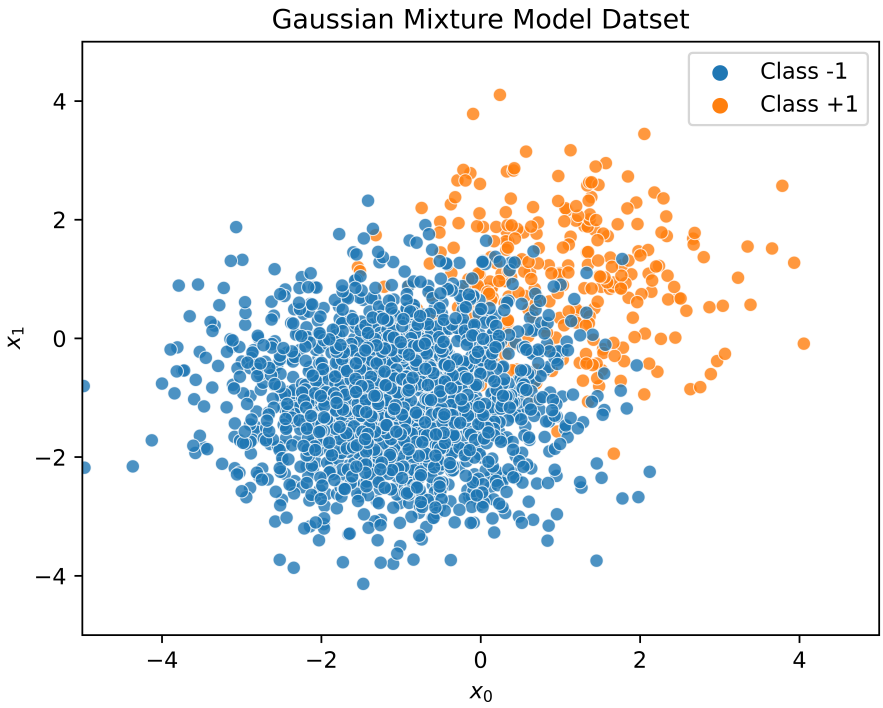


Figure 29: Sample dataset generated with Gaussian class-conditional distributions with $P(Y = 1) = 0.14$ and $\mu_1 = [1, 1]^T, \mu_2 = [-1, -1]^T$; we use a spherical covariance with $\sigma = 1$ for both classes.

B.2.2 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 1. The learning rate was selected as $1\mathrm{e}{-3}$ and models were trained until convergence. 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.

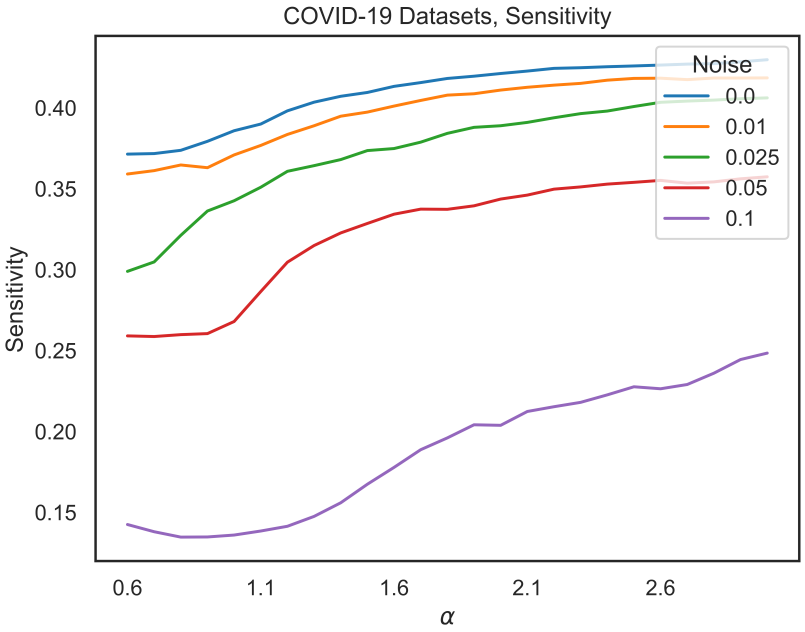


Figure 30: Sensitivity of the classifiers trained on noisy COVID-19 data. We see that $\alpha > 1$ yields gains in sensitivity. This is important to note as the MSE results do not come at the cost of sensitivity. Recall that sensitivity $= \frac{\text{TP}}{\text{TP+FN}}$.