Being Properly Improper

Tyler Sypherd [1] **Richard Nock** [2] **Lalitha Sankar** [1]

Abstract

Properness for supervised losses stipulates that the loss function shapes the learning algorithm towards the true posterior of the data generating distribution. Unfortunately, data in modern machine learning can be corrupted or *twisted* in many ways. Hence, optimizing a proper loss function on twisted data could perilously lead the learning algorithm towards the twisted posterior, rather than to the desired clean posterior. Many papers cope with specific twists (e.g., label/feature/adversarial noise), but there is a growing need for a unified and actionable understanding atop properness. Our chief theoretical contribution is a generalization of the properness framework with a notion called *twist-properness*, which delineates loss functions with the ability to "untwist" the twisted posterior into the clean posterior. Notably, we show that a nontrivial extension of a loss function called α-loss, which was first introduced in information theory, is twist-proper. We study the twist-proper α-loss under a novel boosting algorithm, called PILBOOST, and provide formal and experimental results for this algorithm. Our overarching practical conclusion is that the twist-proper α-loss outperforms the proper $\log$-loss on several variants of twisted data.

1. Introduction

The loss function is a cornerstone of machine learning (ML). The founding theory of properness for supervised losses stipulates that the loss function shapes the learning algorithm towards the true posterior (Reid & Williamson, 2011). Consequently, a model trained with a proper loss function will try to closely approximate the Bayes rule of the data generating distribution. Historically, properness draws its roots from classical work in normative economics for *class probability estimation* (CPE) (Reid & Williamson, 2011; Savage, 1971; Shuford et al., 1966) and Fisher consistency (Fisher, 1922); some of the most famous losses in supervised learning are proper, e.g., log, square, Matusita (Matusita, 1956), to name a few. Unfortunately, in many modern applications data can be corrupted or *twisted* in various ways (see Section 2); examples of twists include label noise, adversarial noise, and feature noise. Thus, optimizing a proper loss function on twisted data could perilously lead the learning algorithm towards the Bayes rule of the twisted posterior, rather than to the desired clean posterior. To ensure that a model trained with a proper loss function on twisted data properly generalizes to the clean distribution, a generalization of properness is clearly required.

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

Recently, α-loss was practically implemented in logistic regression and in deep neural networks (Sypherd et al., 2022). In both settings, it was shown to be more robust to symmetric label noise for fixed $\alpha > 1$ than the proper

[1]School of Electrical, Computer and Energy Engineering, Arizona State University; [2]Google Research. Correspondence to: Tyler Sypherd <tsypherd@asu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

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

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

In Section 6, we implement PILBOOST with the approximate inverse canonical link of α-loss on several tabular datasets, each suffering from various twists (label, feature, and adversarial noise), and compare against AdaBoost (Freund & Schapire, 1997) and XGBoost (Chen & Guestrin, 2016). In general, we find improved algorithmic robustness to all twists through using simple (fixed) hyperparameter corrections via the α-loss, which aligns with our theoretical contributions (see Section 4).

2. Related Work

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

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

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

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

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

3. Losses for Class-Probability Estimation

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

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

where $[\![\cdot]\!]$ is Iverson's bracket (Knuth, 1992). Functions ℓ_1, ℓ_{-1} are called *partial* losses, minimally assumed to satisfy $\mathrm{dom}(\ell_1) = \mathrm{dom}(\ell_{-1}) = [0, 1]$ and $|\ell_1(u)| \ll \infty, |\ell_{-1}(u)| \ll \infty, \forall u \in (0, 1)$ to be useful for ML. Key additional properties of partial losses are:
(M) Monotonicity: ℓ_1, ℓ_{-1} are non-increasing and non-decreasing, respectively;
(D) Differentiability: ℓ_1 and ℓ_{-1} are differentiable;
(S) Symmetry: $\ell_1(u) = \ell_{-1}(1 - u), \forall u \in [0, 1]$.
Commonly used proper losses such as log, square and Matusita all satisfy the above three assumptions. The pointwise conditional risk of the local guess $u \in [0, 1]$ with respect to a *ground truth* $v \in [0, 1]$ is

$$\begin{aligned} L(u, v) &\doteq \mathbb{E}_{\mathsf{Y} \sim \mathrm{B}(v)} \left[\ell(\mathsf{Y}, u)\right] \\ &= v \cdot \ell_1(u) + (1 - v) \cdot \ell_{-1}(u), \end{aligned} \quad (2)$$

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

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

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

$$F(z) \doteq (-\underline{L})^\star(-z), \forall z \in \mathbb{R}. \quad (3)$$

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

Lemma 1 $\forall \ell$ CPE *loss,* $\underline{L}$ *is concave and continuous;* F *is convex, continuous and non-increasing.*

A proof is provided for completeness in Appendix A.2.

4. Twist-Proper Losses

With the classical setting of properness in hand, we now provide fundamental definitions of *twists* and *twist-properness*, and study the twist-properness of several hyperparameterized loss functions. When it comes to correcting (or untwisting) twists, one needs a loss with the property that its minimizer in (2) is different from the now twisted value $\tilde{v}$ *and* recovers the "hidden" ground truth v.

Bayes tilted estimates We first characterize the minimizers of (2) when the CPE loss is not necessarily proper. We define the set-valued (pointwise) Bayes *tilted estimate* t_ℓ as

$$t_\ell(\tilde{v}) \doteq \arg \inf_{u \in [0,1]} L(u, \tilde{v}). \quad (4)$$

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

Lemma 2 *If the partial losses* ℓ_1 *and* ℓ_{-1} *of a given* CPE *loss* $\ell : \mathcal{Y} \times [0, 1] \to \mathbb{R}$ *satisfy* ***(M)***, ***(D)***, *and* ***(S)*** *and are also strictly convex, then* $|t_\ell(\tilde{v})| = 1$ *for every* $\tilde{v} \in [0, 1]$.

In Appendix A.3, we provide an extended version of Lemma 2, denoted Lemma 12, where we prove properties of Bayes tilted estimates for when t_ℓ is set-valued (e.g., set-valued monotonicity and symmetry, and analysis of extreme values). As a consequence, we show that strict monotonicity of the partial losses is not sufficient to guarantee that t_ℓ is a singleton; in fact, strict convexity is required as in Lemma 2.

An important class of twists We now adopt more conventional ML notations and instead of a hidden ground truth

v and twisted ground truth $\tilde{v}$, we use η_c and η_t to denote the "clean" and "twisted" posterior probabilities that $Y = 1$ given $X = x$, respectively. Further, a "**twist**" refers to a general mapping $\eta_c \mapsto \eta_t$, which could be a consequence of label/feature/adversarial noise. The following delineates a fundamental class of twists, important in the sequel.

Definition 3 *A twist $\eta_c \mapsto \eta_t$ is said to be Bayes blunting iff $(\eta_c \leq \eta_t \leq 1/2) \vee (\eta_c \geq \eta_t \geq 1/2)$.*

The term "blunting" is inspired by adversarial training (Cranko et al., 2019). Intuitively, a Bayes blunting twist does not change the *maximum a posteriori* guess for the label given the observation, but it does reduce algorithmic confidence in learned posterior estimates, which is particularly damaging in practice where the learning algorithm only has a finite number of (twisted) training examples. Furthermore, Bayes blunting twists capture a very important twist (see Section 2): *symmetric label noise* (SLN). Under symmetric label noise with flip probability $p \in [0, 1]$, the twisted posterior η_t is given by $\eta_t = \eta_c(1 - p) + (1 - \eta_c)p$ (Reid & Williamson, 2010). The following result readily follows via Definition 3 from consideration of p for fixed η_c.

Lemma 4 *SLN is Bayes blunting for $p < 1/2$.*

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

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

Definition 5 *A loss ℓ is twist-proper (respectively, strictly twist-proper) iff for any twist, there exists hyperparameter(s) such that $\eta_c \in t_\ell(\eta_t)$ (respectively, $\{\eta_c\} = t_\ell(\eta_t)$).*

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

Ostensibly, "optimal" hyperparameters requires explicit knowledge of the distribution and twist, *and* each example in the training set requires a different hyperparameter to untwist its twisted posterior. However, in the sequel, we show that a twist-proper loss, namely α-loss, with a *fixed* hyperparameter (α) can untwist a large class of twists, i.e. Bayes blunting twists (such as SLN), better than $\log$-loss. Thus, we posit through our experimental results in Section 6 that the practitioner only needs peripheral, rather than explicit, knowledge of a Bayes blunting twist in the data.

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

Lemma 6 *Define the **focal loss** via the following partial losses: $\ell_1^{FL}(u) \doteq -(1-u)^\gamma \log u$ and $\ell_{-1}^{FL}(u) \doteq \ell_1^{FL}(1-u)$, with $\gamma \geq 0$. Then the focal loss is **not** twist proper.*

In the proof (see Appendix A.4), we also provide a proof that a loss which acts as a general "wrapper" of a loss, the Super Loss (Castells et al., 2020), is *not* twist proper. Concerning the focal loss, Lemma 6 is not necessarily an impediment for this loss function, which was designed to deal with a specific twist, class imbalance, and it does not prevent *generalizations* of the focal loss that would be twist proper. However, our proof suggests that the Bayes tilted estimate (4) of such generalizations risks not being in a simple analytical form. Intuitively, twist-properness requires more than a trivial extension of the hyperparameter of the loss; it also seems to require a certain *symmetry*, which we observe with the following twist-proper loss, α-loss.

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

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

$$\ell_1^\alpha(u) \doteq \frac{\alpha}{\alpha - 1} \cdot \left(1 - u^{\frac{\alpha-1}{\alpha}}\right), \tag{5}$$

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

$$\ell_1^\alpha(u) \doteq \ell_{-1}^{-\alpha}(u) = \ell_1^{-\alpha}(1 - u). \tag{6}$$

For a plot of (5), see Figure 4 in the appendix. Note that the α-loss is **(S)**ymmetric by construction, and that

it continuously interpolates the log-loss ($\alpha = 1$) which is proper (Reid & Williamson, 2010). Our definition extends the previous definitions with (6), which induces a fundamental symmetry that is required for twist-properness and is utilized in the following result. For any $u \in [0,1]$, we let $\iota(u) \doteq \log(u/(1-u))$ denote the logit of u.

Lemma 8 *The following four properties, labeled **(a)-(d)**, all hold for α-loss: **(a)** **(M)**, **(D)**, **(S)** all hold, $\forall \alpha \in \mathbb{R} \setminus \{0\}$; **(b)** if $(\alpha = 0) \vee (\alpha = \pm\infty \wedge \eta_t = 1/2)$, then $t_{\ell^\alpha}(\eta_t) = [0,1]$, if $\alpha \in \mathbb{R} \setminus \{0, \pm\infty\}$, then $t_{\ell^\alpha}(\eta_t) = \left\{ \frac{\eta_t^\alpha}{\eta_t^\alpha + (1-\eta_t)^\alpha} \right\}$, and if $\alpha \to \pm\infty$, then $t_{\ell^{\pm\infty}}(\eta_t) = \pm 1$ or ∓ 1, depending on the sign of $\eta_t - 1/2$; **(c)** hence, α-loss is twist-proper with $\alpha^* = \iota(\eta_c)/\iota(\eta_t)$; **(d)** for any Bayes blunting twist, $\alpha^* \geq 1$.*

The proof of Lemma 8 can be found in Appendix A.5. Note that **(a)** readily follows from Definition 7. The Bayes tilted estimate in **(b)**, i.e. t_{ℓ^α} for $\alpha \in \mathbb{R} \setminus \{0, \pm\infty\}$, is known in the literature as the α-tilted distribution (Arimoto, 1971; Liao et al., 2018; Sypherd et al., 2022). We observe that the α-tilted distribution is symmetric upon permuting (η_t, α) and $(1-\eta_t, -\alpha)$. Hence, our nontrivial extension of the α-loss induces a *symmetry*, particularly useful for untwisting malevolent twists, which thereby yields twist-properness, **(c)**. Lastly, **(d)** indicates that $\alpha^* \geq 1$ for any Bayes blunting twist (e.g., SLN with $p < 1/2$); however, note that this holds merely for a given x, not over the whole domain $\mathcal{X}$.

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

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

$$\begin{aligned} \mathrm{CE}(\eta_c, \eta_t; \alpha) \doteq \mathbb{E}_{\mathsf{X} \sim \mathrm{M}}[\eta_c(\mathsf{X}) \cdot -\log t_{\ell^\alpha}(\eta_t(\mathsf{X})) \\ + (1 - \eta_c(\mathsf{X})) \cdot -\log t_{\ell^\alpha}(1 - \eta_t(\mathsf{X}))], \end{aligned} \quad (7)$$

where for convenience we used the symmetry property of t_ℓ from Appendix A.3, i.e., $t_{\ell^\alpha}(1 - \eta_t(\mathsf{X})) = 1 - t_{\ell^\alpha}(\eta_t(\mathsf{X}))$. Following (Schapire & Freund, 2012), we denote the binary entropy as $H_b(u) \doteq -u \cdot \log(u) - (1-u) \cdot \log(1-u)$, for $u \in [0,1]$. We let $H(\eta_c)$ represent an averaged binary entropy of the η_c-mapping, given by

$$H(\eta_c) \doteq \mathbb{E}_{\mathsf{X} \sim \mathrm{M}}[H_b(\eta_c(X))]. \quad (8)$$

With (7) and (8), we obtain (cf. Thomas & Joy (2006)),

$$D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha) \doteq \mathrm{CE}(\eta_c, \eta_t; \alpha) - H(\eta_c), \quad (9)$$

that is, a KL-divergence between the α-Bayes tilted estimate of the twisted posterior and the clean posterior mappings. Intuitively, $D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha)$ aggregates a series of information-trajectories, strictly dependent on α (either fixed or a mapping), each tracing a path on the probability *simplex* between the two posterior mappings for every $x \in \mathcal{X}$. Slightly more restrictive than Definition 3, we define a *strictly* Bayes blunting twist as a Bayes blunting twist where $(\eta_c < \eta_t \leq 1/2) \vee (\eta_c > \eta_t \geq 1/2)$; we state one of our main results whose proof is in Appendix A.6.

Theorem 9 *For any strictly Bayes blunting twist $\eta_c \mapsto \eta_t$, there exists a fixed $\alpha_0 > 1$ and an optimal $\alpha^\star$-mapping, $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$, which induces the following ordering*

$$D_{\mathrm{KL}}(\eta_c, \eta_t; 1) > D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha_0) \geq D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha^\star). \quad (10)$$

This result answers in the affirmative that untwisting $\mathcal{X}$ for a large class of twists with a fixed hyperparameter $\alpha_0 > 1$ is *strictly* better than simply using the proper choice, i.e., $\alpha = 1$ (log-loss). Specifically, Theorem 9 holds for SLN, which is a strictly Bayes blunting twist for flip probability $0 < p < 1/2$ (Lemma 4). The result also states that there exists a mapping $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$ which optimally untwists the strictly Bayes blunting twist; indeed, $\alpha^\star$ can be recovered from Lemma 8**(c)**, i.e., $\alpha^\star(x) := \iota(\eta_c(x))/\iota(\eta_t(x))$, for every $x \in \mathcal{X}$. Thus, by the twist-properness of α-loss, $D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha^\star) = 0$ (more details in Appendix A.6). Regarding the search for a fixed $\alpha_0 > 1$ in practice, Sypherd et al. (2022) showed via optimization landscape analysis and experiments on SLN for logistic regression and neural-networks that the search space for α_0 is bounded (due to saturation), typically $\alpha_0 \in [1.1, 8]$. In Section 6, we report experimental results for several α; we also incorporate a method inspired by (Menon et al., 2015) to estimate the amount of SLN in training data and thus estimate α_0 using Lemma 8**(c)** as motivated by Theorem 9.

Theorem 9 gave a *nonconstructive* indication for the optimal regime of α for strictly Bayes blunting twists. Our next result gives a *constructive* formula for a fixed α for any twist. Given $B > 0$, let $\mathrm{M}(B)$ denote the distribution restricted to the support over $\mathcal{X}$ for which we have almost surely

$$(1 + \exp(B))^{-1} \leq \eta_t(x) \leq (1 + \exp(-B))^{-1}, \quad (11)$$

and let $p(B) \in [0,1]$ be the weight of this support in M. We

let $\mathrm{D}(B)$ denote the product distribution on examples ($\mathcal{X} \times \mathcal{Y}$) induced by marginal $\mathrm{M}(B)$ and posterior η_c (see Reid & Williamson (2011, Section 4)). We define the logit-*edge* as

$$\mathsf{e} \doteq (1/B) \cdot \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim \mathrm{D}(B)} \left[\mathsf{Y} \cdot \iota(\eta_\iota(\mathsf{X}))\right], \tag{12}$$

where we note that $\mathsf{e} \in [-1, 1]$ due to the assumption in (11). Finally, we let $q \doteq (1 + \mathsf{e})/2 \in [0, 1]$.

Theorem 10 *Let $B > 0$. If $p(B) = 1$ and we fix $\alpha = \alpha^{**}$ with $\alpha^{**} \doteq \iota(q)/B$, then the following bound holds*

$$D_{\mathrm{KL}}(\eta_c, \eta_\iota; \alpha^{**}) \leq H_b(q) - H(\eta_c). \tag{13}$$

The proof of Theorem 10 is in Appendix A.7, where we also prove an extended version of the result when $p(B) < 1$. In addition, we provide a simple example where $D_{\mathrm{KL}}(\eta_c, \eta_\iota; \alpha^{**})$ can vanish with respect to $D_{\mathrm{KL}}(\eta_c, \eta_\iota; 1)$ (the "proper" choice). Intuitively, the difference on the right-hand-size of (13) in Theorem 10 is reminiscent of a Jensen's gap. Also in the proof of Theorem 10, we find that if $|\alpha^{**}|$ is large, there is more "flatness" in the bounded terms near α^{**}. Hence, this suggests that a choice of α_0 "close-enough" to α^{**} could yield similar performance.

5. Sideways Boosting a Surrogate Loss

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

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

$$H_{\boldsymbol{\beta}} \doteq \textstyle\sum_j \beta_j h_j, \tag{14}$$

where $\{h_\cdot : \mathcal{X} \to \mathbb{R}\}$ are WL (weak learning) classifiers with slightly better than random classification accuracy. The oracle WL returns an index $j \in \mathbb{N}$, and the task for the boosting algorithm is to learn the coordinates of $\boldsymbol{\beta}$, initialized to the null vector. In our general framework, the losses we consider are the surrogates F in Lemma 1, essentially convex and non-increasing functions, adding the condition that they are twice differentiable. We compute weights using the blueprint of (Friedman, 2001), which uses the full $H_{\boldsymbol{\beta}}$,

$$w_i \doteq -F'(y_i H_{\boldsymbol{\beta}}(\boldsymbol{x}_i)), \forall i \in [m]. \tag{15}$$

Algorithm 1 PILBOOST

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

Via Lemma 1, weights w_i are non-negative and tend to increase for an example given the wrong class by the current weak classifier h_j, thus, weighting puts emphasis on "hard" examples. For an underlying CPE loss ℓ, we have that (see Appendix A.8 for a derivation)

$$-F'(z) = (\ell_{-1} \circ t_\ell - \ell_1 \circ t_\ell)^{-1}(-z). \tag{16}$$

We thus need to invert the *difference* of the partial losses to recover $-F'$. The inversion is easy for the log-loss because of properties of the $\log$ function and for the square loss because its partial losses are quadratic functions. However, for hyperparameterized losses, such as the α-loss, the inversion in (16) is nontrivial. We circumvent this difficulty by proposing a novel boosting algorithm, PILBOOST, given in Algorithm 1. Rather than using $-F'$ as in (15) for the weight update in Step 2.1, PILBOOST uses an approximation function $\widetilde{f}$, which is non-negative and increasing, that we dub *pseudo-inverse link* (PIL), which is studied in general in Appendix A.8. Specifically, in Lemma 18, we provide $\widetilde{f}_\ell$ for α-loss, given in (160). Furthermore in Lemma 19, we show that there exists $K > 0$ such that, for almost all $z \in \mathbb{R}$, $|(\widetilde{f}_\ell - (-\underline{L}')^{-1})(z)| \lesssim K/\alpha$. We now theoretically analyze PILBOOST, and we make two classical assumptions on WL (Schapire & Singer, 1999; Nock & Williamson, 2019).

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

Let $\tilde{\mathsf{e}}_j \doteq m \cdot \mathsf{e}_j / (\mathbf{1}^\top \boldsymbol{w}_j) \in [-M, M]$ be the normalized edge of the j-th weak classifier, where with a slight abuse of notation of (12), e_j is the (unnormalized) edge (Step 2.3).

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

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

$$\Delta_j(F) \doteq \left|\mathbb{E}_{i\sim\mathcal{S}} \left[y_i h_j(\boldsymbol{x}_i) F'(y_i H_{\boldsymbol{\beta}}(\boldsymbol{x}_i))\right] - \mathsf{e}_j\right|, \tag{17}$$

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

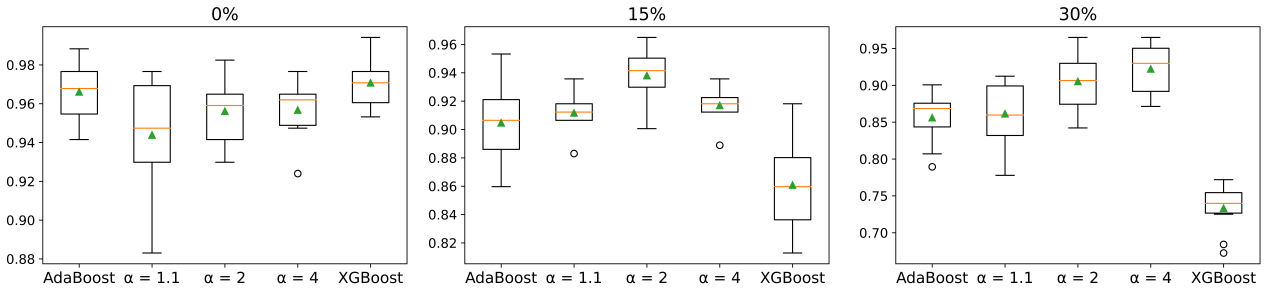

Figure 1: Box and whisker plots reporting the *classification accuracy* of AdaBoost, PILBOOST (for $\alpha \in \{1.1, 2, 4\}$), and XGBoost on the cancer dataset affected by the class noise twister with 0%, 15%, and 30% twist. Note that the orange line is the median, the green triangle is the mean, the box is the interquartile range, and the circles outside of the whiskers are outliers. All three algorithms were trained with decision stumps (depth 1 regression trees). For $\alpha = 1.1, 2$, and 4, we set $a_f = 7, 2$, and 4, respectively. Numeric values corresponding to the box and whisker plots are provided in Table 2 in Section B.3. We find that PILBOOST has gains over AdaBoost and XGBoost when there is twist present, and α^* (of our set) increases as the amount of twist increases, which follows theoretical intuition (Lemma 8).

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

(E) *the edge discrepancy is bounded* $\forall t$: $\Delta_j(F) \leq \zeta \cdot e_j$, *where* j *is returned by* WL *at iteration* t*;*

(C) *the curvature of* F *is bounded:* $F^* \doteq \sup_z F''(z) \leq (1 - \zeta)(1 + \pi)/(a_f M^2)$.

Note that **(C)** is quite mild for specific sets of functions, e.g., proper canonical losses are Lipschitz (Reid & Williamson, 2010), so **(C)** can in general be ensured by a simple renormalization of the loss. On the other hand, **(E)** can become progressively harder to ensure as the number of iterations increases because the choices of the WL will become restricted; nevertheless, it is not prohibitory in practice as our experiments in Section 6 suggest (also see the remark in Appendix A.10 for further commentary on this assumption). Let $\tilde{w}_t \doteq \mathbf{1}^\top \boldsymbol{w}_t$, the total weight at iteration t in PILBOOST.

Theorem 11 *Suppose* ***(R, WLA)*** *hold on* WL *and* ***(E, C)*** *hold on* F*, for each iteration of* PILBOOST*. Denote* $Q(F) \doteq 2F^*/(\gamma^2(1-\zeta)^2(1-\pi^2))$*. The following holds:*

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

$$\mathbb{E}_{i \sim \mathcal{S}} \left[F(y_i H_{\boldsymbol{\beta}}(x_i)) \right] \leq F(z^*). \tag{18}$$

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

$$\mathbb{P}_{i \sim \mathcal{S}} \left[y_i H_{\boldsymbol{\beta}}(x_i) \leq \theta \right] \leq \varepsilon. \tag{19}$$

Thus, Theorem 11 gives boosting compliant convergence on training, and the synthesis of (18) and (19) provides a very strong convergence guarantee. When classical assumptions about the loss of interest are satisfied, such as it being Lipschitz (ensured for proper canonical losses (Reid & Williamson, 2010)), there is a natural extension to generalization following standard approaches (Bartlett & Mendelson, 2002; Schapire et al., 1998). See Appendix A.10 for the proof of Theorem 11, and for additional remarks regarding its *optimality* and further application to addressing discrepancies due to *machine type approximations*.

6. Experiments

We provide experimental results on PILBOOST (for $\alpha \in \{1.1, 2, 4\}$) and compare with AdaBoost (Freund & Schapire, 1997) and XGBoost (Chen & Guestrin, 2016) on four binary classification datasets, namely, cancer (Wolberg et al., 1995), xd6 (Buntine & Niblett, 1992), diabetes (Smith et al., 1988), and online shoppers intention (Sakar et al., 2019). We performed 10 runs per algorithm with randomization over the train/test split and the twisters. All experiments use regression decision trees (of varying depths 1-3) in order to align with XGBoost. Hyperparameters of XGBoost were kept to default to maintain the fairest comparison between the three algorithms; for more of these experimental details, please refer to Appendix B.5. In order to demonstrate the *twist-properness* of α-loss as implemented in PILBOOST, we corrupt the training examples of these datasets according to three different (malicious) twisters.

Class Noise Twister (all datasets): This twister is equivalent to SLN in the training sample. Results on this twister for the cancer dataset are presented in Figure 1 and see Appendix B.3 for further results. In general, we find that PILBOOST is more robust to the Class Noise Twister than AdaBoost and XGBoost, and we find that α^* increases as the amount of twist increases, which complies with our theory (Lemma 8 and Theorem 9). We also present an *adaptive* α *experiment* in Figure 2. We denote the adaptive method Menon PILBOOST, since we take inspiration from (Menon et al., 2015), where they show that one can estimate the level

Dataset	Algorithm	Feature Noise Twister			
		$p = 0$	0.15	0.25	0.5
xd6	AdaBoost	1.000 ± 0.000	0.988 ± 0.013	0.966 ± 0.013	0.884 ± 0.019
	us ($\alpha = 1.1$)	1.000 ± 0.000	0.998 ± 0.004	0.994 ± 0.006	0.905 ± 0.020
	us ($\alpha = 2.0$)	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	0.910 ± 0.026
	us ($\alpha = 4.0$)	1.000 ± 0.000	0.997 ± 0.006	0.999 ± 0.002	$\mathbf{0.958 \pm 0.017}$
	XGBoost	1.000 ± 0.000	0.970 ± 0.016	0.962 ± 0.009	0.833 ± 0.027

Table 1: Accuracies of AdaBoost, PILBOOST (for $\alpha \in \{1.1, 2, 4\}$), and XGBoost on the xd6 dataset affected by the feature noise twister with the flipping probability $p = \{0, 0.15, 0.25, 0.5\}$. All three algorithms were trained with depth 3 regression trees. For each value of α, we set $a_f = 8$. Note that the xd6 dataset is perfectly classified (when there is no twist) by a Boolean formula on the features, given in (Buntine & Niblett, 1992), which explains the performance when $p = 0$.

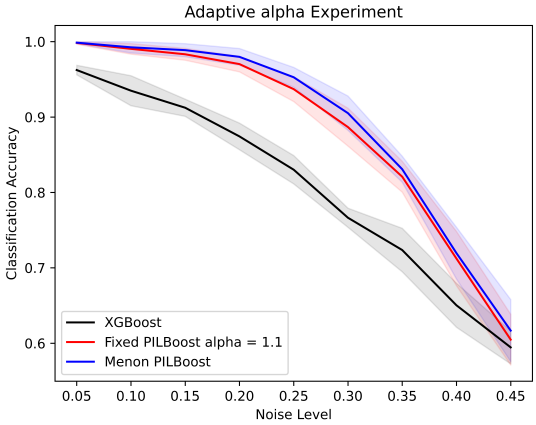

Figure 2: Adaptive α experiment on the xd6 dataset with depth 3 regression trees. Solid curves correspond to mean classification accuracy and shaded areas are the associated 95% confidence intervals obtained from a t-test. For each label noise value, we train three algorithms: 1) vanilla XGBoost; 2) PILBOOST with fixed $\alpha = 1.1$; 3) and, an adaptive α PILBOOST (we refer to as Menon PILBOOST). For details regarding Menon PILBOOST, refer to Class Noise Twister in the main body. The result suggests that a fixed value of $\alpha = 1.1$ in PILBOOST is good, but approximating α_0 does induce slightly better model performance. For general twists, we suggest this heuristic (or some variant) as inspired by (Menon et al., 2015) could be used to learn α_0. Further experimental consideration is given in Appendix B.6.

of label noise (see their Appendix D.1) from the minimum and maximum posterior values. Using a single decision tree classifier with $O(\log(m))$ leaves and $O(\sqrt{m})$ samples per leaf ($m \approx 681$ examples for xd6 dataset with 70/30 train/test-split), and information gain as the splitting criterion, we estimate the minimum and maximum posterior values directly from the training data with local counts of number of samples classified such that $Y = 1$ at each leaf. Once we obtain $\eta_{\min}$ and $\eta_{\max}$ in this way, we estimate the symmetric noise value $p \in [0, 1]$ with the geometric mean $p = \sqrt{\eta_{\min}(1 - \eta_{\max})}$. Finally, to estimate α_0 for each noise level, we apply the formula in Lemma 8**(c)** and the SLN formula given just before Lemma 4 where we estimate η_c with the average posterior from the decision tree classifier. Further experimental consideration is given in Appendix B.6.

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

Insider Twister (online shoppers intention dataset): This twister assumes more knowledge about the model than the previous two twisters. In essence, the insider twister adds noise to a few of the most informative features for predicting the class. Specifically for the online shoppers intention dataset, the insider twister adds noise to feature 8 (*page values* - numeric type with range in $[-250, 435]$), feature 10 (*month*), and feature 15 (*visitor type* - ternary alphabet). For *page values*, the insider twister adds i.i.d. $\mathcal{N}(0, 60)$ to the entries; for both *month* and *visitor type*, the insider twister independently increments (with probability $1/2$) the symbol according to their respective alphabets such that about 50% of each of these features are perturbed. Results on this twister are presented in Figure 3 and further discussion in Appendix B.4 (Figure 13); post-twister, the feature importance profile of XGBoost is almost uniform, displaying damages to the algorithm's discriminative abilities (Figure 3, right), while the feature importance profile of PILBOOST is much less perturbed overall.

7. Conclusion

In this work, we have introduced a generalization of the properness framework via the notion of *twist-properness*,

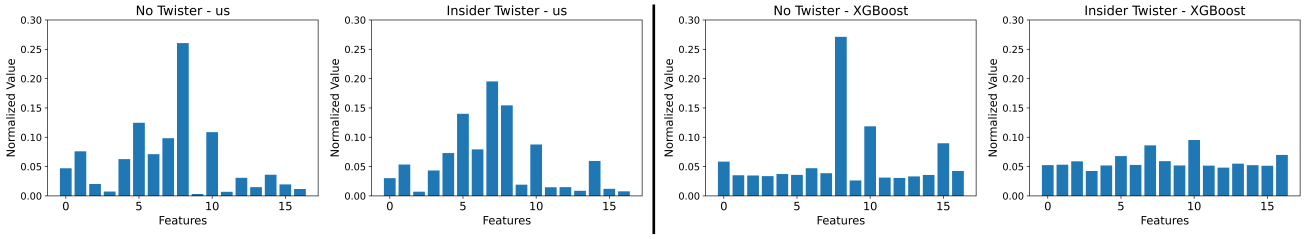

Figure 3: Normalized feature importance profiles for PILBOOST with $\alpha = 1.1$ and $a_f = 7$ (*left*) and for XGBoost (*right*) on the online shoppers intention dataset (both for depth 3 trees) with and without the insider twister. We find that the insider twister significantly perturbs the feature importance of XGBoost as evidenced in the plot (*far right*), and hence significantly reduces the inferential capacity of the learned model. More details can be found in Insider Twister (main body) and Appendix B.4.

which allows delineating loss functions with the ability to "untwist" the twisted posterior into the clean posterior. Notably, we have shown in Lemma 8 that a nontrivial extension of a loss function called α-loss, first introduced in information theory, is twist-proper. For a large class of twists (Definition 3), we have shown in Theorem 9 that while a pointwise estimation of α is optimal, a fixed choice of $\alpha_0 > 1$ readily outperforms the proper choice (log-loss). We then studied the twist-proper α-loss under a novel boosting algorithm, called PILBOOST, which uses the pseudo-inverse link for weight updates, for which we proved convergence guarantees in Theorem 11. Lastly, we have presented experimental results for PILBOOST optimizing α-loss on several twists, and also a method inspired by Menon et al. (2015) to estimate α_0 using *only* training data in Figure 2. A possible next step for this line of work would be to design losses that are both twist-proper *and* algorithmically convenient.

8. Acknowledgments

We thank the anonymous reviewers for their comments and suggestions. This work is supported in part by NSF grants CIF-1901243, CIF-1815361, CIF-2007688, CIF-2134256, SaTC-2031799, and a Google AI for Social Good grant.

References

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

Alon, N., Gonen, A., Hazan, E., and Moran, S. Boosting simple learners. In *STOC'21*, 2021.

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

Andriushchenko, M. and Hein, M. Provably robust boosted decision stumps and trees against adversarial attacks. In *NeurIPS*32*, 2019.

Arimoto, S. Information-theoretical considerations on estimation problems. *Information and control*, 19:181–194, 1971.

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

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

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

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

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

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

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

Castells, T., Weinzaepfel, P., and Revaud, J. SuperLoss: A generic loss for robust curriculum learning. In *NeurIPS*33*, 2020.

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

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

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

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

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

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

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

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

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

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

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

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.-Q. On calibration of modern neural networks. In *34th ICML*, pp. 1321–1330, 2017.

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

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

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

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

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

Liao, J., Kosut, O., Sankar, L., and du Pin Calmon, F. A tunable measure for information leakage. In *2018 IEEE International Symposium on Information Theory, ISIT 2018*, pp. 701–705. IEEE, 2018.

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

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

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

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

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th ICLR*, 2018.

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

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

Menon, A. K., van Rooyen, B., Ong, C.-S., and Williamson, R.-C. Learning from corrupted labels via class probability estimation. In *32nd ICML*, pp. 125–134, 2015.

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

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

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

Nock, R. and Williamson, R.-C. Lossless or quantized boosting with integer arithmetic. In *36th ICML*, pp. 4829–4838, 2019.

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

Reid, M.-D. and Williamson, R.-C. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

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

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

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

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

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

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

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

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

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

Sypherd, T., Diaz, M., Cava, J. K., Dasarathy, G., Kairouz, P., and Sankar, L. A tunable loss function for robust classification: Calibration, landscape, and generalization. *IEEE Transactions on Information Theory*, pp. 1–1, 2022. doi: 10.1109/TIT.2022.3169440.

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

Thekumparampil, K.-K., Jain, P., Netrapalli, P., and Oh, S. Projection efficient subgradient method and optimal nonsmooth Frank-Wolfe method. In *NeurIPS*33*, 2020.

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

Truong, L., Jones, C., Hutchinson, B., August, A., Praggastis, B., Jasper, R., Nichols, N., and Tuor, A. Systematic evaluation of backdoor data poisoning attacks on image classifiers. In *CVPR'20*, pp. 3422–3431, 2020.

Valiant, L. G. A theory of the learnable. *Communications of the ACM*, 27:1134–1142, 1984.

Valiant, L. G. Learning disjunctions of conjunctions. In *Proc. of the 9 th International Joint Conference on Artificial Intelligence*, pp. 560–566, 1985.

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

Zhang, H., Cisse, M., Dauphin, Y.-D., and Lopez-Paz, D. *mixup*: beyond empirical risk minimization. In *6th ICLR*, 2018.

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

Zhang, Y., Niu, G., and Sugiyama, M. Learning noise transition matrix from only noisy labels via total variation regularization. In *38th ICML*, pp. 12501–12512, 2021.

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

Appendices for "Being Properly Improper"

A. Proofs, Further Theoretical Results, and Additional Commentary

A.1. Illustration of Proper Loss to Surrogate through the Convex Conjugate

In this subsection, we provide a worked-out example for how picking the log-loss as ℓ gives the binary entropy for $\underline{L}$ and the logistic loss for F.

From (Reid & Williamson, 2010), we have that the log-loss has partial losses $\ell_1(u) = -\log u$, $\ell_{-1}(u) = -\log(1-u)$ and is a proper loss. In order to compute the (pointwise) *Bayes* risk $\underline{L}$ for the log-loss, we first obtain from (2),

$$L(u, v) = v \cdot \ell_1(u) + (1-v) \cdot \ell_{-1}(u) = v \cdot -\log u + (1-v) \cdot -\log(1-u). \tag{20}$$

Recall that $\underline{L}(v) \doteq \inf_u L(u, v)$. In (20), taking the derivative with respect to u and setting the expression equal to zero, i.e., $\frac{d}{du} L(u, v) = 0$, and solving for u, obtains that $u = v$, in other words, the log-loss is indeed proper. Plugging $u = v$ back into (20), we find that the pointwise Bayes risk of the log-loss is

$$\underline{L}(v) = -v \log v - (1-v) \cdot \log(1-v), \tag{21}$$

which is indeed the binary (Shannon) entropy (Thomas & Joy, 2006). Finally, to obtain the logistic loss as the surrogate, we compute the convex conjugate of (21). Formally, we have from (3) that $\forall z \in \mathbb{R}$,

$$F(z) = (-\underline{L})^\star(-z) = (v \log v + (1-v) \cdot \log(1-v))^\star(-z). \tag{22}$$

Indeed, we have that $\forall z \in \mathbb{R}$,

$$(v \log v + (1-v) \cdot \log(1-v))^\star = \sup_v \{z \cdot v - v \log v - (1-v) \cdot \log(1-v)\}, \tag{23}$$

which is similarly obtained by setting the derivative equal to zero and solving, i.e.,

$$\frac{d}{dv} \left[z \cdot v - v \log v - (1-v) \cdot \log(1-v)\right] = 0 \tag{24}$$

$$z + \log(1-v) - \log v = 0 \tag{25}$$

$$z = \log\left(\frac{v}{1-v}\right) \tag{26}$$

$$v = \frac{1}{1+e^{-z}}, \tag{27}$$

which is obtained after a few steps of algebra. Plugging (27) back into (23), we obtain that

$$\sup_v \{z \cdot v - v \log v - (1-v) \cdot \log(1-v)\} = \frac{z}{1+e^{-z}} + \frac{\log(1+e^{-z})}{1+e^{-z}} + \frac{\log(1+e^{z})}{1+e^{z}}. \tag{28}$$

Noticing that $\log(1+e^{-z}) = \log((e^{-z}) \cdot (1+e^{z}))$, we have from (28) that

$$\sup_v \{z \cdot v - v \log v - (1-v) \cdot \log(1-v)\} = \log(1+e^{z}). \tag{29}$$

Plugging this back into (22), we have that for the log-loss, the surrogate is given by $\forall z \in \mathbb{R}$,

$$F(z) = (-\underline{L})^\star(-z) = \log(1+e^{-z}), \tag{30}$$

which is indeed the logistic loss (cf. (Bartlett et al., 2006)), useful in the margin setting.

A.2. Proof of Lemma 1

We study $U \doteq (-\underline{L})^\star$, which is convex by definition, and show that it is non-decreasing. Monotonicity follows from the non-negativity of the argument of the partial losses and the definition of the convex conjugate: suppose $z' \geq z$ and let $u^* \in \arg\sup_u zu + \underline{L}(u)$. We have

$$\begin{aligned} U(z') &\doteq \sup_{u\in[0,1]} z'u + \underline{L}(u) && (31)\\ &= \sup_{u\in[0,1]} (z'-z)u + zu + \underline{L}(u) && (32)\\ &\geq (z'-z)u^* + zu^* + \underline{L}(u^*) && (33)\\ &= (z'-z)u^* + U(z) && (34)\\ &\geq U(z), && (35)\end{aligned}$$

which completes the proof that U is non-decreasing and therefore $F(z) \doteq U(-z)$ non-increasing.

Concavity of $\underline{L}$ follows from definition. We show continuity of $\underline{L}$, the continuity of F then following from the definition of the convex conjugate F (Boyd & Vandenberghe, 2004). Let $a, u \in (0,1)$, let $u^* \in t_\ell(u), a^* \in t_\ell(a)$. We get:

$$\begin{aligned} \underline{L}(u) &\doteq u\ell_1(u^*) + (1-u)\ell_{-1}(u^*) && (36)\\ &\leq u\ell_1(a^*) + (1-u)\ell_{-1}(a^*) && (37)\\ &= \underline{L}(a) + (u-a)(\ell_1(a^*) - \ell_{-1}(a^*)), && (38)\end{aligned}$$

(the inequality holds since otherwise $u^* \notin t_\ell(u)$) Permuting the roles of u and a, we also get

$$\underline{L}(a) \leq \underline{L}(u) + (a-u)(\ell_1(u^*) - \ell_{-1}(u^*)), \quad (39)$$

from which we get

$$|\underline{L}(a) - \underline{L}(u)| \leq Z \cdot |a-u|, \quad (40)$$

with $Z \doteq \max_{v\in\{a,u\}} \sup|\ell_1(t_\ell(v)) - \ell_{-1}(t_\ell(v))|$ (where we use set differences if t_ℓs are not singletons). Since $Z \ll \infty$, (40) is enough to show the continuity of $\underline{L}$ (we have by assumption $\mathrm{dom}(\underline{L}) = [0,1]$).

A.3. Bayes Tilted Estimates

The proof of Lemma 2 readily follows from Definition 4 and standard properties of convex functions, see e.g., (Boyd & Vandenberghe, 2004).

Below, we also provide analysis of the properties of Bayes tilted estimates for more general losses which induce set-valued functions. Following convention, we denote the set valued inequality $A \leq B$, such that, $\forall a \in A, \exists b \in B, a \leq b$ and the set-valued (Minkowski) difference $A - B \doteq \{a - b : a \in A, b \in B\}$.

Lemma 12 *The following properties of t_ℓ follow from assumptions M, D or S on partial losses:*
***(M)** implies set-valued monotonicity: $\forall u_1 < u_3 \in [0,1]$, we have $t_\ell(u_1) \leq t_\ell(u_3)$ and $t_\ell(u_1) \cap t_\ell(u_3) \subseteq t_\ell(u_2), \forall u_2 \in (u_1, u_3)$;*
***(D)** and t_ℓ differentiable imply monotonicity: $\forall u \in [0,1]$, $\ell_1'(t_\ell(u)) \leq \ell_{-1}'(t_\ell(u)) \Leftrightarrow t_\ell'(u) \geq 0$;*
***(S)** implies set-valued symmetry: $t_\ell(1-u) = \{1\} - t_\ell(u), \forall u \in [0,1]$;*
***(E)** Extreme values: $\ell_1(1) = \ell_{-1}(0) = 0$, $\ell_1([0,1]) \subseteq \mathbb{R}_+, \ell_{-1}([0,1]) \subseteq \mathbb{R}_+$. Further, this implies properness on extreme values, as $0 \in t_\ell(0), 1 \in t_\ell(1)$.*

Case (M) – Suppose $t_\ell(a) \cap t_\ell(a') \neq \emptyset$ for some $a \neq a'$ and let $v^* \in t_\ell(a) \cap t_\ell(a')$. It means $\forall v \in [0,1]$,

$$\begin{aligned} a\ell_1(v^*) + (1-a)\ell_{-1}(v^*) &\leq a\ell_1(v) + (1-a)\ell_{-1}(v), && (41)\\ a'\ell_1(v^*) + (1-a')\ell_{-1}(v^*) &\leq a'\ell_1(v) + (1-a')\ell_{-1}(v), && (42)\end{aligned}$$

and so $\forall \delta \in [0,1]$, if we let $a_\delta \doteq a + \delta(a'-a)$, a $1-\delta, \delta$ convex combination of both inequalities yields $\forall v \in [0,1]$,

$$a_\delta\ell_1(v^*) + (1-a_\delta)\ell_{-1}(v^*) \leq a_\delta\ell_1(v) + (1-a_\delta)\ell_{-1}(v), \forall v \in [0,1], \quad (43)$$

which implies $v^* \in t_\ell(a_\delta)$ and shows the right part of **Case (M)**.
To show show the left part of **Case (M)**; we add to (41) and (42) we now add the inequality:

$$a\ell_1(v^\circ) + (1-a)\ell_{-1}(v^\circ) \quad \leq \quad a\ell_1(v) + (1-a)\ell_{-1}(v), \tag{44}$$

with therefore $v^\circ \in t_\ell(a)$, implying $a\ell_1(v^\circ) + (1-a)\ell_{-1}(v^\circ) = a\ell_1(v^*) + (1-a)\ell_{-1}(v^*)$ as otherwise one of v°, v^* would not be in $t_\ell(a)$. We then get

$$\begin{aligned} a'\ell_1(v^\circ) + (1-a')\ell_{-1}(v^\circ) &= a\ell_1(v^\circ) + (1-a)\ell_{-1}(v^\circ) + (a'-a)\cdot(\ell_1(v^\circ) - \ell_{-1}(v^\circ)) \\ &= a\ell_1(v^*) + (1-a)\ell_{-1}(v^*) + (a'-a)\cdot(\ell_1(v^\circ) - \ell_{-1}(v^\circ)) \\ &= a'\ell_1(v^*) + (1-a')\ell_{-1}(v^*) + (a'-a)\cdot\Delta, \end{aligned} \tag{45}$$

with $\Delta \doteq \ell_1(v^\circ) - \ell_{-1}(v^\circ) - (\ell_1(v^*) - \ell_{-1}(v^*))$. Considering (45), we deduce from (42) that to have $v^\circ \in t_\ell(a')$, we equivalently need $(a'-a)\cdot\Delta \leq 0$. We also know by assumption that ℓ_1 is non-increasing and ℓ_{-1} is non-decreasing, so $g(u) \doteq \ell_1(u) - \ell_{-1}(u)$ is non-increasing. We thus have $(a'-a)\cdot\Delta \leq 0$ iff one of the two possibilities hold:

- $a' \geq a$ and $v^\circ \geq v^*$, or
- $a' \leq a$ and $v^\circ \leq v^*$,

which shows the right part of **Case (M)**.

Case (E) – we have $\underline{L}(0) = \inf_{v\in[0,1]} \ell_{-1}(v) = 0$ for $v = 0$, hence $0 \in t_\ell(0)$. Similarly, $\underline{L}(1) = \inf_{v\in[0,1]} \ell_1(v) = 0$ for $v = 1$, hence $1 \in t_\ell(1)$.

Case (D) – we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}u}\underline{L}(u) &= \ell_1(t_\ell(u)) + u\ell_1'(t_\ell(u))t_\ell'(u) - \ell_{-1}(t_\ell(u)) + (1-u)\ell_{-1}'(t_\ell(u))t_\ell'(u) \\ &= \ell_1(t_\ell(u)) - \ell_{-1}(t_\ell(u)) + t_\ell'(u)\cdot(u\ell_1'(t_\ell(u)) + (1-u)\ell_{-1}'(t_\ell(u))), \end{aligned} \tag{46}$$

but since $v = t_\ell(u)$ is the solution to (41) it satisfies $u\ell_1'(t_\ell(u)) + (1-u)\ell_{-1}'(t_\ell(u)) = 0$, so that (46) simplifies to

$$\frac{\mathrm{d}}{\mathrm{d}u}\underline{L}(u) \quad = \quad \ell_1(t_\ell(u)) - \ell_{-1}(t_\ell(u)), \tag{47}$$

and since $\underline{L}$ is concave and the partial losses are differentiable,

$$\frac{\mathrm{d}^2}{\mathrm{d}u^2}\underline{L}(u) \quad = \quad t_\ell'(u)\cdot(\ell_1'(t_\ell(u)) - \ell_{-1}'(t_\ell(u))) \leq 0, \forall u, \tag{48}$$

which proves the statement of the Lemma.

Case (S) – Suppose $v^* \in t_\ell(a)$, which implies

$$a\ell_1(v^*) + (1-a)\ell_{-1}(v^*) \quad \leq \quad a\ell_1(v) + (1-a)\ell_{-1}(v), \forall v \in [0,1]. \tag{49}$$

We also note that since symmetry holds, $a\ell_1(v^*) + (1-a)\ell_{-1}(v^*) = (1-a)\ell_1(1-v^*) + a\ell_{-1}(1-v^*)$, which implies because of (49) $1 - v^* \in t_\ell(1-a)$.

Remark: even if we assume the partial losses to be strictly monotonic, the tilted estimate can still be set valued. To see this, craft the partial losses such that $v \in t_\ell(u)$ and then for some $w > v$, replace the partial losses in the interval $[v, w]$ by affine parts w/ slope $-a < 0$ for ℓ_1, $b > 0$ for ℓ_{-1} and such that $b/a = u/(1-u)$ which guarantees $L(u, v) = L(u, w)$ and thus $w \in t_\ell(u)$;

A.4. Proof of Lemma 6

We recall the focal loss' corresponding pointwise conditional risk in lieu of (2):

$$L_\gamma(u,v) \doteq -v\cdot(1-u)^\gamma\log u-(1-v)\cdot u^\gamma\log(1-u), \tag{50}$$

and if it is twist proper, then for any $\eta_{\rm t},\eta_{\rm c}\in[0,1]$, there exists $\gamma\geq 0$ such that

$$\frac{\partial}{\partial u}L_\gamma(u,\eta_{\rm t})\Big|_{u=\eta_{\rm c}} = 0. \tag{51}$$

Equivalently, we must find $\gamma\geq 0$ such that (keeping notations $u\doteq\eta_{\rm c}, v\doteq\eta_{\rm t}$ for clarity):

$$(1-v)u^\gamma\cdot(\gamma(1-u)\log(1-u)-u) = v(1-u)^\gamma\cdot(\gamma u\log u+u-1)\,, \tag{52}$$

and we see that twist properness implies the statement that for any $K\geq 0$ (note that $K=\frac{v}{1-v}$) and any $u\in[0,1)$, there exists $\gamma\geq 0$ such that

$$\frac{f(u,\gamma)}{f(1-u,\gamma)} = K, \tag{53}$$

$$f(u,\gamma) \doteq u^\gamma\cdot(\gamma(1-u)\log(1-u)-u)\,. \tag{54}$$

We study the ratio for $u\in[0,1/2]$. We have $f(u,\gamma)\leq 0, \forall u\in[0,1], \forall\gamma\geq 0$ and

$$\frac{\partial}{\partial\gamma}f(u,\gamma) = u^\gamma(\gamma\cdot a(u)-b(u)), \tag{55}$$

with $a(u)\doteq(1-u)\log(1-u)\log(u)\geq 0$ and $b(u)\doteq u\log u-(1-u)\log(1-u)$, satisfying $b(1-u)=-b(u)$ and $ua(1-u)=(1-u)a(u)$. Hence, we arrive after some derivations to

$$\frac{\partial}{\partial\gamma}\frac{f(u,\gamma)}{f(1-u,\gamma)} = \frac{(u(1-u))^\gamma}{f^2(1-u,\gamma)}\cdot\left(A(u)\gamma^2+B(u)\gamma+C(u)\right), \tag{56}$$

$$A(u) \doteq u(1-u)\log(u)\log(1-u)\log(u(1-u)), \tag{57}$$

$$B(u) \doteq -(u^2\log^2 u+(1-u)^2\log^2(1-u)+(1-2u)^2\log u\log(1-u)), \tag{58}$$

$$C(u) \doteq (1-2u)b(u). \tag{59}$$

All functions A,B,C are non positive for any fixed $u\in[0,1/2]$, so the ratio in (53) is non-increasing over $\gamma\geq 0$ and as a consequence, for any fixed $u\in[0,1/2]$,

$$\frac{f(u,\gamma)}{f(1-u,\gamma)} \leq \frac{f(u,0)}{f(1-u,0)} \tag{60}$$

$$= \frac{u}{1-u}, \forall\gamma\geq 0, \tag{61}$$

so we see that (53) cannot be satisfied when $K>u/(1-u)$ and as a consequence, the focal loss is not twist-proper.

Twist-improperness of the Super Loss The Super Loss (Castells et al., 2020) works as a "wrapper" of a loss, its partial losses being defined as

$$L_{b,\lambda}(\ell,\sigma_i)=(\ell_b-\tau)\sigma_i+\lambda(\log\sigma_i)^2, \tag{62}$$

where $b\in\{-1,1\}$ indicates the partial loss of a loss of interest, $\tau\in{\rm Im}\ell_b, \lambda>0$ are user-defined parameters. σ_i is a functional computed to minimize the partial losses, and we get the optimal expression:

$$\sigma^*(\ell_b)=\exp\left(-W(1/2\max\left(-2/e,\beta\right))\right), \tag{63}$$

with $\beta=\frac{\ell_b-\tau}{\lambda}$ (notice this is also a function via the partial loss). W is called Lambert's function. It does not have an analytical form.

Lemma 13 *Suppose loss ℓ in the Super Loss is such that its partial loss ℓ_1 is strictly decreasing and ℓ_{-1} is strictly increasing. Then the corresponding Super Loss with partial losses $L_{b,\lambda}(\ell, \sigma^*(\ell_b))$ ($b \in \{-1, 1\}$) is not twist proper.*

Remark: the assumptions about the partial losses are very weak and would be satisfied by all popular losses (e.g. log, square, Matusita, etc.).

Proof The notable facts about W, useful for our proof are:

$$e^{W(z)} = \frac{z}{W(z)} \quad \text{and} \quad \frac{d}{dz}W(z) = \frac{1}{z + e^{W(z)}} \quad \text{and} \quad \sup \exp(-W(z)) = e. \tag{64}$$

Simplifying notations above, we end up studing a loss with partial losses defined as

$$L^*_\lambda(\ell_b) = (\ell_b - \tau)\sigma_i + \lambda(\log \sigma^*(\ell_b))^2. \tag{65}$$

Recall that a loss ℓ is twist-proper iff for any twist, there exists hyperparameters such that $\eta_c \in t_\ell(\eta_t)$. Examining this for the Super Loss, we obtain

$$\begin{aligned}
t_L(v) &= \text{arginf}_{u\in[0,1]} L(u, v) && (66)\\
&= \text{arginf}_{u\in[0,1]} v \cdot L^*_\lambda(\ell_1(u)) + (1 - v) \cdot L^*_\lambda(\ell_{-1}(u)) && (67)\\
&= \text{arginf}_{u\in[0,1]} \left\{ \begin{array}{l} v \cdot \left[(\ell_1(u) - \tau)\sigma^*(\ell_1) + \lambda(\log \sigma^*(\ell_1))^2\right] \\ +(1 - v) \cdot \left[(\ell_{-1}(u) - \tau)\sigma^*(\ell_{-1}) + \lambda(\log \sigma^*(\ell_{-1}))^2\right] \end{array} \right. && (68)
\end{aligned}$$

We note that if ℓ is proper, then $v \in t_\ell(v)$. Computing the minimum in (68), we obtain

$$0 = \frac{d}{du} v \cdot \left[(\ell_1(u) - \tau)\sigma^*(\ell_1) + \lambda(\log \sigma^*(\ell_1))^2\right] + (1 - v) \cdot \left[(\ell_{-1}(u) - \tau)\sigma^*(\ell_{-1}) + \lambda(\log \sigma^*(\ell_{-1}))^2\right]. \tag{69}$$

Similar to the computation of the focal loss, we need

$$\frac{\frac{d}{du}(\ell_1(u) - \tau)\sigma^*(\ell_1) + \lambda(\log \sigma^*(\ell_1))^2}{\frac{d}{du}(\ell_{-1}(u) - \tau)\sigma^*(\ell_{-1}) + \lambda(\log \sigma^*(\ell_{-1}))^2} = -\frac{(1 - v)}{v} = -K, \tag{70}$$

and this needs to hold (via the choice of parameters τ, λ) for any $u \in [0, 1)$ and $K > 0$. To save notations, define

$$\beta_b(u) \quad = \quad \frac{\ell_b(u) - \tau}{2\lambda}.$$

Remark that *if* $\ell_b(u) > \tau - (2\lambda)/e$, we have $\sigma^*(\ell_b(u)) = \exp(-W(\beta_b(u)))$ and so

$$\begin{aligned}
&\frac{d}{du}(\ell_b(u) - \tau)\sigma^*(\ell_b(u)) + \lambda(\log \sigma^*(\ell_b(u)))^2 \\
&= 2\lambda \cdot \frac{d}{du}\left[\beta_b(u) \exp(-W(\beta_b(u))) + \frac{W^2(\beta_b(u))}{2}\right] \\
&= 2\lambda \cdot \left[\beta'_b(u) \exp(-W(\beta_b(u))) - \beta_b(u)\beta'_b(u) \cdot \frac{\exp(-W(\beta_b(u)))}{\beta_b(u) + \exp(W(\beta_b(u)))} + \frac{\beta'_b(u)W(\beta_b(u))}{\beta_b(u) + \exp(W(\beta_b(u)))}\right] \\
&= 2\lambda\beta'_b(u) \cdot \frac{1 + \beta_b(u) \exp(-W(\beta_b(u))) - \beta_b(u) \exp(-W(\beta_b(u))) + W(\beta_b(u))}{\beta_b(u) + \exp(W(\beta_b(u)))} \\
&= \ell'_b(u) \cdot \frac{1 + W(\beta_b(u))}{\beta_b(u) + \exp(W(\beta_b(u)))} && (71)\\
&= \ell'_b(u) \cdot \exp(-W(\beta_b(u))), && (72)
\end{aligned}$$

since indeed it comes from (64),

$$\frac{1 + W(z))}{z + \exp(W(z))} \quad = \quad \exp(-W(z)); \tag{73}$$

also, *if* $\ell_b(u) \leq \tau - (2\lambda)/e$, we have $\sigma^*(\ell_b(u)) = \exp(-W(-1/e)) = e$ and so

$$\frac{d}{du}(\ell_b(u) - \tau)\sigma^*(\ell_b(u)) + \lambda(\log \sigma^*(\ell_b(u)))^2 \quad = \quad \ell_b'(u) \cdot e. \tag{74}$$

Since $\lim_{z \to -1/e^+} \exp(-W(z)) = e$, we can summarize both (72) and (74) as

$$\frac{d}{du}(\ell_b(u) - \tau)\sigma^*(\ell_b(u)) + \lambda(\log \sigma^*(\ell_b(u)))^2 \quad = \quad \ell_b'(u) \cdot \exp(-W(\max\{-1/e, \beta_b(u)\})). \tag{75}$$

Now consider a loss ℓ satisfying ℓ_{-1} strictly increasing and ℓ_1 strictly decreasing. Pick u so that we have simultaneously

$$\ell_1(u) \quad > \quad \tau - \frac{2\lambda}{e}, \tag{76}$$

$$\ell_{-1}(u) \quad \leq \quad \tau - \frac{2\lambda}{e}, \tag{77}$$

which, assuming both inequalities fit in the range of the respective partial loss, that $u \in [0, \gamma]$ for some $\gamma > 0$. Rewriting (70), we need to show that for any such $\gamma > 0$ and $u \in [0, \gamma]$ and $K > 0$, there exists a choice of τ, λ such that

$$\frac{\ell_1'(u) \cdot \exp(-W(\beta_1(u)))}{\ell_{-1}'(u) \cdot e} \quad = \quad -K, \tag{78}$$

which rewrites conveniently as

$$\exp(-W(\beta_1(u))) \quad = \quad -Ke \cdot \frac{\ell_{-1}'(u)}{\ell_1'(u)}, \tag{79}$$

or,

$$\exp(-W(\beta_1(u))) \quad = \quad K', \tag{80}$$

for any $K' > 0$ (the RHS of (79) is indeed always strictly positive). But $\sup \exp(-W(z)) = e$ (64), so (80) cannot hold and the Super Loss is not twist proper. ■

A.5. Proof of Lemma 8

For part **(a)**: we cite (Sypherd et al., 2022) which demonstrates **(M)**, **(D)**, **(S)** for $\alpha > 0$. With our extension of α-loss, these can also be readily shown for $\alpha < 0$, since they are mapped back to the $\alpha > 0$ losses.

For part **(b)**: we know from Lemma 2 that α-loss, for $\alpha \in \mathbb{R} \setminus \{0, \pm\infty\}$, due to strict convexity, returns a singleton, i.e., $|t_{\ell^\alpha}(\eta_t)| = 1$. With regards to that singleton, we know from (Sypherd et al., 2022; Liao et al., 2018) for $\alpha > 0$ that $t_{\ell^\alpha}(\eta_t) = \frac{\eta_t^\alpha}{\eta_t^\alpha + (1-\eta_t)^\alpha}$. A very similar calculation recovers $t_{\ell^\alpha}(\eta_t) = \frac{\eta_t^{-\alpha}}{\eta_t^{-\alpha} + (1-\eta_t)^{-\alpha}}$ for $\alpha < 0$. Multiplying the numerator and denominator of this expression by $(1-\eta_t)^\alpha$, we can simply write both expressions using $\frac{\eta_t^\alpha}{\eta_t^\alpha + (1-\eta_t)^\alpha}$. Regarding the limit as $\alpha \to \pm\infty$ yielding $t_{\ell^{\pm\infty}}(\eta_t) = \pm 1$ or ∓ 1, this was also already shown by (Sypherd et al., 2022) for $+\infty$ and is similarly (readily) extended for the $\alpha \to -\infty$ case.

For part **(c)**: here, we break entirely new ground. Let $\alpha > 0$. To obtain twist-properness as stipulated in Definition 5, we seek to know for what α the following holds

$$\eta_c = \frac{\eta_t^\alpha}{\eta_t^\alpha + (1-\eta_t)^\alpha}. \tag{81}$$

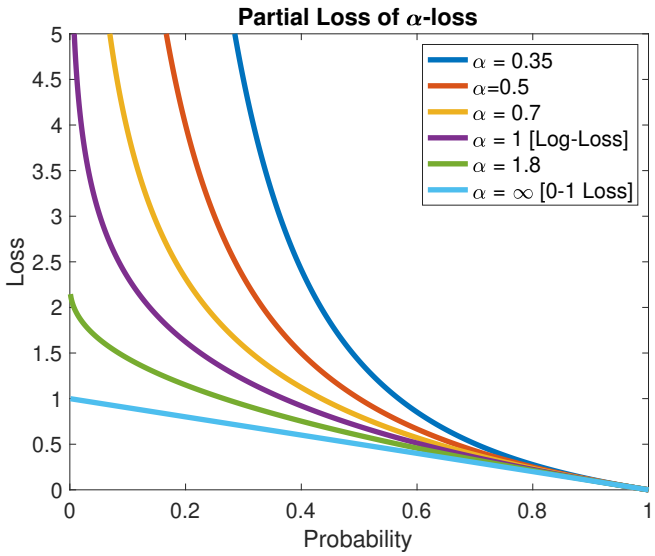

Figure 4: A plot of $\ell_1^{\alpha}(u)$ for $\alpha > 0$ as given in Definition 7.

Solving for α, we obtain

$$\eta_c = \frac{\eta_t^{\alpha}}{\eta_t^{\alpha} + (1-\eta_t)^{\alpha}} \tag{82}$$

$$\frac{1}{\eta_c} = 1 + \left(\frac{1}{\eta_t} - 1\right)^{\alpha} \tag{83}$$

$$\frac{1}{\eta_c} - 1 = \left(\frac{1}{\eta_t} - 1\right)^{\alpha} \tag{84}$$

$$\log\left(\frac{1}{\eta_c} - 1\right) = \alpha \log\left(\frac{1}{\eta_t} - 1\right) \tag{85}$$

$$\alpha^* = \frac{\log\left(\frac{1-\eta_c}{\eta_c}\right)}{\log\left(\frac{1-\eta_t}{\eta_t}\right)}. \tag{86}$$

After multiplying the numerator and denominator by -1, we obtain the desired result. Namely, α-loss is twist-proper for

$$\alpha^* = \frac{\iota(\eta_c)}{\iota(\eta_t)}. \tag{87}$$

Interestingly, (87) is the ratio of the logits (which is the link function $-\underline{L}'$ of the log-loss, $\alpha = 1$) evaluated at the clean posterior and twisted posterior, in essence, a kind of ratio test.

For part **(d)**: we recall the definition of Bayes blunting twist from Definition 3: a twist $\eta_c \mapsto \eta_t$ is Bayes blunting iff $(\eta_c \leq \eta_t \leq 1/2) \vee (\eta_c \geq \eta_t \geq 1/2)$. Also, recall that α^* is given in (87), and see Figure 5 for a plot of the logit function. Let $\eta_c \geq 1/2$. The Bayes blunting twist can take η_t from $\eta_c \geq \eta_t \geq 1/2$. If $\eta_t = \eta_c$, then $\alpha^* = 1$. If $\eta_t \to 1/2$, as can be seen in the figure, the sign crossover point is $1/2$, so $\alpha^* \to \infty$. Thus, by continuity, we have that $\alpha^* \geq 1$. Finally, the case where $\eta_c < 1/2$ follows, *mutatis mutandis*.

A.6. Proof of Theorem 9

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

$$D_{\mathrm{KL}}(\eta_c, \eta_t; 1) > D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha_0) \geq D_{\mathrm{KL}}(\eta_c, \eta_t; \alpha^\star). \tag{88}$$

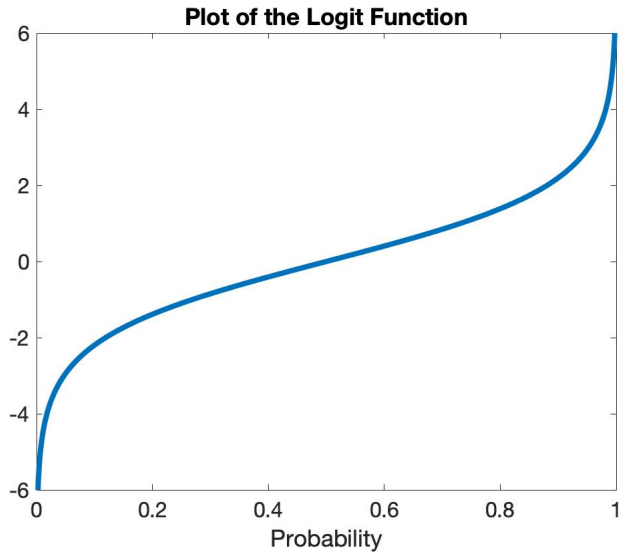

Figure 5: A plot of the logit $\iota(u) \doteq \log(u/(1-u))$.

Recalling (9), which is the identity

$$D_{\text{KL}}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha) = \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha) - H(\eta_{\text{c}}) \tag{89}$$

$$= \mathbb{E}_{\text{X}\sim\text{M}}\left[\eta_{\text{c}}(X)\log\left(\frac{\eta_{\text{c}}(X)}{t_{\ell^\alpha}(\eta_{\text{t}}(X))}\right) + (1-\eta_{\text{c}}(X))\log\left(\frac{1-\eta_{\text{c}}(X)}{1-t_{\ell^\alpha}(\eta_{\text{t}}(X))}\right)\right], \tag{90}$$

by subtracting $H(\eta_{\text{c}})$ from both sides, we rewrite the desired statement (10) (also given here in (88)) as

$$\text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; 1) > \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha_0) \geq \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha^\star). \tag{91}$$

In essence, we want to show that

$$\text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha) < \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; 1) \quad \text{OR} \quad 0 < \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; 1) - \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha), \tag{92}$$

for some $\alpha_0 > 1$. Continuing with the right-hand-side of (92), we have

$$\text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; 1) - \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha) \tag{93}$$

$$= \mathbb{E}_{\text{X}\sim\text{M}}[\eta_{\text{c}}(X)\cdot -\log\eta_{\text{t}}(X) + (1-\eta_{\text{c}}(X))\cdot -\log(1-\eta_{\text{t}}(X))]$$

$$- \mathbb{E}_{\text{X}\sim\text{M}}[\eta_{\text{c}}(X)\cdot -\log t_{\ell^\alpha}(\eta_{\text{t}}(X)) + (1-\eta_{\text{c}}(X))\cdot -\log(1-t_{\ell^\alpha}(\eta_{\text{t}}(X))] \tag{94}$$

$$= \mathbb{E}_{\text{X}\sim\text{M}}\left[\eta_{\text{c}}(X)\log\left(\frac{t_{\ell^\alpha}(\eta_{\text{t}}(X))}{\eta_{\text{t}}(X)}\right) + (1-\eta_{\text{c}}(X))\log\left(\frac{1-t_{\ell^\alpha}(\eta_{\text{t}}(X))}{1-\eta_{\text{t}}(X)}\right)\right] \tag{95}$$

$$= \mathbb{E}_{\text{X}\sim\text{M}}\left[\eta_{\text{c}}(X)\log\left(\frac{\eta_{\text{t}}(X)^{\alpha-1}}{\eta_{\text{t}}(X)^\alpha + (1-\eta_{\text{t}}(X))^\alpha}\right) + (1-\eta_{\text{c}}(X))\log\left(\frac{(1-\eta_{\text{t}}(X))^{\alpha-1}}{(1-\eta_{\text{t}}(X))^\alpha + \eta_{\text{t}}(X)^\alpha}\right)\right], \tag{96}$$

where we used the linearity of the expectation and some algebra to combine the expressions. We want to show that the expression in brackets in (96) is strictly positive as this implies that $0 < \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; 1) - \text{CE}(\eta_{\text{c}}, \eta_{\text{t}}; \alpha)$, in words, that the α-Bayes tilted estimate untwists the Bayes blunting twist. Continuing, we examine the expression in brackets in (96)

$$f_{\alpha,\eta_{\text{c}}}(\eta_{\text{t}}) = \eta_{\text{c}}\log\frac{\eta_{\text{t}}^{\alpha-1}}{\eta_{\text{t}}^\alpha + (1-\eta_{\text{t}})^\alpha} + (1-\eta_{\text{c}})\log\frac{(1-\eta_{\text{t}})^{\alpha-1}}{\eta_{\text{t}}^\alpha + (1-\eta_{\text{t}})^\alpha} \tag{97}$$

$$= (\alpha-1)\eta_{\text{c}}\log\eta_{\text{t}} + (\alpha-1)(1-\eta_{\text{c}})\log(1-\eta_{\text{t}}) - \log(\eta_{\text{t}}^\alpha + (1-\eta_{\text{t}})^\alpha), \tag{98}$$

where we implicitly fix $X = x$ and consider scalar-valued $\eta_{\text{c}}, \eta_{\text{t}} \in [0,1]$ and $\alpha \in \mathbb{R}_+/\{1\}$. We note that $\alpha < 0$ does not need to be considered in this analysis since that regime of α is primarily useful for very strong twists due to its symmetry

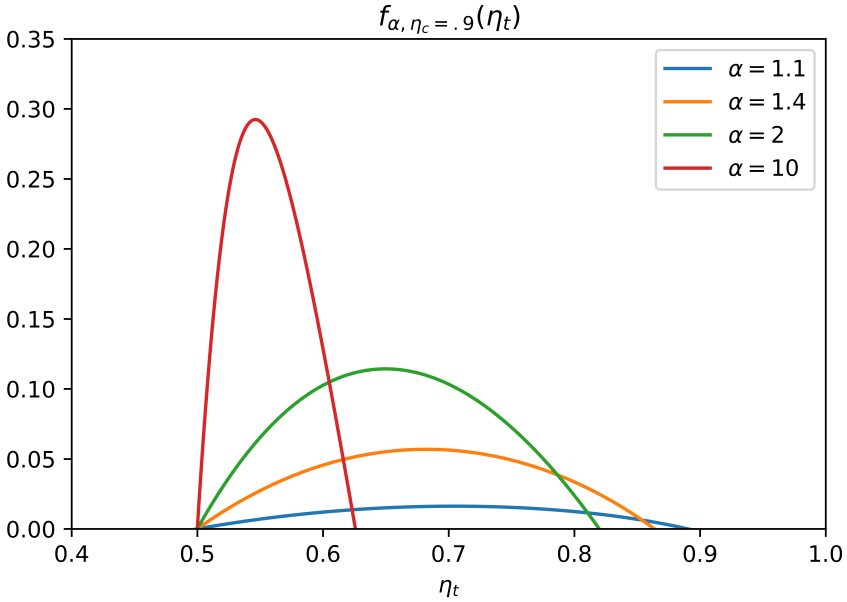

Figure 6: Characteristic plot of the non-negative part of $f_{\alpha,\eta_c}(\eta_t)$, where $\eta_c = 0.9$, as a function of η_t for several values of α. Recall that the non-negative region of f indicates where using Bayes tilted α-estimate, as measured with the cross entropy for α given in (7), is strictly less than the $\alpha = 1$ cross entropy. Also recall that a Bayes blunting twist has the capability to shift η_t anywhere in $[.5, \eta_c = 0.9]$. We see that for small α, more twisted probabilities are "covered", whereas for large α, less twisted probabilities are "covered", however, the large α's induce a large positive magnitude (ultimately measured by the KL-divergence) increase over the proper $\alpha = 1$. A key takeaway is that a fixed α (small enough) can correct a Bayes blunting twist for almost all $x \in \mathcal{X}$. However, it is not necessarily optimal as a perfectly tuned α-mapping will use larger α's to optimally correct strongly twisted posteriors, inducing more gains over the $\alpha = 1$ cross entropy.

property (recall in Lemma 8 that t_{ℓ^α} is symmetric upon permuting (η_t, α) and $(1-\eta_t, -\alpha)$), i.e., *not* useful for Bayes blunting twists which reduce confidence in the posterior but do not flip its sign across $\eta_t - 1/2$. To build intuition of f, see Figure 6 for a plot of this function. Formally, we take note of the following observations/properties of f:

1. **CONTINUITY.** From (98), it can be readily shown that for any fixed $\eta_c, \eta_t \in [0,1]$, $f_{\alpha,\eta_c}(\eta_t)$ is continuous in $\alpha \geq 1$.

2. **CONCAVITY.** For arbitrarily fixed η_c and for any $\alpha > 1$, $f_{\alpha,\eta_c}(\cdot)$ is concave in η_t, since (from (97)) the composition of a concave function with a non-decreasing concave function yields a concave function. As a side note, observe that $f_{\alpha,\eta_c}(\cdot)$ is convex for $0 < \alpha < 1$, thus this regime of α does not untwist Bayes blunting twists. Regarding (increa/decrea)sing concavity of $f_{\alpha,\eta_c}(\cdot)$ for any fixed $\eta_c \in [0,1]$ as a function of α, traditionally a second derivative argument could indicate whether concavity is increasing or decreasing as a function of α. Unfortunately, $\frac{d^2}{d\eta_t^2} f_{\alpha,\eta_c}(\eta_t)$ is an unwieldy analytical expression. However, using a Taylor series approximation of $\frac{d^2}{d\eta_t^2} f_{\alpha,\eta_c}(\eta_t)$ near $\eta_t = 1/2$, we find that the dominating term is $\approx -\alpha^2$. Thus, while not a proof, this *indicates* that concavity of $f_{\alpha,\eta_c}(\cdot)$ increases as α increases greater than 1, which is sufficient for our purposes in the sequel.

3. **ZEROES.** It can be readily shown that for every $\eta_c \in [0,1]$, $f_{\alpha,\eta_c}(1/2) = 0$ for any $\alpha > 1$. Further, it can be shown that for any $\eta_c \in [0,1]$, $\lim_{\alpha \to 1^+} f_{\alpha,\eta_c}(\eta_c) \to 0^-$. Thus, the exact values of η_t for the other zero of $f_{\alpha,\eta_c}(\cdot)$ (not $\eta_t = 1/2$), for each $\alpha > 1$, are given by the solution to the following transcendental equation:

$$\eta_t = \left(\left((1-\eta_t)^{\alpha-1} \right)^{1-\frac{1}{\eta_c}} \left(\eta_t^\alpha + (1-\eta_t)^\alpha \right)^{\frac{1}{\eta_c}} \right)^{\frac{1}{\alpha-1}}, \tag{99}$$

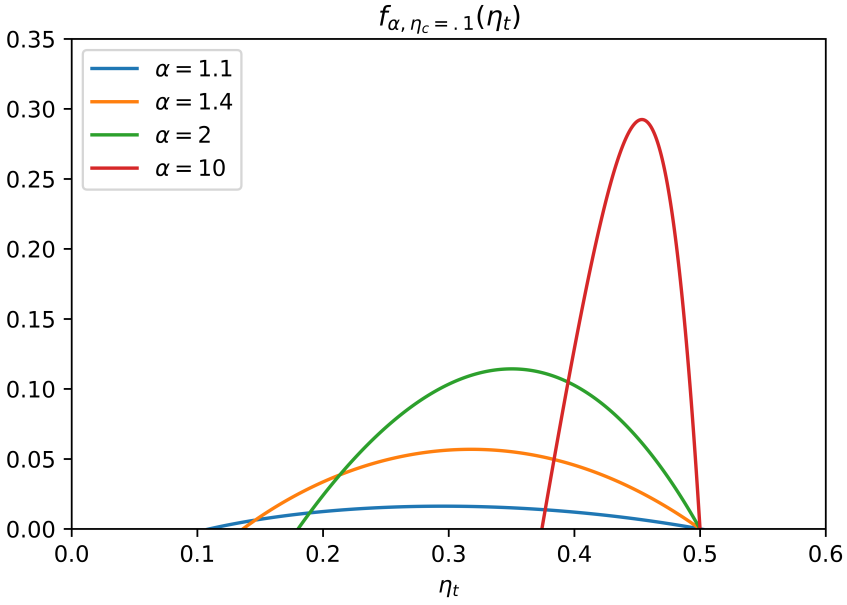

Figure 7: Symmetric image of Figure 6 for $\eta_c = 0.1$.

which can be rewritten as

$$\log\left(\frac{\eta_t}{(1-\eta_t)^{1-\frac{1}{\eta_c}}}\right) = \frac{\alpha}{\alpha-1}\log\left(\eta_t^{\frac{1}{\eta_c}}\right) + \frac{1}{\eta_c(\alpha-1)}\log\left(1+\left(\frac{1}{\eta_t}-1\right)^{\alpha}\right), \tag{100}$$

and can also be rewritten as

$$\left(\frac{\eta_t}{1-\eta_t}\right)^{\eta_c(\alpha-1)} = \eta_t\left(\frac{\eta_t}{1-\eta_t}\right)^{\alpha-1} + (1-\eta_t). \tag{101}$$

Suppose $\eta_c > 1/2$, then since we have a Bayes blunting twist, $\eta_c \geq \eta_t \geq 1/2$. Letting $\alpha \to \infty$, note that the second term on the right-hand-side is 0. Thus, we can solve for the zeroes from (100) when $\alpha = \infty$ by examining

$$\log\left(\frac{\eta_t}{(1-\eta_t)^{1-1/\eta_c}}\right) = \log\left(\eta_t^{\frac{1}{\eta_c}}\right). \tag{102}$$

After some manipulations, we obtain $\log\left(\frac{1}{\eta_t}-1\right) = 0$, which is only satisfied when $\eta_t = 1/2$. Thus, for $\eta_c > 1/2$ and $\alpha \to \infty$, both zeroes of (97) converge at $\eta_t = 1/2$. For $\eta_c < 1/2$, the same argument holds, *mutatis mutandis*. Lastly, from (101), it can be shown that given fixed η_c and η_t under a Bayes blunting twist, a solution $\alpha > 1$ must exist, through reasoning about the rate of increase of $\left(\frac{\eta_t}{1-\eta_t}\right)^{\alpha-1}$ as a function of α, which is common to both sides.

4. **MAXIMUM.** It can also be shown that the maximum of $f_{\alpha,\eta_c}(\cdot)$ for each $\alpha > 1$ as a function of η_t is given by the following transcendental equation

$$\frac{\alpha(1-\eta_c)+\eta_c-\eta_t}{\alpha\eta_c-\eta_c+\eta_t} = \left(\frac{1}{\eta_t}-1\right)^{\alpha}. \tag{103}$$

One key observation we can make from (103) is that as α increases, the term on the right-hand-side grows (or decays) exponentially with α, whereas the term on the left-hand-side is linear in α. With case-by-case analysis, i.e., for $\eta_c > 1/2$ or $\eta_c < 1/2$, it can be reasoned that as α increases, the solution to (103), η_t, approaches $1/2$. A second key observation we can make from (103) is that as $\alpha \to 1^+$, the solution to (103), η_t, approaches $\eta_c/2 + 1/4$. This is readily observed by setting $\alpha = 1+\epsilon$, for some $\epsilon > 0$, and $\eta_t = \frac{\eta_c - 1/2}{2} + 1/2 = \eta_c/2 + 1/4$, along with a Taylor series approximation of $(1/\eta_t - 1)^{1+\epsilon}$ for ϵ near 0.

Intuitively, the remainder of the proof consists of a "covering" argument. In words, we choose the least twisted η_c, i.e. η_c^*, under $\eta_c \to \eta_t$, via its associated $\alpha_0 > 1$ (as given in Lemma 8**(d)**), then we notice that this induces non-negativity of $f_{\alpha_0,\eta_c^*}(\eta_t)$ given in (97). Next, we argue that this choice of $\alpha_0 > 1$ implies that all η_c are "covered" - in the sense of inducing non-negativity of (97), i.e., $f_{\alpha_0,\eta_c}(\eta_t) > 0$ for all η_c under $\eta_c \to \eta_t$. Finally, we use the non-negativity of the expectation to achieve the desired result, i.e., the left-hand-side of (88).

Continuing, let $\eta_c \to \eta_t$ be a *strictly* Bayes blunting twist. Thus, we have that either $(\eta_c < \eta_t \leq 1/2)$ or $(\eta_c > \eta_t \geq 1/2)$ for all η_c. By ZEROES of f, we have that for every $\eta_c \in [0,1]$, $f_{\alpha,\eta_c}(1/2) = 0$ for any $\alpha > 1$ and $\lim_{\alpha\to 1^+} f_{\alpha,\eta_c}(\eta_c) \to 0^-$. We also have that as $\alpha \to \infty$, both zeroes of (97) converge at $\eta_t = 1/2$. Thus, for every η_c, the second zero (the first one is at $\eta_t = 1/2$) continuously shifts (CONTINUITY) from being located at $\eta_t = \eta_c$ (as $\alpha \to 1$) to being located at $\eta_t = 1/2$ (as $\alpha \to \infty$). In order to identify the least twisted η_c under $\eta_c \to \eta_t$, let

$$\alpha^*(\eta_c, \eta_t) := \frac{\iota(\eta_c)}{\iota(\eta_t)}, \tag{104}$$

as stated in Lemma 8**(c)**. By Lemma 8**(d)**, we know that $\alpha^*(\eta_c, \eta_t) \geq 1$; furthermore, due to *strictness* of the Bayes blunting twist $\eta_c \to \eta_t$, we indeed have a *strict inequality*, i.e., $\alpha^*(\eta_c, \eta_t) > 1$ for all η_c under $\eta_c \to \eta_t$. Choose $\alpha_0 := \inf\{\alpha^*(\eta_c, \eta_t) : \forall \eta_c \text{ under } \eta_c \to \eta_t\}$, where we break ties arbitrarily, and note that $\alpha_0 > 1$, again by strictness. Also note that there exists η_c^* associated with α_0 such that $\alpha_0 = \iota(\eta_c^*)/\iota(\eta_t)$ under $\eta_c \to \eta_t$, in other words, $f_{\alpha_0,\eta_c^*}(\eta_t) > 0$ (due to $t_{\ell^\alpha}(\eta_t)$ reversing the effects of the Bayes blunting twist and tuning back to η_c^*). Thus, by CONTINUITY, CONCAVITY, and ZEROES of f above and this choice of $\alpha_0 > 1$, we have that for all η_c, $f_{\alpha_0,\eta_c}(\eta_t) > 0$ under $\eta_c \to \eta_t$, i.e., that all η_c are "covered", due to the ordering of the relative zeroes of f induced by identifying η_c^* through α_0. Therefore, we obtain from (96) and (97) (i.e., the non-negativity of the expectation) that for the chosen $\alpha_0 > 1$, we have that $0 < \text{CE}(\eta_c, \eta_t; 1) - \text{CE}(\eta_c, \eta_t; \alpha_0)$, i.e.,

$$\text{CE}(\eta_c, \eta_t; 1) > \text{CE}(\eta_c, \eta_t; \alpha_0), \tag{105}$$

as desired.

We now show that $\text{CE}(\eta_c, \eta_t; \alpha_0) \geq \text{CE}(\eta_c, \eta_t; \alpha^\star)$, where $\alpha^\star$ is a mapping such that $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$, i.e., returning an $\alpha > 1$ for every $x \in \mathcal{X}$. By MAXIMUM above, we note that for a given η_c, the maximum of $f_{\alpha,\eta_c}(\cdot)$ moves from being achieved at $\eta_t = \eta_c/2 + 1/4$, to being achieved at $\eta_t = 1/2$ in the limit as α increases greater than 1. By CONCAVITY above, we also observe that $f_{\alpha,\eta_c}(\cdot)$ for a fixed η_c *appears* (which is sufficient for the inequality) to become more strongly convex in general as α increases greater than 1. We also note by CONTINUITY of f above that the maximums are continuous in $\alpha > 1$. Thus, under the *strictly* Bayes blunting twist $\eta_c \to \eta_t$, for every η_c, there may exist an $\alpha > 1$ which induces a larger (positive) magnitude in $f_{\alpha,\eta_c}(\eta_t)$ than for the fixed $\alpha_0 > 1$ we found previously (for (105)). Thus, there exists an optimal mapping $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$, such that

$$\text{CE}(\eta_c, \eta_t; \alpha_0) \geq \text{CE}(\eta_c, \eta_t; \alpha^\star). \tag{106}$$

Note that in the degenerate case, $\alpha^\star = \alpha_0$ for every $x \in \mathcal{X}$. Therefore, combining (105) and (106), we obtain

$$\text{CE}(\eta_c, \eta_t; 1) > \text{CE}(\eta_c, \eta_t; \alpha_0) \geq \text{CE}(\eta_c, \eta_t; \alpha^\star), \tag{107}$$

which is the desired result. Lastly, note from Lemma 8**(c)** and (90) that $\alpha^\star : \mathcal{X} \to \mathbb{R}_{>1}$ is indeed given by $\alpha^\star(x) := \iota(\eta_c(x))/\iota(\eta_t(x))$, for every $x \in \mathcal{X}$, and hence note that $\text{CE}(\eta_c, \eta_t; \alpha^\star) = H(\eta_c)$, i.e., from (9) that $D_{\text{KL}}(\eta_c, \eta_t; \alpha^\star) = 0$.

A.7. Proof of Theorem 10

As explained in the main body, we prove a result more general than Theorem 10. However, briefly note that (13), like the statement provided in Theorem 9 in (10), is proved for CE, but the statement provided in the main body as KL is readily obtained by subtracting $H(\eta_c)$ from both sides of the inequality.

First, we need a simple technical Lemma.

Lemma 14 *For any $B > 0$, $\forall |z| \leq B, \forall \alpha \in \mathbb{R}$,*

$$\log(1 + \exp(\alpha z)) \quad \leq \quad \log(1 + \exp(\alpha B)) - \frac{B - z}{2} \cdot \alpha. \tag{108}$$

Proof We first note that $\forall |z| \leq 1, \forall \alpha \in \mathbb{R}$,

$$\log(1+\exp(\alpha z)) \leq \frac{1+z}{2} \cdot \log(1+\exp(\alpha)) + \frac{1-z}{2} \cdot \log(1+\exp(-\alpha)) \tag{109}$$

$$= \log(1+\exp(\alpha)) - \frac{1-z}{2} \cdot \alpha, \tag{110}$$

which indeed holds as the LHS of (109) is convex and the RHS is the equation of a line passing through the points $(-1, \log(1+\exp(-\alpha)))$ and $(1, \log(1+\exp(\alpha)))$. In (110), we use $\log(1+\exp(-\alpha)) = \log(\exp(-\alpha)\cdot(1+\exp(\alpha)))$ on the second term in the RHS. Hence if instead $|z| \leq B$, then

$$\log(1+\exp(\alpha z)) = \log\left(1+\exp\left(\alpha B \cdot \frac{z}{B}\right)\right) \tag{111}$$

$$\leq \log(1+\exp(\alpha B)) - \frac{B-z}{2} \cdot \alpha, \tag{112}$$

as claimed. ■

We now show another Lemma which bounds the log quantities appearing in the cross-entropy in (7), recalling that $\iota(u) \doteq \log(u/(1-u))$.

Lemma 15 *Fix $B > 0$. For any $\boldsymbol{x} \in \mathcal{X}$ such that*

$$\frac{1}{1+\exp B} \leq \eta_t(\boldsymbol{x}) \leq \frac{\exp B}{1+\exp B}, \tag{113}$$

the following properties hold for the Bayes tiltes estimate t_ℓ of α-loss:

$$-\log t_{\ell^\alpha}(\eta_t(\boldsymbol{x})) \leq \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B+\iota(\eta_t(\boldsymbol{x}))}{2},$$

$$-\log(1-t_{\ell^\alpha}(\eta_t(\boldsymbol{x}))) \leq \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B-\iota(\eta_t(\boldsymbol{x}))}{2}.$$

Proof We note, using $z \doteq -\log\left(\frac{1-\eta_t(\boldsymbol{x})}{\eta_t(\boldsymbol{x})}\right)$, which satisfies $|z| \leq B$ from (113) and Lemma 14,

$$-\log t_{\ell^\alpha}(\eta_t(\boldsymbol{x})) = -\log\left(\frac{\eta_t(\boldsymbol{x})^\alpha}{\eta_t(\boldsymbol{x})^\alpha + (1-\eta_t(\boldsymbol{x}))^\alpha}\right)$$

$$= -\log\left(\frac{1}{1+\left(\frac{1-\eta_t(\boldsymbol{x})}{\eta_t(\boldsymbol{x})}\right)^\alpha}\right)$$

$$= \log\left(1+\left(\frac{1-\eta_t(\boldsymbol{x})}{\eta_t(\boldsymbol{x})}\right)^\alpha\right)$$

$$= \log\left(1+\exp\left(\alpha\log\left(\frac{1-\eta_t(\boldsymbol{x})}{\eta_t(\boldsymbol{x})}\right)\right)\right) \tag{114}$$

$$\leq \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B - \log\left(\frac{1-\eta_t(\boldsymbol{x})}{\eta_t(\boldsymbol{x})}\right)}{2} \tag{115}$$

$$= \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B+\iota(\eta_t(\boldsymbol{x}))}{2}, \tag{116}$$

and similarly,

$$-\log(1-t_{\ell^\alpha}(\eta_t(\boldsymbol{x}))) = \log\left(1+\exp\left(-\alpha\log\left(\frac{1-\eta_t(\boldsymbol{x})}{\eta_t(\boldsymbol{x})}\right)\right)\right) \tag{117}$$

$$\leq \log(1+\exp(-\alpha B)) + \alpha \cdot \frac{B+\iota(\eta_t(\boldsymbol{x}))}{2}$$

$$= \log(1+\exp(\alpha B)) - \alpha B + \alpha \cdot \frac{B+\iota(\eta_t(\boldsymbol{x}))}{2}$$

$$= \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B-\iota(\eta_t(\boldsymbol{x}))}{2}, \tag{118}$$

as claimed. ■

Denote $\mathrm{M}(B)$ the distribution restricted to the support for which we have a.s. (113) and let $p(B)$ be the weight of this support in M. Let $\mathrm{M}(\overline{B})$ denote the restriction of M to the complement of this support. We let $\mathrm{D}(B)$ is the product distribution on examples $(\mathcal{X} \times \mathcal{Y})$ over the support of $\mathrm{M}(B)$ induced by marginal $\mathrm{M}(B)$ and posterior η_c (see Reid & Williamson (2011, Section 4)). We are now in a position to show our generalization to Theorem 10.

Theorem 16 *For any fixed $B > 0$, let*

$$e(B) \doteq \frac{\mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim \mathrm{D}(B)}\left[\mathsf{Y} \cdot \iota(\eta_t(\mathsf{X}))\right]}{B} \in [-1, 1]. \tag{119}$$

and suppose we fix the scalar $\alpha \doteq \alpha^{**}$ with

$$\alpha^{**} \doteq \frac{\iota\left(\frac{1+e(B)}{2}\right)}{B}. \tag{120}$$

then the following bound holds on the cross-entropy of the Bayes tilted estimate of the α-loss:

$$\begin{aligned}\mathrm{CE}(\eta_c, \eta_t; \alpha) \leq\ & p(B) \cdot H\left(\frac{1+e(B)}{2}\right) \\ & +(1-p(B)) \cdot \left(e(\overline{B}) \cdot \log\left(\frac{1+|e(B)|}{1-|e(B)|}\right) + \frac{1-|e(B)|}{1+|e(B)|}\right),\end{aligned}$$

where $e(\overline{B}) \doteq \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim \mathrm{D}(\overline{B})}\left[\max\left\{0, -\mathrm{sign}(\alpha^{**}) \cdot \mathsf{Y} \cdot \iota(\eta_t(\mathsf{X}))\right\}\right]/B$ and $\mathrm{D}(\overline{B})$ is is defined analogously to $\mathrm{D}(B)$ with respect to $\mathrm{M}(\overline{B})$.

Remark: we notice this is indeed a generalization of Theorem 10, which corresponds to case $p(B) = 1$. We also note $|e(\overline{B})| \geq 1$.

Proof We remark that the cross-entropy (7) can be split as:

$$\begin{aligned}\mathrm{CE}(\eta_c, \eta_t; \alpha) &\doteq \mathbb{E}_{\mathsf{X}\sim \mathrm{M}}\left[\begin{array}{c}\eta_c(\mathsf{X}) \cdot -\log t_{\ell^\alpha}(\eta_t(\mathsf{X})) \\ +(1-\eta_c(\mathsf{X})) \cdot -\log(1 - t_{\ell^\alpha}(\eta_t(\mathsf{X})))\end{array}\right] \\ &= p(B) \cdot K(\alpha) + (1-p(B)) \cdot L(B),\end{aligned} \tag{121}$$

with

$$K(\alpha) \doteq \mathbb{E}_{\mathsf{X}\sim \mathrm{M}(B)}\left[\begin{array}{c}\eta_c(\mathsf{X}) \cdot -\log t_{\ell^\alpha}(\eta_t(\mathsf{X})) \\ +(1-\eta_c(\mathsf{X})) \cdot -\log(1 - t_{\ell^\alpha}(\eta_t(\mathsf{X})))\end{array}\right], \tag{122}$$

$$J(B) \doteq \mathbb{E}_{\mathsf{X}\sim \mathrm{M}(\overline{B})}\left[\begin{array}{c}\eta_c(\mathsf{X}) \cdot -\log t_{\ell^\alpha}(\eta_t(\mathsf{X})) \\ +(1-\eta_c(\mathsf{X})) \cdot -\log(1 - t_{\ell^\alpha}(\eta_t(\mathsf{X})))\end{array}\right]. \tag{123}$$

We now focus on a bound on $K(\alpha)$, which we achieve via Lemma 15:

$$\begin{aligned}K(\alpha) &\leq \mathbb{E}_{\mathsf{X}\sim \mathrm{M}(B)}\left[\begin{array}{c}\eta_c(\mathsf{X}) \cdot \left(\log(1+\exp(\alpha B)) - \alpha \cdot \frac{B+\iota(\eta_t(\mathsf{X}))}{2}\right) \\ +(1-\eta_c(\mathsf{X})) \cdot \left(\log(1+\exp(\alpha B)) - \alpha \cdot \frac{B-\iota(\eta_t(\mathsf{X}))}{2}\right)\end{array}\right] \\ &= \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B + \mathbb{E}_{\mathsf{X}\sim \mathrm{M}(B)}\left[\eta_c(\mathsf{X})\iota(\eta_t(\mathsf{X})) + (1-\eta_c(\mathsf{X})) \cdot -\iota(\eta_t(\mathsf{X}))\right]}{2} \\ &= \log(1+\exp(\alpha B)) - \alpha \cdot \frac{B + \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim \mathrm{D}(B)}\left[\mathsf{Y} \cdot \iota(\eta_t(\mathsf{X}))\right]}{2} \\ &\doteq \underbrace{\log(1+\exp(\alpha B)) - \alpha \cdot \frac{B + B \cdot e(B)}{2}}_{\doteq L(\alpha)},\end{aligned} \tag{124}$$

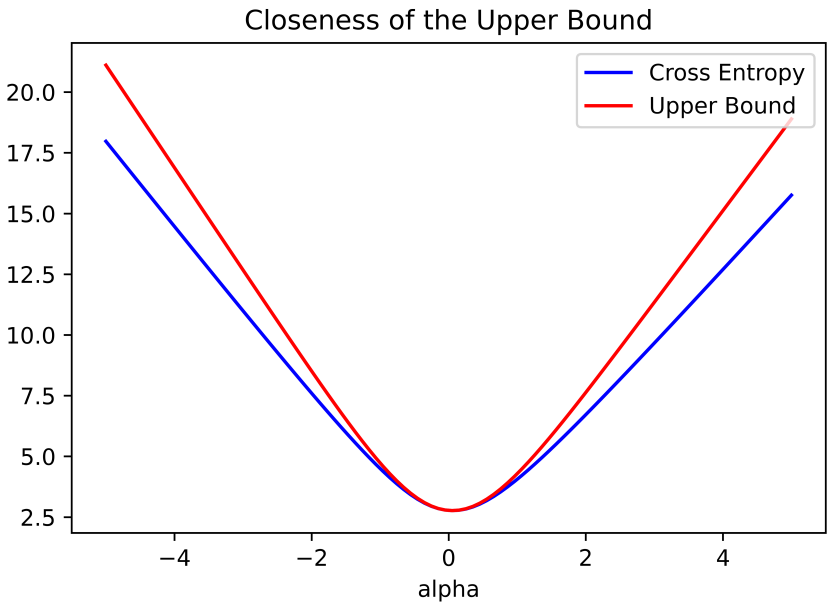

Figure 8: A plot illustrating the closeness of (124) where $K(\alpha)$ is given in blue and $L(\alpha)$ is given in red for a toy distribution: $x \sim \text{Uniform}([-B, B])$ (recall $B > 0$ is the clipping threshold and we set $B = 2$ here) where $\eta_c(x) = (1 + \exp(-x/a))^{-1}$ and $\eta_\iota(x) = (1 + \exp(-x/b))^{-1}$ such that $a = 10$ and $b = 0.6$.

where we recall

$$\mathrm{e}(B) \doteq \frac{\mathbb{E}_{\mathsf{X} \sim \mathrm{D}(B)}\left[\mathsf{Y} \cdot \iota(\eta_\iota(\mathsf{X}))\right]}{B} \in [-1, 1]. \tag{125}$$

Notice the change in distribution in (124), where $\mathrm{D}(B)$ is the product distribution on examples $(\mathcal{X} \times \mathcal{Y})$ over the support of $\mathrm{M}(B)$ induced by marginal $\mathrm{M}(B)$ and posterior η_c (see Reid & Williamson (2011, Section 4)). We have

$$L'(\alpha) = B \cdot \left(\frac{\exp(B\alpha)}{1 + \exp(B\alpha)} - \frac{1 + \mathrm{e}(B)}{2}\right), \tag{126}$$

which zeroes for

$$\alpha^{**} = \frac{1}{B} \cdot \log\left(\frac{1 + \mathrm{e}(B)}{1 - \mathrm{e}(B)}\right) = \frac{\iota(q(B))}{B}. \tag{127}$$

Further, we have that

$$L''(\alpha) = \frac{d}{d\alpha} L'(\alpha) = \frac{B^2 \exp(\alpha B)}{(\exp(\alpha B) + 1)^2}, \tag{128}$$

and plugging in (127) yields

$$L''(\alpha^{**}) = B^2 \frac{1 - \mathrm{e}^2}{4}. \tag{129}$$

Note that for fixed $B > 0$, as $|\alpha|$ increases in (128), $L''(\alpha)$ decreases. Thus, when the magnitude of α^* is large (due to the distribution and twist), this implies that there is more "flatness" near the choice of α^*. Hence, in these regimes, a choice of α_0 "close-enough" to α^* should have similar performance in practice. Continuing with the main line, plugging in (127)

into (124) yields

$$\begin{aligned} K(\alpha^{**}) &\leq \log(1+\exp(B\alpha^{**})) - B\cdot\frac{1+\mathrm{e}(B)}{2}\cdot\alpha^{**} && (130)\\ &= -\log\left(\frac{1-\mathrm{e}(B)}{2}\right) - \frac{1+\mathrm{e}(B)}{2}\cdot\log\left(\frac{1+\mathrm{e}(B)}{1-\mathrm{e}(B)}\right) && (131)\\ &= -\frac{1+\mathrm{e}(B)}{2}\log\left(\frac{1+\mathrm{e}(B)}{2}\right) - \frac{1-\mathrm{e}(B)}{2}\log\left(\frac{1-\mathrm{e}(B)}{2}\right) && (132)\\ &= H\left(\frac{1+\mathrm{e}(B)}{2}\right), && (133)\end{aligned}$$

which is the statement of Theorem 10. We now focus on $J(B)$. Since $\log(1+\exp(-z)) \leq \exp(-z), \forall z$ via an order-1 Taylor expansion, it follows that if $z \geq C$ for some $C > 0$, then $\log(1+\exp(-z)) \leq \exp(-C)$. Equivalently, we get

$$z \geq C \quad \Rightarrow \quad \log(1+\exp(z)) \leq z + \exp(-C). \tag{134}$$

By symmetry, we have

$$z \leq -C \quad \Rightarrow \quad \log(1+\exp(z)) \leq \exp(-C), \tag{135}$$

so we get

$$|z| \geq C \quad \Rightarrow \quad \log(1+\exp(z)) \leq \max\{0, z\} + \exp(-C). \tag{136}$$

By definition, we have for any $\boldsymbol{x}$ in the support of $\mathrm{M}(\overline{B})$,

$$\left|\log\left(\frac{1-\eta_{\mathrm{t}}(\boldsymbol{x})}{\eta_{\mathrm{t}}(\boldsymbol{x})}\right)\right| \quad \geq \quad B, \tag{137}$$

so have, considering $C \doteq B\cdot|\alpha^*|$, from (114) and (117),

$$\begin{aligned} J(B) &\doteq \mathbb{E}_{\mathsf{X}\sim\mathrm{M}(\overline{B})}\left[\begin{array}{c}\eta_{\mathrm{c}}(\mathsf{X})\cdot -\log t_{\ell^\alpha}(\eta_{\mathrm{t}}(\mathsf{X}))\\ +(1-\eta_{\mathrm{c}}(\mathsf{X}))\cdot -\log(1-t_{\ell^\alpha}(\eta_{\mathrm{t}}(\mathsf{X})))\end{array}\right] && (138)\\ &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathrm{D}(\overline{B})}\left[\log\left(1+\exp\left(\mathsf{Y}\alpha^*\log\left(\frac{1-\eta_{\mathrm{t}}(\mathsf{X})}{\eta_{\mathrm{t}}(\mathsf{X})}\right)\right)\right)\right]\\ &\leq \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathrm{D}(\overline{B})}\left[\max\left\{0, \mathsf{Y}\alpha^{**}\log\left(\frac{1-\eta_{\mathrm{t}}(\mathsf{X})}{\eta_{\mathrm{t}}(\mathsf{X})}\right)\right\}\right] + \exp\left(-B\cdot|\alpha^{**}|\right)\\ &= |\alpha^{**}|\cdot\mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim\mathrm{D}(\overline{B})}\left[\max\left\{0, -\mathrm{sign}(\alpha^{**})\cdot\mathsf{Y}\cdot\iota(\eta_{\mathrm{t}}(\mathsf{X}))\right\}\right] + \frac{1-|\mathrm{e}(B)|}{1+|\mathrm{e}(B)|}\\ &= \mathrm{e}(\overline{B})\log\left(\frac{1+|\eta(B)|}{1-|\mathrm{e}(B)|}\right) + \frac{1-|\mathrm{e}(B)|}{1+|\mathrm{e}(B)|}, && (139)\end{aligned}$$

which completes the proof of Theorem 16 after replacing the expression of α^*. ■

Remarks: Theorem 16 calls for several remarks:

Gains with respect to the "proper" choice $\alpha = 1$: the case we develop is simplistic but allows a graphical comparison of the gains that Theorem allow to get compared to the choice $\alpha = 1$, which we recall corresponds to the (proper) logistic loss. Suppose $p(B) = 1$ so the cross-entropy $\mathrm{CE}(\eta_{\mathrm{c}}, \eta_{\mathrm{t}}; \alpha)$ in (121) reduces to $K(.)$, $B = 1$ and all logits take ± 1 value a.e.,

$$z(\boldsymbol{x}) \doteq \log\left(\frac{\eta_{\mathrm{t}}(\boldsymbol{x})}{1-\eta_{\mathrm{t}}(\boldsymbol{x})}\right) \quad = \quad \pm 1, \tag{140}$$

which can be achieved by clamping, and $p \doteq \mathbb{P}_{(\mathsf{X},\mathsf{Y})\sim\mathrm{M}(1)}[\mathsf{Y}z(\mathsf{X}) = 1]$, which gives $\mathrm{e}(1) = 2p - 1$,

$$\alpha^{**} \quad = \quad \log\left(\frac{p}{1-p}\right),$$

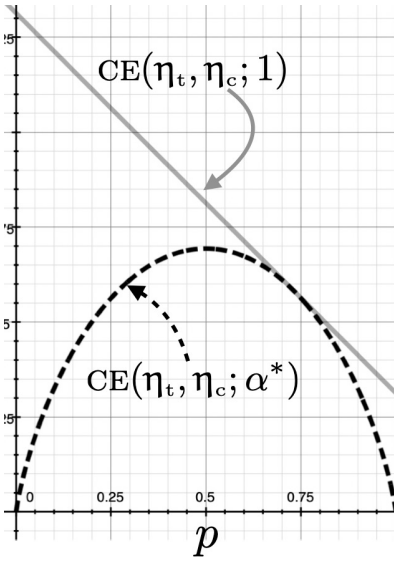

Figure 9: Comparison between the cross-entropy of the logistic loss ($\alpha = 1$) and that of the α-loss for the scalar correction in (141) in Theorem 16.

and

$$\mathrm{CE}(\eta_c, \eta_t; \alpha^{**}) \quad \leq \quad H(p) \tag{141}$$

from Theorem 16. The properness choice $\alpha^* = 1$ however gives

$$\begin{aligned} \mathrm{CE}(\eta_c, \eta_t; 1) = K(1) &= \mathbb{E}_{(\mathsf{X},\mathsf{Y})\sim \mathrm{M}(1)} \left[\log\left(1 + \exp\left(-\mathsf{Y}z(\mathsf{X})\right)\right)\right] &(142)\\ &= p\log(1+\exp(-1)) + (1-p)\log(1+\exp(1)). &(143)\\ &= \log(1+e) - p. &(144) \end{aligned}$$

Figure 9 plots $\mathrm{CE}(\eta_c, \eta_t; \alpha^{**})$ (141) vs $\mathrm{CE}(\eta_c, \eta_t; 1)$ (144). We remark that $\mathrm{CE}(\eta_c, \eta_t; \alpha^{**}) \leq \mathrm{CE}(\eta_c, \eta_t; 1)$, and the difference is especially large as $p \to \{0, 1\}$, for which $\mathrm{CE}(\eta_c, \eta_t; \alpha^{**}) \to 0$ while we always have $\mathrm{CE}(\eta_c, \eta_t; 1) > 0.3, \forall p$.

Incidence of computing α^* on an *estimate* of $\mathsf{e}(B)$: Theorem 16 can be refined if, instead of the true value $\mathrm{e}(B)$ we have access to an estimate $\hat{\mathrm{e}}(B)$. In this case, we can refine the proof of the Theorem from the series of eqs in (133). We remark that

$$H'\left(\frac{1+z}{2}\right) \quad = \quad \frac{1}{2}\cdot\log\left(\frac{1-z}{1+z}\right), \tag{145}$$

so since H is concave, we have for any $\mathrm{e}(B), \hat{\mathrm{e}}(B)$,

$$\begin{aligned} H\left(\frac{1+\mathrm{e}(B)}{2}\right) &\leq H\left(\frac{1+\hat{\mathrm{e}}(B)}{2}\right) + \left(\frac{1+\mathrm{e}(B)}{2} - \frac{1+\hat{\mathrm{e}}(B)}{2}\right)\cdot\frac{1}{2}\cdot\log\left(\frac{1-\hat{\mathrm{e}}(B)}{1+\hat{\mathrm{e}}(B)}\right)\\ &= H\left(\frac{1+\hat{\mathrm{e}}(B)}{2}\right) + \frac{\mathrm{e}(B)-\hat{\mathrm{e}}(B)}{4}\cdot\log\left(\frac{1-\hat{\mathrm{e}}(B)}{1+\hat{\mathrm{e}}(B)}\right)\\ &\leq H\left(\frac{1+\hat{\mathrm{e}}(B)}{2}\right) + \frac{|\mathrm{e}(B)-\hat{\mathrm{e}}(B)|}{4}\cdot\log\left(\frac{1+|\hat{\mathrm{e}}(B)|}{1-|\hat{\mathrm{e}}(B)|}\right)\\ &= H\left(\frac{1+\hat{\mathrm{e}}(B)}{2}\right) + \frac{|\mathrm{e}(B)-\hat{\mathrm{e}}(B)|}{4}\cdot\log\left(1+\frac{2|\hat{\mathrm{e}}(B)|}{1-|\hat{\mathrm{e}}(B)|}\right)\\ &\leq H\left(\frac{1+\hat{\mathrm{e}}(B)}{2}\right) + \frac{|\mathrm{e}(B)-\hat{\mathrm{e}}(B)||\hat{\mathrm{e}}(B)|}{2(1-|\hat{\mathrm{e}}(B)|)}, \end{aligned} \tag{146}$$

where we have used $\log(1+z) \leq z$ for the last inequality.

Polarity of α^{**}: as presented in the main body, the state of the art defines the α-loss only for $\alpha \geq 0$. The proof of Theorem 16, and more specifically its proof, hints at why alleviating this constraint is important and corresponds to especially difficult cases. We have the general rule $\alpha^{**} \leq 0$ iff $\mathrm{e}(B) \leq 0$, which indicates that the twisted posterior tends to be small when the clean posterior tends to be large. Since the Bayes tilted estimate is symmetric if we switch the couple $(\alpha, \eta_{\mathrm{t}})$ for $(-\alpha, 1-\eta_{\mathrm{t}})$, $\alpha^{**} \leq 0$ provokes a change of polarity in the Bayes tilted estimate compared to the twisted posterior. It thus corrects the twisted posterior. We emphasize that such a situation happens for especially damaging twists (in particular, *not* Bayes blunting).

A.8. Pseudo-Inverse Link

Derivation of (16): From the definition of F in (3), we have that

$$F(z) := (-\underline{L})^\star(-z), \forall z \in \mathbb{R}. \tag{147}$$

Given $\underline{L}(u)$ associated with an underlying CPE loss, we have that

$$(-\underline{L})^\star(z) = \sup_{u \in [0,1]} [zu + \underline{L}(u)]. \tag{148}$$

Taking the derivative of the expression in brackets and solving for u, we obtain (assuming strictly concave $\underline{L}$)

$$z + \underline{L}'(u) = 0 \tag{149}$$

$$u^* = (\underline{L}')^{-1}(-z). \tag{150}$$

Plugging (150) back into the expression in brackets in (148), we obtain

$$(-\underline{L})^\star(z) = z(\underline{L}')^{-1}(-z) + \underline{L}((\underline{L}')^{-1}(-z)), \tag{151}$$

and

$$F(z) = (-\underline{L})^\star(-z) = -z(\underline{L}')^{-1}(z) + \underline{L}((\underline{L}')^{-1}(z)). \tag{152}$$

Then, we have by the chain rule that

$$\frac{d}{dz}F(z) = \frac{d}{dz}\left[-z(\underline{L}')^{-1}(z) + \underline{L}((\underline{L}')^{-1}(z))\right] \tag{153}$$

$$= -(\underline{L}')^{-1}(z) - z\left((\underline{L}')^{-1}(z)\right)' + \underline{L}'((\underline{L}')^{-1}(z))\left((\underline{L}')^{-1}(z)\right)' \tag{154}$$

$$= -(\underline{L}')^{-1}(z), \tag{155}$$

where the last step is obtained since $\underline{L}'((\underline{L}')^{-1}(z))\left((\underline{L}')^{-1}(z)\right)' = z\left((\underline{L}')^{-1}(z)\right)'$. Thus, we have that

$$F'(z) := -(\underline{L}')^{-1}(z) = -(-\underline{L}')^{-1}(-z). \tag{156}$$

From property **(D)** of Lemma 12 in Appendix A.3 in (47), namely that $\underline{L}'(u) = \ell_1(t_\ell(u)) - \ell_{-1}(t_\ell(u))$, and from (156), we get that

$$-F'(z) = (\ell_{-1} \circ t_\ell - \ell_1 \circ t_\ell)^{-1}(-z), \tag{157}$$

as desired.

PIL Approximation: Let $\alpha \in [-\infty, \infty]$, and define the conjugate α^{c} such that $1/\alpha^{\mathrm{c}} + 1/\alpha = 1$, using by extension $\alpha^{\mathrm{c}}(\infty) = 1, \alpha^{\mathrm{c}}(1) = \infty$. If we were to exactly implement a boosting algorithm for the α-loss, we would have to find the *exact* inverse of (16), which would require inverting $-\underline{L}'(v) \doteq \alpha^{\mathrm{c}} \cdot t_\ell(v)^{\alpha^{\mathrm{c}}} - \alpha^{\mathrm{c}} \cdot t_\ell(1-v)^{\alpha^{\mathrm{c}}}$. Owing to the difficulty to carry out this step, we choose a sidestep that makes inversion straightforward and can fall in the conditions to apply Theorem 11, thus making PILBOOST a boosting algorithm for the α-loss of interest. The trick does not just hold for the α-loss, so we describe it for a general loss ℓ assuming for simplicity that $\ell_1(1) = \ell_{-1}(0) = 0$ and $t_\ell, \ell_1, \ell_{-1}$ are invertible with ℓ_1, ℓ_{-1} non-negative,

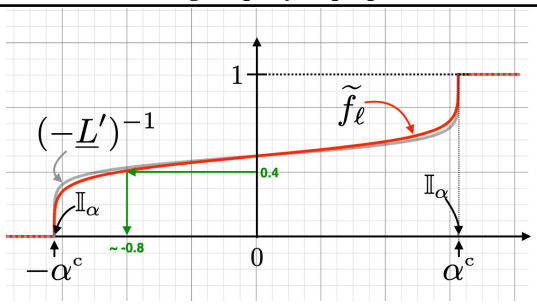

Figure 10: CIL link $\widetilde{f}_\ell$ vs inverse link $-\underline{L}'$ for ($\alpha = 5$)-loss. Notice the quality of the approximation.

conditions that would hold for many popular losses (log, square, Matusita, etc.), and the α-loss. We then approximate the link $-\underline{L}'$ by using just one of ℓ_{-1} or ℓ_1 depending on their argument, while ensuring functions match in $0, 1/2, 1$. We name $\widetilde{f}_\ell$ the *clipped inverse link*, CIL. Letting $a_\ell^- \doteq \ell_1(0)/(\ell_1(0) - \ell_1(1/2))$ and $a_\ell^+ \doteq \ell_{-1}(1)/(\ell_{-1}(1) - \ell_{-1}(1/2))$, our link approximation makes use of the following function: $f_\ell(u) \doteq f_\ell^-(u)$ if $u \leq 1/2$ and $f_\ell(u) \doteq f_\ell^+(u)$ otherwise, with the shorthands $f_\ell^-(u) \doteq a_\ell^- \cdot (\ell_1(1/2) - \ell_1(t_\ell(u)))$, $f_\ell^+(u) \doteq a_\ell^+ \cdot (\ell_{-1}(t_\ell(u)) - \ell_{-1}(1/2))$. The following Lemma shows, in addition to properties of f_ℓ, the expression obtained for the clipped inverse link for a general CPE loss.

Lemma 17 *$f_\ell(u) = -\underline{L}'(u), \forall u \in \{0, 1/2, 1\}$; furthermore, the clipped inverse link $\widetilde{f}_\ell \doteq f_\ell^{-1}$ is: (i) $\widetilde{f}_\ell(z) = 0$ if $z < -\ell_1(0)$; (ii) $\widetilde{f}_\ell(z) = t_\ell^{-1} \circ \ell_1^{-1}\left(\frac{\ell_1(1/2)-\ell_1(0)}{\ell_1(0)} \cdot z + \ell_1(1/2)\right)$ if $-\ell_1(0) \leq z < 0$; (iii) $\widetilde{f}_\ell(z) = t_\ell^{-1} \circ \ell_{-1}^{-1}\left(\frac{\ell_{-1}(1)-\ell_{-1}(1/2)}{\ell_{-1}(1)} \cdot z + \ell_{-1}(1/2)\right)$ if $0 \leq z < \ell_{-1}(1)$; (iv) $\widetilde{f}_\ell(z) = 1$ if $z \geq \ell_{-1}(1)$. Furthermore, $\widetilde{f}_\ell$ is continuous and if **(S)** and **(D)** hold, then $\widetilde{f}_\ell$ is derivable on $\mathbb{R}$ (with the only possible exceptions of $\{-\ell_1(0), \ell_{-1}(1)\}$).*

The proof is immediate once we remark that $\ell_1(1) = \ell_{-1}(0) = 0$ bring "properness for the extremes", *i.e.* $0 \in t_\ell(0), 1 \in t_\ell(1)$. We now give the expression of the formulas of interest regarding Lemma 17 for the α-loss.

Lemma 18 *We have for the α-loss,*

$$f_\ell(u) = \alpha^c \cdot \begin{cases} \left(\frac{2 \cdot u^\alpha}{u^\alpha + (1-u)^\alpha}\right)^{\frac{1}{\alpha^c}} - 1 & \text{if} \quad u \leq 1/2, \\ 1 - \left(\frac{2 \cdot (1-u)^\alpha}{u^\alpha + (1-u)^\alpha}\right)^{\frac{1}{\alpha^c}} & \text{if} \quad u \geq 1/2 \end{cases}, \tag{158}$$

$$\widetilde{f}(z) = \begin{cases} 0 & \text{if} \quad z \leq -\alpha^c, \\ \frac{(\alpha^c + z)^{\frac{\alpha^c}{\alpha}}}{(\alpha^c + z)^{\frac{\alpha^c}{\alpha}} + \left(2\alpha^{c\alpha^c} - (\alpha^c + z)^{\alpha^c}\right)^{\frac{1}{\alpha}}} & \text{if} \quad -\alpha^c \leq z \leq 0, \\ \frac{\left(2\alpha^{c\alpha^c} - (\alpha^c - z)^{\alpha^c}\right)^{\frac{1}{\alpha}}}{(\alpha^c - z)^{\frac{\alpha^c}{\alpha}} + \left(2\alpha^{c\alpha^c} - (\alpha^c - z)^{\alpha^c}\right)^{\frac{1}{\alpha}}} & \text{if} \quad 0 \leq z \leq \alpha^c, \\ 1 & \text{if} \quad z \geq \alpha^c. \end{cases} . \tag{159}$$

Rewritten, we have that

$$\widetilde{f}(z) = \begin{cases} 0 & z \leq -\frac{\alpha}{\alpha-1} \\ \frac{\left(\frac{\alpha}{\alpha-1} + z\right)^{\frac{1}{\alpha-1}}}{\left(\frac{\alpha}{\alpha-1} + z\right)^{\frac{1}{\alpha-1}} + \left(2\left(\frac{\alpha}{\alpha-1}\right)^{\frac{\alpha}{\alpha-1}} - \left(\frac{\alpha}{\alpha-1} + z\right)^{\frac{\alpha}{\alpha-1}}\right)^{\frac{1}{\alpha}}} & -\frac{\alpha}{\alpha-1} \leq z \leq 0 \\ \frac{\left(2\left(\frac{\alpha}{\alpha-1}\right)^{\frac{\alpha}{\alpha-1}} - \left(\frac{\alpha}{\alpha-1} - z\right)^{\frac{\alpha}{\alpha-1}}\right)^{\frac{1}{\alpha}}}{\left(\frac{\alpha}{\alpha-1} - z\right)^{\frac{1}{\alpha-1}} + \left(2\left(\frac{\alpha}{\alpha-1}\right)^{\frac{\alpha}{\alpha-1}} - \left(\frac{\alpha}{\alpha-1} - z\right)^{\frac{\alpha}{\alpha-1}}\right)^{\frac{1}{\alpha}}} & 0 \leq z \leq \frac{\alpha}{\alpha-1} \\ 1 & z \geq \frac{\alpha}{\alpha-1} \end{cases} \tag{160}$$

Figure 11 plots (160) for several values of α.

Remark. It could be tempting to think that the clipped inverse link trivially comes from clipping the partial losses themselves such as replacing $\ell_1(u)$ by 0 if $u \geq 1/2$ and symmetrically for $\ell_{-1}(u)$. This is not the case as it would lead to $\underline{L}$

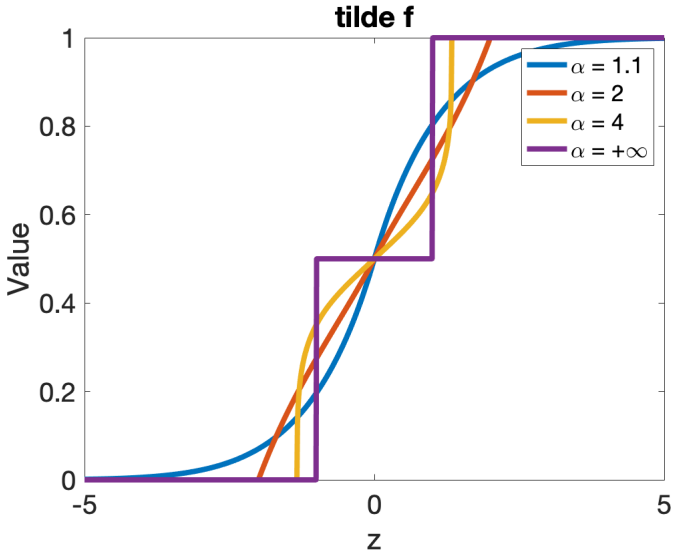

Figure 11: A plot of $\tilde{f}(z)$ as a function of α as given in (160).

piecewise constant and therefore $-\underline{L}' = 0$ when defined. ■

We turn to a result that authorizes us to use Thm 11 while virtually not needing **(E)** and **(C)** for α-loss. Denote $\mathbb{I}_\alpha \doteq \pm\alpha^c \cdot [1 - (1/\alpha^4), 1]$ (See Fig. 10).

Lemma 19 *Suppose $\alpha \geq 1.2$. For $\widetilde{f}_\ell$ defined as in Lemma 17, $\exists K \geq 0.133$ such that α-loss satisfies:*

$$\forall z \notin \mathbb{I}_\alpha, |(\widetilde{f}_\ell - (-\underline{L}')^{-1})(z)| \quad \lesssim \quad K/\alpha. \tag{161}$$

Remark the necessity of a trick as we do not compute $(-\underline{L}')^{-1}$ in (161). The proof, in Section A.9, bypasses the difficulty by bounding the *horizontal* distance between the *inverses*. The Lemma can be read as: with the exception of an interval vanishing rapidly with α, the difference between $\widetilde{f}_\ell$ (that we can easily compute for the α-loss) and $(-\underline{L}')^{-1}$ (that we do not compute for the α-loss), in order or just pointwise (typically for $\alpha < 10$) is at most $0.14/\alpha$. We now show how we can virtually "get rid of" (**E**) and (**C**) in such a context to apply Theorem 11. Consider the following assumptions: (i) no edge falls in $\mathbb{I}_\alpha$, (ii) the weak learner guarantees $\gamma = 0.14$, (iii) the average weights, $\overline{w}_j \doteq \mathbf{1}^\top \boldsymbol{w}_j/m$, satisfies $\overline{w}_j \geq 0.4$. Looking at Figure 10, we see that (i) is virtually not limiting at all; (ii) is a reasonable assumption on WL; remembering that a weight has the form $w = \widetilde{f}_\ell(-yH(\boldsymbol{x}))$, we see that (iii) requires H to be not "too good", see for example Figure 10 in which case $w = 0.4$ implies an edge $yH \leq 0.8$. We now observe that given (i), it is trivial to find a_f to satisfy (**C**) since we focus only on one α-loss. Suppose $\alpha \geq 2.7$, which approaches the average value of the αs in our experiments, and finally let $\zeta \doteq 2.5/2.7 \approx 0.926$. Then we get the chain of inequalities:

$$\begin{aligned}\Delta(F) &\underset{\text{Lem. 19}}{\leq} M \cdot \frac{0.14}{\alpha} \underset{\text{(ii)}}{=} \gamma M \cdot \frac{1}{\alpha} \underset{\textbf{(WLA)}}{\leq} \frac{1}{\alpha} \cdot \tilde{\mathsf{e}}_j \\ &\doteq \frac{1}{\alpha \overline{w}_j} \cdot \mathsf{e}_j \underset{\text{(iii)}}{\leq} \frac{2.5}{\alpha} \cdot \mathsf{e}_j \leq \frac{2.5}{2.7} \cdot \mathsf{e}_j \doteq \zeta \cdot \mathsf{e}_j,\end{aligned} \tag{162}$$

and so (**E**) is implied by the weak learning assumption. To summarise, PILBOOST boosts the convex surrogate of the α-loss without either computing it or its derivative, and achieves boosting compliant convergence using only the classical assumptions of boosting, **(R, WLA)**. The proof of Lemma 19 being very conservative, we can expect that the smallest value of K of interest is smaller than the one we use, indicating that (162) should hold for substantially smaller limit values in (ii, iii).

A.9. Proof of Lemma 19

Define for short

$$F(u) \doteq \left(\frac{u^\alpha}{u^\alpha+(1-u)^\alpha}\right)^{\alpha^{\mathrm{c}}} - \left(\frac{(1-u)^\alpha}{u^\alpha+(1-u)^\alpha}\right)^{\alpha^{\mathrm{c}}} \tag{163}$$

$$G(u) \doteq 1 - \left(\frac{2\cdot(1-u)^\alpha}{u^\alpha+(1-u)^\alpha}\right)^{\frac{1}{\alpha^{\mathrm{c}}}}, \tag{164}$$

that we study for $u \geq 1/2$ (the bound also holds by construction for $u < 1/2$). Define the following functions:

$$g(u) \doteq 1-(2u)^{\frac{1}{\alpha^{\mathrm{c}}}}, \tag{165}$$

$$h(z) \doteq \frac{1}{1+z}, \tag{166}$$

$$i_\alpha(u) \doteq u^\alpha, \tag{167}$$

and $u_\alpha \doteq h\circ i_{-\alpha}\circ h^{-1}(u)$, $f(u) \doteq i_{\alpha^{\mathrm{c}}}(1-u) - i_{\alpha^{\mathrm{c}}}(u)$. We remark that g is convex if $\alpha \geq 1$ while f is concave. Both derivatives match in $1/2$ if

$$(\alpha^{\mathrm{c}})^2 2^{1-\alpha^{\mathrm{c}}} = 1, \tag{168}$$

whose roots are $\alpha^{\mathrm{c}} < 6$. It means if $\alpha \geq 6/5 = 1.2$, then $(g-f)' \geq 0$, and so if we measure

$$k^* \doteq \arg\sup_k \sup_{x,x':g(x)=f(x')=k} |x-x'|,$$

then k^* is obtained for $x=1$, for which $g(x) = 1-2^{\frac{1}{\alpha^{\mathrm{c}}}} = k^*$. We then need to lowerbound x' such that $f(x') = 1-2^{\frac{1}{\alpha^{\mathrm{c}}}}$, which amounts to finding x^* such that $f(x^*) \geq 1-2^{\frac{1}{\alpha^{\mathrm{c}}}}$, since f is strictly decreasing. Fix

$$x^* \doteq 1-\frac{K}{\alpha}, \tag{169}$$

A series expansion reveals that for $x = x^*$ and $K = \log 2$,

$$f(x^*) = g(x^*) + O\left(\frac{1}{\alpha^2}\right), \tag{170}$$

and we thus get that there exists $K \geq \log 2$ such that

$$\sup_k \sup_{x,x':g(x)=f(x')=k} |x-x'| \leq \frac{K}{\alpha}, \tag{171}$$

or similarly for any ordinate value, the difference between the abscissae giving the value for f and g are distant by at most K/α. The exact value of the constant is not so much important than the dependence in $1/\alpha$: we now plug this in the u_αs notation and ask the following question: suppose $f(u_\alpha) = g(v_\alpha) = k$. Since $|u_\alpha - v_\alpha| \leq K/\alpha$, what is the maximum difference $|u-v|$ as a function of α ? We observe

$$\frac{\partial}{\partial u}u_\alpha = -\frac{\alpha(u(1-u))^{\alpha-1}}{(u^\alpha+(1-u)^\alpha)^2}, \tag{172}$$

$$\frac{\partial^2}{\partial u^2}u_\alpha = \alpha\cdot\frac{(u(1-u))^{\alpha-2}((\alpha-2u+1)u^\alpha - (\alpha+2u-1)(1-u)^\alpha)}{(u^\alpha+(1-u)^\alpha)^3}, \tag{173}$$

which shows the convexity of u_α as long as

$$\left(\frac{u}{1-u}\right)^\alpha \geq \frac{\alpha+2u-1}{\alpha-2u+1}, \tag{174}$$

a sufficient condition for which – given the RHS increases with u – is

$$u \geq \frac{\left(\frac{4}{\alpha-1}\right)^{\frac{1}{\alpha}}}{1+\left(\frac{4}{\alpha-1}\right)^{\frac{1}{\alpha}}}. \tag{175}$$

Since $u \geq 1/2$, we note the constraint quickly vanishes. In particular, if $\alpha \geq 5$, the RHS is $\leq 1/2$, so u_α is strictly convex. Otherwise, scrutinising the maximal values of the derivative for $\alpha \geq 1$ reveals that if we suppose $v \leq \delta$ for some δ, then $|u-v|$ is maximal for $v = \delta$. So, suppose $v_\alpha = \epsilon$ and we solve for $u_\alpha = K/\alpha + \varepsilon$, which yields

$$u = \frac{\left(1-\frac{K}{\alpha}-\varepsilon\right)^{\frac{1}{\alpha}}}{\left(\frac{K}{\alpha}+\varepsilon\right)^{\frac{1}{\alpha}}+\left(1-\frac{K}{\alpha}-\varepsilon\right)^{\frac{1}{\alpha}}} \tag{176}$$

$$= \frac{((1-\varepsilon)\alpha-K)^{\frac{1}{\alpha}}}{(K+\varepsilon\alpha)^{\frac{1}{\alpha}}+((1-\varepsilon)\alpha-K)^{\frac{1}{\alpha}}}, \tag{177}$$

while the v producing the largest $|u-v|$ is:

$$v = \frac{(1-\varepsilon)^{\frac{1}{\alpha}}}{\varepsilon^{\frac{1}{\alpha}}+(1-\varepsilon)^{\frac{1}{\alpha}}}, \tag{178}$$

so

$$|v-u| = (v-u)(\varepsilon) = \frac{(1-\varepsilon)^{\frac{1}{\alpha}}}{\varepsilon^{\frac{1}{\alpha}}+(1-\varepsilon)^{\frac{1}{\alpha}}} - \frac{((1-\varepsilon)\alpha-K)^{\frac{1}{\alpha}}}{(K+\varepsilon\alpha)^{\frac{1}{\alpha}}+((1-\varepsilon)\alpha-K)^{\frac{1}{\alpha}}}. \tag{179}$$

If we fix

$$\varepsilon = \frac{1}{\alpha^4}, \tag{180}$$

then we get after separate series are computed in $\alpha \to +\infty$,

$$|v-u| = (v-u)(\varepsilon) = \frac{\log(1+\log K)}{4\alpha} + O\left(\frac{1}{\alpha^2}\right) \tag{181}$$

$$\lesssim \frac{0.133}{\alpha}. \tag{182}$$

The "forbidden interval" for v is then

$$\left[\frac{(\alpha^4-1)^{\frac{1}{\alpha}}}{1+(\alpha^4-1)^{\frac{1}{\alpha}}}, 1\right] \approx \left[\frac{1}{2}+\frac{\log\alpha}{\alpha}, 1\right]; \tag{183}$$

what is more interesting for us is the corresponding forbidden images for $g(v_\alpha)$, which are thus

$$g \notin \alpha^{\mathrm{c}} \cdot \left[1-\frac{1}{\alpha^4}, 1\right] \doteq \mathbb{I}_\alpha, \tag{184}$$

where we use shorthand $z \cdot [a,b] \doteq [az, bz]$. This, we note, translates directly into observable edges since g is the function we invert. Summarizing, we have shown that if (i) $\alpha \geq 1.2$ then for any u, v such that $F(u) = G(v) \notin \mathbb{I}_\alpha$, then $|u-v| \lesssim 0.133/\alpha$. It suffices to remark that $\mathbb{I}_\alpha$ represents the set of forbidden weights to get the statement of the Lemma.

A.10. Proof of Theorem 11

Remark. Consider the following setup: $h_\cdot \in [-1,1]$, use the logistic loss surrogate for F and the derivative of Matusita's loss (just for illustration as the pseudo-inverse-link approximation of the logistic loss) for the weights (see e.g., (Nock &

Nielsen, 2008) Table 1), we get *in the worst case* that $\Delta_j(F) \leq 2 \cdot \sup \left| \frac{1}{2} - \frac{x}{2\sqrt{1+x^2}} - \frac{1}{1+\exp(x)} \right| < 0.24707$. So, all we need is $|\mathsf{e}_j| > 0.24707$ to get $\zeta < 1$ constant. Since $|\mathsf{e}_j| \in [0,1]$, this constraint is more than reasonable and turns into a very reasonable penalty in $Q(F)$ (Theorem 11). ■

Remark. PILBOOST and its convergence analysis bring a side contribution of ours: it is impossible to exactly encode in standard machine types the inverse link of losses like the log loss, so the implementation of classical boosting algorithms (Friedman, 2001; Friedman et al., 2000) can only rely on approximations of the inverse link function. Our results yield convergence guarantees for the *implementations* of such algorithms, and **(E)** can be interpreted and checked in the context of machine encoding. Two additional remarks hold regarding convergence rate: first, the $1/\gamma^2$ dependence meets the general optimum for boosting (Alon et al., 2021); second, the $1/\varepsilon^2$ dependence parallels classical training convergence of convex optimization (Thekumparampil et al., 2020) (and references therein). There is however a major difference with such work: PILBOOST requires *no* function oracles for F (function values, (sub)gradients, etc.). This "*sideways*" fork to minimizing F pays (only) a $1/(1-\zeta)^2$ factor in convergence. ■

We proceed in two steps, assuming **(WLA)** holds for WL and **(R)** holds for the weak classifiers.

In Step 1, we show that for any loss defined by F twice differentiable, convex and non-increasing, for any $z^* \in \mathbb{R}$, as long as F satisfies assumptions **(E)** and **(C)** for T iterations such that

$$\sum_{t=0}^{T} \tilde{w}_t^2 \quad \geq \quad \frac{2F^*(F(0) - F(z^*))}{\gamma^2(1-\zeta)^2(1-\pi^2)}, \tag{185}$$

we have the guarantee on the risk defined by F:

$$\mathbb{E}_{i \sim \mathcal{S}} \left[F(y_i H_T(\boldsymbol{x}_i)) \right] \quad \leq \quad F(z^*). \tag{186}$$

Let F be any twice differentiable, convex and non-increasing function. We wish to find a lowerbound $\triangle$ on the decrease of the expected loss computed using F:

$$\mathbb{E}_{i \sim \mathcal{S}} \left[F(y_i H_t(\boldsymbol{x}_i)) \right] - \mathbb{E}_{i \sim \mathcal{S}} \left[F(y_i H_{t+1}(\boldsymbol{x}_i)) \right] \quad \geq \quad \triangle, \tag{187}$$

where with a slight abuse of notation we let H_t denote the learned real-valued classifier at iteration t. We make use of a similar proof technique as in Nock & Williamson (2019, Theorem 7). Suppose

$$H_{t+1} \quad = \quad H_t + \beta_j \cdot h_j, \tag{188}$$

index j being returned by WL at iteration t. For any such index j, any $g : \mathbb{R} \to \mathbb{R}_+$ and any $H \in \mathbb{R}^{\mathcal{X}}$, let

$$\mathsf{e}(j, g, H) \quad \doteq \quad \mathbb{E}_{i \sim \mathcal{S}} \left[y_i h_j(\boldsymbol{x}_i) \cdot g(y_i H(\boldsymbol{x}_i)) \right], \tag{189}$$

denote the expected edge of h_j on weights defined by the couple (g, H). Furthermore, let

$$\Delta(g_1, g_2) \quad \doteq \quad \left| \mathsf{e}(j, g_1, H_t) - \mathsf{e}(j, g_2, H_t) \right|, \tag{190}$$

denote the discrepancy between two expected edges defined by g_1, g_2, respectively.

There are two quantities we consider. First, let

$$\begin{aligned}
X \quad &\doteq \quad \mathbb{E}_{i \sim \mathcal{S}} \left[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i)) F'(y_i H_t(\boldsymbol{x}_i)) \right] && (191)\\
&= \quad \beta_j \cdot \mathbb{E}_{i \sim \mathcal{S}} \left[y_i h_j(\boldsymbol{x}_i) \cdot -F'(y_i H_t(\boldsymbol{x}_i)) \right] && (192)\\
&= \quad \beta_j \cdot \mathsf{e}(j, -F', H_t) && (193)\\
&\geq \quad \beta_j \cdot \mathbb{E}_{i \sim \mathcal{S}} \left[y_i h_j(\boldsymbol{x}_i) \cdot \widetilde{f}(-y_i H_t(\boldsymbol{x}_i)) \right] - \beta_j \cdot \Delta(-F', \widetilde{f}_s) && (194)\\
&= \quad a_f \mathsf{e}^2(j, \widetilde{f}_s, H_t) - a_f \mathsf{e}(j, \widetilde{f}_s, H_t) \cdot \Delta(-F', \widetilde{f}_s) && (195)\\
&\geq \quad a_f (1-\zeta) \mathsf{e}^2(j, \widetilde{f}_s, H_t), && (196)
\end{aligned}$$

where $\widetilde{f}_s(z) \doteq \widetilde{f}_s(-z)$ and finally (196) makes use of assumption **(E)**. The second quantity we define is:

$$Y(\mathcal{Z}) \doteq \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i))^2 F''(z_i)\right], \tag{197}$$

where $\mathcal{Z} \doteq \{z_1, z_2, ..., z_m\} \subset \mathbb{R}^m$. Using assumption **(R)** and letting F^* be any real such that $F^* \geq \sup F''(z)$, we obtain:

$$\begin{aligned} Y(\mathcal{Z}) &\leq F^* \cdot \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i))^2\right] \\ &= F^* \cdot \beta_j^2 \cdot \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i h_j(\boldsymbol{x}_i))^2\right] \\ &\leq F^* \cdot \beta_j^2 \cdot M^2 \\ &= F^* a^2 M^2 \cdot \mathrm{e}^2(j, \widetilde{f}_s, H_t). \end{aligned} \tag{198}$$

A second order Taylor expansion on F brings that there exists $\mathcal{Z} \doteq \{z_1, z_2, ..., z_m\} \subset \mathbb{R}^m$ such that:

$$\begin{aligned} \mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_t(\boldsymbol{x}_i))\right] &= \mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_{t+1}(\boldsymbol{x}_i))\right] + \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i))F'(y_i H_t(\boldsymbol{x}_i))\right] \\ &\quad + \mathbb{E}_{i\sim\mathcal{S}}\left[(y_i H_t(\boldsymbol{x}_i) - y_i H_{t+1}(\boldsymbol{x}_i))^2 \cdot \frac{F''(z_i)}{2}\right], \end{aligned} \tag{199}$$

So,

$$\begin{aligned} \mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_t(\boldsymbol{x}_i))\right] - \mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_{t+1}(\boldsymbol{x}_i))\right] &= X - \frac{Y(\mathcal{Z})}{2} \\ &\geq \underbrace{\left(1 - \zeta - \frac{F^* a M^2}{2}\right) a}_{\doteq Z(a)} \cdot \mathrm{e}^2(j, \widetilde{f}_s, H_t). \end{aligned} \tag{200}$$

Suppose we fix $\pi \in [0, 1]$ and choose any

$$a \in \frac{1-\zeta}{F^* M^2} \cdot [1 - \pi, 1 + \pi]. \tag{201}$$

We can check that

$$Z(a) \geq \frac{(1-\zeta)^2(1-\pi^2)}{2F^* M^2}, \tag{202}$$

and so

$$\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_t(\boldsymbol{x}_i))\right] - \mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_{t+1}(\boldsymbol{x}_i))\right] \geq \frac{(1-\zeta)^2(1-\pi^2)}{2F^* M^2} \cdot \mathrm{e}^2(j, \widetilde{f}_s, H_t). \tag{203}$$

So, taking into account that for the first classifier, we have $\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_0(\boldsymbol{x}_i))\right] = F(0)$, if we take any $z^* \in \mathbb{R}$ and we boost for a number of iterations T satisfying (we use notation e_t as a summary for $\mathrm{e}^2(j, \widetilde{f}_s, H_t)$ with respect to PILBOOST):

$$\sum_{t=1}^{T} \mathrm{e}_t^2 \geq \frac{2F^* M^2 (F(0) - F(z^*))}{(1-\zeta)^2(1-\pi^2)}, \tag{204}$$

then $\mathbb{E}_{i\sim\mathcal{S}}\left[F(y_i H_T(\boldsymbol{x}_i))\right] \leq F(z^*)$. We now assume **(WLA)** holds, the LHS of (204) is $\geq T\gamma^2$. Given that we choose $a = a_f$ in PILBOOST, we need to make sure (201) is satisfied for the loss defined by F, which translates to

$$F^* \in \frac{1-\zeta}{a_f M^2} \cdot [1 - \pi, 1 + \pi], \tag{205}$$

and defines assumption **(C)**. To complete Step 1, we normalize the edge. Letting $\tilde{w}_i \doteq w_i / \sum_k w_k$, $\tilde{w}_t \doteq \mathbf{1}^\top \boldsymbol{w}_t / m$ and

$$\tilde{\mathrm{e}}_t \doteq \frac{\mathrm{e}_t}{\tilde{w}_t} \in [-M, M], \tag{206}$$

which is then properly normalized and such that (204) becomes equivalently:

$$\sum_{t=0}^{T} \tilde{w}_t^2 \tilde{\mathrm{e}}_t^2 \quad \geq \quad \frac{2F^* M^2 (F(0) - F(z^*))}{(1-\zeta)^2(1-\pi^2)}, \tag{207}$$

and so under the (weak learning) assumption on $\tilde{\mathrm{e}}_t$ that $|\tilde{\mathrm{e}}_t| \geq \gamma \cdot M$, a sufficient condition for (207) is then

$$\sum_{t=0}^{T} \tilde{w}_t^2 \quad \geq \quad \frac{2F^* (F(0) - F(z^*))}{\gamma^2 (1-\zeta)^2(1-\pi^2)}, \tag{208}$$

completing step 1 of the proof.

In Step 2, we show a result on the distribution of edges, *i.e.* margins. (208) contains all the intuition about how the rest of the proof unfolds, as we have two major steps: in step 2.1, we translate the guarantee of (208) on margins, and in step 2.2, we translate the "margin" based (208) in a readable guarantee in the boosting framework (we somehow "get rid" of the $\tilde{w}_t^2$ in the LHS of (208)).

Step 2.1. Let $\mathcal{Z} \doteq \{z_1, z_2, ..., z_m\} \subset \mathbb{R}$ a set of reals. Since F is non-increasing, we have $\forall u \in [0,1], \forall \theta \geq 0$,

$$\begin{aligned} \mathbb{P}_i[z_i \leq \theta] > u \Rightarrow \mathbb{E}_i[F(z_i)] \quad &> \quad (1-u) \inf_z F(z) + uF(\theta) \\ &\doteq (1-u)F^\circ + uF(\theta), \end{aligned} \tag{209}$$

so if we pick z^* in (208) such that

$$F(z^*) \quad \doteq \quad (1-u)F^\circ + uF(\theta), \tag{210}$$

then (208) implies $\mathbb{E}_{i \sim \mathcal{S}}\left[F(y_i H_T(\boldsymbol{x}_i))\right] \leq (1-u)F^\circ + uF(\theta)$ and so by the contraposition of (209) yields:

$$\mathbb{P}_{i \sim \mathcal{S}}\left[y_i H_T(\boldsymbol{x}_i) \leq \theta\right] \quad \leq \quad u, \tag{211}$$

which yields our margin based guarantee.

Step 2.2. At this point, the key (in)equalities are (208) (for boosting) and (211) (for margins). Fix $\kappa > 0$. We have two cases:

- Case 1: $\tilde{w}_t$ never gets too small, say $\tilde{w}_t \geq \kappa, \forall t \geq 0$. In this case, granted the weak learning assumption holds on $\tilde{\mathrm{e}}_t$, (208) yields a direct lowerbound on iteration number T to get $\mathbb{P}_{i \sim \mathcal{S}}\left[y_i H_T(\boldsymbol{x}_i) \leq \theta\right] \leq u$;
- Case 2: $\tilde{w}_t \leq \kappa$ at some iteration t. Since the smaller it is, the better classified are the examples, if we pick κ small enough, then we can get $\mathbb{P}_{i \sim \mathcal{S}}\left[y_i H_T(\boldsymbol{x}_i) \leq \theta\right] \leq u$ "straight".

This suggests our use of the notion of "denseness" for weights (Bun et al., 2020).

Definition 20 *The weights at iteration t is called κ-dense iff $\tilde{w}_t \geq \kappa$.*

We now have the following Lemma.

Lemma 21 *For any $t \geq 0, \theta \in \mathbb{R}, \kappa > 0$, if weights produced in Step 2.1 of* PILBOOST *fail to be κ-dense, then*

$$\mathbb{P}_{i \sim \mathcal{S}}\left[y_i H_T(\boldsymbol{x}_i) \leq \theta\right] \quad \leq \quad \frac{\kappa}{\widetilde{f}(-\theta)}. \tag{212}$$

Proof Let $\mathcal{Z} \doteq \{z_1, z_2, ..., z_m\} \subset \mathbb{R}$ a set of reals. Since $\widetilde{f}$ is non-decreasing, we have $\forall \theta \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}_i[\widetilde{f}(z_i)] &\geq \mathbb{P}_i[z_i < -\theta] \cdot \inf_z \widetilde{f}(z) + \mathbb{P}_i[z_i \geq -\theta] \cdot \widetilde{f}(-\theta) \\ &= \mathbb{P}_i[z_i \geq -\theta] \cdot \widetilde{f}(-\theta) \end{aligned} \tag{213}$$

since by assumption $\inf \widetilde{f} = 0$. Pick $z_i \doteq -y_i H_T(\boldsymbol{x}_i)$. We get that if $\mathbb{P}_{i\sim\mathbb{S}}[-y_i H_T(\boldsymbol{x}_i) \geq -\theta] = \mathbb{P}_{i\sim\mathbb{S}}[y_i H_T(\boldsymbol{x}_i) \leq \theta] \geq \xi$, then $\tilde{w}_t \doteq \mathbb{E}_{i\sim\mathbb{S}}[\widetilde{f}(-y_i H_T(\boldsymbol{x}_i))] \geq \xi \cdot \widetilde{f}(-\theta)$. If we fix

$$\xi = \frac{\kappa}{\widetilde{f}(-\theta)}, \tag{214}$$

then $\tilde{w}_t < \kappa$ implies (212), which ends the proof of Lemma 21. ■

From Lemma 21, we let $\kappa \doteq \xi_* \cdot \widetilde{f}(-\theta)$ and $u \doteq \xi_*$ in (211). If at any iteration, H_T fails to be κ-dense, then $\mathbb{P}_{i\sim\mathbb{S}}[y_i H_{\boldsymbol{\beta}}(\boldsymbol{x}_i) \leq \theta] \leq \xi_*$ and classifier $H_{\boldsymbol{\beta}}$ satisfies the conditions of Theorem 11 (this is our Case 2 above).

Otherwise, suppose it is always κ-dense (this is our Case 1 above). We then have at any iteration T $\sum_{t<T} \tilde{w}_t^2 \geq T\xi_*^2 \cdot \widetilde{f}^2(-\theta)$ and so a sufficient condition to get (208) is then $T \geq \frac{2F^*(F(0)-F(z^*))}{\xi_*^2 \widetilde{f}^2(-\theta)\gamma^2(1-\zeta)^2(1-\pi^2)}$, where we recall z^* is chosen so that $F(z^*) = (1-\xi_*)F^\circ + \xi_* F(\theta)$. This ends the proof of Theorem 11 (with the change of notation $\xi_* \leftrightarrow \varepsilon$).

B. Additional Experimental Results

In this section, we provide additional experimental results and discussion to accompany Section 6 in the main text. The code for all of our experiments (including the implementation of PILBOOST) can be found at the following github repository link:

`https://github.com/SankarLab/Being-Properly-Improper`

B.1. General Details

Most of the experiments were performed over the course of a month on a 2015 MacBook Pro with a 2.2 GHz Quad-Core Intel Core $i7$ processor and 16GB of memory. The Adaptive α experiments were performed on a computing cluster and each required about 30 minutes of compute time. Code can be found in *PILBoostExperiments.py*, *AdaptiveAlphaMenon.py*, *AdaptiveAlphaALL.py*. Averaged experiments employed 10-fold cross validation, and when twisters were present, randomization occurred over the twisted samples as well. All algorithms across all experiments ran for 1000 iterations.

B.2. Discussion of a_f and α

In general, we found that for most experiments, $0.1 \leq a_f \leq 15$. From the theory, we know that if a_f is too small, boosting needs to occur for a very long time, and if a_f is too large, almost no loss fits to (**C**) (equivalently, (**C**) fails for us). We also generally found that PILBOOST was not particularly sensitive to the choice of a_f as long as it was in the "right ballpark", hence our use of integer or rational values of a_f for all experiments. When there is twist present, we found that $\alpha > 1$ performed best, where α^* increased as the amount of twist increased (both observations are conistent with our theory, see for example Lemma 3.4). Regarding the relationship between a_f and α, this appeared to depend on the dataset and depth of the decision trees.

B.3. Random Class Noise Twister

Dataset	Algorithm	Random Class Noise Twister		
		$p = 0$	0.15	0.3
cancer	AdaBoost	0.966 ± 0.015	0.905 ± 0.027	0.856 ± 0.033
	us ($\alpha = 1.1$)	0.944 ± 0.029	0.912 ± 0.013	0.861 ± 0.042
	us ($\alpha = 2.0$)	0.956 ± 0.018	0.938 ± 0.017	0.905 ± 0.039
	us ($\alpha = 4.0$)	0.957 ± 0.014	0.917 ± 0.012	0.922 ± 0.032
	XGBoost	0.971 ± 0.012	0.861 ± 0.033	0.733 ± 0.031

Table 2: cancer feature random class noise. Accuracies reported for each algorithm and level of twister. Depth one trees. For $\alpha = 1.1$, $a_f = 7$, for $\alpha = 2$, $a_f = 2$, and for $\alpha = 4$, $a_f = 1$.

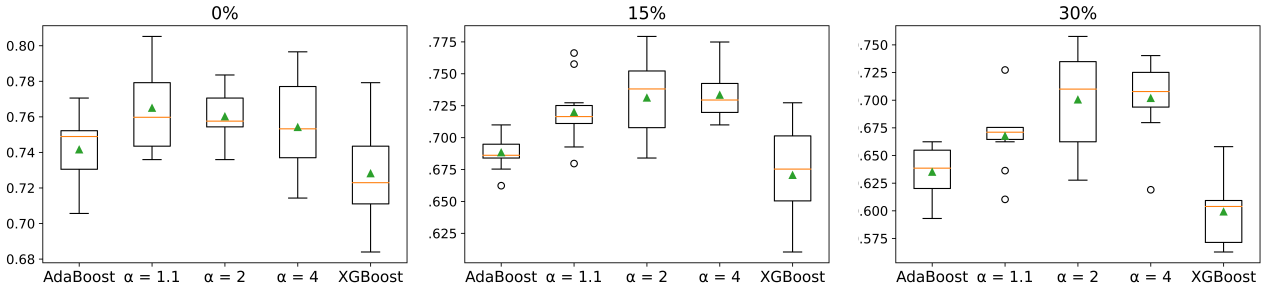

Figure 12: Random class noise twister on the diabetes dataset. Depth 3 trees. $a_f = 0.1$ for all α.

Dataset	Algorithm	Random Class Noise Twister		
		$p = 0$	0.15	0.3
xd6	AdaBoost	1.000 ± 0.000	0.949 ± 0.016	0.830 ± 0.043
	us ($\alpha = 1.1$)	1.000 ± 0.000	0.981 ± 0.013	0.886 ± 0.033
	us ($\alpha = 2.0$)	1.000 ± 0.000	0.992 ± 0.009	0.900 ± 0.027
	us ($\alpha = 4.0$)	1.000 ± 0.000	0.999 ± 0.003	0.927 ± 0.023
	XGBoost	1.000 ± 0.000	0.912 ± 0.016	0.776 ± 0.041

Table 3: xd6 random class noise. Accuracies reported for each algorithm and level of twister. Depth three trees. $a_f = 8$ for all α. Note that for 0% noise $\alpha = 4$ used $a_f = 0.1$.

Dataset	Algorithm	Random Class Noise Twister			
		$p = 0$	0.10	0.20	0.30
Online Shopping	AdaBoost	0.902 ± 0.002	0.900 ± 0.004	0.898 ± 0.005	0.894 ± 0.004
	us ($\alpha = 1.1$)	0.901 ± 0.005	0.899 ± 0.003	0.897 ± 0.004	0.890 ± 0.004
	us ($\alpha = 2.0$)	0.901 ± 0.004	0.895 ± 0.004	0.895 ± 0.003	0.894 ± 0.004
	us ($\alpha = 4.0$)	0.898 ± 0.003	0.873 ± 0.009	0.892 ± 0.005	0.889 ± 0.005
	XGBoost	0.893 ± 0.005	0.874 ± 0.002	0.842 ± 0.006	0.782 ± 0.008

Table 4: Accuracies reported for each algorithm and level of twister. Random training sample selected with probability p. Then, for selected training sample, boolean feature flipped with probability p for each feature, independently. Depth three trees. For $\alpha = 1.1$, $a_f = 7$, for $\alpha = 2$, $a_f = 8$, and for $\alpha = 4$, $a_f = 15$.

B.4. Insider Twister

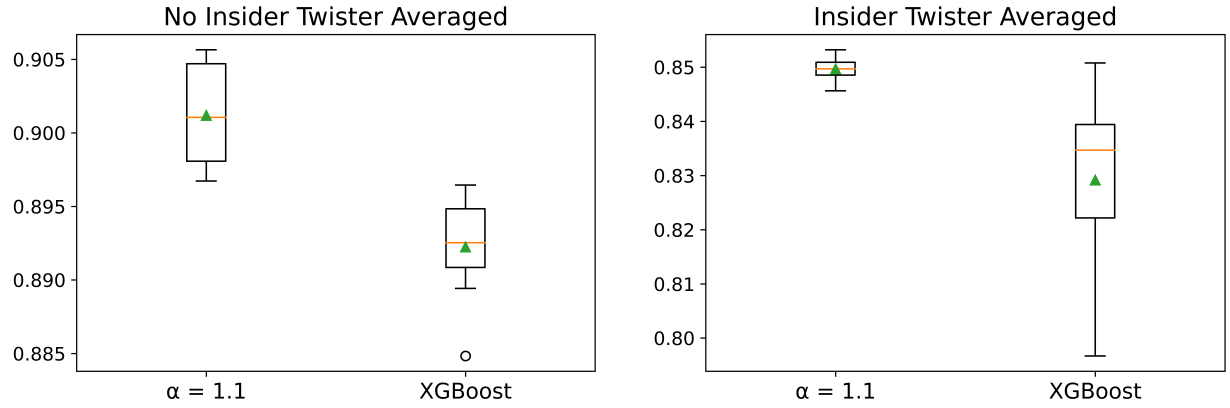

Figure 13: Box and whisker visualization of scores associated with Figure 3. For all insider twister results, we fixed $a_f = 7$. Under no twister, $\alpha = 1.1$, has accuracy 0.901 ± 0.003, and XGBoost has accuracy 0.892 ± 0.003. Under the insider twister, $\alpha = 1.1$, has accuracy 0.850 ± 0.002, and XGBoost has accuracy 0.829 ± 0.016; under the Welch t-test, the results have a p-value of 0.004.

B.5. Discussion of XGBoost

Algorithm	Average Compute Times			
	cancer	xd6	diabetes	shoppers
AdaBoost	1.41	0.75	1.11	13.68
us ($\alpha = 1.1$)	2.19	2.01	2.19	30.88
us ($\alpha = 2.0$)	1.11	0.79	2.09	21.85
us ($\alpha = 4.0$)	0.96	1.35	1.82	13.01
XGBoost	0.29	0.28	0.46	3.16

Table 5: Average compute times per run (10 runs) in seconds across the datasets. Note that the values of a_f are chosen identically to choices in Section B.3.

XGBoost is a very fast, very well engineered boosting algorithm. It employs many different hyperparameters and customizations. In order to report the fairest comparison between AdaBoost, PILBOOST , and XGBoost, we opted to keep as many hyperparameters fixed (and similar, e.g., depth of decision trees) as possible. That being said, it appears that XGBoost inherently uses pruning, so the algorithm pruned while the other two did not. Further details regarding three other important points related to XGBoost:

1. Please refer to Table 5 for averaged compute times for the three different algorithms. In general, XGBoost had the far faster computation time among the three. However, note that PILBOOST was not particularly engineered for speed. Indeed, we estimate that the computation of $\tilde{f}$ accounts for $40 - 50\%$ of the total computation time, which we believe can be improved. Thus, we leave the further computational optimization of PILBOOST for future work.

2. For details regarding regularization, refer to Figure 14, where we report a comparison of regularized XGBoost and PILBOOST such that the training data suffers from the insider twister. We find that regularization improves the ability of XGBoost to combat the twister, but it is not as effective as PILBOOST.

3. For details regarding early stopping, refer to Figure 16, where we report a comparison of early-stopped XGBoost (on un-twisted validation data, i.e., cheating) and PILBOOST such that the training data suffers from the insider twister. We find that even early-stopping does not improve XGBoost's ability to combat the insider twister as effectively as PILBOOST.

Early stopping - on an untwisted hold-out set contradicts our experiment. With early stopping enabled on a twisted hold-out set, XGBoost generally did not early stop.

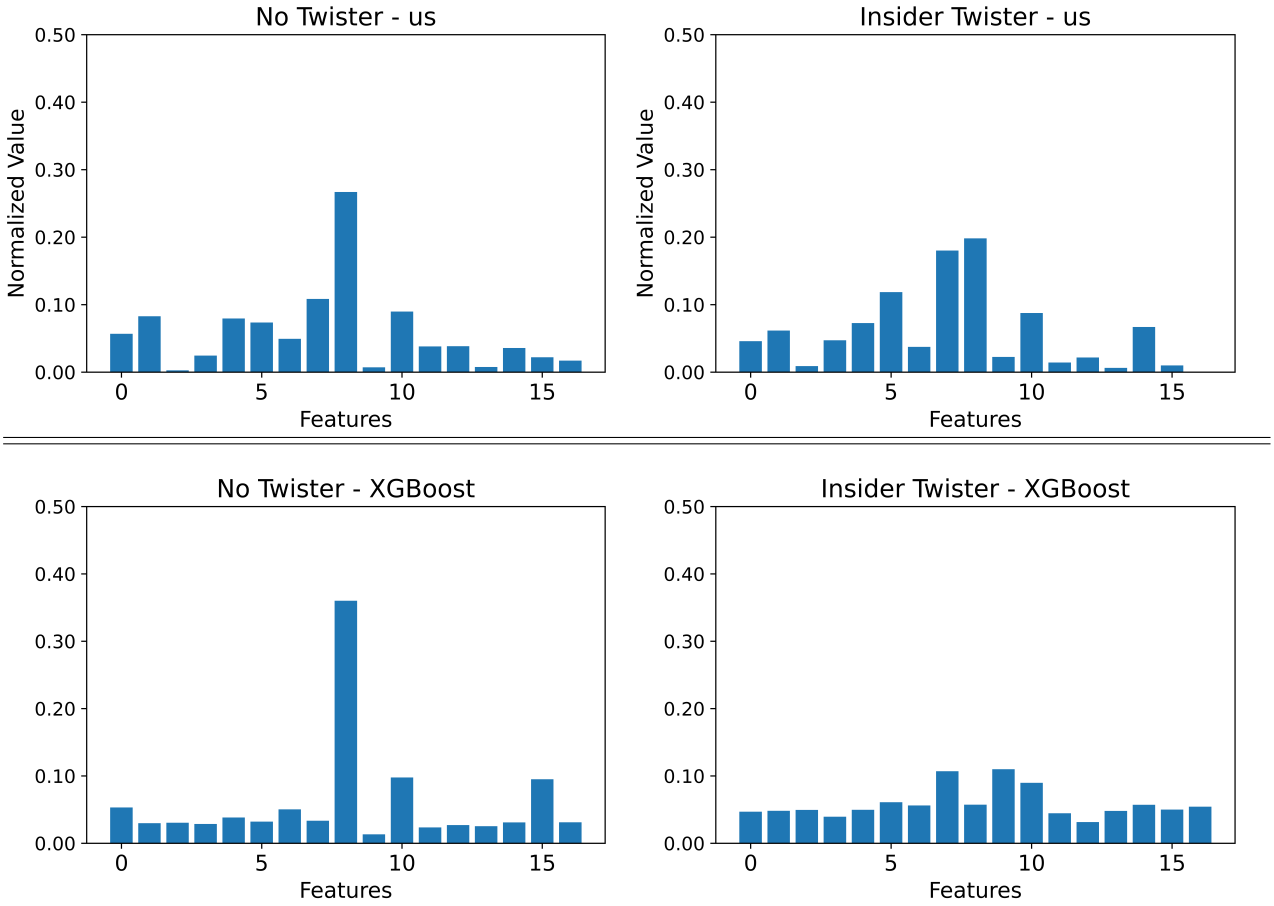

Figure 14: With regularization (where $\lambda = 20$), we still observe that the feature importance of XGBoost is perturbed. Note that PILBOOST is not regularized.

B.6. Adaptive α Experiment

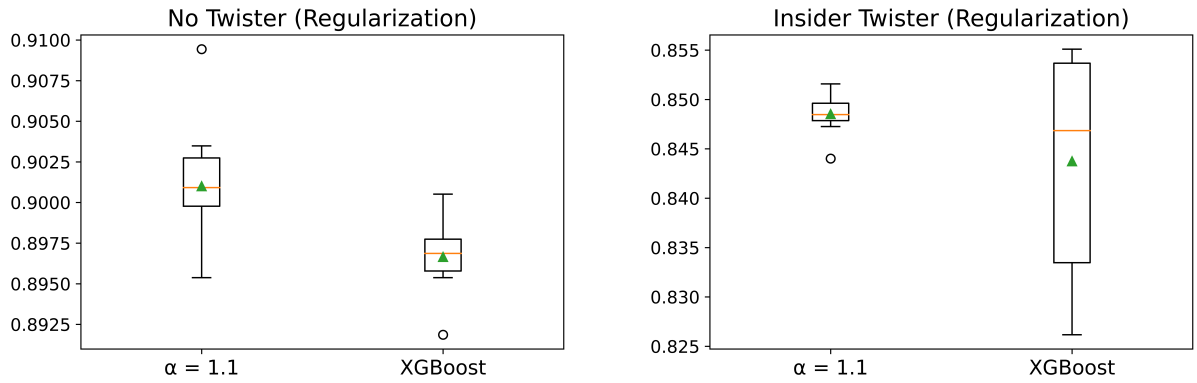

Figure 15: Scores associated with Figure 14.

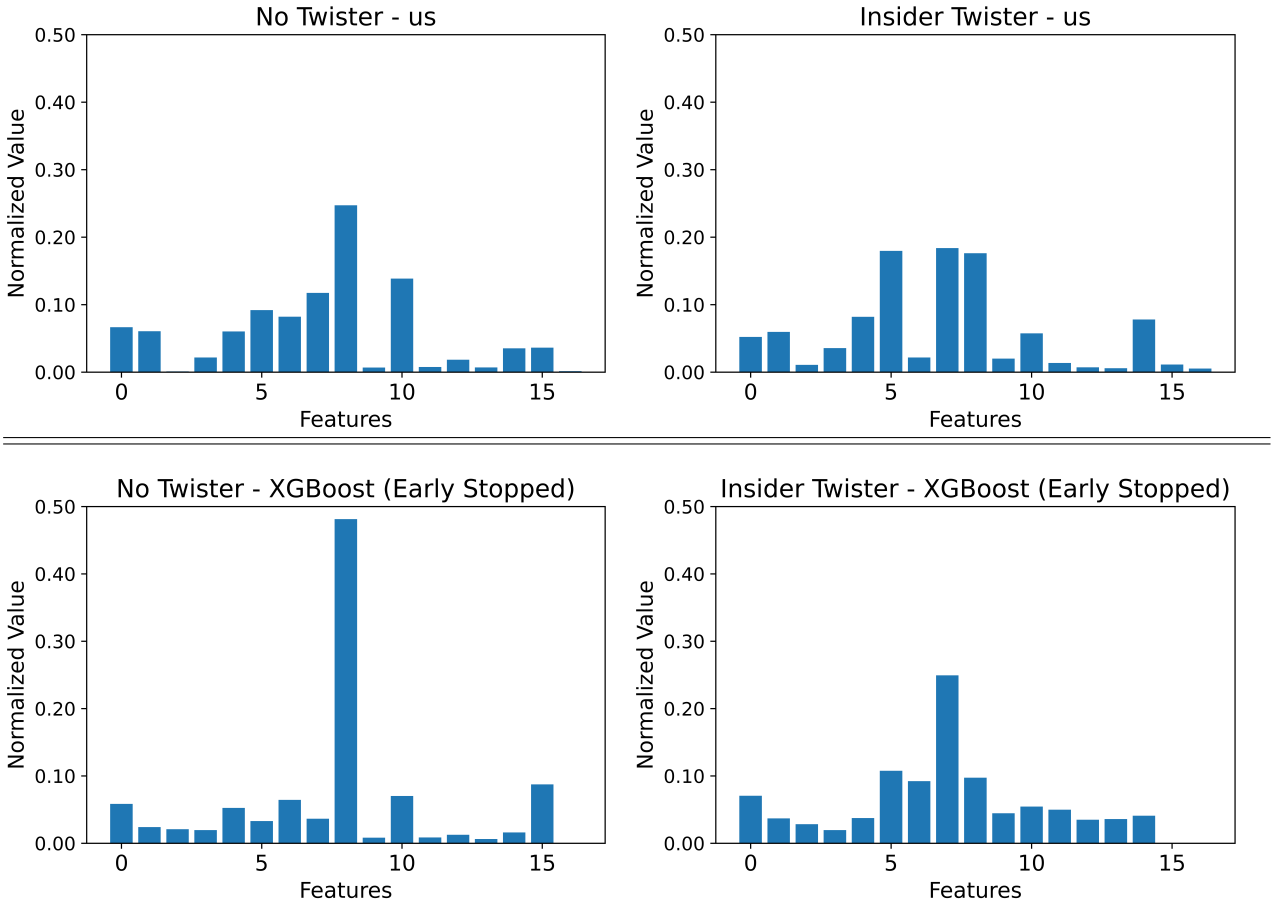

Figure 16: With early stopping (where XGBoost has access to clean validation data - cheating scenario), we still observe that the feature importance of XGBoost is perturbed. Note that PILBOOST is not early stopped.

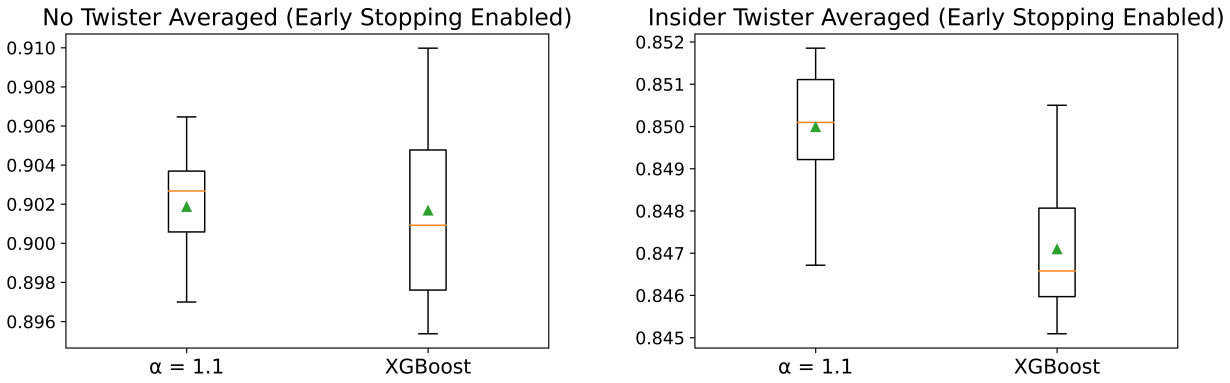

Figure 17: Scores associated with Figure 16.

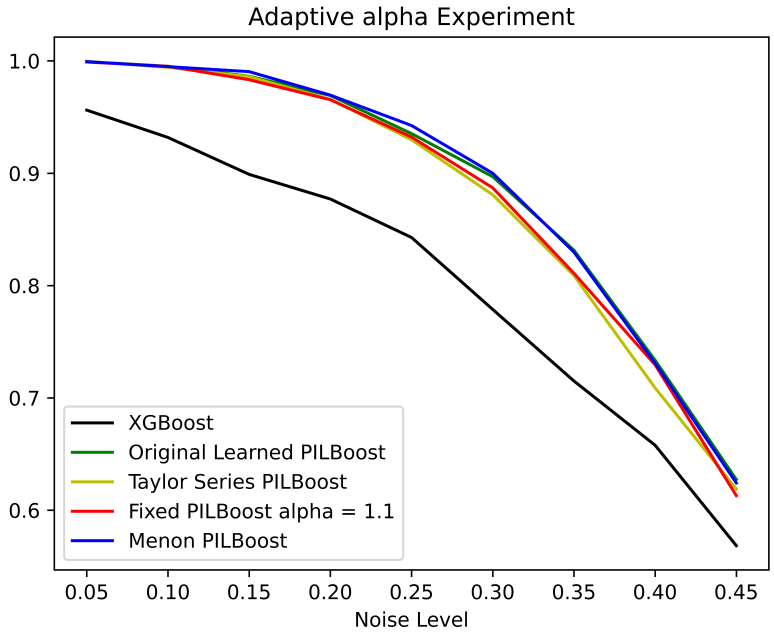

Figure 18: Extended version of Figure 2 with two additional adaptive α methods. Original Learned PILBOOST estimates its choice of α by using XGBoost as an oracle. That is, the method trains XGBoost on the noisy data, then computes its confusion matrix on a *clean* validation set. From the confusion matrix, the label flip probability p is estimated using $p = \text{avg}\left(\frac{\text{FP}}{\text{TP+FP}}, \frac{\text{FN}}{\text{FN+TN}}\right)$. Next, we estimate η_c and η_t with $\eta_c = \frac{\text{FN+TP}}{\text{FP+TP+FN+TN}}$ and $\eta_t = \frac{\text{FP+TP}}{\text{FP+TP+FN+TN}}$, respectively. Lastly similar to Menon PILBOOST, using the estimates of p, η_c, and η_t, we apply the formula in Lemma 8**(c)** and the SLN formula given just before Lemma 4 to obtain an estimate for α^*. Taylor Series PILBOOST is identical to Original Learned PILBOOST except at the last step, where a Taylor series approximation of the formula in Lemma 8**(c)** is used instead. We find that Menon's Method also outperforms both of these methods on the xd6 dataset, except for when Original Learned PILBOOST slightly outperforms Menon's Method in the very high noise regime. Even stronger, note that both Original Learned PILBOOST and Taylor Series PILBOOST assume more information than "Menon's Method", which only uses the noisy training data, not a clean validation set.

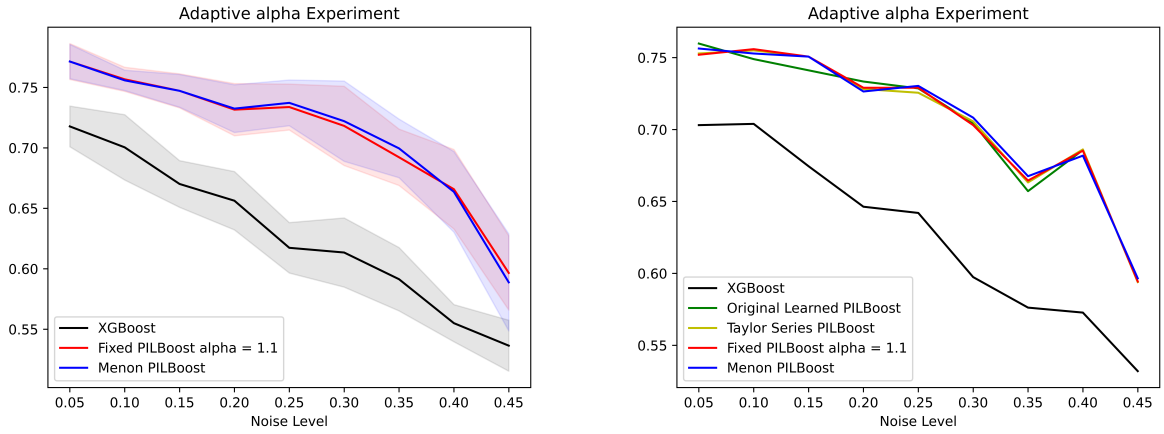

Figure 19: Companion Figure to Figures 2 and 18 on the *diabetes* dataset.

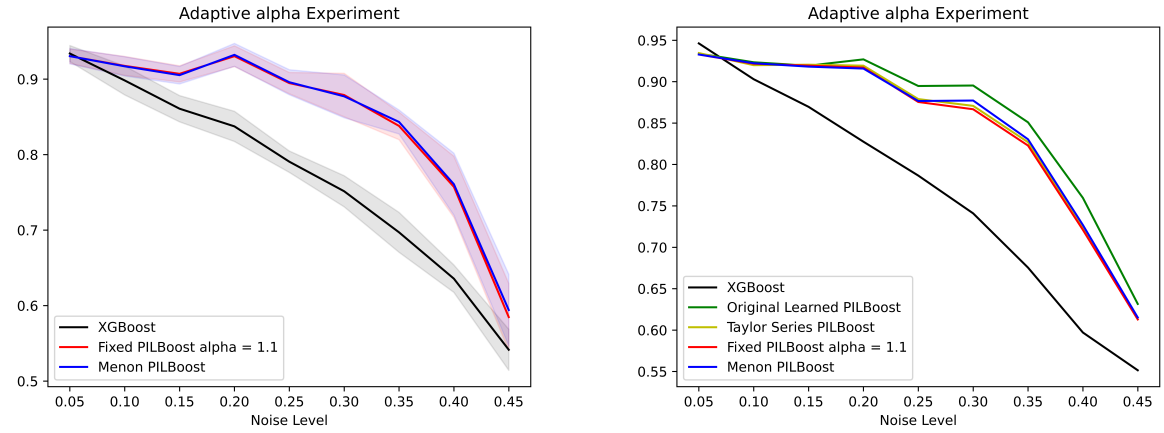

Figure 20: Companion Figure to Figures 2 and 18 on the *cancer* dataset.