Learning with Symmetric Label Noise: The Importance of Being Unhinged

Brendan van Rooyen[∗,†] **Aditya Krishna Menon**[†,∗] **Robert C. Williamson**[∗,†]

[∗]The Australian National University [†]National ICT Australia
{ brendan.vanrooyen, aditya.menon, bob.williamson }@nicta.com.au

Abstract

Convex potential minimisation is the *de facto* approach to binary classification. However, Long and Servedio [2010] proved that under symmetric label noise (SLN), minimisation of *any* convex potential over a linear function class can result in classification performance equivalent to random guessing. This ostensibly shows that convex losses are not SLN-robust. In this paper, we propose a convex, classification-calibrated loss and prove that it *is* SLN-robust. The loss avoids the Long and Servedio [2010] result by virtue of being *negatively unbounded*. The loss is a modification of the hinge loss, where one does not clamp at zero; hence, we call it the *unhinged loss*. We show that the optimal unhinged solution is equivalent to that of a strongly regularised SVM, and is the limiting solution for *any* convex potential; this implies that strong ℓ_2 regularisation makes most standard learners SLN-robust. Experiments confirm the unhinged loss' SLN-robustness. Thus, with apologies to Wilde [1895], the truth is rarely pure, but it *can* be simple.

1 Learning with symmetric label noise

Binary classification is the canonical supervised learning problem. Given an instance space $\mathcal{X}$, and samples from some distribution D over $\mathcal{X} \times \{\pm 1\}$, the goal is to learn a scorer $s \colon \mathcal{X} \to \mathbb{R}$ with low *misclassification error* on future samples drawn from D. Our interest is in the more realistic scenario where the learner observes samples from some corruption $\bar{D}$ of D, where labels have some constant probability of being flipped, and the goal is still to perform well with respect to D. This problem is known as learning from symmetric label noise (SLN learning) [Angluin and Laird, 1988].

Long and Servedio [2010] showed that there exist linearly separable D where, when the learner observes some corruption $\bar{D}$ with symmetric label noise of *any nonzero rate*, minimisation of *any convex potential* over a linear function class results in classification performance on D that is equivalent to random guessing. Ostensibly, this establishes that convex losses are not "SLN-robust" and motivates the use of non-convex losses [Stempfel and Ralaivola, 2009, Masnadi-Shirazi et al., 2010, Ding and Vishwanathan, 2010, Denchev et al., 2012, Manwani and Sastry, 2013].

In this paper, we propose a convex loss and prove that it *is* SLN-robust. The loss avoids the result of Long and Servedio [2010] by virtue of being *negatively unbounded*. The loss is a modification of the hinge loss where one does not clamp at zero; thus, we call it the *unhinged loss*. This loss has several appealing properties, such as being the unique convex loss satisfying a notion of "strong" SLN-robustness (Proposition 5), being classification-calibrated (Proposition 6), consistent when minimised on $\bar{D}$ (Proposition 7), and having an simple optimal solution that is the difference of two kernel means (Equation 8). Finally, we show that this optimal solution is equivalent to that of a strongly regularised SVM (Proposition 8), and *any* twice-differentiable convex potential (Proposition 9), implying that strong ℓ_2 regularisation endows most standard learners with SLN-robustness.

The classifier resulting from minimising the unhinged loss is not new [Devroye et al., 1996, Chapter 10], [Schölkopf and Smola, 2002, Section 1.2], [Shawe-Taylor and Cristianini, 2004, Section 5.1]. However, establishing this classifier's (strong) SLN-robustness, uniqueness thereof, and its equivalence to a highly regularised SVM solution, to our knowledge *is* novel.

2 Background and problem setup

Fix an instance space $\mathcal{X}$. We denote by D a distribution over $\mathcal{X} \times \{\pm 1\}$, with random variables $(\mathsf{X}, \mathsf{Y}) \sim D$. Any D may be expressed via the *class-conditionals* $(P, Q) = (\mathbb{P}(\mathsf{X} \mid \mathsf{Y} = 1), \mathbb{P}(\mathsf{X} \mid \mathsf{Y} = -1))$ and *base rate* $\pi = \mathbb{P}(\mathsf{Y} = 1)$, or via the *marginal* $M = \mathbb{P}(\mathsf{X})$ and *class-probability function* $\eta \colon x \mapsto \mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x)$. We interchangeably write D as $D_{P,Q,\pi}$ or $D_{M,\eta}$.

2.1 Classifiers, scorers, and risks

A *scorer* is any function $s \colon \mathcal{X} \to \mathbb{R}$. A *loss* is any function $\ell \colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$. We use ℓ_{-1}, ℓ_1 to refer to $\ell(-1, \cdot)$ and $\ell(1, \cdot)$. The *ℓ-conditional risk* $L_\ell \colon [0,1] \times \mathbb{R} \to \mathbb{R}$ is defined as $L_\ell \colon (\eta, v) \mapsto \eta \cdot \ell_1(v) + (1 - \eta) \cdot \ell_{-1}(v)$. Given a distribution D, the *ℓ-risk* of a scorer s is defined as

$$\mathbb{L}_\ell^D(s) \doteq \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right], \tag{1}$$

so that $\mathbb{L}_\ell^D(s) = \mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[L_\ell(\eta(\mathsf{X}), s(\mathsf{X})) \right]$. For a set $\mathcal{S}$, $\mathbb{L}_\ell^D(\mathcal{S})$ is the set of ℓ-risks for all scorers in $\mathcal{S}$.

A *function class* is any $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$. Given some $\mathcal{F}$, the set of *restricted Bayes-optimal scorers* for a loss ℓ are those scorers in $\mathcal{F}$ that minimise the ℓ-risk:

$$\mathcal{S}_\ell^{D,\mathcal{F},*} \doteq \mathop{\mathrm{Argmin}}_{s \in \mathcal{F}} \mathbb{L}_\ell^D(s).$$

The set of (unrestricted) Bayes-optimal scorers is $\mathcal{S}_\ell^{D,*} = \mathcal{S}_\ell^{D,\mathcal{F},*}$ for $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$. The *restricted ℓ-regret* of a scorer is its excess risk over that of any restricted Bayes-optimal scorer:

$$\mathrm{regret}_\ell^{D,\mathcal{F}}(s) \doteq \mathbb{L}_\ell^D(s) - \inf_{t \in \mathcal{F}} \mathbb{L}_\ell^D(t).$$

Binary classification is concerned with the *zero-one loss*, $\ell^{01} \colon (y, v) \mapsto [\![yv < 0]\!] + \frac{1}{2} [\![v = 0]\!]$. A loss ℓ is *classification-calibrated* if all its Bayes-optimal scorers are also optimal for zero-one loss: $(\forall D)\, \mathcal{S}_\ell^{D,*} \subseteq \mathcal{S}_{01}^{D,*}$. A *convex potential* is any loss $\ell \colon (y, v) \mapsto \phi(yv)$, where $\phi \colon \mathbb{R} \to \mathbb{R}_+$ is convex, non-increasing, differentiable with $\phi'(0) < 0$, and $\phi(+\infty) = 0$ [Long and Servedio, 2010, Definition 1]. All convex potentials are classification-calibrated [Bartlett et al., 2006, Theorem 2.1].

2.2 Learning with symmetric label noise (SLN learning)

The problem of learning with *symmetric label noise* (*SLN learning*) is the following [Angluin and Laird, 1988, Kearns, 1998, Blum and Mitchell, 1998, Natarajan et al., 2013]. For some notional "clean" distribution D, which we would like to observe, we instead observe samples from some corrupted distribution $\mathrm{SLN}(D, \rho)$, for some $\rho \in [0, 1/2)$. The distribution $\mathrm{SLN}(D, \rho)$ is such that the marginal distribution of instances is unchanged, but each label is independently flipped with probability ρ. The goal is to learn a scorer from these corrupted samples such that $\mathbb{L}_{01}^D(s)$ is small.

For any quantity in D, we denote its corrupted counterparts in $\mathrm{SLN}(D, \rho)$ with a bar, e.g. $\bar{M}$ for the corrupted marginal distribution, and $\bar{\eta}$ for the corrupted class-probability function; additionally, when ρ is clear from context, we will occasionally refer to $\mathrm{SLN}(D, \rho)$ by $\bar{D}$. It is easy to check that the corrupted marginal distribution $\bar{M} = M$, and [Natarajan et al., 2013, Lemma 7]

$$(\forall x \in \mathcal{X})\, \bar{\eta}(x) = (1 - 2\rho) \cdot \eta(x) + \rho. \tag{2}$$

3 SLN-robustness: formalisation

We consider learners $(\ell, \mathcal{F})$ for a loss ℓ and a function class $\mathcal{F}$, with learning being the search for some $s \in \mathcal{F}$ that minimises the ℓ-risk. Informally, $(\ell, \mathcal{F})$ is "robust" to symmetric label noise (SLN-robust) if minimising ℓ over $\mathcal{F}$ gives the same classifier on both the clean distribution D, which the learner would *like* to observe, and $\mathrm{SLN}(D, \rho)$ for *any* $\rho \in [0, 1/2)$, which the learner *actually* observes. We now formalise this notion, and review what is known about SLN-robust learners.

3.1 SLN-robust learners: a formal definition

For some fixed instance space $\mathcal{X}$, let Δ denote the set of distributions on $\mathcal{X} \times \{\pm 1\}$. Given a notional "clean" distribution D, $\mathcal{N}_{\mathrm{sln}} \colon \Delta \to 2^{\Delta}$ returns the *set* of possible corrupted versions of D the learner may observe, where labels are flipped with unknown probability ρ:

$$\mathcal{N}_{\mathrm{sln}} \colon D \mapsto \left\{ \mathrm{SLN}(D, \rho) \mid \rho \in \left[0, \frac{1}{2}\right) \right\}.$$

Equipped with this, we define our notion of SLN-robustness.

Definition 1 (SLN-robustness)**.** *We say that a learner* $(\ell, \mathcal{F})$ *is* SLN-robust *if*

$$(\forall D \in \Delta)\, (\forall \bar{D} \in \mathcal{N}_{\mathrm{sln}}(D))\, \mathbb{L}_{01}^{D}(\mathcal{S}_{\ell}^{D, \mathcal{F}, *}) = \mathbb{L}_{01}^{D}(\mathcal{S}_{\ell}^{\bar{D}, \mathcal{F}, *}). \tag{3}$$

That is, SLN-robustness requires that for *any* level of label noise in the observed distribution $\bar{D}$, the classification performance (wrt D) of the learner is the same as if the learner directly observes D. Unfortunately, a widely adopted class of learners is *not* SLN-robust, as we will now see.

3.2 Convex potentials with linear function classes are not SLN-robust

Fix $\mathcal{X} = \mathbb{R}^d$, and consider learners with a convex potential ℓ, and a function class of linear scorers

$$\mathcal{F}_{\mathrm{lin}} = \{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^d\}.$$

This captures e.g. the linear SVM and logistic regression, which are widely studied in theory and applied in practice. Disappointingly, these learners are *not* SLN-robust: Long and Servedio [2010, Theorem 2] give an example where, when learning under symmetric label noise, for *any* convex potential ℓ, the corrupted ℓ-risk minimiser over $\mathcal{F}_{\mathrm{lin}}$ has classification performance equivalent to random guessing on D. This implies that $(\ell, \mathcal{F}_{\mathrm{lin}})$ is not SLN-robust[1] as per Definition 1.

Proposition 1 (Long and Servedio [2010, Theorem 2])**.** *Let* $\mathcal{X} = \mathbb{R}^d$ *for any* $d \geq 2$*. Pick any convex potential* ℓ*. Then,* $(\ell, \mathcal{F}_{\mathrm{lin}})$ *is not SLN-robust.*

3.3 The fallout: what learners *are* SLN-robust?

In light of Proposition 1, there are two ways to proceed in order to obtain SLN-robust learners: either we change the class of losses ℓ, or we change the function class $\mathcal{F}$.

The first approach has been pursued in a large body of work that embraces non-convex losses [Stempfel and Ralaivola, 2009, Masnadi-Shirazi et al., 2010, Ding and Vishwanathan, 2010, Denchev et al., 2012, Manwani and Sastry, 2013]. While such losses avoid the conditions of Proposition 1, this does not automatically imply that they are SLN-robust when used with $\mathcal{F}_{\mathrm{lin}}$. In Appendix B, we present evidence that some of these losses are in fact *not* SLN-robust when used with $\mathcal{F}_{\mathrm{lin}}$.

The second approach is to consider suitably rich $\mathcal{F}$ that contains the Bayes-optimal scorer for $\bar{D}$, e.g. by employing a universal kernel. With this choice, one can still use a convex potential loss, and in fact, owing to Equation 2, *any* classification-calibrated loss.

Proposition 2. *Pick any classification-calibrated* ℓ*. Then,* $(\ell, \mathbb{R}^{\mathcal{X}})$ *is SLN-robust.*

Both approaches have drawbacks. The first approach has a computational penalty, as it requires optimising a non-convex loss. The second approach has a statistical penalty, as estimation rates with a rich $\mathcal{F}$ will require a larger sample size. Thus, it appears that SLN-robustness involves a computational-statistical tradeoff. However, there is a variant of the first option: pick a loss that is convex, *but not a convex potential*. Such a loss would afford the computational and statistical advantages of minimising convex risks with linear scorers. Manwani and Sastry [2013] demonstrated that square loss, $\ell(y, v) = (1 - yv)^2$, is one such loss. We will show that there is a simpler loss that is convex and SLN-robust, but is not in the class of convex potentials by virtue of being *negatively unbounded*. To derive this loss, we first re-interpret robustness via a noise-correction procedure.

[1]Even if we were content with a difference of $\epsilon \in [0, 1/2]$ between the clean and corrupted minimisers' performance, Long and Servedio [2010, Theorem 2] implies that in the worst case $\epsilon = 1/2$.

4 A noise-corrected loss perspective on SLN-robustness

We now re-express SLN-robustness to reason about optimal scorers on the *same distribution*, but with two *different losses*. This will help characterise a set of "strongly SLN-robust" losses.

4.1 Reformulating SLN-robustness via noise-corrected losses

Given any $\rho \in [0, 1/2)$, Natarajan et al. [2013, Lemma 1] showed how to associate with a loss ℓ a *noise-corrected* counterpart $\bar{\ell}$ such that $\mathbb{L}_\ell^D(s) = \mathbb{L}_{\bar{\ell}}^{\bar{D}}(s)$. The loss $\bar{\ell}$ is defined as follows.

Definition 2 (Noise-corrected loss)**.** *Given any loss ℓ and $\rho \in [0, 1/2)$, the noise-corrected loss $\bar{\ell}$ is*

$$(\forall y \in \{\pm 1\})\, (\forall v \in \mathbb{R})\, \bar{\ell}(y, v) = \frac{(1-\rho) \cdot \ell(y, v) - \rho \cdot \ell(-y, v)}{1 - 2\rho}. \tag{4}$$

Since $\bar{\ell}$ depends on the unknown parameter ρ, it is not directly usable to design an SLN-robust learner. Nonetheless, it is a useful theoretical device, since, by construction, for *any* $\mathcal{F}$, $\mathcal{S}_\ell^{D,\mathcal{F},*} = \mathcal{S}_{\bar{\ell}}^{\bar{D},\mathcal{F},*}$. This means that a sufficient condition for $(\ell, \mathcal{F})$ to be SLN-robust is for $\mathcal{S}_\ell^{\bar{D},\mathcal{F},*} = \mathcal{S}_{\bar{\ell}}^{\bar{D},\mathcal{F},*}$. Ghosh et al. [2015, Theorem 1] proved a *sufficient* condition on ℓ such that this holds, namely,

$$(\exists C \in \mathbb{R})(\forall v \in \mathbb{R})\, \ell_1(v) + \ell_{-1}(v) = C. \tag{5}$$

Interestingly, Equation 5 is *necessary* for a *stronger* notion of robustness, which we now explore.

4.2 Characterising a stronger notion of SLN-robustness

As the first step towards a stronger notion of robustness, we rewrite (with a slight abuse of notation)

$$\mathbb{L}_\ell^D(s) = \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D} [\ell(\mathsf{Y}, s(\mathsf{X}))] = \mathop{\mathbb{E}}_{(\mathsf{Y},\mathsf{S})\sim R(D,s)} [\ell(\mathsf{Y}, \mathsf{S})] \doteq \mathbb{L}_\ell(R(D, s)),$$

where $R(D, s)$ is a distribution over labels and *scores*. Standard SLN-robustness requires that label noise does not change the ℓ-risk minimisers, i.e. that if s is such that $\mathbb{L}_\ell(R(D, s)) \leq \mathbb{L}_\ell(R(D, s'))$ for all s', the same relation holds with $\bar{D}$ in place of D. Strong SLN-robustness strengthens this notion by requiring that label noise does not affect the ordering of *all* pairs of joint distributions over labels and scores. (This of course trivially implies SLN-robustness.) As with the definition of $\bar{D}$, given a distribution R over labels and scores, let $\bar{R}$ be the corresponding distribution where labels are flipped with probability ρ. Strong SLN-robustness can then be made precise as follows.

Definition 3 (Strong SLN-robustness)**.** *Call a loss ℓ strongly SLN-robust if for every $\rho \in [0, 1/2)$,*

$$(\forall R, R')\, \mathbb{L}_\ell(R) \leq \mathbb{L}_\ell(R') \iff \mathbb{L}_\ell(\bar{R}) \leq \mathbb{L}_\ell(\bar{R}').$$

We now re-express strong SLN-robustness using a notion of *order equivalence* of loss pairs, which simply requires that two losses order all distributions over labels and scores identically.

Definition 4 (Order equivalent loss pairs)**.** *Call a pair of losses $(\ell, \tilde{\ell})$ order equivalent if*

$$(\forall R, R')\, \mathbb{L}_\ell(R) \leq \mathbb{L}_\ell(R') \iff \mathbb{L}_{\tilde{\ell}}(R) \leq \mathbb{L}_{\tilde{\ell}}(R').$$

Clearly, order equivalence of $(\ell, \bar{\ell})$ implies $\mathcal{S}_\ell^{D,\mathcal{F},*} = \mathcal{S}_{\bar{\ell}}^{D,\mathcal{F},*}$, which in turn implies SLN-robustness. It is thus not surprising that we can relate order equivalence to strong SLN-robustness of ℓ.

Proposition 3. *A loss ℓ is strongly SLN-robust iff for every $\rho \in [0, 1/2)$, $(\ell, \bar{\ell})$ are order equivalent.*

This connection now lets us exploit a classical result in decision theory about order equivalent losses being affine transformations of each other. Combined with the definition of $\bar{\ell}$, this lets us conclude that the sufficient condition of Equation 5 is also *necessary* for strong SLN-robustness of ℓ.

Proposition 4. *A loss ℓ is strongly SLN-robust if and only if it satisfies Equation 5.*

We now return to our original goal, which was to find a convex ℓ that is SLN-robust for $\mathcal{F}_{\text{lin}}$ (and ideally more general function classes). The above suggests that to do so, it is reasonable to consider those losses that satisfy Equation 5. Unfortunately, it is evident that if ℓ is convex, non-constant, and bounded below by zero, then it cannot possibly be admissible in this sense. But we now show that removing the boundedness restriction allows for the existence of a convex admissible loss.

5 The unhinged loss: a convex, strongly SLN-robust loss

Consider the following simple, but non-standard convex loss:

$$\ell_1^{\mathrm{unh}}(v) = 1 - v \text{ and } \ell_{-1}^{\mathrm{unh}}(v) = 1 + v.$$

Compared to the hinge loss, the loss does not clamp at zero, i.e. it does not have a hinge. (Thus, peculiarly, it is negatively unbounded, an issue we discuss in §5.3.) Thus, we call this the *unhinged loss*[2]. The loss has a number of attractive properties, the most immediate being is its SLN-robustness.

5.1 The unhinged loss is strongly SLN-robust

Since $\ell_1^{\mathrm{unh}}(v) + \ell_{-1}^{\mathrm{unh}}(v) = 2$, Proposition 4 implies that ℓ^{unh} is strongly SLN-robust, and thus that $(\ell^{\mathrm{unh}}, \mathcal{F})$ is SLN-robust for *any* $\mathcal{F}$. Further, the following uniqueness property is not hard to show.

Proposition 5. *Pick any convex loss ℓ. Then,*

$$(\exists C \in \mathbb{R})\, \ell_1(v) + \ell_{-1}(v) = C \iff (\exists A, B, D \in \mathbb{R})\, \ell_1(v) = -A \cdot v + B, \ell_{-1}(v) = A \cdot v + D.$$

That is, up to scaling and translation, ℓ^{unh} is the only convex loss that is strongly SLN-robust.

Returning to the case of linear scorers, the above implies that $(\ell^{\mathrm{unh}}, \mathcal{F}_{\mathrm{lin}})$ is SLN-robust. This does not contradict Proposition 1, since ℓ^{unh} is not a convex potential as it is negatively unbounded. Intuitively, this property allows the loss to offset the penalty incurred by instances that are misclassified with high margin by awarding a "gain" for instances that correctly classified with high margin.

5.2 The unhinged loss is classification calibrated

SLN-robustness is by itself insufficient for a learner to be useful. For example, a loss that is uniformly zero is strongly SLN-robust, but is useless as it is not classification-calibrated. Fortunately, the unhinged loss is classification-calibrated, as we now establish. For technical reasons (see §5.3), we operate with $\mathcal{F}_B = [-B, +B]^{\mathcal{X}}$, the set of scorers with range bounded by $B \in [0, \infty)$.

Proposition 6. *Fix $\ell = \ell^{\mathrm{unh}}$. For any $D_{M,\eta}$, $B \in [0, \infty)$, $\mathcal{S}_\ell^{D, \mathcal{F}_B, *} = \{x \mapsto B \cdot \mathrm{sign}(2\eta(x) - 1)\}$.*

Thus, for every $B \in [0, \infty)$, the restricted Bayes-optimal scorer over $\mathcal{F}_B$ has the same sign as the Bayes-optimal classifier for 0-1 loss. In the limiting case where $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$, the optimal scorer is attainable if we operate over the extended reals $\mathbb{R} \cup \{\pm\infty\}$, so that ℓ^{unh} is classification-calibrated.

5.3 Enforcing boundedness of the loss

While the classification-calibration of ℓ^{unh} is encouraging, Proposition 6 implies that its (unrestricted) Bayes-risk is $-\infty$. Thus, the regret of every non-optimal scorer s is identically $+\infty$, which hampers analysis of consistency. In orthodox decision theory, analogous theoretical issues arise when attempting to establish basic theorems with unbounded losses [Ferguson, 1967, pg. 78].

We can side-step this issue by restricting attention to bounded scorers, so that ℓ^{unh} is effectively bounded. By Proposition 6, this does not affect the classification-calibration of the loss. In the context of linear scorers, boundedness of scorers can be achieved by regularisation: instead of working with $\mathcal{F}_{\mathrm{lin}}$, one can instead use $\mathcal{F}_{\mathrm{lin},\lambda} = \{x \mapsto \langle w, x \rangle \mid ||w||_2 \leq 1/\sqrt{\lambda}\}$, where $\lambda > 0$, so that $\mathcal{F}_{\mathrm{lin},\lambda} \subseteq \mathcal{F}_{R/\sqrt{\lambda}}$ for $R = \sup_{x \in \mathcal{X}} ||x||_2$. Observe that as $(\ell^{\mathrm{unh}}, \mathcal{F})$ is SLN-robust for *any* $\mathcal{F}$, $(\ell^{\mathrm{unh}}, \mathcal{F}_{\mathrm{lin},\lambda})$ is SLN-robust for any $\lambda > 0$. As we shall see in §6.3, working with $\mathcal{F}_{\mathrm{lin},\lambda}$ also lets us establish SLN-robustness of the hinge loss when λ is large.

5.4 Unhinged loss minimisation on corrupted distribution is consistent

Using bounded scorers makes it possible to establish a surrogate regret bound for the unhinged loss. This shows classification consistency of unhinged loss minimisation on the *corrupted* distribution.

[2] This loss has been considered in Sriperumbudur et al. [2009], Reid and Williamson [2011] in the context of maximum mean discrepancy; see Appendix E.4. The analysis of its SLN-robustness is to our knowledge novel.

Proposition 7. *Fix $\ell = \ell^{\mathrm{unh}}$. Then, for any $D, \rho \in [0, 1/2)$, $B \in [1, \infty)$, and scorer $s \in \mathcal{F}_B$,*

$$\mathrm{regret}_{01}^{D}(s) \leq \mathrm{regret}_{\ell}^{D,\mathcal{F}_B}(s) = \frac{1}{1-2\rho} \cdot \mathrm{regret}_{\ell}^{\bar{D},\mathcal{F}_B}(s).$$

Standard rates of convergence via generalisation bounds are also trivial to derive; see Appendix D.1.

6 Learning with the unhinged loss and kernels

We now show that the optimal solution for the unhinged loss when employing regularisation and kernelised scorers has a simple form. This sheds further light on SLN-robustness and regularisation.

6.1 The centroid classifier optimises the unhinged loss

Consider minimising the unhinged risk over the class of kernelised scorers $\mathcal{F}_{\mathcal{H},\lambda} = \{s \colon x \mapsto \langle w, \Phi(x) \rangle_{\mathcal{H}} \mid ||w||_{\mathcal{H}} \leq 1/\sqrt{\lambda}\}$ for some $\lambda > 0$, where $\Phi \colon \mathcal{X} \to \mathcal{H}$ is a feature mapping into a reproducing kernel Hilbert space $\mathcal{H}$ with kernel k. Equivalently, given a distribution[3] D, we want

$$w^*_{\mathrm{unh},\lambda} = \underset{w \in \mathcal{H}}{\mathrm{argmin}} \underset{(\mathsf{X},\mathsf{Y}) \sim D}{\mathbb{E}} [1 - \mathsf{Y} \cdot \langle w, \Phi(\mathsf{X}) \rangle] + \frac{\lambda}{2} \langle w, w \rangle_{\mathcal{H}}. \tag{6}$$

The first-order optimality condition implies that

$$w^*_{\mathrm{unh},\lambda} = \frac{1}{\lambda} \cdot \underset{(\mathsf{X},\mathsf{Y}) \sim D}{\mathbb{E}} [\mathsf{Y} \cdot \Phi(\mathsf{X})], \tag{7}$$

which is the *kernel mean map* of D [Smola et al., 2007], and thus the optimal unhinged scorer is

$$s^*_{\mathrm{unh},\lambda} \colon x \mapsto \frac{1}{\lambda} \cdot \underset{(\mathsf{X},\mathsf{Y}) \sim D}{\mathbb{E}} [\mathsf{Y} \cdot k(\mathsf{X}, x)] = x \mapsto \frac{1}{\lambda} \cdot \left(\pi \cdot \underset{\mathsf{X} \sim P}{\mathbb{E}} [k(\mathsf{X}, x)] - (1 - \pi) \cdot \underset{\mathsf{X} \sim Q}{\mathbb{E}} [k(\mathsf{X}, x)] \right). \tag{8}$$

From Equation 8, the unhinged solution is equivalent to a *nearest centroid classifier* [Manning et al., 2008, pg. 181] [Tibshirani et al., 2002] [Shawe-Taylor and Cristianini, 2004, Section 5.1]. Equation 8 gives a simple way to understand the SLN-robustness of $(\ell^{\mathrm{unh}}, \mathcal{F}_{\mathcal{H},\lambda})$, as the optimal scorers on the clean and corrupted distributions only differ by a scaling (see Appendix C):

$$(\forall x \in \mathcal{X}) \underset{(\mathsf{X},\mathsf{Y}) \sim D}{\mathbb{E}} [\mathsf{Y} \cdot k(\mathsf{X}, x)] = \frac{1}{1 - 2\rho} \cdot \underset{(\mathsf{X},\bar{\mathsf{Y}}) \sim \bar{D}}{\mathbb{E}} \left[\bar{\mathsf{Y}} \cdot k(\mathsf{X}, x) \right]. \tag{9}$$

Interestingly, Servedio [1999, Theorem 4] established that a nearest centroid classifier (which they termed "AVERAGE") is robust to a general class of label noise, but required the assumption that M is uniform over the unit sphere. Our result establishes that SLN robustness of the classifier holds without any assumptions on M. In fact, Ghosh et al. [2015, Theorem 1] lets one quantify the unhinged loss' performance under a more general noise model; see Appendix D.2 for discussion.

6.2 Practical considerations

We note several points relating to practical usage of the unhinged loss with kernelised scorers. First, cross-validation is not required to select λ, since changing λ only changes the magnitude of scores, *not their sign*. Thus, for the purposes of classification, one can simply use $\lambda = 1$.

Second, we can easily extend the scorers to use a bias regularised with strength $0 < \lambda_b \neq \lambda$. Tuning λ_b is equivalent to computing $s^*_{\mathrm{unh},\lambda}$ as per Equation 8, and tuning a threshold on a holdout set.

Third, when $\mathcal{H} = \mathbb{R}^d$ for d small, we can store $w^*_{\mathrm{unh},\lambda}$ explicitly, and use this to make predictions. For high (or infinite) dimensional $\mathcal{H}$, we can either make predictions directly via Equation 8, or use random Fourier features [Rahimi and Recht, 2007] to (approximately) embed $\mathcal{H}$ into some low-dimensional $\mathbb{R}^d$, and then store $w^*_{\mathrm{unh},\lambda}$ as usual. (The latter requires a translation-invariant kernel.)

We now show that under some assumptions, $w^*_{\mathrm{unh},\lambda}$ coincides with the solution of two established methods; Appendix E discusses some further relationships, e.g. to the maximum mean discrepancy.

[3] Given a training sample $\mathsf{S} \sim D^n$, we can use plugin estimates as appropriate.

6.3 Equivalence to a highly regularised SVM and other convex potentials

There is an interesting equivalence between the unhinged solution and that of a *highly regularised SVM*. This has been noted in e.g. Hastie et al. [2004, Section 6], which showed how SVMs approach a nearest centroid classifier, which is of course the optimal unhinged solution.

Proposition 8. *Pick any D and $\Phi\colon \mathcal{X} \to \mathcal{H}$ with $R = \sup_{x \in \mathcal{X}} ||\Phi(x)||_{\mathcal{H}} < \infty$. For any $\lambda > 0$, let*

$$w^*_{\text{hinge},\lambda} = \underset{w \in \mathcal{H}}{\operatorname{argmin}} \underset{(\mathsf{X},\mathsf{Y}) \sim D}{\mathbb{E}} [\max(0, 1 - \mathsf{Y} \cdot \langle w, \Phi(x) \rangle_{\mathcal{H}})] + \frac{\lambda}{2} \langle w, w \rangle_{\mathcal{H}}$$

be the soft-margin SVM solution. Then, if $\lambda \geq R^2$, $w^*_{\text{hinge},\lambda} = w^*_{\text{unh},\lambda}$.

Since $(\ell^{\text{unh}}, \mathcal{F}_{\mathcal{H},\lambda})$ is SLN-robust, it follows that for $\ell^{\text{hinge}}\colon (y, v) \mapsto \max(0, 1 - yv)$, $(\ell^{\text{hinge}}, \mathcal{F}_{\mathcal{H},\lambda})$ is similarly SLN-robust *provided λ is sufficiently large*. That is, strong ℓ_2 regularisation (and a bounded feature map) endows the hinge loss with SLN-robustness[4]. Proposition 8 can be generalised to show that $w^*_{\text{unh},\lambda}$ is the limiting solution of *any* twice differentiable convex potential. This shows that *strong ℓ_2 regularisation endows most learners with SLN-robustness*. Intuitively, with strong regularisation, one only considers the behaviour of a loss near zero; since a convex potential ϕ has $\phi'(0) < 0$, it will behave similarly to its linear approximation around zero, viz. the unhinged loss.

Proposition 9. *Pick any D, bounded feature mapping $\Phi\colon \mathcal{X} \to \mathcal{H}$, and twice differentiable convex potential ϕ with $\phi''([-1, 1])$ bounded. Let $w^*_{\phi,\lambda}$ be the minimiser of the regularised ϕ risk. Then,*

$$\lim_{\lambda \to \infty} \left|\left| \frac{w^*_{\phi,\lambda}}{||w^*_{\phi,\lambda}||_{\mathcal{H}}} - \frac{w^*_{\text{unh},\lambda}}{||w^*_{\text{unh},\lambda}||_{\mathcal{H}}} \right|\right|^2_{\mathcal{H}} = 0.$$

6.4 Equivalence to Fisher Linear Discriminant with whitened data

For binary classification on $D_{M,\eta}$, the Fisher Linear Discriminant (FLD) finds a weight vector proportional to the minimiser of square loss $\ell^{\text{sq}}\colon (y, v) \mapsto (1 - yv)^2$ [Bishop, 2006, Section 4.1.5],

$$w^*_{\text{sq},\lambda} = (\mathbb{E}_{\mathsf{X} \sim M}[\mathsf{X}\mathsf{X}^T] + \lambda I)^{-1} \cdot \mathbb{E}_{(\mathsf{X},\mathsf{Y}) \sim D}[\mathsf{Y} \cdot \mathsf{X}]. \quad (10)$$

By Equation 9, and the fact that the corrupted marginal $\bar{M} = M$, $w^*_{\text{sq},\lambda}$ is only changed by a scaling factor under label noise. This provides an alternate proof of the fact that $(\ell^{\text{sq}}, \mathcal{F}_{\text{lin}})$ is SLN-robust[5] [Manwani and Sastry, 2013, Theorem 2]. Clearly, the unhinged loss solution $w^*_{\text{unh},\lambda}$ is equivalent to the FLD and square loss solution $w^*_{\text{sq},\lambda}$ when the input data is whitened i.e. $\underset{\mathsf{X} \sim M}{\mathbb{E}} [\mathsf{X}\mathsf{X}^T] = I$. With a well-specified $\mathcal{F}$, e.g. with a universal kernel, both the unhinged and square loss asymptotically recover the optimal classifier, but the unhinged loss does not require a matrix inversion. With a misspecified $\mathcal{F}$, one cannot in general argue for the superiority of the unhinged loss over square loss, or vice-versa, as there is no universally good surrogate to the 0-1 loss [Reid and Williamson, 2010, Appendix A]; Appendices F, G illustrate examples where both losses may underperform.

7 SLN-robustness of unhinged loss: empirical illustration

We now illustrate that the unhinged loss' SLN-robustness is empirically manifest. We reiterate that with high regularisation, the unhinged solution is equivalent to an SVM (and in the limit any classification-calibrated loss) solution. Thus, we do *not* aim to assert that the unhinged loss is "better" than other losses, but rather, to demonstrate that its SLN-robustness is not *purely* theoretical.

We first show that the unhinged risk minimiser performs well on the example of Long and Servedio [2010] (henceforth `LS10`). Figure 1 shows the distribution D, where $\mathcal{X} = \{(1, 0), (\gamma, 5\gamma), (\gamma, -\gamma)\} \subset \mathbb{R}^2$, with marginal distribution $M = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\}$ and all three instances are deterministically positive. We pick $\gamma = 1/2$. The unhinged minimiser perfectly classifies all three points, regardless of the level of label noise (Figure 1). The hinge minimiser is perfect when there is no noise, but with even a small amount of noise, achieves a 50% error rate.

[4] Long and Servedio [2010, Section 6] show that ℓ_1 regularisation does not endow SLN-robustness.

[5] Square loss escapes the result of Long and Servedio [2010] since it is not monotone decreasing.

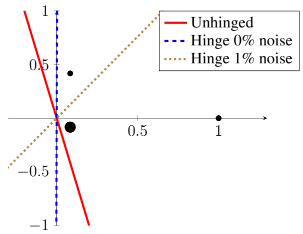

Figure 1: LS10 dataset.

	Hinge	t-logistic	Unhinged
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.1$	0.15 ± 0.27	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.2$	0.21 ± 0.30	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.3$	0.38 ± 0.37	0.22 ± 0.08	0.00 ± 0.00
$\rho = 0.4$	0.42 ± 0.36	0.22 ± 0.08	0.00 ± 0.00
$\rho = 0.49$	0.47 ± 0.38	0.39 ± 0.23	0.34 ± 0.48

Table 1: Mean and standard deviation of the 0-1 error over 125 trials on LS10. Grayed cells denote the best performer at that noise rate.

We next consider empirical risk minimisers from a random training sample: we construct a training set of 800 instances, injected with varying levels of label noise, and evaluate classification performance on a test set of 1000 instances. We compare the hinge, t-logistic (for $t = 2$) [Ding and Vishwanathan, 2010] and unhinged minimisers using a linear scorer *without* a bias term, and regularisation strength $\lambda = 10^{-16}$. From Table 1, even at 40% label noise, the unhinged classifier is able to find a perfect solution. By contrast, both other losses suffer at even moderate noise rates.

We next report results on some UCI datasets, where we additionally tune a threshold so as to ensure the best training set 0-1 accuracy. Table 2 summarises results on a sample of four datasets. (Appendix H contains results with more datasets, performance metrics, and losses.) Even at noise close to 50%, the unhinged loss is often able to learn a classifier with some discriminative power.

	Hinge	t-Logistic	Unhinged
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.1$	0.01 ± 0.03	0.01 ± 0.03	0.00 ± 0.00
$\rho = 0.2$	0.06 ± 0.12	0.04 ± 0.05	0.00 ± 0.01
$\rho = 0.3$	0.17 ± 0.20	0.09 ± 0.11	0.02 ± 0.07
$\rho = 0.4$	0.35 ± 0.24	0.24 ± 0.16	0.13 ± 0.22
$\rho = 0.49$	0.60 ± 0.20	0.49 ± 0.20	0.45 ± 0.33

(a) iris.

	Hinge	t-Logistic	Unhinged
$\rho = 0$	0.05 ± 0.00	0.05 ± 0.00	0.05 ± 0.00
$\rho = 0.1$	0.06 ± 0.01	0.07 ± 0.02	0.05 ± 0.00
$\rho = 0.2$	0.06 ± 0.01	0.08 ± 0.03	0.05 ± 0.00
$\rho = 0.3$	0.08 ± 0.04	0.11 ± 0.05	0.05 ± 0.01
$\rho = 0.4$	0.14 ± 0.10	0.24 ± 0.13	0.09 ± 0.10
$\rho = 0.49$	0.45 ± 0.26	0.49 ± 0.16	0.46 ± 0.30

(b) housing.

	Hinge	t-Logistic	Unhinged
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.1$	0.10 ± 0.08	0.11 ± 0.02	0.00 ± 0.00
$\rho = 0.2$	0.19 ± 0.11	0.15 ± 0.02	0.00 ± 0.00
$\rho = 0.3$	0.31 ± 0.13	0.22 ± 0.03	0.01 ± 0.00
$\rho = 0.4$	0.39 ± 0.13	0.33 ± 0.04	0.02 ± 0.02
$\rho = 0.49$	0.50 ± 0.16	0.48 ± 0.04	0.34 ± 0.21

(c) usps0v7.

	Hinge	t-Logistic	Unhinged
$\rho = 0$	0.05 ± 0.00	0.04 ± 0.00	0.19 ± 0.00
$\rho = 0.1$	0.15 ± 0.03	0.24 ± 0.00	0.19 ± 0.01
$\rho = 0.2$	0.21 ± 0.03	0.24 ± 0.00	0.19 ± 0.01
$\rho = 0.3$	0.25 ± 0.03	0.24 ± 0.00	0.19 ± 0.03
$\rho = 0.4$	0.31 ± 0.05	0.24 ± 0.00	0.22 ± 0.05
$\rho = 0.49$	0.48 ± 0.09	0.40 ± 0.24	0.45 ± 0.08

(d) splice.

Table 2: Mean and standard deviation of the 0-1 error over 125 trials on UCI datasets.

8 Conclusion and future work

We proposed a convex, classification-calibrated loss, proved that is robust to symmetric label noise (SLN-robust), showed it is the unique loss that satisfies a notion of strong SLN-robustness, established that it is optimised by the nearest centroid classifier, and showed that most convex potentials, such as the SVM, are also SLN-robust when highly regularised. So, with apologies to Wilde [1895]:

While the truth is rarely pure, it *can* be simple.

Acknowledgments

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. The authors thank Cheng Soon Ong for valuable comments on a draft of this paper.

Proofs for "Learning with Symmetric Label Noise: The Importance of Being Unhinged"

A Proofs of results in main body

We now present proofs of all results in the main body.

Proof of Proposition 1. This result is implicit in Long and Servedio [2010, Theorem 2]; the aim of this proof is simply to make the result explicit, and to cast it in our terminology.

Let $\mathcal{X} = \{(1,0), (\gamma, 5\gamma), (\gamma, -\gamma), (\gamma, -\gamma)\} \subset \mathbb{R}^2$, for some $\gamma < 1/6$. Define a distribution D as follows: let the marginal distribution over $\mathcal{X}$ be uniform, and let $\eta \colon x \mapsto 1$, i.e. every example is deterministically positive.

Now suppose we observe some $\mathrm{SLN}(D, \rho)$, for $\rho \in [0, 1/2)$. We minimise the ℓ-risk some convex potential $\ell \colon (y, v) \mapsto \phi(y, v)$ using a linear function class[6] $\mathcal{F}_{\mathrm{lin}}$. Then, Long and Servedio [2010, Theorem 2] establishes that

$$(\forall s \in \mathcal{S}_\ell^{\bar{D}, \mathcal{F}_{\mathrm{lin}}, *})\, \mathbb{L}_{01}^D(s) = \frac{1}{2}.$$

On the other hand, since D is linearly separable and a convex potential ℓ is classification-calibrated, we must have $\mathbb{L}_{01}^D(\mathcal{S}_\ell^{D, \mathcal{F}_{\mathrm{lin}}, *}) = 0$. Consequently, for any convex potential ℓ, $(\ell, \mathcal{F}_{\mathrm{lin}})$ is not SLN-robust. □

Proof of Proposition 2. Let $\bar{\eta}$ be the class-probability function of $\bar{D}$. By [Natarajan et al., 2013, Lemma 7],

$$(\forall x \in \mathcal{X})\ \mathrm{sign}(2\bar{\eta}(x) - 1) = \mathrm{sign}(2\eta(x) - 1),$$

so that the optimal classifiers on the clean and corrupted distributions coincide. Therefore, intuitively, if the Bayes-optimal solution for loss recovers $\mathrm{sign}(2\bar{\eta}(x) - 1)$, it will also recover $\mathrm{sign}(2\eta(x) - 1)$. Formally, since ℓ is classification-calibrated, for any $D \in \Delta$, and $s \in \mathcal{S}_\ell^{D,*}$

$$(\forall x \in \mathcal{X})\ \mathrm{sign}(s(x)) = \mathrm{sign}(2\eta(x) - 1),$$

and similarly, for any $\bar{D} \in \mathcal{N}_{\mathrm{sln}}(D)$, and $\bar{s} \in \mathcal{S}_\ell^{\bar{D},*}$

$$(\forall x \in \mathcal{X})\ \mathrm{sign}(\bar{s}(x)) = \mathrm{sign}(2\bar{\eta}(x) - 1).$$

Thus, for any $D, \bar{D}$, since the 0-1 risk of a scorer depends only on its sign,

$$\begin{aligned}
\mathbb{L}_{01}^D(s) &= \mathbb{L}_{01}^D(\mathrm{sign}(s)) \\
&= \mathbb{L}_{01}^D(\mathrm{sign}(2\eta - 1)) \\
&= \mathbb{L}_{01}^D(\mathrm{sign}(2\bar{\eta} - 1)) \\
&= \mathbb{L}_{01}^D(\mathrm{sign}(\bar{s})) \\
&= \mathbb{L}_{01}^D(\bar{s}).
\end{aligned}$$

Consequently, $(\ell, \mathbb{R}^{\mathcal{X}})$ is SLN-robust. □

Proof of Proposition 3. For brevity, we will write $R \preceq_\ell R'$ to mean that $\mathbb{L}_\ell(R) \leq \mathbb{L}_\ell(R')$, so that strong SLN-robustness is

$$(\forall \rho \in [0, 1/2))\, (\forall R, R')\, R \preceq_\ell R' \iff \bar{R} \preceq_\ell \bar{R}'.$$

while order equivalence is

$$(\forall R, R')\, R \preceq_\ell R' \iff R \preceq_{\tilde{\ell}} R'.$$

[6] The specific choice of D *requires* that one not include a bias term; with a bias term, it can be checked that the example as-stated has a trivial solution.

Observe also that by definition of $\bar{\ell}$ and $\bar{R}$,

$$\mathbb{L}_{\ell}(R) = \mathop{\mathbb{E}}_{(\mathsf{Y},\mathsf{S})\sim R}[\ell(\mathsf{Y},\mathsf{S})] = \mathop{\mathbb{E}}_{(\mathsf{Y},\bar{\mathsf{S}})\sim \bar{R}}\left[\bar{\ell}(\mathsf{Y},\mathsf{S})\right] = \mathbb{L}_{\bar{\ell}}(\bar{R}).$$

($\Longleftarrow$). Pick any $\rho \in [0, 1/2)$. Suppose that $(\ell, \bar{\ell})$ are order equivalent. We have

$$\begin{aligned}(\forall R, R')\, R \preceq_{\ell} R' &\iff \bar{R} \preceq_{\bar{\ell}} \bar{R}' \\ &\iff \bar{R} \preceq_{\ell} \bar{R}',\end{aligned}$$

where the first line is from definition of $\bar{\ell}$, and the second is by assumed order equivalence of $(\ell, \bar{\ell})$. As ρ is arbitrary, we conclude that ℓ is strongly SLN-robust.

($\Longrightarrow$). Pick any $\rho \in [0, 1/2)$. For this direction, we will need to define the following "inverse" noise-corrected loss,

$$(\forall y \in \{\pm 1\})\, (\forall v \in \mathbb{R})\, \tilde{\ell}(y, v) = (1 - \rho) \cdot \ell(y, v) + \rho \cdot \ell(-y, v)$$

which is so named because $\bar{\tilde{\ell}} = \ell$.

Now suppose that ℓ is strongly SLN-robust. We have by definition

$$\begin{aligned}(\forall R, R')\, R \preceq_{\ell} R' &\iff \bar{R} \preceq_{\ell} \bar{R}' \\ &\iff R \preceq_{\tilde{\ell}} R',\end{aligned}$$

where the second line by the fact that $\mathbb{L}^{\bar{D}}_{\ell}(s) = \mathbb{L}^{D}_{\tilde{\ell}}(s)$. This means that $(\ell, \tilde{\ell})$ are order equivalent for any $\rho \in [0, 1/2)$. Thus, by Lemma 10,

$$(\forall \rho \in [0, 1/2))\, (\forall y \in \{\pm 1\})\, (\forall v \in \mathbb{R})\, \tilde{\ell}(y, v) = \alpha \cdot \ell(y, v) + \beta.$$

But now using the fact that $\bar{\tilde{\ell}} = \ell$, we get

$$(\forall \rho \in [0, 1/2))\, (\forall y \in \{\pm 1\})\, (\forall v \in \mathbb{R})\, \ell(y, v) = \alpha \cdot \bar{\ell}(y, v) + \beta.$$

Applying Lemma 10 once more, we conclude that $(\ell, \bar{\ell})$ are order equivalent for any $\rho \in [0, 1/2)$. □

Proof of Proposition 4. ($\Longleftarrow$). For standard SLN-robustness, this is shown in Ghosh et al. [2015, Theorem 1]; for strong SLN-robustness, the same basic proof strategy is adopted. Suppose that Equation 5 holds. Then,

$$\begin{aligned}\bar{\ell}(y, v) &= \frac{(1 - \rho) \cdot \ell(y, v) - \rho \cdot \ell(-y, v)}{1 - 2\rho} \\ &= \frac{(1 - \rho) \cdot \ell(y, v) - \rho \cdot (C - \ell(y, v))}{1 - 2\rho} \\ &= \frac{\ell(y, v) - \rho \cdot C}{1 - 2\rho},\end{aligned}$$

at which stage we appeal to Lemma 10 to conclude that $(\ell, \bar{\ell})$ are order equivalent for any ρ. Thus by Proposition 3, ℓ is strongly SLN-robust.

($\Longrightarrow$). We have shown in Proposition 3 that strong SLN-robustness is equivalent to order equivalence of $(\ell, \bar{\ell})$ for every $\rho \in [0, 1/2)$. Thus, by Lemma 10,

$$(\forall \rho \in [0, 1/2))\, (\forall y \in \{\pm 1\})\, (\forall v \in \mathbb{R})\, \ell(y, v) = \alpha \cdot \bar{\ell}(y, v) + \beta.$$

By the definition of the noise-corrected loss (Equation 4), the given statement is that there exist $\alpha, \beta \colon [0, 1/2) \to \mathbb{R}$ with

$$(\forall \rho \in [0, 1/2))\, (\forall v \in \mathbb{R})\, \begin{bmatrix}\ell_1(v) \\ \ell_{-1}(v)\end{bmatrix} = \frac{\alpha(\rho)}{1 - 2\rho} \cdot \begin{bmatrix}1 - \rho & -\rho \\ -\rho & 1 - \rho\end{bmatrix} \cdot \begin{bmatrix}\ell_1(v) \\ \ell_{-1}(v)\end{bmatrix} + \beta(\rho).$$

Adding together the two sets of equations (i.e. multiplying both sides by the all ones vector),

$$\begin{aligned}
&(\forall \rho \in [0, 1/2))\, (\forall v \in \mathbb{R})\, \ell_1(v) + \ell_{-1}(v) = \alpha(\rho) \cdot (\ell_1(v) + \ell_{-1}(v)) + 2\beta(\rho) \\
\iff &(\forall \rho \in [0, 1/2))\, (\forall v \in \mathbb{R})\, (1 - \alpha(\rho)) \cdot (\ell_1(v) + \ell_{-1}(v)) = 2\beta(\rho) \\
\iff &(\forall \rho \mid \alpha(\rho) \neq 1)\, (\forall v \in \mathbb{R})\, \ell_1(v) + \ell_{-1}(v) = \frac{2\beta(\rho)}{1 - \alpha(\rho)},
\end{aligned}$$

which is a constant independent of v. If $\alpha(\rho) \equiv 1$, then clearly $\beta(\rho) \equiv 0$, and we have

$$(\forall v \in \mathbb{R})\, (1 - \rho) \cdot \ell_1(v) - \rho \cdot \ell_{-1}(v) = (1 - 2\rho) \cdot \ell_1(v),$$

thus implying that $\ell_1(v) = \ell_{-1}(v)$. Such a loss is not interesting since it cannot possibly be classification calibrated. ☐

Proof of Proposition 5. ($\Longleftarrow$). Clearly for an ℓ satisfying the given condition, $\ell_1(v) + \ell_{-1}(v) = B + C$, a constant.

($\Longrightarrow$). By assumption, ℓ_1 is convex. By the given condition, equivalently, $(\exists C \in \mathbb{R})\, C - \ell_1$ is convex. But this is in turn equivalent to $-\ell_1$ also being convex. The only possibility for both ℓ_1 and $-\ell_1$ being convex is that ℓ_1 is affine, hence showing the desired implication. ☐

Proof of Proposition 6. Fix $\ell = \ell^{\mathrm{unh}}$. It is easy to check that

$$(\forall \eta \in [0, 1])\, (\forall v \in \mathbb{R})\, L_\ell(\eta, v) = (1 - 2\eta) \cdot v + 1, \tag{11}$$

and so

$$(\forall \eta \in [0, 1])\, \underset{v \in [-B, +B]}{\operatorname{argmin}}\, L_\ell(\eta, v) = \begin{cases} +B & \text{if } \eta > \frac{1}{2} \\ -B & \text{else.} \end{cases}$$

It is not a coincidence that the above is a scaled version of the minimiser for the hinge loss. Trivially, for any $v \in [-B, +B]$, we have that

$$\begin{aligned}
\ell(y, v) &= 1 - yv \\
&= B \cdot \left(1 - \frac{yv}{B}\right) + (1 - B) \\
&= B \cdot \max\left(0, 1 - \frac{yv}{B}\right) + (1 - B) \text{ as } yv \leq B \\
&= B \cdot \ell^{\mathrm{hinge}}\left(y, \frac{v}{B}\right) + (1 - B) \qquad (12)
\end{aligned}$$

where $\ell^{\mathrm{hinge}}(y, v) = \max(0, 1 - yv)$. It follows that

$$(\forall \eta \in [0, 1])\, (\forall v \in [-B, +B])\, L_\ell(\eta, v) = B \cdot L_{\mathrm{hinge}}\left(\eta, \frac{v}{B}\right) + (1 - B),$$

and so

$$\begin{aligned}
(\forall \eta \in [0, 1])\, \underset{v \in [-B, +B]}{\operatorname{argmin}}\, L_\ell(\eta, v) &= \underset{v \in [-B, +B]}{\operatorname{argmin}}\, L_{\mathrm{hinge}}\left(\eta, \frac{v}{B}\right) \\
&= B \cdot \underset{\tilde{v} \in [-1, +1]}{\operatorname{argmin}}\, L_{\mathrm{hinge}}(\eta, \tilde{v}) \\
&= B \cdot \begin{cases} +1 & \text{if } \eta > \frac{1}{2} \\ -1 & \text{else.} \end{cases}
\end{aligned}$$

☐

Proof of Proposition 7. The basic idea of the surrogate regret bound can be seen visually. Figure 2 compares the zero-one, unhinged and hinge losses. If we restrict attention to $[-1, 1]$, the unhinged and hinge losses are identical, and surrogate regret bounds for the latter apply to the former. For general $B > 1$, we simply need to consider an appropriately scaled version of the hinge loss, and proceed identically.

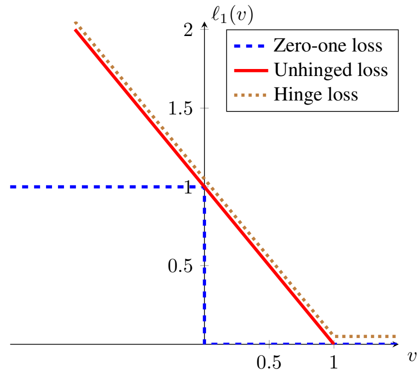

Figure 2: Relationship between zero-one, unhinged and hinge loss.

Now we make this more formal. Fix $\ell = \ell^{\text{unh}}$. Clearly, for any $v \in [-B, B]$,

$$\begin{aligned}\ell^{01}(y, v) &\leq \ell^{\text{hinge}}\left(y, \frac{v}{B}\right) \\ &= \frac{1}{B} \cdot \ell(y, v) + 1 - \frac{1}{B} \text{ by Equation 12.}\end{aligned}$$

Therefore, for any $s \in \mathcal{F}_B$,

$$\mathbb{L}_{01}^{D}(s) \leq \frac{1}{B} \cdot \mathbb{L}_{\ell}^{D}(s) + 1 - \frac{1}{B}.$$

Further, it is not hard to check that $\mathbb{L}_{01}^{D,\mathcal{F}_B,*} = \mathbb{L}_{\text{hinge}}^{D,\mathcal{F}_B,*}$. Thus,

$$\begin{aligned}\mathbb{L}_{01}^{D,\mathcal{F}_B,*} &= \mathbb{L}_{\text{hinge}}^{D,\mathcal{F}_B,*} \\ &= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\ell^{\text{hinge}}(\mathsf{Y}, \operatorname{sign}(2\eta(\mathsf{X}) - 1))\right] \\ &= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\frac{1}{B} \cdot \ell(\mathsf{Y}, B \cdot \operatorname{sign}(2\eta(\mathsf{X}) - 1)) + 1 - \frac{1}{B}\right] \text{ by Equation 12} \\ &= \frac{1}{B} \cdot \mathbb{L}_{\ell}^{D,\mathcal{F}_B,*} + 1 - \frac{1}{B} \text{ by Proposition 6.}\end{aligned}$$

Now, since the scorer $x \mapsto \operatorname{sign}(2\eta(x)-1) \in \mathcal{S}_{01}^{D,*} \cap \mathcal{F}_B$, we have that $\text{regret}_{01}^{D,\mathcal{F}_B}(s) = \text{regret}_{01}^{D}(s)$. Thus, for any $s \in \mathcal{F}_B$,

$$\begin{aligned}\text{regret}_{01}^{D}(s) &= \mathbb{L}_{01}^{D,\mathcal{F}_B}(s) - \mathbb{L}_{01}^{D,\mathcal{F}_B,*} \\ &\leq \frac{1}{B} \cdot \left(\mathbb{L}_{\ell}^{D,\mathcal{F}_B}(s) - \mathbb{L}_{\ell}^{D,\mathcal{F}_B,*}\right) \\ &= \frac{1}{B} \cdot \text{regret}_{\ell}^{D}(s) \\ &\leq \text{regret}_{\ell}^{D}(s) \text{ since } B \geq 1.\end{aligned}$$

To show the relation to the corrupted regret, by Equation 4, for $\ell = \ell^{\text{unh}}$,

$$(\forall y \in \{\pm 1\})\, (\forall v \in \mathbb{R})\, \bar{\ell}(y, v) = \frac{1}{1 - 2\rho} \cdot \ell(y, v),$$

i.e. the unhinged loss is its own noise-corrected loss, with a scaling factor of $\frac{1}{1-2\rho}$. Thus, since the ℓ-regret on D and $\bar{\ell}$-regret on $\bar{D}$ coincide,

$$\text{regret}_{\ell}^{D,\mathcal{F}_B}(s) = \text{regret}_{\bar{\ell}}^{\bar{D},\mathcal{F}_B}(s) = \frac{1}{1 - 2\rho} \cdot \text{regret}_{\ell}^{\bar{D},\mathcal{F}_B}(s).$$

$\square$

Proof of Proposition 8. While this has effectively been shown in Hastie et al. [2004, Section 6] by virtue of the connection between the unhinged loss and nearest centroid classification, we will prove this using a different technique. On a distribution D, a soft-margin SVM solves

$$\min_{w \in \mathcal{H}} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\max(0, 1 - \mathsf{Y} \cdot \langle w, \Phi(x) \rangle_{\mathcal{H}})] + \frac{\lambda}{2} \langle w, w \rangle_{\mathcal{H}}^2.$$

Let $w^*_{\text{hinge},\lambda}$ denote the optimal solution to this objective. Now, by Shalev-Shwartz et al. [2007, Theorem 1],

$$||w^*_{\text{hinge},\lambda}||_{\mathcal{H}} \leq \frac{1}{\sqrt{\lambda}}.$$

Now suppose that $R = \sup_{x \in \mathcal{X}} ||\Phi(x)||_{\mathcal{H}} < \infty$. Then, by the Cauchy-Schwartz inequality,

$$(\forall x \in \mathcal{X})\, |\langle w^*_{\text{hinge},\lambda}, \Phi(x) \rangle_{\mathcal{H}}| \leq ||w^*_{\text{hinge},\lambda}||_{\mathcal{H}} \cdot ||\Phi(x)||_{\mathcal{H}} \leq \frac{R}{\sqrt{\lambda}}.$$

It follows that if $\lambda \geq R^2$, then

$$(\forall x \in \mathcal{X})\, |\langle w^*_{\text{hinge},\lambda}, \Phi(x) \rangle_{\mathcal{H}}| \leq 1.$$

But this means that we never activate the flat portion of the hinge loss. Thus, for $\lambda \geq R^2$, the SVM objective is equivalent to

$$\min_{w \in \mathcal{H}} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [1 - \mathsf{Y} \cdot \langle w, \Phi(x) \rangle_{\mathcal{H}}] + \frac{\lambda}{2} \langle w, w \rangle_{\mathcal{H}}^2.$$

which means the optimal solution will coincide with that of the regularised unhinged loss. Therefore, we can view unhinged loss minimisation as corresponding to learning a highly regularised SVM[7]. □

Proof of Proposition 9. Fix some distribution D. Let

$$\mu = \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\mathsf{Y} \cdot \Phi(\mathsf{X})]$$

be the optimal unhinged solution with regularisation strength $\lambda = 1$. Observe that $||\mu||_{\mathcal{H}} \leq R = \sup_{x \in \mathcal{X}} ||\Phi(x)||_{\mathcal{H}} < \infty$. For some $r > 0$, let

$$w^*_\phi = \mathop{\text{argmin}}_{||w||_{\mathcal{H}} \leq r} \mathbb{L}^D_\phi(w)$$

be the optimal ϕ solution with norm bounded by r. Similarly, let

$$w^*_{\text{unh}} = ||w^*_\phi|| \cdot \frac{\mu}{||\mu||_{\mathcal{H}}}$$

be the optimal unhinged solution with the same norm as the optimal ϕ solution. We will show that these two vectors have similar unhinged risks, and use this to show that the corresponding unit vectors must be close.

By definition, a convex potential has $\phi'(0) < 0$. As scaling of a loss does not affect its optimal solution, without loss of generality, we can assume $\phi'(0) = -1$. Then, since ϕ is convex, it is lower bounded by the linear approximation at zero:

$$(\forall v \in \mathbb{R})\, \phi(v) + 1 - \phi(0) \geq 1 - v.$$

Observe that the RHS is the unhinged loss. Thus, the unhinged risk of a candidate solution can be bounded by its ϕ counterpart. Now we compare the unhinged and ϕ optimal solutions in terms of their unhinged risks:

$$\begin{aligned}
\mathbb{L}^D_{\text{unh}}(w^*_\phi) - \mathbb{L}^D_{\text{unh}}(w^*_{\text{unh}}) &\leq \mathbb{L}_\phi(w^*_\phi) - \mathbb{L}_{\text{unh}}(w^*_{\text{unh}}) + 1 - \phi(0) \\
&\leq \mathbb{L}^D_\phi(w^*_{\text{unh}}) - \mathbb{L}^D_{\text{unh}}(w^*_{\text{unh}}) + 1 - \phi(0) \text{ by optimality of } w^*_\phi \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} \left[\tilde{\phi}(\mathsf{Y} \langle w^*_{\text{unh}}, \Phi(\mathsf{X}) \rangle_{\mathcal{H}}) \right], \qquad (13)
\end{aligned}$$

[7]The result also holds if we add a regularised bias term. With an unregularised bias term, Bedo et al. [2006] showed that the limiting solution of a soft-margin SVM is distribution dependent.

where $\tilde{\phi}\colon v \mapsto \phi(v) - \phi(0) + v$, the remainder term from the linear approximation to ϕ. By Cauchy-Schwartz, we can restrict attention in Equation 13 to the behaviour of $\tilde{\phi}$ in the interval

$$I = [-||w^*_{\text{unh}}||_{\mathcal{H}} \cdot R, ||w^*_{\text{unh}}||_{\mathcal{H}} \cdot R],$$

where $R = \sup_{x \in \mathcal{X}} ||\Phi(x)||_{\mathcal{H}} < \infty$.

Now, by Taylor's remainder theorem,

$$(\forall v \in (-1, 1))\, \tilde{\phi}(v) \leq \frac{a}{2} v^2, \tag{14}$$

where $a = \max_{v \in [-1,1]} \phi''(v) < +\infty$. Therefore, if $r \leq \frac{1}{R}$, $I \subseteq [-1, 1]$ and so

$$\begin{aligned}
\mathbb{L}^D_{\text{unh}}(w^*_\phi) - \mathbb{L}^D_{\text{unh}}(w^*_{\text{unh}}) &\leq \frac{a}{2} \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} \left[\langle w^*_{\text{unh}}, \Phi(\mathsf{X}) \rangle^2_{\mathcal{H}} \right] \text{ by Equation 14} \\
&\leq \frac{a}{2} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim M} \left[||w^*_{\text{unh}}||^2_{\mathcal{H}} \cdot ||\Phi(\mathsf{X})||^2_{\mathcal{H}} \right] \text{ by Cauchy-Schwartz} \\
&\leq \frac{aR^2}{2} \cdot ||w^*_{\text{unh}}||^2_{\mathcal{H}}.
\end{aligned}$$

Now, the unhinged risk is

$$\mathbb{L}^D_{\text{unh}}(w) = 1 - \langle w, \mu \rangle_{\mathcal{H}}.$$

Thus,

$$-\langle w^*_\phi, \mu \rangle_{\mathcal{H}} + \langle w^*_{\text{unh}}, \mu \rangle_{\mathcal{H}} \leq \frac{aR^2}{2} \cdot ||w^*_{\text{unh}}||^2_{\mathcal{H}}.$$

Rearranging the above,

$$\begin{aligned}
\langle w^*_\phi, \mu \rangle_{\mathcal{H}} &\geq \langle w^*_{\text{unh}}, \mu \rangle_{\mathcal{H}} - \frac{aR^2}{2} \cdot ||w^*_{\text{unh}}||^2_{\mathcal{H}} \\
&= ||w^*_\phi||_{\mathcal{H}} \cdot ||\mu||_{\mathcal{H}} - \frac{aR^2}{2} \cdot ||w^*_\phi||^2_{\mathcal{H}} \text{ by definition of } w^*_{\text{unh}} \\
&= ||w^*_\phi||_{\mathcal{H}} \cdot ||\mu||_{\mathcal{H}} \cdot \left(1 - \frac{aR^2}{2||\mu||_{\mathcal{H}}} \cdot ||w^*_\phi||_{\mathcal{H}} \right) \\
&\geq ||w^*_\phi||_{\mathcal{H}} \cdot ||\mu||_{\mathcal{H}} \cdot \left(1 - \frac{aR^2}{2||\mu||_{\mathcal{H}}} \cdot r \right) \text{ since } ||w^*_\phi||_{\mathcal{H}} \leq r.
\end{aligned}$$

Thus, for $\epsilon = \frac{aR^2}{2||\mu||_{\mathcal{H}}}$,

$$\left\langle \frac{w^*_\phi}{||w^*_\phi||_{\mathcal{H}}}, \frac{\mu}{||\mu||_{\mathcal{H}}} \right\rangle_{\mathcal{H}} \geq 1 - \epsilon.$$

It follows that the two unit vectors can be made arbitrarily close to each other by decreasing r. Since this corresponds to increasing the strength of regularisation (by Lagrange duality), and since $\frac{\mu}{||\mu||_{\mathcal{H}}}$ corresponds to the normalised unhinged solution for any regularisation strength, we may conclude that

$$(\forall \epsilon > 0)\, (\exists \lambda_0 > 0)\, (\forall \lambda > \lambda_0) \left|\left| \frac{w^*_{\phi,\lambda}}{||w^*_{\phi,\lambda}||_{\mathcal{H}}} - \frac{w^*_{\text{unh},\lambda}}{||w^*_{\text{unh},\lambda}||_{\mathcal{H}}} \right|\right|^2_{\mathcal{H}} \leq \epsilon,$$

and the result follows. ∎

A.1 Additional helper lemmas

Lemma 10. *A pair of losses $(\ell, \tilde{\ell})$ are order equivalent if and only if*

$$(\exists \alpha \in R_+, \beta \in \mathbb{R})\, \tilde{\ell}(y, v) = \alpha \cdot \ell(y, v) + \beta.$$

Proof of Lemma 10. Recall that order equivalence of $(\ell, \tilde{\ell})$ is (Definition 4)

$$(\forall R, R') \underset{(\mathsf{Y},\mathsf{S})\sim R}{\mathbb{E}} [\ell(\mathsf{Y},\mathsf{S})] \leq \underset{(\mathsf{Y},\mathsf{S})\sim R'}{\mathbb{E}} [\ell(\mathsf{Y},\mathsf{S})] \iff \underset{(\mathsf{Y},\mathsf{S})\sim R}{\mathbb{E}} \left[\tilde{\ell}(\mathsf{Y},\mathsf{S})\right] \leq \underset{(\mathsf{Y},\mathsf{S})\sim R'}{\mathbb{E}} \left[\tilde{\ell}(\mathsf{Y},\mathsf{S})\right].$$

Now define the utility functions

$$U\colon (y, s) \mapsto -\ell(y, s)$$

and

$$V\colon (y, s) \mapsto -\tilde{\ell}(y, s).$$

Then, order equivalence is trivially equivalent to

$$(\forall R, R') \underset{(\mathsf{Y},\mathsf{S})\sim R}{\mathbb{E}} [U(\mathsf{Y},\mathsf{S})] \geq \underset{(\mathsf{Y},\mathsf{S})\sim R'}{\mathbb{E}} [U(\mathsf{Y},\mathsf{S})] \iff \underset{(\mathsf{Y},\mathsf{S})\sim R}{\mathbb{E}} [V(\mathsf{Y},\mathsf{S})] \geq \underset{(\mathsf{Y},\mathsf{S})\sim R'}{\mathbb{E}} [V(\mathsf{Y},\mathsf{S})].$$

That is, the utility functions U, V specify the same ordering over distributions. By DeGroot [1970, Section 7.9, Theorem 2], this is possible if and only if U, V are affinely related:

$$(\exists \alpha \in \mathbb{R}_+, \beta \in \mathbb{R})\, (\forall y, s)\, U(y, s) = \alpha \cdot V(y, s) + \beta.$$

The result follows by re-expressing this in terms of ℓ and $\tilde{\ell}$. $\square$

Additional Discussion for "Learning with Symmetric Label Noise: The Importance of Being Unhinged"

B Evidence that non-convex losses and linear scorers may not be SLN-robust

We now present preliminary evidence that for ℓ being the TangentBoost loss,

$$\ell(y, v) = (2 \tan^{-1}(yv) - 1)^2,$$

or the t-logistic regression loss for $t = 2$,

$$\ell(y, v) = \log(1 - yv + \sqrt{1 + v^2}),$$

$(\ell, \mathcal{F}_{\text{lin}})$ may not be SLN-robust. We do this by looking at the minimisers of these losses on the 2D example of Long and Servedio [2010]. Of course, as these losses are non-convex, exact minimisation of the risk is challenging. However, as the search space is $\mathbb{R}^2$, we construct a grid of resolution 0.025 over $[-10, 10]^2$. We then exhaustively compute the objective for all grid points, and seek the minimiser.

We apply this procedure to the Long and Servedio [2010] dataset with $\gamma = \frac{1}{60}$, and with a 30% noise rate. Figure 3 plots the results of the objective for the TangentBoost loss. We find that the minimiser is at $w^* = (0.2, 1.3)$. This results in a classifier with error rate of $\frac{1}{2}$ on D. Similarly, from Figure 4, we find that the minimiser is $w^* = (1.025, 5.1)$, which also results in a classifier with error rate of $\frac{1}{2}$.

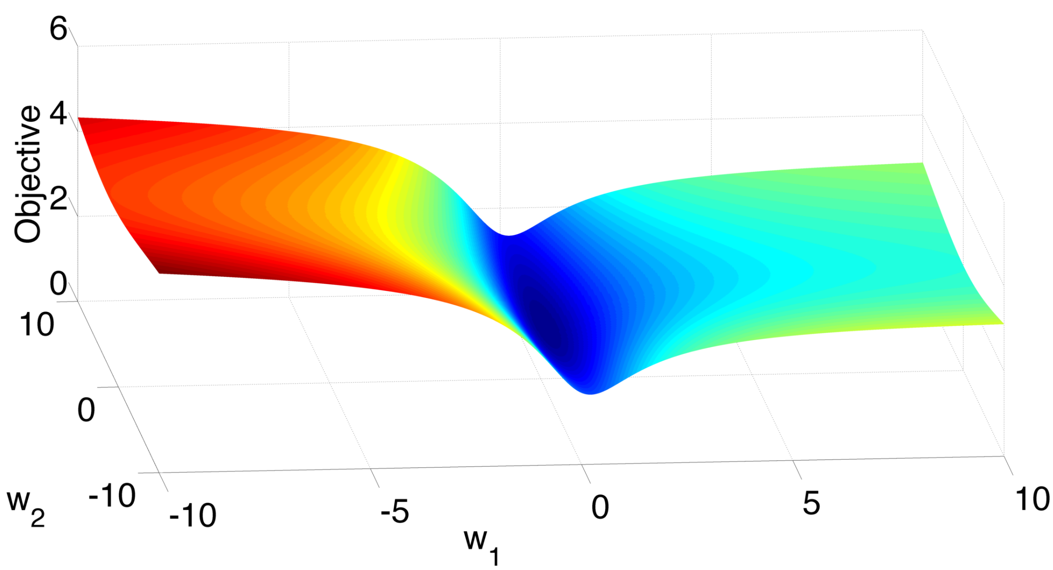

Figure 3: Risk values for various weight vectors $w = (w_1, w_2)$, TangentBoost, Long and Servedio [2010] dataset.

The shape of these plots suggests that the minimiser is indeed found in the interval $[-10, 10]^2$. To further verify this, we performed L-BFGS minimisation of these losses using 100 different random initialisations, uniformly from $[-100, 100]^2$. We find that in each trial, the TangentBoost solution converges to $w^* = (0.2122, 1.3031)$, while the t-logistic solution converges to $w^* = (1.0372, 5.0873)$, both of which result in accuracy of 50% on D.

B.1 Conjecture: (most) strictly proper composite losses are not SLN-robust

Recall that a loss ℓ is *strictly proper composite* [Reid and Williamson, 2010] if its (unique) Bayes-optimal scorer is some strictly monotone transformation ψ of the class-probability function:

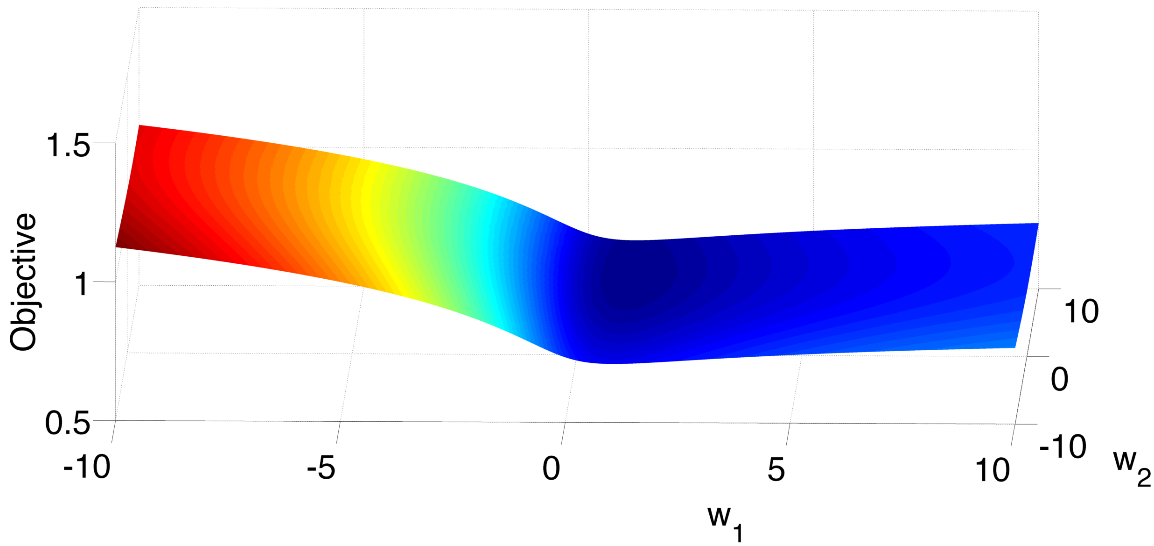

Figure 4: Risk values for various weight vectors $w = (w_1, w_2)$, t-logistic regression, Long and Servedio [2010] dataset.

$(\forall D)\, \mathcal{S}_\ell^{D,*} = \{\psi \circ \eta\}$. It is easy to check that both the TangentBoost and t-logistic losses are proper composite. We conjecture the above is a manifestation of the following phenomenon.

Conjecture 1. *Pick any strictly proper composite (but not necessarily convex) ℓ whose link function has range $\mathbb{R}$. Then, $(\ell, \mathcal{F}_{\text{lin}})$ is not SLN-robust.*

We believe the above is true for the following reason. Suppose D is some linearly separable distribution, with $\eta \colon x \mapsto [\![\langle w^*, x\rangle > 0]\!]$ for some w^*. Then, minimising ℓ with $\mathcal{F}_{\text{lin}}$ will be well-specified: the Bayes-optimal scorer is $\psi([\![\langle w^*, x\rangle > 0]\!])$. If the range of ψ is $\mathbb{R}$, then this is equivalent to $\infty \cdot (2[\![\langle w^*, x\rangle > 0]\!] - 1)$, which is in $\mathcal{F}_{\text{lin}}$ if we allow for the extended reals. The resulting classifier will thus have 100% accuracy. However, by injecting any non-zero label noise, minimising ℓ with $\mathcal{F}_{\text{lin}}$ will no longer be well-specified, as $\bar{\eta}$ takes on the values $\{1 - \rho, \rho\}$, which cannot be the sole set of output scores for any linear scorer if $|\mathcal{X}| > 3$. We believe it unlikely that every such misspecified solution have 100% accuracy on D. We further believe it likely that one can exhibit a scenario, possibly the same as the Long and Servedio [2010] example, where the resulting solution has accuracy 50%.

Two further comments are in order. First, if a loss is strictly proper composite, then it cannot satisfy Equation 5, and hence it cannot be strongly SLN-robust. (However, this does leave open the possibility that with $\mathcal{F}_{\text{lin}}$, the loss is SLN-robust.) Second, observe that the restriction that ψ have range $\mathbb{R}$ is necessary to rule out cases such as square loss, where the link function has range $[-1, 1]$.

B.2 In defence of non-convex losses: beyond SLN-robustness

The above illustrates the possible non SLN-robustness of two non-convex losses. However, there may be *other* notions under which these losses are robust. For example, Ding and Vishwanathan [2010] defines robustness to be a stability of the asymptotic maximum likelihood solution when adding a new labelled instance (chosen *arbitrarily* from $\mathcal{X} \times \{\pm 1\}$), based on a definition in O'Hagan [1979]. Intuitively, this captures robustness to outliers in the instance space, so that e.g. an adversarial mislabelling of an instance far from the true decision boundary does not adversely affect the learned model. Such a notion of robustness is clearly of practical interest, and future study of such alternate notions would be of value. (Appendix D.2 highlights some guarantees that are possible instance dependent noise, but still in the regime where instances are drawn from the true marginal M.)

C Preservation of mean maps

Pick any D, and $\rho \in [0, 1/2)$. Then,

$$(\forall x \in \mathcal{X})\, 2\bar{\eta}(x) - 1 = 2 \cdot ((1 - 2\rho) \cdot \eta(x) + \rho) - 1 \\ = (1 - 2\rho) \cdot (2\eta(x) - 1).$$

Thus, for any feature mapping $\Phi \colon \mathcal{X} \to \mathcal{H}$, the kernel mean map of the clean distribution is

$$\begin{aligned} \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} [\mathsf{Y} \cdot \Phi(\mathsf{X})] &= \mathop{\mathbb{E}}_{\mathsf{X} \sim M} [(2\eta(\mathsf{X}) - 1) \cdot \Phi(\mathsf{X})] \\ &= \frac{1}{(1 - 2\rho} \cdot \mathop{\mathbb{E}}_{\mathsf{X} \sim M} [(2\bar{\eta}(\mathsf{X}) - 1) \cdot \Phi(\mathsf{X})] \\ &= \frac{1}{(1 - 2\rho)} \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim \mathrm{SLN}(D,\rho)} [\mathsf{Y} \cdot \Phi(\mathsf{X})], \end{aligned}$$

which is a scaled version of the kernel mean map of the noisy distribution. That is, the kernel mean map is preserved under symmetric label noise. Instantiating the above with a specific $x \in \mathcal{X}$ gives Equation 9.

D Additional theoretical considerations

We discuss some further theoretical properties of the unhinged loss.

D.1 Generalisation bounds

Generalisation bounds are readily derived for the unhinged loss. For a training sample $\mathsf{S} \sim D^n$, define the ℓ-deviation of a scorer $s \colon \mathcal{X} \to \mathbb{R}$ to be the difference in its population and empirical ℓ-risk,

$$\mathrm{dev}_{\ell}^{D,\mathsf{S}}(s) = \mathbb{L}_{\ell}^{D}(s) - \mathbb{L}_{\ell}^{\mathsf{S}}(s).$$

This quantity is of interest because a standard result says that for the empirical risk minimiser s_n over some function class $\mathcal{F}$, $\mathrm{regret}_{\ell}^{D,\mathcal{F}}(s_n) \leq 2 \cdot \sup_{s \in \mathcal{F}} |\mathrm{dev}_{\ell}^{D,\mathsf{S}}(s)|$ [Boucheron et al., 2005, Equation 2]. For unhinged loss, we have the following Rademacher based bound.

Proposition 11. *Pick any D and $n \in \mathbb{N}_+$. Let $\mathsf{S} \sim D^n$ denote an empirical sample. For some $B \in \mathbb{R}_+$, let $s \in \mathcal{F}_B$. Then, with probability at least $1 - \delta$ over the choice of S, for $\ell = \ell^{\mathrm{unh}}$,*

$$\mathrm{dev}_{\ell}^{D,\mathsf{S}}(s) \leq 2 \cdot \mathcal{R}_n(\mathcal{F}_B, \mathsf{S}) + B \cdot \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

where $\mathcal{R}_n(\mathcal{F}_B, \mathsf{S})$ is the empirical Rademacher complexity of $\mathcal{F}_B$ on sample S.

Proof of Proposition 11. The standard Rademacher-complexity generalisation bound [Bartlett and Mendelson, 2002, Theorem 7], [Boucheron et al., 2005, Theorem 4.1] states that with probability at least $1 - \delta$ over the choice of S,

$$\mathrm{dev}_{\ell}^{D,\mathsf{S}}(s) \leq 2 \cdot ||(\ell)'||_{\infty} \cdot \mathcal{R}_n(\mathcal{F}_B, \mathsf{S}) + ||\ell||_{\infty} \cdot \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

For the unhinged loss, $||(\ell^{\mathrm{unh}})'||_{\infty} = 1$. Further, since we work over bounded scorers, $||\ell^{\mathrm{unh}}||_{\infty} = B$. The result follows. ☐

Proposition 11 holds equally when learning from a corrupted sample $\bar{\mathsf{S}} \sim \bar{D}^n$. Since $\mathrm{regret}_{\ell^{\mathrm{unh}}}^{D,\mathcal{F}}(s_n) = \frac{1}{1-2\rho} \cdot \mathrm{regret}_{\ell^{\mathrm{unh}}}^{\bar{D},\mathcal{F}}(s_n)$ by Proposition 7, by minimising the unhinged loss on the corrupted sample, we can bound the regret on the clean distribution.

D.2 The unhinged loss and instance dependent noise

The SLN model is a special case of the following (more realistic) *non uniform* noise model: given a notional clean distribution D, one observes samples from a corrupted distribution $\bar{D}$, where labels are flipped with some *instance dependent* probability $\rho(x) \in [0, 1/2)$. It is not hard to check that the unhinged solution will no longer be perfectly robust under this noise model. However, Servedio [1999] showed that if one further assumes M is uniform over the unit sphere, a nearest centroid classifier will be robust. More generally, one might hope that one can at least *bound* the degradation in performance under the noise model. Ghosh et al. [2015, Remark 1] showed that this is indeed possible for losses ℓ satisfying Equation 5: a simple argument reveals a bound on the ℓ-risk of the minimiser on the *corrupted* distribution in terms of the risk of the minimiser on the *clean* distribution, and a scaling factor that depends on the highest noise rate over all instances,

$$\mathbb{L}_\ell^D(\mathcal{S}_\ell^{\bar{D},\mathcal{F},*}) = \frac{\mathbb{L}_\ell^D(\mathcal{S}_\ell^{D,\mathcal{F},*})}{1 - 2\max_{x \in \mathcal{X}} \rho(x)}.$$

As the unhinged loss satisfies Equation 5, this immediately implies a guarantee for the more realistic instance dependent noise case. Further study of this subject will be the matter of future work.

D.3 On balanced error and area under the ROC immunity

Menon et al. [2015] recently showed that with rich function classes, class-probability estimation techniques will be robust to a general class of corruptions in terms of ranking performance (as measured by the area under the ROC curve or AUC) as well as classification performance (as measured by the balanced error or BER). In particular, they showed an affine relationship between the BER for an arbitrary classifier on the clean and corrupted distributions, and similarly for the AUC. This means that minimisers of the BER on the clean and corrupted distributions coincide for *any* function class $\mathcal{F}$; however, this does not mean that the minimisers for a *surrogate* to the BER coincide on the clean and corrupted distributions. In order to establish that surrogate minimisation is sensible for the BER, Menon et al. [2015] rely on the choice of $\mathcal{F} = \mathbb{R}^{\mathcal{X}}$. When $\mathcal{F} = \mathcal{F}_{\text{lin}}$, the example of Long and Servedio [2010] shows that a class-probability estimation technique such as logistic regression may perform poorly. By contrast, we emphasise that the unhinged loss' robustness holds as-is for any choice of $\mathcal{F}$.

E Additional relations to existing methods

We discuss some further connections of the unhinged loss to existing methods.

E.1 Unhinging the SVM

We can motivate the unhinged loss intuitively by studying the noise-corrected versions of the hinge loss, as per Equation 4. Figure 5 shows the noise corrected hinge loss for $\rho \in \{0, 0.2, 0.4\}$. We see that as the noise rate increases, the effect is to slightly *unhinge* the original loss, by removing its flat portion[8]. Thus, if we knew the noise rate ρ, we could use these *slightly unhinged* losses to learn.

Of course, in general we do not know the noise rate. Further, the slightly unhinged losses are non-convex. So, in order to be robust to an *arbitrary* noise rate ρ, we can *completely unhinge* the loss, yielding

$$\ell_1^{\text{unh}}(v) = 1 - v \text{ and } \ell_{-1}^{\text{unh}}(v) = 1 + v.$$

E.2 Relation to centroid classifiers

As established in §6.1, the optimal unhinged classifier (Equation 8) is equivalent to a centroid classifier, where one replaces the positive and negative classes by their centroids, and performs classification based on the distance of an instance to the two centroids. Such a classifier has been proposed

[8] Another interesting observation is that these noise-corrected losses are negatively unbounded – that is, minimising hinge loss on D is equivalent to minimising a negatively unbounded loss on $\bar{D}$. This is another justification for studying negatively unbounded losses.

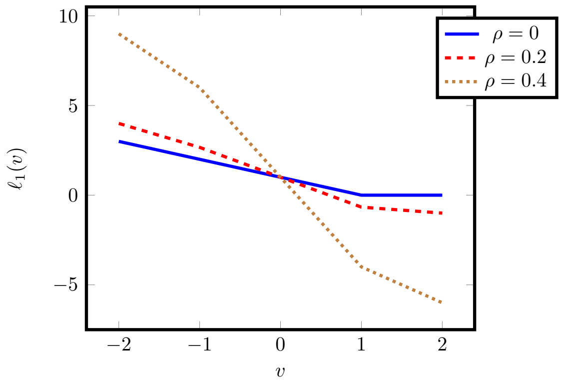

Figure 5: Noise-corrected versions of hinge loss, $\ell_1(v) = \max(0, 1 - v)$. Best viewed in colour.

as a prototypical example of a simple kernel-based classifier [Schölkopf and Smola, 2002, Section 1.2], [Shawe-Taylor and Cristianini, 2004, Section 5.1] Balcan et al. [2008, Definition 4] considers such classification rules using general similarity functions in place of kernels corresponding to an RKHS.

The optimal unhinged classifier is also closely related to the Rocchio classifier in information retrieval [Manning et al., 2008, pg. 181], and the nearest centroid classifier in computational genomics [Tibshirani et al., 2002]. The optimal kernelised scorer for these approaches is [Doloc-Mihu et al., 2003]

$$s^* \colon x \mapsto \left(\mathop{\mathbb{E}}_{\mathsf{X} \sim P} [k(\mathsf{X}, x)] - \mathop{\mathbb{E}}_{\mathsf{X} \sim Q} [k(\mathsf{X}, x)] \right),$$

i.e. it does not weight each of the kernel means.

E.3 Relation to kernel density estimation

When working with an RKHS with a translation invariant kernel[9], the optimal unhinged scorer (Equation 8) can be interpreted as follows: perform kernel density estimation on the positive and negative classes, and then classify instances according to Bayes' rule. For example, with a Gaussian RBF kernel, the classifier is equivalent to using a Gaussian kernel to compute density estimates of P, Q, and using these to classify. This is known as a kernel classification rule [Devroye et al., 1996, Chapter 10].

This perspective suggests that in computing $s^*_{\mathrm{unh},\lambda}$, we may also estimate the corrupted class-probability function. In particular, observe that if we compute $\frac{\pi}{1-\pi} \cdot \frac{\mathbb{E}_{\mathsf{X} \sim P}[k(\mathsf{X},x)]}{\mathbb{E}_{\mathsf{X} \sim Q}[k(\mathsf{X},x)]}$, similar to the Nadaraya-Watson estimator [Bishop, 2006, pg. 300], then this provides an estimate of $\frac{\eta(x)}{1-\eta(x)}$. Of course, such an approach will succumb to the curse of dimensionality[10].

An alternative is to use the Probing reduction [Langford and Zadrozny, 2005], by computing an ensemble of cost-sensitive classifiers at varying cost ratios. To this end, observe that the following weighted unhinged (or *whinge*) loss,

$$\ell_1^{\mathrm{whinge}}(v) = c_1 \cdot -v$$
$$\ell_{-1}^{\mathrm{whinge}}(v) = c_{-1} \cdot v$$

for some $c_{-1} \in [0,1]$ and $c_1 = 1 - c_{-1}$, will have a restricted Bayes-optimal scorer of $B \cdot \mathrm{sign}(\eta(x) - c_{-1})$ over $\mathcal{F}_B$. Further, it will result in an optimal scorer that simply weights each of the kernel

[9]For a general (not necessarily translation invariant) kernel, this is known as a potential function rule [Devroye et al., 1996, §10.3]. The use of "potential" here is distinct from that of a "convex potential".

[10]This refers to the rate of convergence of the estimate of η to the true η. By contrast, generalisation bounds establish that the rate of convergence of the estimate of the corresponding *classifier* to the Bayes-optimal classifier $\mathrm{sign}(2\eta(x) - 1)$ is independent of the dimension of the feature space.

means,

$$s^*_{\text{whinge},\lambda} \colon x \mapsto \frac{1}{\lambda} \cdot \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})\sim D} \left[c_{\mathsf{Y}} \cdot \mathsf{Y} \cdot k(\mathsf{X}, x) \right],$$

making it trivial to compute as c is varied.

E.4 Relation to the MMD witness

The optimal weight vector for unhinged loss (Equation 7) can be expressed as

$$w^*_{\text{unh},\lambda} = \frac{1}{\lambda} \cdot (\pi \cdot \mu_P - (1 - \pi) \cdot \mu_Q),$$

where μ_P and μ_Q are the *kernel mean maps* with respect to $\mathcal{H}$ of the positive and negative class-conditionals distributions,

$$\mu_P = \mathop{\mathbb{E}}_{\mathsf{X}\sim P} [\Phi(\mathsf{X})]$$
$$\mu_Q = \mathop{\mathbb{E}}_{\mathsf{X}\sim Q} [\Phi(\mathsf{X})].$$

When $\pi = \frac{1}{2}$, $||w_1^*||_{\mathcal{H}}$ is precisely the *maximum mean discrepancy* (*MMD*) [Gretton et al., 2012] between P and Q, using all functions in the unit ball of $\mathcal{H}$. The mapping $x \mapsto \langle w_1^*, x\rangle_{\mathcal{H}}$ itself is referred to as the *witness function* [Gretton et al., 2012, §2.3]. While the motivation of MMD is to perform hypothesis testing so as to distinguish between two distributions P, Q, rather than constructing a suitable scorer, the fact that it arises from the optimal scorer for the unhinged loss has been previously noted [Sriperumbudur et al., 2009, Theorem 1].

F Example of poor classification with square loss

We illustrate that square loss with a linear function class may perform poorly even when the underlying distribution is linearly separable. We consider the dataset of Long and Servedio [2010], with *no* label noise. That is, we have $\mathcal{X} = \{(1, 0), (\gamma, 5\gamma), (\gamma, -\gamma), (\gamma, -\gamma)\} \subset \mathbb{R}^2$, and $\eta \colon x \mapsto 1$. Let $X \in \mathbb{R}^{4\times 2}$ be the feature matrix of the four data points. Then, the optimal weight vector for square loss is

$$w^* = (X^T X)^{-1} X^T \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{8\gamma+3}{8\gamma^2+3} \\ -\frac{\gamma+1}{3\gamma\cdot(8\gamma^2+3)} \end{bmatrix}.$$

It is easy to check that the predicted scores are then

$$s^* = \begin{bmatrix} \frac{8\gamma+3}{8\gamma^2+3} \\ \frac{\gamma\cdot(8\gamma+3)}{8\gamma^2+3-\frac{5\cdot(\gamma-1)\gamma}{24\gamma^3+9\gamma}} \\ \frac{(\gamma-1)\cdot\gamma}{24\gamma^3+9\gamma+\frac{\gamma\cdot(8\gamma+3)}{8\gamma^2+3}} \\ \frac{(\gamma-1)\cdot\gamma}{24\gamma^3+9\gamma+\frac{\gamma\cdot(8\gamma+3)}{8\gamma^2+3}} \end{bmatrix}.$$

But for $\gamma < \frac{1}{12}$, this means that the predicted scores for the last two examples are negative. That is, the resulting classifier will have 50% accuracy. (This does not contradict the robustness of square loss, as robustness simply requires that performance is the *same* with and without noise.)

It is initially surprising that square loss fails in this example, as we are employing a linear function class, and the true η is expressible as a linear function. However, recall that the Bayes-optimal scorer for square loss is

$$\mathcal{S}^{D,*}_{\ell} = \{s \colon x \mapsto 2\eta(x) - 1\}.$$

In this case, the Bayes-optimal scorer is

$$s^* \colon x \mapsto 2[\![x_1 > 0]\!] - 1.$$

The application of a threshold means the that scorer is *not expressible as a linear model*. Therefore, the combination of loss and function class is in fact *not* well-specified for the problem. By contrast, consider the use of the squared hinge loss, $\ell(y, v) = \max(0, 1 - yv)^2$. This loss induces a *set* of Bayes-optimal scorers, which are:

$$\mathcal{S}_\ell^{D,*} = \left\{ s \mid (\forall x \in \mathcal{X}) \begin{cases} \eta(x) = 1 & \implies s(x) \in [1, \infty) \\ \eta(x) \in (0, 1) & \implies s(x) = 2\eta(x) - 1 \\ \eta(x) = 0 & \implies s(x) \in (-\infty, 1]. \end{cases} \right\}$$

Crucially, we *can* find a linear scorer that is in this set: for, say, $v = (\frac{1}{\gamma}, 0)$, we clearly have $\langle v, x \rangle \geq 1$ for every $x \in \mathcal{X}$, and so this is a Bayes-optimal scorer. Thus, minimising the square hinge loss on this distribution will indeed find a classifier with 100% accuracy.

G Example of poor classification with unhinged loss

We illustrate that the unhinged loss with a linear function class may perform poorly even when the underlying distribution is linearly separable. (For another example where instances are on the unit ball, see Balcan et al. [2008, Figure 1].) Consider a distribution $D_{M,\eta}$ uniformly concentrated on $\mathcal{X} = \{x_1, x_2, x_3\}$ with $x_1 = (1, 2), x_2 = (1, -4), x_3 = (-1, 1)$, with $\eta(x_1) = \eta(x_2) = 1$ and $\eta(x_3) = 0$, i.e. the first two instances are positive, and the third instance negative. Then it is evident that the optimal unhinged hyperplane, with regularisation strength 1, is $w^* = (1, -1)$. This will misclassify the first instance as being negative. Figure 6 illustrates.

It is easy to check that for this particular distribution, the optimal weight for square loss is $w^* = (1, 0)$. This results in perfect classification. Thus, we have a reversal of the scenario of the previous section – here, square loss classifies perfectly, while the unhinged loss classifies no better than random guessing.

It may appear that the above contradicts the classification-calibration of the unhinged loss: there certainly is a linear scorer that is Bayes-optimal over $\mathcal{F}_B$, namely, $w^* = (B, 0)$. The subtlety is that in this case, minimisation over the unit ball $||w||_2 \leq 1$ (as implied by ℓ_2 regularisation) is unable to restrict attention to the desired scorer.

There are two ways to rectify examples such as the above. First, as in general, we can employ a suitably rich kernel, e.g. a Gaussian RBF kernel. It is not hard to verify that on this dataset, such a kernel will find a perfect classifier. Second, we can look to explicitly enforce that minimisation is over all w satisfying $|\langle w, x_n \rangle| \leq 1$. This will result in a linear program (LP) that may be solved easily, but does not admit a closed form solution as in the case of minimising over the unit ball. It may be checked that the resulting LP will recover the optimal weight $w^* = (1, 0)$. While this approach is suitable for this particular example, issues arise when dealing with infinite dimensional feature mappings (as we lose the existence of a representer theorem without regularisation based on the norm in the Hilbert space [Yu et al., 2013]).

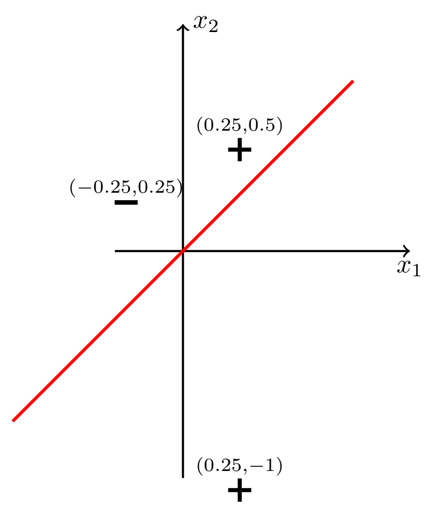

Figure 6: Example of linearly separable distribution where, when learning with the unhinged loss and a linear function class, the resulting hyperplane (in red) misclassifies one of the instances.

Additional Experiments for "Learning with Symmetric Label Noise: The Importance of Being Unhinged"

H Additional experimental results

Table 4 reports the 0-1 error for a range of losses on the Long and Servedio [2010] dataset. TanBoost refers to the loss of Masnadi-Shirazi et al. [2010]. As before, we find the unhinged loss to generally find a good classifier. Observe that the relatively poor performance of the square and TanBoost loss can be attributed to the findings of Appendix B, F.

We next report the 0-1 error and one minus the AUC for a range of datasets. We begin with a dataset of Mease and Wyner [2008], where $\mathcal{X} = [0,1]^{20}$, and M is the uniform distribution. Further, we have $\eta \colon x \mapsto [\![\langle w^*, x \rangle > 2.5]\!]$ for $w^* = [\mathbf{1}_5 \quad \mathbf{0}_{15}]$, i.e. there is a sparse separating hyperplane. Table 5 reports the results on this dataset injected with various levels of symmetric noise. On this dataset, the t-logistic loss generally performs the best.

Finally, we report the 0-1 error and one minus the AUC on some UCI datasets in Tables 6 – 7. Table 3 summarises statistics of the UCI data. Several datasets are imbalanced, meaning that 0-1 error is not the ideal measure of performance (as it can be made small with a trivial majority classifier). The AUC is thus arguably a better indication of performance for these datasets. We generally find that at high noise rates (40%), the AUC of the unhinged loss is superior to that of other losses.

Dataset	N	D	$\mathbb{P}(\mathsf{Y} = 1)$
Iris	150	4	0.3333
Ionosphere	351	34	0.3590
Housing	506	13	0.0692
Car	1,728	8	0.0376
USPS 0v7	2,200	256	0.5000
Splice	3,190	61	0.2404
Spambase	4,601	57	0.3940

Table 3: Summary of UCI datasets. Here, N denotes the total number of samples, and D the dimensionality of the feature space.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.25 ± 0.00	0.00 ± 0.00	0.25 ± 0.00	0.00 ± 0.00
$\rho = 0.1$	0.15 ± 0.27	0.24 ± 0.05	0.25 ± 0.00	0.00 ± 0.00	0.25 ± 0.00	0.00 ± 0.00
$\rho = 0.2$	0.21 ± 0.30	0.25 ± 0.00	0.25 ± 0.00	0.00 ± 0.00	0.25 ± 0.00	0.00 ± 0.00
$\rho = 0.3$	0.38 ± 0.37	0.25 ± 0.03	0.25 ± 0.02	0.22 ± 0.08	0.25 ± 0.03	0.00 ± 0.00
$\rho = 0.4$	0.42 ± 0.36	0.22 ± 0.08	0.22 ± 0.08	0.22 ± 0.08	0.22 ± 0.08	0.00 ± 0.00
$\rho = 0.49$	0.46 ± 0.38	0.39 ± 0.23	0.39 ± 0.23	0.39 ± 0.23	0.39 ± 0.23	0.34 ± 0.48

Table 4: Results on `LS10` dataset. Reported is the mean and standard deviation of the 0-1 error over 125 trials. Grayed cells denote the best performer at that noise rate.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.02 ± 0.00	0.01 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.05 ± 0.00
$\rho = 0.1$	0.13 ± 0.01	0.05 ± 0.01	0.06 ± 0.01	0.03 ± 0.01	0.05 ± 0.01	0.06 ± 0.01
$\rho = 0.2$	0.14 ± 0.01	0.09 ± 0.02	0.09 ± 0.02	0.06 ± 0.02	0.08 ± 0.02	0.08 ± 0.02
$\rho = 0.3$	0.15 ± 0.01	0.13 ± 0.03	0.13 ± 0.03	0.12 ± 0.03	0.12 ± 0.03	0.12 ± 0.02
$\rho = 0.4$	0.17 ± 0.05	0.24 ± 0.08	0.24 ± 0.08	0.23 ± 0.07	0.23 ± 0.08	0.23 ± 0.08
$\rho = 0.49$	0.47 ± 0.24	0.46 ± 0.11	0.47 ± 0.11	0.48 ± 0.10	0.47 ± 0.12	0.48 ± 0.12

(a) 0-1 Error.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
$\rho = 0.1$	0.25 ± 0.10	0.02 ± 0.01	0.02 ± 0.01	0.00 ± 0.00	0.02 ± 0.01	0.02 ± 0.01
$\rho = 0.2$	0.34 ± 0.10	0.05 ± 0.02	0.05 ± 0.02	0.02 ± 0.01	0.04 ± 0.02	0.05 ± 0.02
$\rho = 0.3$	0.41 ± 0.11	0.11 ± 0.04	0.11 ± 0.04	0.09 ± 0.04	0.11 ± 0.04	0.10 ± 0.04
$\rho = 0.4$	0.44 ± 0.12	0.24 ± 0.08	0.24 ± 0.08	0.24 ± 0.08	0.24 ± 0.08	0.23 ± 0.08
$\rho = 0.49$	0.50 ± 0.13	0.47 ± 0.11	0.47 ± 0.11	0.47 ± 0.11	0.47 ± 0.11	0.46 ± 0.11

(b) 1 - AUC.

Table 5: Results on `mease` dataset. Reported is the mean and standard deviation of performance over 125 trials. Grayed cells denote the best performer at that noise rate.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.1$	0.01 ± 0.03	0.01 ± 0.01	0.01 ± 0.02	0.01 ± 0.03	0.01 ± 0.02	0.00 ± 0.00
$\rho = 0.2$	0.06 ± 0.12	0.02 ± 0.05	0.03 ± 0.04	0.04 ± 0.05	0.03 ± 0.05	0.00 ± 0.01
$\rho = 0.3$	0.17 ± 0.20	0.09 ± 0.10	0.08 ± 0.09	0.09 ± 0.11	0.09 ± 0.10	0.02 ± 0.07
$\rho = 0.4$	0.35 ± 0.24	0.24 ± 0.17	0.24 ± 0.17	0.24 ± 0.16	0.24 ± 0.17	0.13 ± 0.22
$\rho = 0.49$	0.60 ± 0.20	0.49 ± 0.20	0.49 ± 0.19	0.49 ± 0.20	0.49 ± 0.19	0.45 ± 0.33

(a) 0-1 Error.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.1$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.2$	0.03 ± 0.11	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.00
$\rho = 0.3$	0.14 ± 0.26	0.02 ± 0.06	0.02 ± 0.05	0.02 ± 0.06	0.02 ± 0.05	0.01 ± 0.06
$\rho = 0.4$	0.36 ± 0.38	0.13 ± 0.18	0.13 ± 0.18	0.14 ± 0.18	0.13 ± 0.18	0.09 ± 0.27
$\rho = 0.49$	0.72 ± 0.34	0.47 ± 0.31	0.48 ± 0.30	0.48 ± 0.30	0.48 ± 0.30	0.45 ± 0.48

(b) 1 - AUC.

Table 6: Results on `iris` dataset. Reported is the mean and standard deviation of performance over 125 trials. Grayed cells denote the best performer at that noise rate.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.11 ± 0.00	0.13 ± 0.00	0.17 ± 0.00	0.24 ± 0.00	0.17 ± 0.00	0.20 ± 0.00
$\rho = 0.1$	0.17 ± 0.04	0.18 ± 0.04	0.16 ± 0.03	0.19 ± 0.05	0.17 ± 0.04	0.19 ± 0.02
$\rho = 0.2$	0.20 ± 0.05	0.19 ± 0.05	0.18 ± 0.04	0.21 ± 0.06	0.18 ± 0.04	0.19 ± 0.02
$\rho = 0.3$	0.23 ± 0.06	0.22 ± 0.05	0.22 ± 0.05	0.24 ± 0.06	0.22 ± 0.05	0.21 ± 0.03
$\rho = 0.4$	0.31 ± 0.11	0.31 ± 0.10	0.29 ± 0.09	0.32 ± 0.09	0.30 ± 0.10	0.27 ± 0.12
$\rho = 0.49$	0.48 ± 0.16	0.47 ± 0.16	0.47 ± 0.16	0.47 ± 0.14	0.45 ± 0.15	0.46 ± 0.22

(a) 0-1 Error.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.12 ± 0.00	0.13 ± 0.00	0.07 ± 0.00	0.20 ± 0.00	0.07 ± 0.00	0.21 ± 0.00
$\rho = 0.1$	0.18 ± 0.07	0.18 ± 0.07	0.12 ± 0.04	0.22 ± 0.07	0.13 ± 0.05	0.21 ± 0.00
$\rho = 0.2$	0.23 ± 0.09	0.22 ± 0.09	0.18 ± 0.07	0.25 ± 0.08	0.19 ± 0.08	0.21 ± 0.01
$\rho = 0.3$	0.31 ± 0.11	0.29 ± 0.09	0.26 ± 0.09	0.30 ± 0.09	0.27 ± 0.09	0.21 ± 0.01
$\rho = 0.4$	0.40 ± 0.11	0.40 ± 0.10	0.38 ± 0.10	0.40 ± 0.10	0.38 ± 0.10	0.25 ± 0.12
$\rho = 0.49$	0.49 ± 0.12	0.50 ± 0.10	0.50 ± 0.10	0.50 ± 0.10	0.50 ± 0.10	0.46 ± 0.25

(b) 1 - AUC.

Table 7: Results on `ionosphere` dataset. Reported is the mean and standard deviation of performance over 125 trials. Grayed cells denote the best performer at that noise rate.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.05 ± 0.00	0.05 ± 0.00	0.07 ± 0.00	0.05 ± 0.00	0.07 ± 0.00	0.05 ± 0.00
$\rho = 0.1$	0.06 ± 0.01	0.06 ± 0.02	0.07 ± 0.02	0.07 ± 0.02	0.07 ± 0.02	0.05 ± 0.00
$\rho = 0.2$	0.06 ± 0.01	0.07 ± 0.03	0.07 ± 0.02	0.08 ± 0.03	0.07 ± 0.02	0.05 ± 0.00
$\rho = 0.3$	0.08 ± 0.04	0.10 ± 0.06	0.11 ± 0.06	0.11 ± 0.05	0.11 ± 0.06	0.05 ± 0.01
$\rho = 0.4$	0.14 ± 0.10	0.21 ± 0.12	0.22 ± 0.12	0.24 ± 0.13	0.22 ± 0.13	0.09 ± 0.10
$\rho = 0.49$	0.45 ± 0.26	0.49 ± 0.16	0.50 ± 0.16	0.49 ± 0.16	0.51 ± 0.17	0.46 ± 0.30

(a) 0-1 Error.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.25 ± 0.00	0.15 ± 0.00	0.17 ± 0.00	0.25 ± 0.00	0.17 ± 0.00	0.69 ± 0.00
$\rho = 0.1$	0.38 ± 0.12	0.27 ± 0.07	0.27 ± 0.07	0.30 ± 0.09	0.27 ± 0.07	0.69 ± 0.00
$\rho = 0.2$	0.41 ± 0.13	0.35 ± 0.10	0.35 ± 0.10	0.35 ± 0.10	0.35 ± 0.10	0.68 ± 0.00
$\rho = 0.3$	0.44 ± 0.12	0.40 ± 0.11	0.40 ± 0.11	0.40 ± 0.11	0.40 ± 0.11	0.69 ± 0.01
$\rho = 0.4$	0.43 ± 0.12	0.45 ± 0.12	0.45 ± 0.12	0.45 ± 0.12	0.45 ± 0.12	0.68 ± 0.02
$\rho = 0.49$	0.45 ± 0.13	0.49 ± 0.13	0.49 ± 0.13	0.49 ± 0.13	0.49 ± 0.13	0.57 ± 0.16

(b) 1 - AUC.

Table 8: Results on `housing` dataset. Reported is the mean and standard deviation of performance over 125 trials. Grayed cells denote the best performer at that noise rate.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.01 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.02 ± 0.00	0.03 ± 0.00
$\rho = 0.1$	0.05 ± 0.00	0.04 ± 0.01	0.04 ± 0.01	0.02 ± 0.01	0.04 ± 0.01	0.04 ± 0.01
$\rho = 0.2$	0.05 ± 0.00	0.05 ± 0.01	0.05 ± 0.01	0.04 ± 0.01	0.05 ± 0.01	0.05 ± 0.01
$\rho = 0.3$	0.05 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.02	0.06 ± 0.01	0.06 ± 0.01
$\rho = 0.4$	0.06 ± 0.02	0.11 ± 0.06	0.11 ± 0.06	0.11 ± 0.06	0.11 ± 0.06	0.10 ± 0.05
$\rho = 0.49$	0.33 ± 0.27	0.46 ± 0.16	0.46 ± 0.16	0.47 ± 0.16	0.47 ± 0.16	0.46 ± 0.16

(a) 0-1 Error.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.02 ± 0.00
$\rho = 0.1$	0.34 ± 0.18	0.03 ± 0.02	0.03 ± 0.02	0.00 ± 0.00	0.03 ± 0.02	0.04 ± 0.02
$\rho = 0.2$	0.40 ± 0.17	0.07 ± 0.05	0.08 ± 0.05	0.04 ± 0.04	0.07 ± 0.05	0.08 ± 0.05
$\rho = 0.3$	0.43 ± 0.17	0.17 ± 0.10	0.17 ± 0.10	0.14 ± 0.10	0.16 ± 0.10	0.16 ± 0.10
$\rho = 0.4$	0.44 ± 0.18	0.30 ± 0.16	0.30 ± 0.16	0.30 ± 0.16	0.30 ± 0.16	0.30 ± 0.16
$\rho = 0.49$	0.51 ± 0.19	0.46 ± 0.17	0.46 ± 0.17	0.46 ± 0.17	0.46 ± 0.17	0.46 ± 0.18

(b) 1 - AUC.

Table 9: Results on `car` dataset. Reported is the mean and standard deviation of performance over 125 trials. Grayed cells denote the best performer at that noise rate.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.1$	0.10 ± 0.08	0.05 ± 0.01	0.01 ± 0.01	0.11 ± 0.02	0.02 ± 0.01	0.00 ± 0.00
$\rho = 0.2$	0.19 ± 0.11	0.09 ± 0.02	0.05 ± 0.02	0.15 ± 0.02	0.06 ± 0.02	0.00 ± 0.00
$\rho = 0.3$	0.31 ± 0.13	0.17 ± 0.03	0.14 ± 0.02	0.22 ± 0.03	0.16 ± 0.03	0.01 ± 0.00
$\rho = 0.4$	0.39 ± 0.13	0.31 ± 0.04	0.30 ± 0.04	0.33 ± 0.04	0.31 ± 0.04	0.02 ± 0.02
$\rho = 0.49$	0.50 ± 0.16	0.48 ± 0.04	0.47 ± 0.04	0.48 ± 0.04	0.48 ± 0.04	0.34 ± 0.21

(a) 0-1 Error.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.1$	0.05 ± 0.06	0.01 ± 0.00	0.00 ± 0.00	0.05 ± 0.01	0.00 ± 0.00	0.00 ± 0.00
$\rho = 0.2$	0.12 ± 0.11	0.03 ± 0.01	0.01 ± 0.00	0.07 ± 0.01	0.02 ± 0.01	0.00 ± 0.00
$\rho = 0.3$	0.26 ± 0.18	0.10 ± 0.02	0.07 ± 0.02	0.14 ± 0.03	0.08 ± 0.02	0.00 ± 0.00
$\rho = 0.4$	0.37 ± 0.19	0.25 ± 0.04	0.24 ± 0.04	0.27 ± 0.04	0.24 ± 0.04	0.00 ± 0.00
$\rho = 0.49$	0.51 ± 0.23	0.47 ± 0.05	0.46 ± 0.05	0.47 ± 0.05	0.47 ± 0.05	0.25 ± 0.29

(b) 1 - AUC.

Table 10: Results on `usps_0_vs_7` dataset. Reported is the mean and standard deviation of performance over 125 trials. Grayed cells denote the best performer at that noise rate.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.05 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.19 ± 0.00
$\rho = 0.1$	0.15 ± 0.03	0.05 ± 0.01	0.04 ± 0.01	0.24 ± 0.00	0.04 ± 0.01	0.19 ± 0.01
$\rho = 0.2$	0.21 ± 0.03	0.08 ± 0.01	0.07 ± 0.01	0.24 ± 0.00	0.07 ± 0.01	0.19 ± 0.01
$\rho = 0.3$	0.25 ± 0.03	0.14 ± 0.02	0.14 ± 0.02	0.24 ± 0.00	0.14 ± 0.02	0.19 ± 0.03
$\rho = 0.4$	0.31 ± 0.05	0.28 ± 0.05	0.28 ± 0.04	0.24 ± 0.00	0.28 ± 0.04	0.22 ± 0.05
$\rho = 0.49$	0.48 ± 0.09	0.47 ± 0.06	0.48 ± 0.05	0.40 ± 0.24	0.48 ± 0.05	0.45 ± 0.08

(a) 0-1 Error.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.09 ± 0.00
$\rho = 0.1$	0.10 ± 0.03	0.01 ± 0.00	0.01 ± 0.00	0.03 ± 0.01	0.01 ± 0.00	0.09 ± 0.01
$\rho = 0.2$	0.20 ± 0.05	0.03 ± 0.01	0.02 ± 0.01	0.04 ± 0.01	0.02 ± 0.01	0.10 ± 0.02
$\rho = 0.3$	0.30 ± 0.06	0.08 ± 0.02	0.08 ± 0.02	0.09 ± 0.02	0.07 ± 0.02	0.11 ± 0.03
$\rho = 0.4$	0.40 ± 0.07	0.22 ± 0.04	0.22 ± 0.04	0.23 ± 0.04	0.22 ± 0.04	0.16 ± 0.07
$\rho = 0.49$	0.49 ± 0.08	0.46 ± 0.05	0.46 ± 0.05	0.46 ± 0.05	0.45 ± 0.05	0.42 ± 0.15

(b) 1 - AUC.

Table 11: Results on `splice` dataset. Reported is the mean and standard deviation of performance over 125 trials. Grayed cells denote the best performer at that noise rate.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.16 ± 0.01	0.08 ± 0.00	0.10 ± 0.00	0.24 ± 0.00	0.09 ± 0.00	0.15 ± 0.00
$\rho = 0.1$	0.14 ± 0.03	0.10 ± 0.02	0.10 ± 0.01	0.13 ± 0.06	0.10 ± 0.01	0.14 ± 0.01
$\rho = 0.2$	0.17 ± 0.03	0.11 ± 0.02	0.11 ± 0.01	0.13 ± 0.05	0.11 ± 0.01	0.14 ± 0.01
$\rho = 0.3$	0.23 ± 0.05	0.13 ± 0.02	0.12 ± 0.01	0.14 ± 0.04	0.13 ± 0.02	0.15 ± 0.01
$\rho = 0.4$	0.33 ± 0.07	0.20 ± 0.04	0.19 ± 0.03	0.21 ± 0.04	0.19 ± 0.03	0.17 ± 0.03
$\rho = 0.49$	0.49 ± 0.10	0.45 ± 0.07	0.44 ± 0.07	0.45 ± 0.07	0.45 ± 0.07	0.43 ± 0.12

(a) 0-1 Error.

	Hinge	**Logistic**	**Square**	**t-logistic**	**TanBoost**	**Unhinged**
$\rho = 0$	0.03 ± 0.00	0.02 ± 0.00	0.05 ± 0.00	0.02 ± 0.00	0.04 ± 0.00	0.07 ± 0.00
$\rho = 0.1$	0.06 ± 0.01	0.04 ± 0.00	0.05 ± 0.00	0.03 ± 0.00	0.04 ± 0.00	0.07 ± 0.00
$\rho = 0.2$	0.10 ± 0.03	0.05 ± 0.00	0.05 ± 0.00	0.04 ± 0.00	0.05 ± 0.00	0.07 ± 0.00
$\rho = 0.3$	0.17 ± 0.06	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	0.07 ± 0.01
$\rho = 0.4$	0.32 ± 0.12	0.12 ± 0.02	0.12 ± 0.02	0.12 ± 0.02	0.12 ± 0.02	0.09 ± 0.02
$\rho = 0.49$	0.49 ± 0.14	0.43 ± 0.08	0.43 ± 0.08	0.43 ± 0.07	0.43 ± 0.08	0.39 ± 0.19

(b) 1 - AUC.

Table 12: Results on `spambase` dataset. Reported is the mean and standard deviation of performance over 125 trials. Grayed cells denote the best performer at that noise rate.

References

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, August 2008. ISSN 0885-6125.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. ISSN 1532-4435.

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

Justin Bedo, Conrad Sanderson, and Adam Kowalczyk. An efficient alternative to SVM based recursive feature elimination with applications in natural language processing and bioinformatics. In Abdul Sattar and Byeong-ho Kang, editors, *AI 2006: Advances in Artificial Intelligence*, volume 4304 of *Lecture Notes in Computer Science*, pages 170–180. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-49787-5.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Conference on Computational Learning Theory (COLT)*, pages 92–100, 1998.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Morris H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, 1970.

Vasil Denchev, Nan Ding, Hartmut Neven, and S.V.N. Vishwanathan. Robust classification with adiabatic quantum optimization. In *International Conference on Machine Learning (ICML)*, pages 863–870, 2012.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Nan Ding and S.V.N. Vishwanathan. t-logistic regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 514–522. Curran Associates, Inc., 2010.

A. Doloc-Mihu, V. V. Raghavan, and P. Bollmann-Sdorra. Color retrieval in vector space model. In *ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, 2003.

Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, 1967.

Aritra Ghosh, Naresh Manwani, and P. S. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93 – 107, 2015.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, March 2012. ISSN 1532-4435.

Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, December 2004. ISSN 1532-4435.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 5(6):392–401, November 1998.

John Langford and Bianca Zadrozny. Estimating class membership probabilities using classifier learners, 2005.

Philip M. Long and Rocco A. Servedio. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304, 2010. ISSN 0885-6125.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

Naresh Manwani and P. S. Sastry. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43(3):1146–1151, June 2013.

Hamed Masnadi-Shirazi, Vijay Mahadevan, and Nuno Vasconcelos. On the design of robust classifiers for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

David Mease and Abraham Wyner. Evidence contrary to the statistical view of boosting. *Journal of Machine Learning Research*, 9:131–156, June 2008. ISSN 1532-4435.

Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning (ICML)*, pages 125–134, 2015.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep D. Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1196–1204, 2013.

Anthony O'Hagan. On outlier rejection phenomena in bayes inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(3):pp. 358–367, 1979. ISSN 00359246.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1177–1184, 2007.

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

Mark D Reid and Robert C Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, Mar 2011.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*, volume 129. MIT Press, 2002.

Rocco A. Servedio. On PAC learning using Winnow, Perceptron, and a Perceptron-like algorithm. In *Conference on Computational Learning Theory (COLT)*, 1999.

Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *International Conference on Machine Learning (ICML)*, pages 807–814, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge Uni. Press, 2004.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, 2007.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert R. G. Lanckriet, and Bernhard Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

Guillaume Stempfel and Liva Ralaivola. Learning SVMs from sloppily labeled data. In *Artificial Neural Networks (ICANN)*, volume 5768, pages 884–893. Springer Berlin Heidelberg, 2009.

Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10): 6567–6572, 2002.

Oscar Wilde. The Importance of Being Earnest, 1895.

Yaoliang Yu, Hao Cheng, Dale Schuurmans, and Csaba Szepesvári. Characterizing the representer theorem. In *International Conference on Machine Learning (ICML)*, volume 28 of *JMLR Proceedings*, pages 570–578. JMLR.org, 2013.