Pessimistic **Limits and Possibilities** of Margin-based Losses in Semi-supervised Learning

Jesse H. Krijthe Radboud University Nijmegen
Marco Loog Delft University of Technology & University of Copenhagen

The Problem of Safe Semi-supervised Learning

Consider a classification problem where we have both labeled and unlabeled data available. Semi-supervised learning (SSL) addresses the question how to use the unlabeled data to improve supervised methods that only use the labeled data. SSL techniques have shown promising results for some problem settings. In other settings, however, using unlabeled data has been shown to lead to a decrease in performance when compared to the supervised solution (Elworthy, 1994; Cozman & Cohen, 2006). For semi-supervised classifiers to be used safely in practice, we may at least want some guarantee that they do not reduce performance compared to their supervised alternatives. This work explores whether and, if so, how we can give such guarantees.

We show that for linear classifiers defined by convex margin-based surrogate losses that are decreasing, it is impossible to construct *any* semi-supervised approach that is able to guarantee an improvement over the supervised classifier measured by this surrogate loss on the labeled and unlabeled data. For convex margin-based loss functions that also increase, we demonstrate safe improvements *are* possible.

Margin-Based Losses

We consider binary linear classifiers in the empirical risk minimization framework and convex margin-based surrogate loss functions, which are loss functions of the form $\phi(y\mathbf{x}^\top\mathbf{w})$, with $y \in \{-1,+1\}$, $\mathbf{x}, \mathbf{w} \in \mathbb{R}^d$, representing the label, feature vector and classifier, respectively. Many well-known classifiers can be described in this way, for instance, support vector machines and logistic regression. In the supervised setting, we can construct classifiers by minimizing these loss functions on the labeled data, plus some (convex) regularization term

$$R_\phi(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{L} \phi(y_i \mathbf{x}_i^\top \mathbf{w}) + \lambda \Omega(\mathbf{w}) \,. \tag{1}$$

In the semi-supervised setting, we have an additional design matrix corresponding to unlabeled objects $\mathbf{X}_\mathrm{u}$, sized $U \times d$, with unknown labels. We therefore consider the corresponding semi-supervised risk function:

$$R_\phi^\mathrm{semi}(\mathbf{w}, \mathbf{X}, \mathbf{y}, \mathbf{X}_\mathrm{u}, \mathbf{q}) = R_\phi(\mathbf{w}, \mathbf{X}, \mathbf{y}) + \sum_{i=1}^{U} q_i \phi(\mathbf{x}_i^\top \mathbf{w}) + (1 - q_i)\phi(-\mathbf{x}_i^\top \mathbf{w}) \,, \tag{2}$$

where $\mathbf{q} \in [0,1]^U$ are what we will refer to as *responsibilities*, indicating the unknown and possibly 'soft' membership of each object to a class. A related concept is the constraint set, the set of all possible classifiers that can be obtained by minimizing the semi-supervised loss for any vector of responsibilities, which will be useful in the proof:

$$\mathcal{C}_\phi = \left\{ \operatorname*{arg\,min}_{\mathbf{w}} R_\phi^\mathrm{semi}(\mathbf{w}, \mathbf{X}, \mathbf{y}, \mathbf{X}_\mathrm{u}, \mathbf{q}) \middle| \mathbf{q} \in [0,1]^U \right\} .$$

Safety and Pessimism

For a particular labeling for the unlabeled objects, as our measure of improvement we take the difference in performance of the supervised and a new classifier, on all data:

$$D_\phi(\mathbf{w}, \mathbf{w}_\mathrm{sup}, \mathbf{X}, \mathbf{y}, \mathbf{X}_\mathrm{u}, \mathbf{q}) = R_\phi^\mathrm{semi}(\mathbf{w}, \mathbf{X}, \mathbf{y}, \mathbf{X}_\mathrm{u}, \mathbf{q}) - R_\phi^\mathrm{semi}(\mathbf{w}_\mathrm{sup}, \mathbf{X}, \mathbf{y}, \mathbf{X}_\mathrm{u}, \mathbf{q}) \,.$$

Note this criterion uses the surrogate loss instead of, for instance, error rate to avoid conflating improvements due to the unlabeled data with improvement due to changes in the loss. The criterion is "transductive" for two reasons: 1. It corresponds to what we would consider if we did have all the labels 2. As the number of unlabeled objects grows, it converges to the difference in inductive performance. For a semi-supervised classifier to be truly safe, we want it to be at least as good as the supervised model, when the true labels are revealed:

$$\max_{\mathbf{q} \in [0,1]^U} D_\phi(\mathbf{w}_\mathrm{semi}, \mathbf{w}_\mathrm{sup}, \mathbf{X}, \mathbf{y}, \mathbf{X}_\mathrm{u}, \mathbf{q}) \leq 0 \,. \tag{3}$$

Can we give such a guarantee, while making sure the semi-supervised model is better for at least some of the possible labelings?

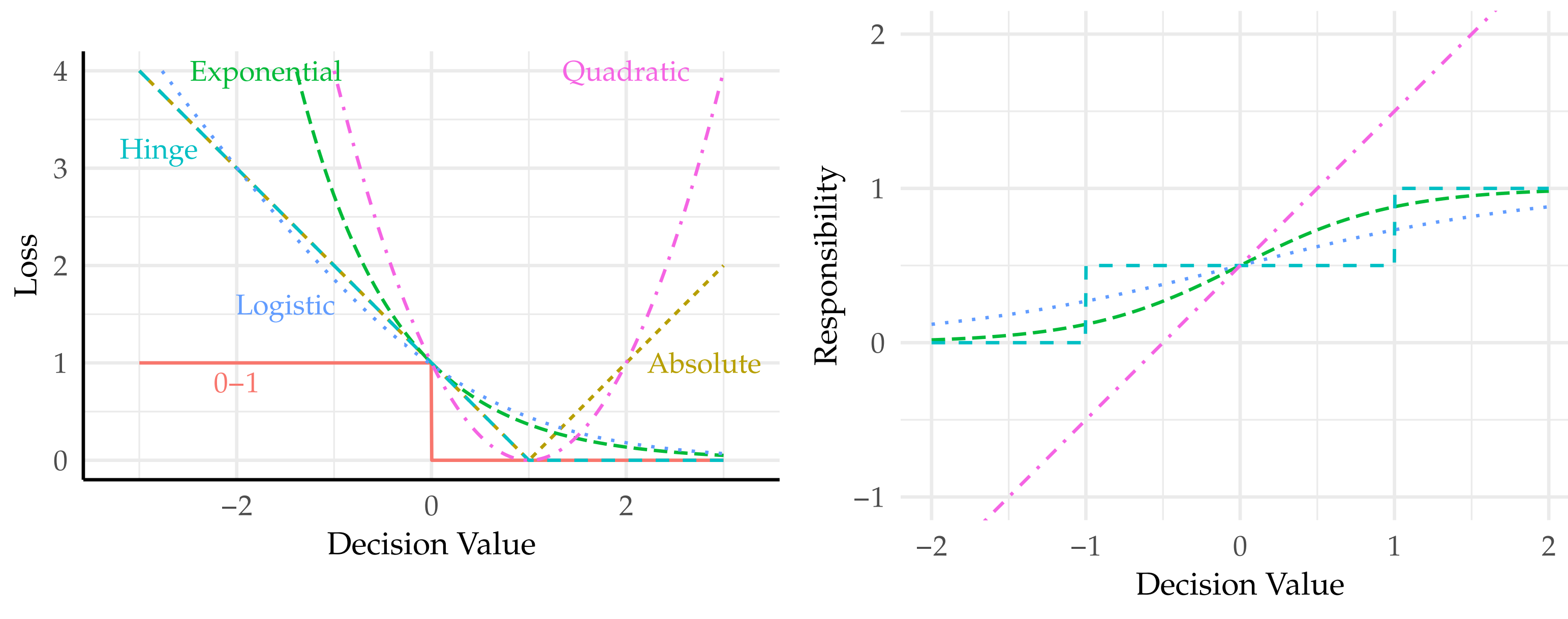

Figure 1: Margin-based loss functions and their corresponding responsibility functions

Impossibilities

For decreasing convex margin-based loss functions, strictly safe semi-supervised learning is impossible.

First, we will discuss the class of decreasing losses for which we can prove it is impossible to get safe performance improvements. The following lemma tells us that no strict improvement is possible if the supervised solution is already part of the constraint set.

Lemma 1. *If $R_\phi(\mathbf{w}, \mathbf{X}, \mathbf{y})$ is strictly convex and $\mathbf{w}_\mathrm{sup} \in \mathcal{C}_\phi$, then there is a soft assignment $\mathbf{q}^*$ such that for every choice of semi-supervised classifier $\mathbf{w}_\mathrm{semi} \neq \mathbf{w}_\mathrm{sup}$, $D_\phi(\mathbf{w}_\mathrm{semi}, \mathbf{w}_\mathrm{sup}, \mathbf{X}, \mathbf{y}, \mathbf{X}_\mathrm{u}, \mathbf{q}^*) > 0$.*

Next, we show that for decreasing losses, it always holds that $\mathbf{w}_\mathrm{sup} \in \mathcal{C}_\phi$, by deriving what q we need to assign to the unlabeled objects to recover $\mathbf{w}_\mathrm{sup}$ by minimizing R^semi.

Lemma 2. *If ϕ is a decreasing convex margin-based loss function where for each unlabeled object $\mathbf{x}$, the derivatives $\phi'(-\mathbf{x}^\top\mathbf{w}_\mathrm{sup})$ and $\phi'(\mathbf{x}^\top\mathbf{w}_\mathrm{sup})$ exist, we can recover $\mathbf{w}_\mathrm{sup}$ by minimizing the semi-supervised loss by assigning responsibilities $\mathbf{q} \in [0,1]^U$ as*

$$q = \frac{\phi'(-\mathbf{x}^\top\mathbf{w}_\mathrm{sup})}{\phi'(\mathbf{x}^\top\mathbf{w}_\mathrm{sup}) + \phi'(-\mathbf{x}^\top\mathbf{w}_\mathrm{sup})} \,, \tag{4}$$

if $\phi'(\mathbf{x}^\top\mathbf{w}_\mathrm{sup}) + \phi'(-\mathbf{x}^\top\mathbf{w}_\mathrm{sup}) \neq 0$, and any $q \in [0,1]$ otherwise.

Combining these two lemmata, we get the following theorem, which specifies our impossibility result.

Theorem 1. *Let ϕ be a decreasing convex margin-based loss function and $\mathbf{w}_\mathrm{sup}$ be the unique minimizer of a strictly convex $R_\phi(\mathbf{w}, \mathbf{X}, \mathbf{y})$ and suppose for each unlabeled object $\mathbf{x}$, the derivatives $\phi'(-\mathbf{x}^\top\mathbf{w}_\mathrm{sup})$ and $\phi'(\mathbf{x}^\top\mathbf{w}_\mathrm{sup})$ exist. There is no semi-supervised classifier $\mathbf{w}_\mathrm{semi}$ for which Equation (3) holds, while having at least one $\mathbf{q}^*$ for which $D_\phi(\mathbf{w}_\mathrm{semi}, \mathbf{w}_\mathrm{sup}, \mathbf{X}, \mathbf{y}, \mathbf{X}_\mathrm{u}, \mathbf{q}^*) < 0$.*

Possibilities

For non-decreasing margin-based loss functions, safe improvements are sometimes possible!

When can we expect safe semi-supervised learning to allow for improvements of its supervised counterpart? We can always find a semi-supervised learner that is at least as good as the supervised one, by simply sticking to the supervised solution. To show that we can do better than that, consider the following.

If R_ϕ^semi is convex in $\mathbf{w}$, then since D_ϕ is linear in $\mathbf{q}$ and $[0,1]^U$ is a compact space we can invoke (Sion, 1958, Corrolary 3.3), which states that the value of our minimax problem of finding a safe classifier is equal to the value of the maximin problem:

$$\max_{\mathbf{q} \in [0,1]^U} \min_{\mathbf{w}} D_\phi(\mathbf{w}, \mathbf{w}_\mathrm{sup}, \mathbf{X}, \mathbf{y}, \mathbf{X}_\mathrm{u}, \mathbf{q}) \,. \tag{5}$$

Now suppose $\mathbf{w}_\mathrm{sup}$ is not in $\mathcal{C}_\phi$. In that case, the inner minimization in Equation (5) is always strictly smaller than 0 for each $\mathbf{q}$ because of the strict convexity of the loss. This means that Equation (5) is strictly smaller than 0 and in turn the same holds for our original minimax problem. So when $\mathbf{w}_\mathrm{sup} \notin \mathcal{C}_\phi$, we can expect improvements. One sufficient condition for this to occur is the following.

Theorem 2. *Let*

$$\phi'(a) \begin{cases} \leq 0, & \text{if } a \leq 1 \\ > 0, & \text{if } a > 1, \end{cases}$$

and R_ϕ^semi be strictly convex. If, for all $\mathbf{x} \in \mathbf{X}_\mathrm{u}$, $|\mathbf{x}^\top\mathbf{w}_\mathrm{sup}|$ is larger than 1, then $\mathbf{w}_\mathrm{semi} \neq \mathbf{w}_\mathrm{sup}$. So, we get an improved semi-supervised estimator if all points in $\mathbf{X}_\mathrm{u}$ are outside of the margin.

While this is sufficient to illustrate our possibility result, stricter conditions are possible.

Discussion

Since we use a rather strict criterion for safety, some may dismiss the fact that for some losses there are no safe procedures as obvious. We also show, however, that in some cases guarantees can actually be given. Overall, our analysis offers a different perspective on the possibilities of (safe) SSL and, as such, on the need for additional assumptions to guarantee strict improvements.