Consistency and Finite Sample Behavior of Binary Class Probability Estimation

Alexander Mey,[1] **Marco Loog,** [1 2]
[1] Delft University of Technology, The Netherlands
[2] University of Copenhagen, Denmark
a.mey@tudelft.nl, m.loog@tudelft.nl

Abstract

We investigate to which extent one can recover class probabilities within the empirical risk minimization (ERM) paradigm. We extend existing results and emphasize the tight relations between empirical risk minimization and class probability estimation. Following previous literature on excess risk bounds and proper scoring rules, we derive a class probability estimator based on empirical risk minimization. We then derive conditions under which this estimator will converge with high probability to the true class probabilities with respect to the L_1-norm. One of our core contributions is a novel way to derive finite sample L_1-convergence rates of this estimator for different surrogate loss functions. We also study in detail which commonly used loss functions are suitable for this estimation problem and briefly address the setting of model-misspecification.

Introduction

In binary classification problems, we try to predict a label $y \in \{-1, 1\} = \mathcal{Y}$ based on an input feature vector $x \in \mathcal{X}$. Since optimizing for the classification accuracy is often computationally too complex, one typically measures performance through a surrogate loss function. Such methods are designed to achieve good classification performance, but often we are also interested in the classifier's confidence or a class probability estimate as such. For instance, we may not only want to classify a tumor as benign or malignant, but also estimate a probability that the predicted label is wrong. Also various methods in active or semi-supervised learning rely on such class probability estimates. For example, in active learning, they are used in uncertainty based rules (Lewis and Catlett 1994; Roy and McCallum 2001), while in semi-supervised learning, they are needed in techniques like entropy regularization (Grandvalet and Bengio 2004).

In this paper, we derive necessary and sufficient conditions under which classifiers, obtained through the minimization of an empirical loss function, allow us to estimate the class probability in a consistent way. More precisely, we present a *general* way to derive finite sample bounds based on those conditions. While the use of class probability estimates, as argued before, finds a broad audience, the necessary tools to understand the behavior, especially the literature on proper scoring rules, is not that broadly known. So next to our contribution on finite sample behavior for class probability estimation, we present a condensed introduction to this, in our opinion, under-appreciated field.

A proper scoring rule is essentially a loss function that can measure the class probability *point-wise*. We investigate in which circumstances those loss functions make use of this potential and lift this point-wise property to the complete space. Next to proper scoring rules we use *excess risk bounds* to come to our results. Excess risk bounds are essentially inequalities that quantify how much an empirical risk minimizer is off from the true risk. Interestingly, our work does not need any specific excess risk bound and is thus very flexible. Any progress in that theory may also translate to this work. Furthermore, if one is willing to make assumptions on the underlying distributions that lead to stronger excess risk bounds, we immediately also get stronger bounds for our results under the same assumptions.

Combining those two areas, our main contributions are the following. Based on the existing literature, we define in Equation (9) a class probability estimate $\hat{\eta}$ derived from an empirical risk minimizer. Based on this, we analyze to which extent commonly used loss functions are suitable for the task of class probability estimation. We then derive conditions that ensure that the estimator $\hat{\eta}$ converges in probability, for an increasing sample size, to the true class probabilities and we also analyze the rate at which this convergence takes place. For ease of exposition all of the previous analysis is done with a well-specification condition that we derive in Theorem 1. We, however, also discuss how this analysis is to be interpreted when this well-specification condition does not hold. As a direct application of our theory we derive error bounds when estimating the class probability with a classification method trained with the squared or the logistic loss. We note already that the rate for the logistic loss has, to the best of our knowledge, not been reported in the literature. Finally we discuss how one can extend this work to asymmetric loss functions and analyze their convergence behavior per class label. To start with, however, the following two sections cover related work and some preliminaries.

Related Work

Many results on class probability estimation in the context of non-parametric regression can be found in Györfi et al.

(2002). The main differences from our results to those type of results is threefold. The first difference is conceptual. While the results presented in Györfi et al. (2002) investigate methods that are specifically designed for class probability estimation, we ask the question if it is possible to obtain consistent class probability estimates with classification methods. Second, to obtain meaningful convergence rate guarantees, the results of Györfi et al. (2002) make assumptions on the distribution. We shift this burden from the distribution to the hypothesis set used. The difference is, that while we always have meaningful finite sample guarantees, our estimation procedure is not consistent in the case of model misspecification. The methods used by Györfi et al. (2002) are always consistent, but may have arbitrarily slow convergence on some distributions. Third, as we assume that the excess risk bounds we use are true with high probability over drawn samples, our convergence results hold with high probability, while Györfi et al. (2002) makes those statements in expectation over the sampling process.

The starting point of our analysis follows closely the notation and concepts as described by Buja, Stuetzle, and Shen (2005), Reid and Williamson (2010) and Reid and Williamson (2011). While Buja, Stuetzle, and Shen (2005) and Reid and Williamson (2010) deal with the inherent structure of proper scoring rules, Reid and Williamson (2011) make connections between the expected loss in prediction problems and divergence measures of two distributions. In contrast to that we investigate under which circumstances proper scoring rules can make use of their full potential in order to estimate class probabilities. Similar to our work Reid and Williamson (2009) gather different sources, in addition to the theory of proper scoring rules, and present general results on regret bounds for class probability estimation. Our work strongly differs in techniques used and thus also in the type of result. Reid and Williamson (2009) use an integral representation of the Bayes risk and derive point-wise regret bounds on the Bregman divergence (as in Theorem 3). We draw from the literature of learning theory and excess risk bounds and derive high-probability L_1 regret bounds.

Telgarsky, Dudík, and Schapire (2015) perform an analysis similar to ours as they also investigate convergence properties of a class probability estimator, their start and end point are very different though. While we start with theory from proper scoring rules, their paper directly adopts the class probability estimator as found in Zhang (2004). The problem is that Zhang (2004) does not evaluate this estimator with respect to any convergence or consistency properties, and it therefore remains unclear if it is the correct choice in any sense. This paper contributes to close this gap and answers this questions They show that the estimator converges to a unique class probability model. In relation to this one can view this paper as an investigation of this unique class probability model and we give necessary and sufficient conditions that lead to convergence to the true class probabilities. Note also that their paper uses convex methods, while our work in comparison draws from the theory of proper scoring rules.

Agarwal and Agarwal (2015) look at the problem in a more general fashion. They connect different surrogate loss functions to certain statistics of the class probability distribution, e.g. the mean, while we focus on the estimation of the full class probability distribution. This allows us to come to more specific results, such as finite sample behavior.

Another general analysis can be found in Steinwart (2007). He presents a general tool to relate convergence in a surrogate risk to the convergence in a target risk, and also presents finite sample rates. As we focus on class probability estimation we are able to derive more specific results, and in particular our Lemma 3 and Corollary 3 tell us when condition (12) of Theorem 2.13 from Steinwart (2007) is true for class probability estimation.

The probability estimator we use also appears in Agarwal (2014) where it is used to derive excess risk bounds, referred to as surrogate risk bounds, for bipartite ranking. The methods used are very similar in the sense that these are also based on proper scoring rules. The difference is again the focus, the conditions used and the conclusions made. They introduce the notion of strongly proper scoring rules which directly allows one to bound the L_2-norm, and thus the L_1-norm, of the estimator in terms of the excess risk. We show that convergence can be achieved already under milder conditions. We then use the concept of modulus of continuity, of which strongly proper scoring rules are a particular case, to analyze the rate of convergence for class probability estimation. Agarwal (2014) on the other hand derives risk bounds for the ranking error, which essentially measures the probability that a randomly drawn positive instance gets assigned a lower value (called score in that context) than a randomly drawn negative instance.

Preliminaries

We work in the classical statistical learning setup for binary classification. We assume that we observe a finite i.i.d. sample $(x_i, y_i)_{1 \leq i \leq n}$ drawn from a distribution P on $\mathcal{X} \times \mathcal{Y}$. Here $\mathcal{X}$ denotes a feature space and $\mathcal{Y} = \{-1, 1\}$ denotes a binary response variable. We then decide upon a hypothesis class $\mathcal{F}$ such that every $f \in \mathcal{F}$ is a map $f : \mathcal{X} \to \mathcal{V}$ for some space $\mathcal{V}$. Given the space $\mathcal{V}$ we call any function $l : \{-1, 1\} \times \mathcal{V} \to [0, \infty)$ a *loss function*. The interpretation of the loss function is that we incur the penalty $l(y, v)$ when we predicted a value v while we actually observed the label y. Our goal is then to find a predictor $f_n \in \mathcal{F}$ based on the finite sample such that $\mathbb{E}[l(Y, f_n(X)]$ is small, where $X \times Y$ is a random variable distributed according to P. In other words, we want to find an estimator f_n that approximates the true risk minimizer f_0 well in terms of the expected loss, where

$$f_0 := \arg\min_{f \in \mathcal{F}} \mathbb{E}[l(Y, f(X))]. \tag{1}$$

The estimator f_n is often chosen to be the empirical risk minimizer

$$f_n = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} l(y_i, f(x_i)). \tag{2}$$

As we show in this paper, finding such an f_n implicitly means to find a good estimate for $p(y \mid x) := P(Y = y \mid$

$X = x$) in many settings. Since we regularly deal with $p(y \mid x)$ and related quantities we introduce the following notation. To start with, we define $\eta(x) := P(Y = 1 \mid X = x)$. Depending on the context we drop the feature x and think of $\eta \in [0, 1]$ as a scalar. Accepting the small risk of overloading the notation we sometimes also think of η as a Bernoulli distribution with outcomes in $\mathcal{Y}$ and parameter η, as in the following definition. We define the *point-wise conditional risk* as

$$L(\eta, v) := \mathbb{E}_{Y\sim\eta}[l(Y, v)] = \eta l(1, v)+(1-\eta)l(-1, v), \quad (3)$$

the *optimal point-wise conditional risk* as

$$L^*(\eta) := \min_{v\in\mathcal{V}} L(\eta, v), \quad (4)$$

and we denote by $v^*(\eta)$ the set of values that optimize the point-wise conditional risk

$$v^*(\eta) := \arg\min_{v\in\mathcal{V}} L(\eta, v). \quad (5)$$

Finally we define the *conditional excess risk* as

$$\Delta L(\eta, v) := L(\eta, v) - L^*(\eta). \quad (6)$$

Proper Scoring Rules

If we chose $\mathcal{V} = [0, 1]$, we say that $l : \{-1, 1\} \times \mathcal{V} \to \mathbb{R}$ is a *CPE loss*, where CPE stands for class probability estimation. The name stems from the fact that if $\mathcal{V} = [0, 1]$ it is already normalized to a value that can be interpreted as a probability. If l is a CPE loss we call it a *proper scoring rule* or *proper loss* if $\eta \in v^*(\eta)$ and we call it a *strictly proper scoring rule* or *strictly proper loss* if $v^*(\eta) = \{\eta\}$. In other words, l is a proper scoring rule if η is *a* minimizer of $L(\eta, \cdot)$ and this is strict if η is the only minimizer. In case l is strict we drop the set notation of v^*, so that $v^*(\eta) = \eta$.

Link Functions

As we will see later strictly proper CPE losses are well suited for class probability estimation. In general, however, we cannot expect that $\mathcal{V} = [0, 1]$, but we may still want to use the corresponding loss function for class probability estimation. To do that we will use the concept of link functions (Buja, Stuetzle, and Shen 2005; Reid and Williamson 2010). A *link function* is a map $\psi : [0, 1] \to \mathcal{V}$, so a function that indeed links the values from $\mathcal{V}$ to something that can be interpreted as a probability. Combining such a link function with a loss $l : \{-1, 1\} \times \mathcal{V} \to [0, \infty)$ one can define a CPE loss l_ψ as follows.

$$l_\psi : \{-1, 1\} \times [0, 1] \to [0, \infty)$$
$$l_\psi(y, q) := l(y, \psi(q))$$

We call the combination of a loss and a link function (l, ψ) a *(strictly) proper composite loss* if l_ψ is (strictly) proper as a CPE loss.

To distinguish between the losses l and l_ψ we subscript the quantities (3)-(6) with a ψ if we talk about l_ψ instead of l. For example we define $L_\psi(\eta, q) := L(\eta, \psi(q))$ for $q \in [0, 1]$ and in the same way we define $v^*_\psi(\eta)$, $L^*_\psi(\eta)$ and $\Delta L_\psi(\eta, q)$. Note that if (l, ψ) is a strictly proper composite loss, we know that $v^*_\psi(\eta)$ are single element sets, but the same does not need to hold for $v^*(\eta)$.

Degenerate Link Functions

To ask a composite loss (l, ψ) to be proper is not a strong requirement, one can check that choosing ψ as constant function already fulfills this. This is because a composite loss (l, ψ) is proper, iff the true class probability η is a minimizer of the conditional risk $L_\psi(\eta, \cdot)$, i.e. $\eta \in v^*_\psi(\eta)$. If ψ is constant, then so is the conditional risk $L_\psi(\eta, \cdot)$ and then every value is a minimizer, so in particular η is a minimizer. We want to avoid this degenerate behavior for the task of probability estimation and will ask ψ to cover enough of $\mathcal{V}$ in the following sense. We call a composite loss (l, ψ) *non-degenerate* if for all $\eta \in [0, 1]$ we have that $\operatorname{Im}\psi \cap v^*(\eta) \neq \emptyset$, where $\operatorname{Im}\psi \subset \mathcal{V}$ is the image of ψ on $[0, 1]$. This does not directly exclude constant link functions for example, but consider the following. If ψ is constant and non-degenerate, then there is a single $v = \operatorname{Im}\psi$ such that $v \in v^*(\eta)$ for all η. Thus v would always minimize the loss, and we would, irrespectively of the input, always predict v. This is of course a property that no reasonable loss function should carry.

Behavior of Proper Composite Losses

For our convergence results we will need a loss function to be a strictly proper CPE loss. In this section we investigate how to characterize those loss functions.

We start by investigating proper CPE loss functions. Our first lemma states that the link functions that turns the loss l into a proper composite loss is already defined by the behavior of v^*. As this lemma and Lemma 2 are straightforward derivations from the definitions, and of no further interest, we refer for the proofs to the supplementary material.

Lemma 1. *Let $l : \{-1, 1\} \times \mathcal{V} \to [0, \infty)$ be a loss function and ψ be a link function. The composite loss function (l, ψ) is then proper and non-degenerate if and only if $\psi \in v^*$, meaning that $\psi(\eta) \in v^*(\eta)$ for all $\eta \in [0, 1]$.*

This lemma gives thus necessary and sufficient condition on our link ψ to lead to a proper loss function. The result is very similar to Corollary 12 and 14 found in Reid and Williamson (2010). Their corollaries state necessary and sufficient conditions on the link function, using the assumption that the loss has differentiable partial losses, which is an assumption we don't require.

Later we show that *strictly* proper losses, together with some additional assumptions, lead to consistent class probability estimates. So it is useful to know how to characterize those functions. The following lemma shows that a link function that turns a loss into strictly proper and non-degenerate CPE loss can be characterized again by the behavior of v^*.

Lemma 2. *Let $l : \{-1, 1\} \times \mathcal{V} \to [0, \infty)$ be a loss function and ψ a link function. A composite loss function (l, ψ) is then strictly proper and non-degenerate if and only if $\psi \in v^*$ and $v^*(\eta_1) \cap v^*(\eta_2) \cap \operatorname{Im}\psi = \emptyset$ for all pairwise different $\eta_1, \eta_2 \in [0, 1]$.*

So if (l, ψ) is a strictly proper composite loss it will fulfill some sort of injectivity condition on the sets $v^*(\eta)$. With this we will be able to define an inverse ψ^{-1} on those sets, and

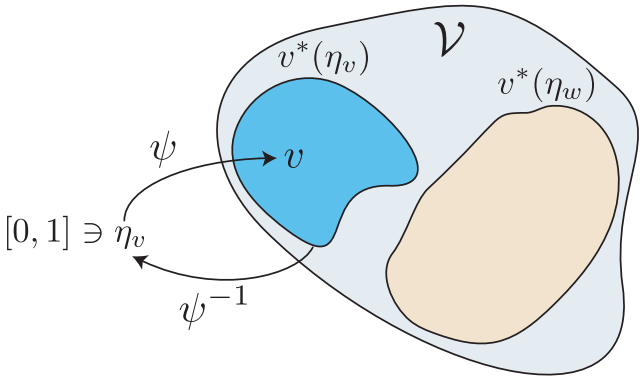


Figure 1: The way we generally think of the mapping ψ, ψ^{-1} and the sets v^* if (l, ψ) is non-degenerate and strictly proper. In those cases we can extend ψ^{-1} to the sets v^*. This is well defined as the sets $v^*(\eta_v)$ and $v^*(\eta_w)$ have empty intersection for different $\eta_v, \eta_w \in [0, 1]$. Note that Lemma 2 guarantees that $\psi(\eta_v) \in v^*(\eta_v)$.

this will be essentially our class probability estimator. With Lemma 2 we can connect every $v \in \mathcal{V}$ to a unique η_v by the unique relation $v \in v^*(\eta_v)$ if we assume that v^* *disjointly covers* $\mathcal{V}$ in the sense that

$$\bigcup_{\eta \in [0,1]} v^*(\eta) = \mathcal{V} \quad \text{and} \tag{7}$$

$$v^*(\eta_1) \cap v^*(\eta_2) = \emptyset \quad \forall\, \eta_1, \eta_2 \in [0, 1], \;\; \eta_1 \neq \eta_2. \tag{8}$$

Note that we know from Lemma 2 that for strict properness it is sufficient for (l, ψ) that the disjoint property (8) only holds on $\operatorname{Im} \psi$, the image of ψ. This is merely a technicality and we will assume from now on that every strictly proper composite loss will satisfy (8). The covering property (7) on the other hand can be violated. This happens for example if we use the squared loss together with $\mathcal{V} = \mathbb{R}$. For the squared loss $v^*(\eta) = 2\eta - 1$, so it only covers the space $[-1, 1]$.

If we assume, however, that the regularity properties (7) and (8) hold for a strictly proper non-degenerate composite loss (l, ψ) we can extend the domain of ψ^{-1} from $\operatorname{Im} \psi$ to the whole of $\mathcal{V}$, see also Figure 1.

Definition 1. *Let (l, ψ) be a strictly proper, non-degenerate composite loss and assume that v^* disjointly covers $\mathcal{V}$. We define, by abuse of notation, the inverse link function $\psi^{-1} : \mathcal{V} \to [0, 1]$ by $\psi^{-1}(v) = \eta_v$, where η_v is the unique element in $[0, 1]$ such that $v \in v^*(\eta_v)$.*

The requirements from the previous definition is what we consider the archetype of a composite loss that is suitable for probability estimation, although not all of the requirements are necessary. This motivates the following definition.

Definition 2. *We call a composite loss (l, ψ) a natural CPE loss if ψ is non-degenerate, v^* fulfills the disjoint cover property* (7) *and* (8) *and (l, ψ) is strictly proper.*

We now have all the necessary work done to make the following observation.

Loss	$l(v, y)$	$v^{*-1}(v)$
Sq	$(1 - yv)^2$	$\frac{v+1}{2}$
Log	$\ln(1 + e^{-vy})$	$\frac{1}{1+e^{-v}}$
SqH	$\max(0, 1 - vy)^2$	$T(\frac{v+1}{2})$
Hinge	$\max(0, 1 - vy)$	$\begin{cases} \frac{1}{2} & v \in (-1, 1) \\ (0, \frac{1}{2}) & v = -1 \\ (\frac{1}{2}, 1) & v = 1 \\ 1, & v > 1 \\ 0, & v < -1 \end{cases}$
0-1	$I_{\{\operatorname{sign}(vy) \neq 1\}}$	$\begin{cases} [\frac{1}{2}, 1] & \text{if } v \in (0, \infty) \\ [0, \frac{1}{2}], & v \in (-\infty, 0) \\ \frac{1}{2}, & v = 0 \end{cases}$

Table 1: The loss functions we consider in this paper. The function $v^{*-1}(v)$ is the function that transforms a real output to a class probability estimate. Here $T(x) := \min(\max(0, x), 1)$.

Corollary 1. *If (l, ψ) is a natural CPE loss, then $\psi^{-1} = v^{*-1}$.*

The corollary tells us that we can optimize our loss function over $\mathcal{V}$ to get $v^*(\eta)$ and then map this back with the inverse link ψ^{-1} to restore the class probability η. For this we once more refer to Figure 1. Remember that the set $v^*(\eta_v)$ is the set of all $v \in \mathcal{V}$ that minimize the loss if the true class probability was η_v. If we use a natural CPE loss (l, ψ) we know then that ψ^{-1} maps all those points back to η_v.

Given a predictor $f : \mathcal{X} \to \mathcal{V}$ this motivates to define an estimator of $\eta(x)$ as

$$\hat{\eta} = \hat{\eta}(x) = \psi^{-1}(f(x)). \tag{9}$$

Later we derive conditions under which $\hat{\eta}(x)$ converges in probability towards $\eta(x)$ when using an empirical risk minimizer f_n as a prediction rule. More formally; Given any $\epsilon > 0$ we show that under certain conditions $\hat{\eta}_n(x) := \psi^{-1}(f_n(x))$ satisfies

$$P(|\hat{\eta}_n(X) - \eta(X)| > \epsilon) \xrightarrow{n \to \infty} 0, \tag{10}$$

where the probability is measured with respect to P. In the next section, however, we want to investigate first v^* and v^{*-1} for some commonly used loss functions.

Analysis of Loss Functions

Throughout this paper we consider the following loss functions: Squared loss (Sq), logistic loss (log), squared hinge loss (SqH), Hinge loss and the 0-1 loss, see the first two columns of Table 1 for specifications. Table 2 shows the link function that turns the loss functions into a strictly proper composite loss, if possible. Note that this can be decided with the help of Lemma 2 and the functions $v^*(\eta)$ which are also shown in Table 2. We note that the behavior of the

Loss	$\psi(\eta)$	$v^*(\eta)$
Sq	$2\eta - 1$	$2\eta - 1$
Log	$\ln \frac{\eta}{1-\eta}$	$\ln \frac{\eta}{1-\eta}$
SqH	$2\eta - 1$	$\begin{cases} 2\eta - 1, & \eta \in (0,1) \\ [1, \infty), & \eta = 1 \\ (-\infty, -1], & \eta = -1 \end{cases}$
Hinge	-	$\begin{cases} \text{sign}(2\eta - 1), & \eta \in (0,1) \setminus \frac{1}{2} \\ [-1, 1] & \eta = \frac{1}{2} \\ [1, \infty), & \eta = 1 \\ (-\infty, -1], & \eta = -1 \end{cases}$
0-1	-	$\begin{cases} (0, \infty), & \eta \in (\frac{1}{2}, 1] \\ (-\infty, 0), & \eta \in [0, \frac{1}{2}) \\ \mathbb{R}, & \eta = \frac{1}{2} \end{cases}$

Table 2: The different loss functions we consider in this paper together with their link functions that turn them into CPE losses (if possible).

squared and squared hinge loss seems to be very similar, expect that from Table 1 we can see that the class probability estimate from the squared loss is not necessarily in $[0, 1]$, and in that sense clipping it to $[0, 1]$, as proposed in Sugiyama (2010), is actually wrong. Instead one would have to make sure that v takes only values in $[-1, 1]$.

As already noted by Buja, Stuetzle, and Shen (2005), also Table 2 shows that the hinge loss is not suitable for class probability estimation. We observe that the intersections of $v^*(\eta)$ for different $\eta \in [0, 1]$ are not disjoint. By Lemma 2 we can conclude that there is no link ψ such that (l, ψ) is strictly proper. One way to fix this, proposed by Duin and Tax (1998) and similar by Platt (1999), is to fit a logistic regressor on top of the support vector machine. Bartlett and Tewari (2004) investigate the behavior of the hinge loss deeper by connecting the class probability estimation task to the sparseness of the predictor. The hinge loss is of course classification calibrated (essentially meaning that we find point-wise the correct label with it), so between our considered surrogate losses it is the only one that really directly solves the classification problem without implicitly estimating the class probability.

Convergence of the Estimator

We now prove that the estimator $\hat{\eta}(x)$ as defined in Equation (9) converges in probability and in the L_1-norm to the true class probability η whenever we use an empirical risk minimizer, for which we have excess risk bounds.

Using the True Risk Minimizer

Before we can investigate under which conditions an empirical risk minimizer can (asymptotically) retrieve $\eta(x)$ we need to investigate under which conditions the true risk minimizer can retrieve it. In this subsection we formulate a theorem that gives necessary and sufficient conditions for that. Not surprisingly we basically require that our hypothesis class is rich enough so as to contain the class probability distribution already. Bartlett, Jordan, and McAuliffe (2006) and similar works often avoid problems caused by restricted classes by assuming from the beginning that the hypothesis class consists of all measurable functions. Having a restricted hypothesis class, however, is crucial for our analysis as that allows us to use the tools from learning theory.

In this setting we assume that we use a hypothesis class $\mathcal{F}$ where $f \in \mathcal{F}$ are functions $f : \mathcal{X} \to \mathcal{V}$. If we want to do class probability estimation we rescale those functions by composing them with the inverse link $\psi^{-1} : \mathcal{V} \to [0, 1]$ so that we effectively use the hypothesis class $\psi^{-1}(\mathcal{F}) := \{\psi^{-1} \circ f \mid f \in \mathcal{F}\}$. We then get the following theorem about the possibility of retrieving the class probability with risk minimization.

Theorem 1. *Assume that (l, ψ) is a natural CPE loss function. Let*

$$f_0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}[l(Y, f(X)].$$

Then $\psi^{-1}(f_0(x)) = \eta(x)$ almost surely if and only if $\eta \in \psi^{-1}(\mathcal{F})$.

Following Theorem 1 we need to assume that our hypothesis class is flexible enough for consistent class probability estimation. We formulate this assumption as follows.

Assumption A Given a natural CPE loss (l, ψ) we assume that $\eta \in \psi^{-1}(\mathcal{F}) = \{\psi^{-1} \circ f \mid f \in \mathcal{F}\}$. Later we will deal with the case of misspecification, i.e. when $\eta \notin \psi^{-1}(\mathcal{F})$.

Using the Empirical Risk Minimizer

In the previous section we considered the possibility of retrieving class probability estimates with the true risk minimizer. To move on to empirical risk minimizers we need the notion of excess risk bounds.

Definition 3. *Let $f_n : \mathcal{X} \to \mathbb{R}$ be any estimator of $f_0 \in \mathcal{F}$, which may depend on a sample of size n. We call*

$$B^{\mathcal{F}}(n, \gamma) : \mathbb{N} \to [0, \infty)$$

an excess risk bound for f_n if for all $\gamma > 0$ we have $B^{\mathcal{F}}(n, \gamma) \to 0$ for $n \to \infty$ and with probability of at least $1 - \gamma$ over the n-sample we have

$$\begin{aligned} &\mathbb{E}_X[\Delta L(\eta(X), f_n(X))] \\ =&\mathbb{E}_{X,Y}[l(Y, f_n(X)) - l(Y, f_0)] \leq B^{\mathcal{F}}(n, \gamma). \end{aligned}$$

Excess risk bounds are typically in the order of $\left(\frac{\text{comp}(\mathcal{F})}{n}\right)^{\beta}$, where $\beta \in [0.5, 1]$ and $\text{comp}(\mathcal{F})$ is a notion of model complexity. Common measures for the model complexity are the VC dimension (Vapnik 1998), Rademacher complexity (Bartlett, Bousquet, and Mendelson 2005) or ϵ-cover (Benedek and Itai 1991). The existence of excess risk bounds is tied to the finiteness of any of those complexity notions. A lot of efforts in this line of research are made to find relations between the exponent β and the statistical learning problem given by $\mathcal{F}$, the loss l and the underlying

distribution P. Conditions that ensure $\beta > \frac{1}{2}$ are often called easiness conditions, such as the Tsybakov condition (Tsybakov 2004) or the Bernstein condition (Audibert 2004). Intuitively those conditions often state that the variance of our estimator gets smaller the closer we are to the optimal solution. For a in-depth discussion and some recent results we refer to the work of Grünwald and Mehta (2016).

Excess risk bounds allow us to bound $\Delta L(\eta(x), f_n(x))$ for a loss l, so in particular we can bound $\Delta L_\psi(\eta(x), \hat{\eta}(x))$ for a composite loss (l, ψ). We will show L_1-convergence by connecting the behavior of $\Delta L_\psi(\eta(x), \hat{\eta}(x))$ to $|\eta(x) - \hat{\eta}(x)|$. The following lemma introduces a condition that allows us to draw this connection.

Lemma 3. *Let (l, ψ) be a natural CPE loss. Assume that for all $\eta \in [0, 1]$ the maps*

$$L^0_\psi(\eta, \cdot) := L_\psi(\eta, \cdot) \restriction_{[0,\eta]} \colon [0, \eta] \to \mathbb{R}$$

and

$$L^1_\psi(\eta, \cdot) := L_\psi(\eta, \cdot) \restriction_{[\eta,1]} \colon [\eta, 1] \to \mathbb{R}$$

are strictly monotonic, where $L_\psi(\eta, \cdot) \restriction_I$ refers to the restriction of the mapping $L_\psi(\eta, \cdot)$ to an interval I. This is the case iff $L_\psi(\eta, \cdot)$ is strictly convex with η as its minimizer. Then there exists for all $\epsilon > 0$ a $\delta = \delta(\epsilon) > 0$ such that for all $\eta, \hat{\eta} \in [0, 1]$

$$|\Delta L_\psi(\eta, \hat{\eta})| < \delta \Rightarrow |\eta - \hat{\eta}| < \epsilon. \tag{11}$$

Proof. With the assumptions on $L^0_\psi(\eta, \cdot)$ and $L^1_\psi(\eta, \cdot)$ we know that ${L^0_\psi}^{-1}(\eta, \cdot)$ and ${L^1_\psi}^{-1}(\eta, \cdot)$ exist and are continuous (Hoffmann 2015). By definition that means that for every $l, \hat{l} \in \operatorname{Im} L^0_\psi(\eta, \cdot)$ and for all $\epsilon > 0$ there exists a $\delta > 0$ such that

$$|\hat{l} - l| < \delta \Rightarrow |{L^0_\psi}^{-1}(\eta, \hat{l}) - {L^0_\psi}^{-1}(\eta, l)| < \epsilon \tag{12}$$

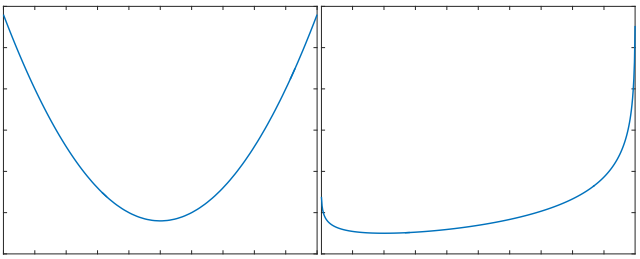
and similar for $L^1_\psi(\eta, \cdot)$. W.l.o.g assume now that $\hat{\eta} < \eta$ so that $\hat{\eta} \in [0, \eta]$. Plugging $l = L^0_\psi(\eta, \eta)$ and $\hat{l} = L^0_\psi(\eta, \hat{\eta})$ into (12) we get the following relation.

$$\begin{aligned} &|\Delta L_\psi(\eta, \hat{\eta})| = |L^0_\psi(\eta, \hat{\eta}) - L^0_\psi(\eta, \eta)| < \delta \\ \Rightarrow &|\hat{\eta} - \eta| = |{L^0_\psi}^{-1}(\eta, \hat{l}) - {L^0_\psi}^{-1}(\eta, l)| < \epsilon \end{aligned}$$

□

The map $L^0_\psi(\eta, \cdot)$ captures the behavior of the loss when η is the true class probability and we predict a class probability less than η. Similarly $L^1_\psi(\eta, \cdot)$ captures the behavior when we predict a class probability bigger than η, see also Figure 2. In Corollary 3, further below, we draw a connection between $\delta(\epsilon)$ and the modulus of continuity of the inverse functions of $L^1_\psi(\eta, \cdot)$ and $L^0_\psi(\eta, \cdot)$. The function $\delta(\epsilon)$ plays an important role in the convergence rate of the estimator $\hat{\eta}(x)$ as described in the next theorem.

Theorem 2. *Let (l, ψ) be a natural CPE loss and assume Assumption A holds. Furthermore let $B^{\mathcal{F}}(n, \gamma)$ be an excess risk bound for f_n and assume that $L_\psi(\eta, \cdot)$ is strictly convex for all η with η as its minimizer. Then there exists a mapping*



(a) The map $L_\psi(\eta, \cdot)$ for $\eta = 0.2$ and l being the squared loss.

(b) The map $L_\psi(\eta, \cdot)$ for $\eta = 0.2$ and l being the logistic loss.

Figure 2: The map $L_\psi(\eta, \cdot)$ for the squared and the logistic loss. The two maps $L^0_\psi(\eta, \cdot)$ and $L^1_\psi(\eta, \cdot)$ split it into the parts left and right of η.

$\delta(\epsilon) : [0, 1] \to \mathbb{R}$ such that for $\hat{\eta}_n(x) := \psi^{-1}(f_n(x))$ we have with probability of at least $1 - \gamma$ that

$$P(|\eta(X) - \hat{\eta}_n(X)| > \epsilon) \leq \frac{B^{\mathcal{F}}(n, \gamma)}{\delta(\epsilon)}. \tag{13}$$

Proof. Using Lemma 3 for the first inequality, Markov's Inequality for the second and the excess risk bound for the third inequality it follows that

$$\begin{aligned} &P(|\eta(X) - \hat{\eta}_n(X)| > \epsilon) \leq P(\Delta L_\psi(\eta(X), \hat{\eta}_n(X)) > \delta) \\ =&P(\Delta L(\eta(X), f_n(X)) > \delta) \\ \leq&\frac{\mathbb{E}[\Delta L(\eta(X), f_n(X))]}{\delta(\epsilon)} \leq \frac{B^{\mathcal{F}}(n, \gamma)}{\delta(\epsilon)}. \end{aligned}$$

□

This theorem gives us directly the earlier claimed asymptotic convergence result.

Corollary 2. *Under the assumptions of Theorem 2 we have that $\hat{\eta}_n(x) = \psi^{-1}(f_n(x))$ converges in probability and L_1-norm to $\eta(x)$ with probability 1.*

We do not have to restrict ourselves to asymptotic results though. Theorem 2 can also be used to derive rate of convergences as we will see later. But before that we briefly want to address the case of misspecification, i.e. the case when Assumption A does not hold.

Misspecification

For ease of exposition we chose to present the previous analysis under the well-specification of Assumption A. More generally one may formulate Theorem 2 and Corollary 2 by replacing $\eta(x)$ with $\psi^{-1}(f_0(x))$, the two quantities that coincide under Assumption A. Moreover, if L^*_ψ has a gradient Reid and Williamson (2010) show the identity $\Delta L_\psi(\eta, \hat{\eta}) = D_{-L^*_\psi}(\eta, \hat{\eta})$ where $D_{-L^*_\psi}(\eta, \hat{\eta})$ is the with $-L^*_\psi$ associated Bregman divergence between η and $\hat{\eta}$. Excess risk bounds on $\Delta L_\psi(\eta, \hat{\eta})$ translate then into bounds on the Bregman divergence between η and $\hat{\eta}$, which means that in the misspecified case we asymptotically approach the best class probability estimate in terms of this divergence.

Rate of Convergence

For the rate of convergence it is crucial to investigate the function $\delta(\epsilon)$ from Inequality (13). One way to analyze this is to study the modulus of continuity of the inverse functions of $L^0_\psi(\eta,\cdot)$ and $L^1_\psi(\eta,\cdot)$:

Definition 4. *Let $\omega : [0,\infty] \to [0,\infty]$ be a monotonically increasing function. Let $I \subset \mathbb{R}$ be an interval. A function $g : I \to \mathbb{R}$ admits ω as a modulus of continuity at $x \in I$ if and only if*

$$|g(x) - g(y)| \leq \omega(|x-y|)$$

for all $y \in I$.

For example Hölder and Lipschitz continuity are particular moduli of continuity. This notion allows us a to draw the following connection between ϵ and $\delta(\epsilon)$.

Corollary 3. *Let (l,ψ) be a natural CPE loss and let $\omega : [0,\infty] \to [0,\infty]$ be a monotonically increasing function. Assume that for all $\eta \in [0,1]$ the mappings ${L^0_\psi}^{-1}(\eta,\cdot)$ and ${L^1_\psi}^{-1}(\eta,\cdot)$ admit ω as a modulus of continuity at η. Then $\delta(\epsilon) := \omega^{-1}(\epsilon)$ is a mapping such that Implication* (11) *holds.*

Proof. W.l.o.g. assume that $\hat{\eta} \in [0,\eta]$. Let $\hat{l} = L^0_\psi(\eta,\hat{\eta})$ and $l = L^0_\psi(\eta,\eta)$. By using that ${L^0_\psi}^{-1}(\eta,\cdot)$ admits ω as a modulus of continuity we have

$$|{L^0_\psi}^{-1}(\eta,l) - {L^0_\psi}^{-1}(\eta,\hat{l})| \leq \omega(|l-\hat{l}|).$$

Plugging in the definition of $\hat{l}$ and l this means that

$$|\hat{\eta} - \eta| \leq \omega(\Delta L_\psi(\eta,\hat{\eta})).$$

Using the monotonicity of ω it follows that if $\Delta L_\psi(\eta,\hat{\eta}) \leq \delta(\epsilon) = \omega^{-1}(\epsilon)$, then

$$|\eta - \hat{\eta}| \leq \omega(\Delta L_\psi(\eta,\hat{\eta})) \leq \omega(\omega^{-1}(\epsilon)) = \epsilon.$$

This is exactly the Implication (11). □

Note that it follows from the proof that finding a modulus of continuity ω for ${L^0_\psi}^{-1}(\eta,\cdot)$ and ${L^1_\psi}^{-1}(\eta,\cdot)$ can be done by showing the bound $|\hat{\eta}-\eta| \leq \omega(\Delta L_\psi(\eta,\hat{\eta}))$. We will use that in the following examples, where we analyze $\delta(\epsilon)$ for the squared (hinge) loss and the logistic loss. We show that those loss functions lead to a modulus of continuity given by the square root times a constant. Agarwal (2014) calls loss functions that admit this modulus of continuity *strongly-proper* loss functions. The following analysis can thus be found there in more detail and for a few more examples. We will use for simplicity versions of the losses that do not need a link function, and are already CPE losses.

Example: Squared Loss and Squared Hinge Loss Let $l(y,\hat{\eta})$ be given by the partial loss functions $l(1,\hat{\eta}) = (1-\hat{\eta})^2$ and $l(-1,\hat{\eta}) = \hat{\eta}^2$. We can derive that $\Delta L(\eta,\hat{\eta}) = (\eta-\hat{\eta})^2$. With this we can directly bound $|\hat{\eta}-\eta| \leq \sqrt{\Delta L(\eta,\hat{\eta})}$ and thus choose $\delta(\epsilon)$ as the inverse of the square-root function, so that $\delta(\epsilon) = \epsilon^2$. The analysis for the squared hinge loss is the same as this version of the squared loss is already a CPE loss.

Example: Logistic Loss Let $l(y,\hat{\eta})$ be given by the partial loss functions $l(1,\hat{\eta}) = -\ln(\hat{\eta})$ and $l(-1,\hat{\eta}) = -\ln(1-\hat{\eta})$. One can derive that $\Delta L(\eta,\hat{\eta}) = -\eta\ln(\frac{\hat{\eta}}{\eta}) - (1-\eta)\ln(\frac{1-\hat{\eta}}{1-\eta})$. In the supplementary we show the bound $|\eta-\hat{\eta}| \leq \sqrt{\frac{1}{2}\Delta L(\eta,\hat{\eta})}$, as well as that $\frac{1}{2}$ is the optimal constant, so that we can choose $\delta(\epsilon) = 2\epsilon^2$.

Discussion and Conclusion

The starting point of this paper is the question if one can retrieve a class probability estimate based on ERM in a consistent way. To answer this question, we draw from earlier work on proper scoring rules and excess risk bounds. Lemmas 1 and 2, our first results, characterize strictly proper composite loss functions in terms of their link function. Based on those lemmas, we subsequently derive fairly general necessary and sufficient conditions for retrieving the true class probability with ERM as formulated in Theorem 1. We show that to retrieve the true probabilities we essentially need that they are already part of our hypothesis class $\mathcal{F}$.

We show that consistency arises whenever we use strictly proper (composite) loss functions, our hypothesis class is flexible enough, and we have excess risk bounds. This is the case, for example, whenever one of the complexity notions mentioned in this paper is finite. Additionally, we discuss the relation between the finite sample size behavior of the excess risk bound and the probability estimate and examine this relation for two loss functions.

In Lemma 3, we introduce conditions under which a composite loss function (l,ψ) leads to a consistent class probability estimator. In particular we have a condition on the conditional risk $L_\psi(\eta,\cdot)$, see also Figure 2. Based on that we derive in Corollary 3 conditions which allow us to analyze the convergence rate for different loss functions. In the corollary we don't distinguish between $L^0_\psi(\eta,\cdot)$ and $L^1_\psi(\eta,\cdot)$, which leads to the same convergence rate for predicting values left and right from η. But the modulus of continuity for those two functions can be really different, especially when using asymmetric proper scoring rules (Winkler 1994). We believe that by analyzing $L^0_\psi(\eta,\cdot)$ and $L^1_\psi(\eta,\cdot)$ individually one can extend our work to analyze the convergence behavior of asymmetric scoring rules in more detail, meaning that one could achieve different rates for over or underestimating a certain class probability level.

Acknowledgments

This work was funded in part by the Netherlands Organisation for Scientific Research (NWO) and carried out under TOP grant project number 612.001.402.

References

Agarwal, A.; and Agarwal, S. 2015. On Consistent Surrogate Risk Minimization and Property Elicitation. In *Proceedings of The 28th Conference on Learning Theory*, 4–22. Paris, France.

Agarwal, S. 2014. Surrogate regret bounds for bipartite

ranking via strongly proper losses. *Journal of Machine Learning Research* 15(1): 1653–1674.

Audibert, J.-Y. 2004. *Une approche PAC-bayésienne de la théorie statistique de l'apprentissage*. Ph.D. thesis, Université Paris 6.

Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2005. Local Rademacher complexities. *The Annals of Statistics* 33(4): 1497–1537.

Bartlett, P. L.; Jordan, M. I.; and McAuliffe, J. D. 2006. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association* 101(473): 138–156.

Bartlett, P. L.; and Tewari, A. 2004. Sparseness Versus Estimating Conditional Probabilities: Some Asymptotic Results. In *17th Annual Conference on Learning Theory*, 564–578. Banff, Canada.

Benedek, G. M.; and Itai, A. 1991. Learnability with Respect to Fixed Distributions. *Theory of Computer Science* 86(2): 377–389.

Buja, A.; Stuetzle, W.; and Shen, Y. 2005. Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications. Technical report, University Washington.

Duin, R. P.; and Tax, D. M. 1998. Classifier conditional posterior probabilities. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, 611–619. Sydney, NSW, Australia.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised Learning by Entropy Minimization. In *Advances in Neural Information Processing Systems 17*, 529–536. Vancouver, BC, Canada.

Grünwald, P. D.; and Mehta, N. A. 2016. Fast Rates with Unbounded Losses. *The Computing Research Repository* abs/1605.00252.

Györfi, L.; Kohler, M.; Krzyzak, A.; and Walk, H. 2002. *A Distribution-Free Theory of Nonparametric Regression.* Springer series in statistics. Springer.

Hoffmann, H. 2015. On the Continuity of the Inverses of Strictly Monotonic Functions. *Bulletin of the Irish Mathematical Society* 75: 45–57.

Lewis, D. D.; and Catlett, J. 1994. Heterogeneous Uncertainty Sampling for Supervised Learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, 148–156. New Brunswick, NJ, USA.

Platt, J. C. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, 61–74. The MIT Press.

Reid, M. D.; and Williamson, R. C. 2009. Surrogate regret bounds for proper losses. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 897–904. Montreal, QC, Canada.

Reid, M. D.; and Williamson, R. C. 2010. Composite Binary Losses. *Journal of Machine Learning Research* 11: 2387–2422.

Reid, M. D.; and Williamson, R. C. 2011. Information, Divergence and Risk for Binary Experiments. *Journal of Machine Learning Research* 12: 731–817.

Roy, N.; and McCallum, A. 2001. Toward Optimal Active Learning Through Sampling Estimation of Error Reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 441–448. Williamstown, MA, USA.

Steinwart, I. 2007. How to Compare Different Loss Functions and Their Risks. *Constructive Approximation* 26(2): 225–287.

Sugiyama, M. 2010. Superfast-Trainable Multi-Class Probabilistic Classifier by Least-Squares Posterior Fitting. *IEICE Transactions* 93-D(10): 2690–2701.

Telgarsky, M.; Dudík, M.; and Schapire, R. 2015. Convex Risk Minimization and Conditional Probability Estimation. In *Proceedings of The 28th Conference on Learning Theory*, 1629–1682. Paris, France.

Tsybakov, A. B. 2004. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics* 32(1): 135–166.

Vapnik, V. N. 1998. *Statistical Learning Theory*. Wiley-Interscience.

Winkler, R. L. 1994. Evaluating Probabilities: Asymmetric Scoring Rules. *Management Science* 40(11): 1395–1405.

Zhang, T. 2004. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics* 32: 56–134.